Differential Equations,
Dynamical Systems,
and Linear Algebra

This is a volume in
PURE AND APPLIED MATHEMATICS

A Series of Monographs and Textbooks

Editors: SAMUEL EILENBERG AND HYMAN BASS

A complete list of titles in this series is available from the Publishers upon request.

Differential Equations, Dynamical Systems, and Linear Algebra

MORRIS W. HIRSCH AND STEPHEN SMALE
University of California, Berkeley

ACADEMIC PRESS, INC.
Harcourt Brace Jovanovich, Publishers
San Diego New York Boston
London Sydney Tokyo Toronto

ACADEMIC PRESS, INC.
San Diego, California 92101

United Kingdom Edition published by
ACADEMIC PRESS LIMITED
24-28 Oval Road, London NW1 7DX

Library of Congress Cataloging in Publication Data

Hirsch, Morris, Date
Differential equations, dynamical systems, and linear algebra.

(Pure and applied mathematics; a series of monographs and textbooks, v.)
1. Differential equations. 2. Algebras, Linear. I. Smale, Stephen, Date joint author. II. Title. III. Series.
QA3.P8 [QA372] 510'.8s [515'.35] 73-18951
ISBN 0-12-349550-4

AMS (MOS) 1970 Subject Classifications: 15-01, 34-01

PRINTED IN THE UNITED STATES OF AMERICA
90 91 92 93 94 9 8

Contents

Preface

This book is about dynamical aspects of ordinary differential equations and the relations between dynamical systems and certain fields outside pure mathematics. A prominent role is played by the structure theory of linear operators on finite-dimensional vector spaces; we have included a self-contained treatment of that subject.

The background material needed to understand this book is differential calculus of several variables. For example, Serge Lang's *Calculus of Several Variables*, up to the chapter on integration, contains more than is needed to understand much of our text. On the other hand, after Chapter 7 we do use several results from elementary analysis such as theorems on uniform convergence; these are stated but not proved. This mathematics is contained in Lang's *Analysis I*, for instance. Our treatment of linear algebra is systematic and self-contained, although the most elementary parts have the character of a review; in any case, Lang's *Calculus of Several Variables* develops this elementary linear algebra at a leisurely pace.

While this book can be used as early as the sophomore year by students with a strong first year of calculus, it is oriented mainly toward upper division mathematics and science students. It can also be used for a graduate course, especially if the later chapters are emphasized.

It has been said that the subject of ordinary differential equations is a collection of tricks and hints for finding solutions, and that it is important because it can solve problems in physics, engineering, etc. Our view is that the subject can be developed with considerable unity and coherence; we have attempted such a development with this book. The importance of ordinary differential equations *vis à vis* other areas of science lies in its power to motivate, unify, and give force to those areas. Our four chapters on "applications" have been written to do exactly this, and not merely to provide examples. Moreover, an understanding of the ways that differential equations relates to other subjects is a primary source of insight and inspiration for the student and working mathematician alike.

Our goal in this book is to develop nonlinear ordinary differential equations in open subsets of real Cartesian space, $\mathbf{R}^n$, in such a way that the extension to manifolds is simple and natural. We treat chiefly autonomous systems, emphasizing qualitative behavior of solution curves. The related themes of stability and physical significance pervade much of the material. Many topics have been omitted, such as Laplace transforms, series solutions, Sturm theory, and special functions.

The level of rigor is high, and almost everything is proved. More important, however, is that *ad hoc* methods have been rejected. We have tried to develop

proofs that add insight to the theorems and that are important methods in their own right.

We have avoided the introduction of manifolds in order to make the book more widely readable; but the main ideas can easily be transferred to dynamical systems on manifolds.

The first six chapters, especially Chapters 3–6, give a rather intensive and complete study of linear differential equations with constant coefficients. This subject matter can almost be identified with linear algebra; hence those chapters constitute a short course in linear algebra as well. The algebraic emphasis is on eigenvectors and how to find them. We go far beyond this, however, to the "semisimple + nilpotent" decomposition of an arbitrary operator, and then on to the Jordan form and its real analogue. Those proofs that are far removed from our use of the theorems are relegated to appendices. While complex spaces are used freely, our primary concern is to obtain results for real spaces. This point of view, so important for differential equations, is not commonly found in textbooks on linear algebra or on differential equations.

Our approach to linear algebra is a fairly intrinsic one; we avoid coordinates where feasible, while not hesitating to use them as a tool for computations or proofs. On the other hand, instead of developing abstract vector spaces, we work with linear subspaces of $\mathbf{R}^n$ or $\mathbf{C}^n$, a small concession which perhaps makes the abstraction more digestible.

Using our algebraic theory, we give explicit methods of writing down solutions to arbitrary constant coefficient linear differential equations. Examples are included. In particular, the $S + N$ decomposition is used to compute the exponential of an arbitrary square matrix.

Chapter 2 is independent from the others and includes an elementary account of the Keplerian planetary orbits.

The fundamental theorems on existence, uniqueness, and continuity of solutions of ordinary differential equations are developed in Chapters 8 and 16. Chapter 8 is restricted to the autonomous case, in line with our basic orientation toward dynamical systems.

Chapters 10, 12, and 14 are devoted to systematic introductions to mathematical models of electrical circuits, population theory, and classical mechanics, respectively. The Brayton–Moser circuit theory is presented as a special case of the more general theory recently developed on manifolds. The Volterra–Lotka equations of competing species are analyzed, along with some generalizations. In mechanics we develop the Hamiltonian formalism for conservative systems whose configuration space is an open subset of a vector space.

The remaining five chapters contain a substantial introduction to the phase portrait analysis of nonlinear autonomous systems. They include a discussion of "generic" properties of linear flows, Liapunov and structural stability, Poincaré–Bendixson theory, periodic attractors, and perturbations. We conclude with an Afterword which points the way toward manifolds.

The following remarks should help the reader decide on which chapters to read and in what order.

Chapters 1 and 2 are elementary, but they present many ideas that recur throughout the book.

Chapters 3–7 form a sequence that develops linear theory rather thoroughly. Chapters 3, 4, and 5 make a good introduction to linear operators and linear differential equations. The canonical form theory of Chapter 6 is the basis of the stability results proved in Chapters 7, 9, and 13; however, this heavy algebra might be postponed at a first exposure to this material and the results taken on faith.

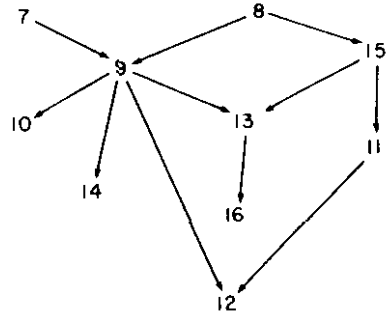
The existence, uniqueness, and continuity of solutions, proved in Chapter 8, are used (often implicitly) throughout the rest of the book. Depending on the reader's taste, proofs could be omitted.

A reader interested in the nonlinear material, who has some background in linear theory, might start with the stability theory of Chapter 9. Chapters 12 (ecology), 13 (periodic attractors), and 16 (perturbations) depend strongly on Chapter 9, while the section on dual vector spaces and gradients will make Chapters 10 (electrical circuits) and 14 (mechanics) easier to understand.

Chapter 12 also depends on Chapter 11 (Poincaré–Bendixson); and the material in Section 2 of Chapter 11 on local sections is used again in Chapters 13 and 16.

Chapter 15 (nonautonomous equations) is a continuation of Chapter 8 and is used in Chapters 11, 13, and 16; however it can be omitted at a first reading.

The logical dependence of the later chapters is summarized in the following chart:



The book owes much to many people. We only mention four of them here. Ikuko Workman and Ruth Suzuki did an excellent job of typing the manuscript. Dick Palais made a number of useful comments. Special thanks are due to Jacob Palis, who read the manuscript thoroughly, found many minor errors, and suggested several substantial improvements. Professor Hirsch is grateful to the Miller Institute for its support during part of the writing of the book.

Chapter 1

First Examples

The purpose of this short chapter is to develop some simple examples of differential equations. This development motivates the linear algebra treated subsequently and moreover gives in an elementary context some of the basic ideas of ordinary differential equations. Later these ideas will be put into a more systematic exposition. In particular, the examples themselves are special cases of the class of differential equations considered in Chapter 3. We regard this chapter as important since some of the most basic ideas of differential equations are seen in simple form.

§1. The Simplest Examples

The differential equation

$$\frac{dx}{dt} = ax \tag{1}$$

is the simplest differential equation. It is also one of the most important. First, what does it mean? Here $x = x(t)$ is an unknown real-valued function of a real variable t and dx/dt is its derivative (we will also use x' or $x'(t)$ for this derivative). The equation tells us that for every value of t the equality

$$x'(t) = ax(t)$$

is true. Here a denotes a constant.

The solutions to (1) are obtained from calculus: if K is any constant (real number), the function $f(t) = Ke^{at}$ is a solution since

$$f'(t) = aKe^{at} = af(t).$$

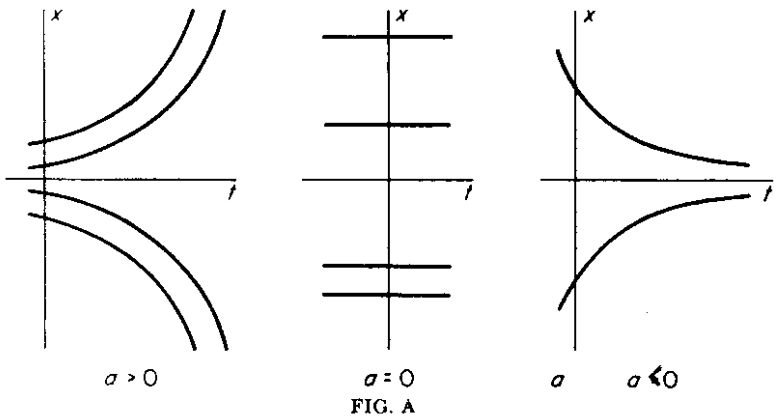


FIG. A

Moreover, *there are no other solutions*. To see this, let $u(t)$ be any solution and compute the derivative of $u(t)e^{-at}$:

$$\frac{d}{dt}\left(u(t)e^{-at}\right) = u'(t)e^{-at} + u(t)(-ae^{-at})$$

$$= au(t)e^{-at} - au(t)e^{-at} = 0.$$

Therefore $u(t)e^{-at}$ is a constant K, so $u(t) = Ke^{at}$. This proves our assertion.

The constant K appearing in the solution is completely determined if the value u_0 of the solution at a single point t_0 is specified. Suppose that a function $x(t)$ satisfying (1) is required such that $x(t_0) = u_0$, then K must satisfy $Ke^{at_0} = u_0$. Thus equation (1) has a unique solution satisfying a specified *initial condition* $x(t_0) = u_0$. For simplicity, we often take $t_0 = 0$; then $K = u_0$. There is no loss of generality in taking $t_0 = 0$, for if $u(t)$ is a solution with $u(0) = u_0$, then the function $v(t) = u(t - t_0)$ is a solution with $v(t_0) = u_0$.

It is common to restate (1) in the form of an *initial value problem*:

$$x' = ax, \qquad x(0) = K. \tag{2}$$

A solution $x(t)$ to (2) must not only satisfy the first condition (1), but must also take on the prescribed initial value K at $t = 0$. We have proved that the initial value problem (2) has a unique solution.

The constant a in the equation $x' = ax$ can be considered as a parameter. If a changes, the equation changes and so do the solutions. Can we describe qualitatively the way the solutions change?

The *sign* of a is crucial here:

if $a > 0$, $\lim_{t\to\infty} Ke^{at}$ equals ∞ when $K > 0$, and equals $-\infty$ when $K < 0$;
if $a = 0$, $Ke^{at} =$ constant;
if $a < 0$, $\lim_{t\to\infty} Ke^{at} = 0$.

The qualitative behavior of solutions is vividly illustrated by sketching the graphs of solutions (Fig. A). These graphs follow a typical practice in this book. The figures are meant to illustrate qualitative features and may be imprecise in quantitative detail.

The equation $x' = ax$ is *stable* in a certain sense if $a \neq 0$. More precisely, if a is replaced by another constant b sufficiently close to a, the qualitative behavior of the solutions does not change. If, for example, $|b - a| < |a|$, then b has the same sign as a. But if $a = 0$, the slightest change in a leads to a radical change in the behavior of solutions. We may also say that $a = 0$ is a *bifurcation point* in the one-parameter family of equations $x' = ax$, a in $\mathbf{R}$.

Consider next a *system* of two differential equations in two unknown functions:

$$x_1' = a_1x_1,$$
$$x_2' = a_2x_2. \tag{3}$$

This is a very simple system; however, many more-complicated systems of two equations can be reduced to this form as we shall see a little later.

Since there is no relation specified between the two unknown functions $x_1(t)$, $x_2(t)$, they are "uncoupled"; we can immediately write down all solutions (as for (1)):

$$x_1(t) = K_1\exp(a_1t), \qquad K_1 = \text{constant},$$
$$x_2(t) = K_2\exp(a_2t), \qquad K_2 = \text{constant}.$$

Here K_1 and K_2 are determined if initial conditions $x_1(t_0) = u_1$, $x_2(t_0) = u_2$ are specified. (We sometimes write $\exp a$ for e^a.)

Let us consider equation (2) from a more geometric point of view. We consider two functions $x_1(t)$, $x_2(t)$ as specifying an unknown *curve* $x(t) = (x_1(t), x_2(t))$ in the (x_1, x_2) plane $\mathbf{R}^2$. That is to say, x is a map from the real numbers $\mathbf{R}$ into $\mathbf{R}^2$, x: $\mathbf{R} \to \mathbf{R}^2$. The right-hand side of (3) expresses the *tangent vector* $x'(t) = (x_1'(t), x_2'(t))$ to the curve. Using vector notation,

$$x' = Ax, \tag{3'}$$

where Ax denotes the vector (a_1x_1, a_2x_2), which one should think of as being based at x.

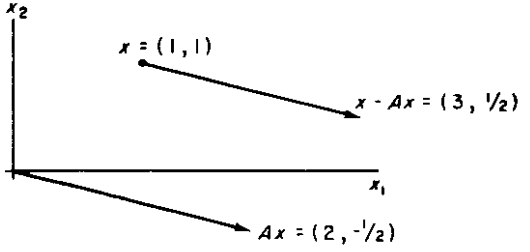


FIG. B

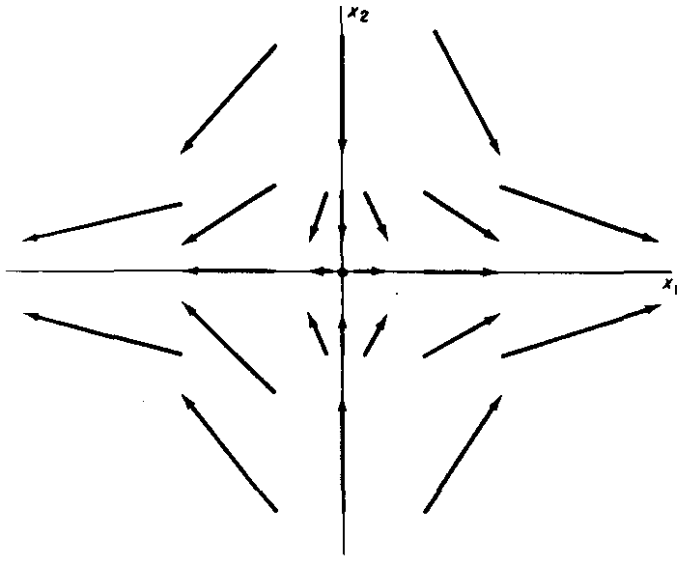


FIG. C. $Ax = (2x_1, -\frac{1}{2}x_2)$.

Initial conditions are of the form $x(t_0) = u$ where $u = (u_1, u_2)$ is a given point of $\mathbf{R}^2$. Geometrically, this means that when $t = t_0$ the curve is required to pass through the given point u.

The map (that is, function) $A: \mathbf{R}^2 \to \mathbf{R}^2$ (or $x \to Ax$) can be considered a *vector field* on $\mathbf{R}^2$. This means that to each point x in the plane we assign the vector Ax. For purposes of visualization, we picture Ax as a vector "based at x"; that is, we assign to x the directed line segment from x to $x + Ax$. For example, if $a_1 = 2$, $a_2 = -\frac{1}{2}$, and $x = (1, 1)$, then at $(1, 1)$ we picture an arrow pointing from $(1, 1)$ to $(1, 1) + (2, -\frac{1}{2}) = (3, \frac{1}{2})$ (Fig. B). Thus if $Ax = (2x_1, -\frac{1}{2}x_2)$, we attach to each point x in the plane an arrow with tail at x and head at $x + Ax$ and obtain the picture in Fig. C.

Solving the differential equation (3) or (3′) with initial conditions (u_1, u_2) at $t = 0$ means finding in the plane a curve $x(t)$ that satisfies (3′) and passes through the point $u = (u_1, u_2)$ when $t = 0$. A few solution curves are sketched in Fig. D.

The trivial solution $(x_1(t), x_2(t)) = (0, 0)$ is also considered a "curve."

The family of all solution curves as subsets of $\mathbf{R}^2$ is called the "phase portrait" of equation (3) (or (3′)).

The one-dimensional equation $x' = ax$ can also be interpreted geometrically: the phase portrait is as in Fig. E, which should be compared with Fig. A. It is clearer to picture the graphs of (1) and the solution curves for (3) since two-dimensional pictures are better than either one- or three-dimensional pictures. The *graphs* of

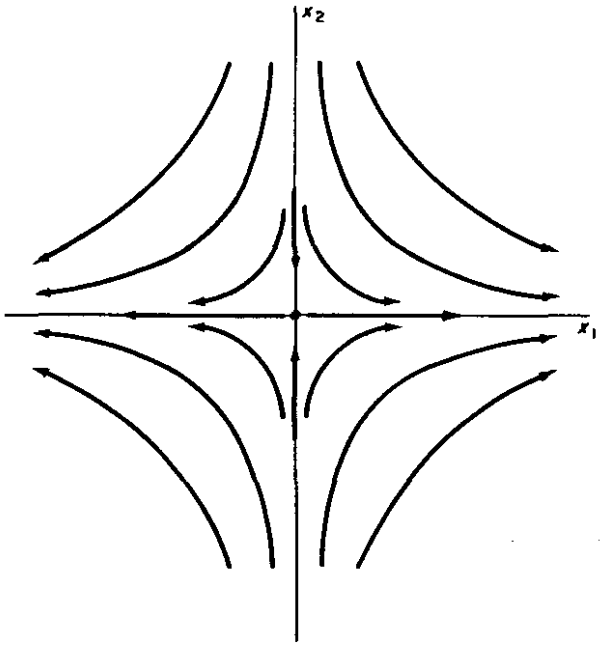


FIG. D. Some solution curves to $x' = Ax$, $A = \begin{bmatrix} 2 & 0 \\ 0 & -\frac{1}{2} \end{bmatrix}$.

solutions to (3) require a three-dimensional picture which the reader is invited to sketch!

Let us consider equation (3) as a *dynamical system*. This means that the independent variable t is interpreted as *time* and the solution curve $x(t)$ could be thought of, for example, as the path of a particle moving in the plane $\mathbf{R}^2$. We can imagine a particle placed at any point $u = (u_1, u_2)$ in $\mathbf{R}^2$ at time $t = 0$. As time proceeds the particle moves along the solution curve $x(t)$ that satisfies the initial condition $x(0) = u$. At any later time $t > 0$ the particle will be in another position $x(t)$. And at an earlier time $t < 0$, the particle was at a position $x(t)$. To indicate the dependence of the position on t and u we denote it by $\phi_t(u)$. Thus

$$\phi_t(u) = (u_1 \exp(a_1 t), u_2 \exp(a_2 t)).$$

We can imagine particles placed at each point of the plane and all moving simultaneously (for example, dust particles under a steady wind). The solution curves are spoken of as trajectories or orbits in this context. For each fixed t in $\mathbf{R}$, we have a transformation assigning to each point u in the plane another point $\phi_t(u)$. This transformation denoted by $\phi_t: \mathbf{R}^2 \to \mathbf{R}^2$ is clearly a *linear* transformation, that is,

FIG. E

$\phi_t(u + v) = \phi_t(u) + \phi_t(v)$ and $\phi_t(\lambda u) = \lambda\phi_t(u)$, for all vectors u, v, and all real numbers λ.

As time proceeds, every point of the plane moves simultaneously along the trajectory passing through it. In this way the collection of maps $\phi_t: \mathbf{R}^2 \to \mathbf{R}^2$, $t \in \mathbf{R}$, is a one-parameter family of transformations. This family is called the *flow* or *dynamical system* or $\mathbf{R}^2$ determined by the vector field $x \to Ax$, which in turn is equivalent to the system (3).

The dynamical system on the real line $\mathbf{R}$ corresponding to equation (1) is particularly easy to describe: if $a < 0$, all points move toward 0 as time goes to ∞; if $a > 0$, all points except 0 move away from 0 toward $\pm\infty$; if $a = 0$, all points stand still.

We have started from a differential equation and have obtained the dynamical system ϕ_t. This process is established through the fundamental theorem of ordinary differential equations as we shall see in Chapter 8.

Later we shall also reverse this process: starting from a dynamical system ϕ_t, a differential equation will be obtained (simply by differentiating $\phi_t(u)$ with respect to t).

It is seldom that differential equations are given in the simple uncoupled form (3). Consider, for example, the system:

$$x_1' = 5x_1 + 3x_2, \tag{4}$$
$$x_2' = -6x_1 - 4x_2$$

or in vector notation

$$x' = (5x_1 + 3x_2, -6x_1 - 4x_2) \equiv Bx. \tag{4'}$$

Our approach is to find a linear *change of coordinates* that will transform equation (4) into uncoupled or diagonal form. It turns out that new coordinates (y_1, y_2) do the job where

$$y_1 = 2x_1 + x_2,$$
$$y_2 = x_1 + x_2.$$

(In Chapter 3 we explain how the new coordinates were found.)

Solving for x in terms of y, we have

$$x_1 = y_1 - y_2,$$
$$x_2 = -y_1 + 2y_2.$$

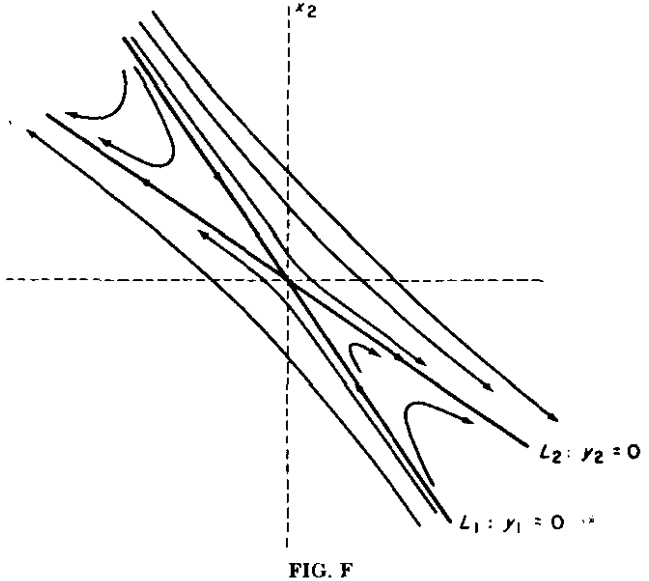


FIG. F

To find y_1', y_2' differentiate the equations defining y_1, y_2 to obtain

$$y_1' = 2x_1' + x_2',$$
$$y_2' = x_1' + x_2'.$$

By substitution

$$y_1' = 2(5x_1 + 3x_2) + (-6x_1 - 4x_2) = 4x_1 + 2x_2,$$
$$y_2' = (5x_1 + 3x_2) + (-6x_1 - 4x_2) = -x_1 - x_2.$$

Another substitution yields

$$y_1' = 4(y_1 - y_2) + 2(-y_1 + 2y_2),$$
$$y_2' = -(y_1 - y_2) - (-y_1 + 2y_2),$$

or

$$y_1' = 2y_1, \tag{5}$$
$$y_2' = -y_2.$$

The last equations are in *diagonal form* and we have already solved this class of systems. The solution $(y_1(t), y_2(t))$ such that $(y_1(0), y_2(0)) = (v_1, v_2)$ is

$$y_1(t) = e^{2t}v_1,$$
$$y_2(t) = e^{-t}v_2.$$

The phase portrait of this system (5) is given evidently in Fig. D. We can find the phase portrait of the original system (4) by simply plotting the new coordinate axes $y_1 = 0$, $y_2 = 0$ in the (x_1, x_2) plane and sketching the trajectories $y(t)$ in these coordinates. Thus $y_1 = 0$ is the line L_1: $x_2 = -2x_1$ and $y_2 = 0$ is the line L_2: $x_2 = -x_1$.

Thus we have the phase portrait of (4) as in Fig. F, which should be compared with Fig. D.

Formulas for the solution to (4) can be obtained by substitution as follows. Let (u_1, u_2) be the initial values $(x_1(0), x_2(0))$ of a solution $(x_1(t), x_2(t))$ to (4). Corresponding to (u_1, u_2) is the initial value (v_1, v_2) of a solution $(y_1(t), y_2(t))$ to (5) where

$$v_1 = 2u_1 + u_2,$$

$$v_2 = u_1 + u_2.$$

Thus

$$y_1(t) = e^{3t}(2u_1 + u_2),$$

$$y_2(t) = e^{-t}(u_1 + u_2)$$

and

$$x_1(t) = e^{3t}(2u_1 + u_2) - e^{-t}(u_1 + u_2),$$

$$x_2(t) = -e^{3t}(2u_1 + u_2) + 2e^{-t}(u_1 + u_2).$$

If we compare these formulas to Fig. F, we see that the diagram instantly gives us the qualitative picture of the solutions, while the formulas convey little geometric information. In fact, for many purposes, it is better to forget the original equation (4) and the corresponding solutions and work entirely with the "diagonalized" equations (5), their solution and phase portrait.

PROBLEMS

1. Each of the "matrices"

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = [a_{ij}]$$

given below defines a vector field on $\mathbf{R}^2$, assigning to $x = (x_1, x_2) \in \mathbf{R}^2$ the vector $Ax = (a_{11}x_1 + a_{12}x_2, a_{21}x_1 + a_{22}x_2)$ based at x. For each matrix, draw enough of the vectors until you get a feeling for what the vector field looks like. Then sketch the phase portrait of the corresponding differential equation $x' = Ax$, guessing where necessary.

(a) $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ (b) $\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 2 \end{bmatrix}$ (c) $\begin{bmatrix} -2 & 0 \\ 0 & 2 \end{bmatrix}$

(d) $\begin{bmatrix} \frac{1}{2} & -2 \\ 2 & 0 \end{bmatrix}$ (e) $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ (f) $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

(g) $\begin{bmatrix} -1 & -1 \\ 1 & -1 \end{bmatrix}$ (h) $\begin{bmatrix} \frac{1}{2} & 1 \\ 0 & \frac{1}{2} \end{bmatrix}$ (i) $\begin{bmatrix} 0 & 0 \\ -3 & 0 \end{bmatrix}$

2. Consider the one-parameter family of differential equations

$$x_1' = 2x_1,$$

$$x_2' = ax_2; \qquad -\infty < a < \infty.$$

(a) Find all solutions $(x_1(t), x_2(t))$.
(b) Sketch the phase portrait for a equal to $-1, 0, 1, 2, 3$. Make some guesses about the stability of the phase portraits.

§2. Linear Systems with Constant Coefficients

This section is devoted to generalizing and abstracting the previous examples. The general problem is stated, but solutions are postponed to Chapter 3.

Consider the following set or "system" of n differential equations:

$$(1) \qquad \begin{aligned} \frac{dx_1}{dt} &= a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n, \\ \frac{dx_2}{dt} &= a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n, \\ &\vdots \\ \frac{dx_n}{dt} &= a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n. \end{aligned}$$

Here the a_{ij} $(i = 1, \ldots, n; j = 1, \ldots, n)$ are n^2 constants (real numbers), while each x_i denotes an unknown real-valued function of a real variable t. Thus (4) of Section 1 is an example of the system (1) with $n = 2$, $a_{11} = 5$, $a_{12} = 3$, $a_{21} = -6$, $a_{22} = -4$.

At this point we are not trying to solve (1); rather, we want to place it in a geometrical and algebraic setting in order to understand better what a solution means.

At the most primitive level, a solution of (1) is a set of n differentiable real-valued functions $x_i(t)$ that make (1) true.

In order to reach a more conceptual understanding of (1) we introduce *real n-dimensional Cartesian space* $\mathbf{R}^n$. This is simply the set of all n-tuples of real numbers. An element of $\mathbf{R}^n$ is a "point" $x = (x_1, \ldots, x_n)$; the number x_i is the ith *coordinate* of the point x. Points x, y in $\mathbf{R}^n$ are added coordinatewise:

$$x + y = (x_1, \ldots, x_n) + (y_1, \ldots, y_n) = (x_1 + y_1, \ldots, x_n + y_n).$$

Also, if λ is a real number we define the *product* of λ and x to be

$$\lambda x = (\lambda x_1, \ldots, \lambda x_n).$$

The *distance* between points x, y in $\mathbf{R}^n$ is defined to be

$$|x - y| = [(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2]^{1/2}.$$

The *length* of x is

$$|x| = (x_1^2 + \cdots + x_n^2)^{1/2}.$$

A vector *based at* $x \in \mathbf{R}^n$ is an ordered pair of points x, y in $\mathbf{R}^n$, denoted by $\overrightarrow{xy}$. We think of this as an arrow or line segment directed from x to y. We say $\overrightarrow{xy}$ is *based* at x.

A vector $\overrightarrow{0x}$ based at the *origin*

$$0 = (0, \ldots, 0) \in \mathbf{R}^n$$

is identified with the point $x \in \mathbf{R}^n$.

To a vector $\overrightarrow{xy}$ based at x is associated the vector $y - x$ based at the origin 0. We call the vectors $\overrightarrow{xy}$ and $y - x$ *translates* of each other.

From now on a vector based at 0 is called simply a vector. Thus an element of $\mathbf{R}^n$ can be considered either as an n-tuple of real numbers or as an arrow issuing from the origin.

It is only for purposes of visualization that we consider vectors based at points other than 0. For computations, all vectors are based at 0 since such vectors can be added and multiplied by real numbers.

We return to the system of differential equations (1). A candidate for a solution is a *curve* in $\mathbf{R}^n$:

$$(*) \qquad x(t) = (x_1(t), \ldots, x_n(t)).$$

By this we mean a map

$$x: \mathbf{R} \to \mathbf{R}^n.$$

Such a map is described in terms of coordinates by (*). If each function $x_i(t)$ is differentiable, then the map x is called differentiable; its derivative is defined to be

$$\frac{dx}{dt} = x'(t) = (x_1'(t), \ldots, x_n'(t)).$$

Thus the derivative, as a function of t, is again a map from $\mathbf{R}$ to $\mathbf{R}^n$.

The derivative can also be expressed in the form

$$x'(t) = \lim_{h \to 0} \frac{1}{h}(x(t + h) - x(t)).$$

It has a natural geometric interpretation as the vector $v(t)$ based at $x(t)$, which is a translate of $x'(t)$. This vector is called the *tangent vector* to the curve at t (or at $x(t)$).

If we imagine t as denoting time, then the length $|x'(t)|$ of the tangent vector is interpreted physically as the speed of a particle describing the curve $x(t)$.

To write (1) in an abbreviated form we call the doubly indexed set of numbers a_{ij} an $n \times n$ *matrix* A, denoted thus:

$$A = [a_{ji}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ & \vdots & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}.$$

Next, for each $x \in \mathbf{R}^n$ we define a vector $Ax \in \mathbf{R}^n$ whose ith coordinate is

$$a_{i1}x_1 + \cdots + a_{in}x_n;$$

note that this is the ith row in the right-hand side of (1). In this way the matrix A is interpreted as a map

$$A: \mathbf{R}^n \to \mathbf{R}^n$$

which to x assigns Ax.

With this notation (1) is rewritten

$$(2) \qquad x' = Ax.$$

Thus the system (1) can be considered as a single "vector differential equation" (2). (The word *equation* is classically reserved for the case of just one variable; we shall call (2) both a system and an equation.)

We think of the map $A: \mathbf{R}^n \to \mathbf{R}^n$ as a *vector field* on $\mathbf{R}^n$: to each point $x \in \mathbf{R}^n$ it assigns the vector based at x which is a translate of Ax. Then a solution of (2) is a curve $x: \mathbf{R} \to \mathbf{R}^n$ whose tangent vector at any given t is the vector $Ax(t)$ (translated to $x(t)$). See Fig. D of Section 1.

In Chapters 3 and 4 we shall give methods of explicitly solving (2), or equivalently (1). In subsequent chapters it will be shown that in fact (2) has a unique solution $x(t)$ satisfying any given initial condition $x(0) = u_0 \in \mathbf{R}^n$. This is the fundamental theorem of linear differential equations with constant coefficients; in Section 1 this was proved for the special case $n = 1$.

PROBLEMS

1. For each of the following matrices A sketch the vector field $x \to Ax$ in $\mathbf{R}^2$. (Missing matrix entries are 0.)

(a) $\begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & & \\ & -2 & \\ & & 0 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & & \\ & -2 & \\ & & 2 \end{bmatrix}$

(d) $\begin{bmatrix} 0 & & \\ & -1 & \\ & & 0 \end{bmatrix}$ (e) $\begin{bmatrix} 0 & -1 & \\ 1 & 0 & \\ & & -\frac{1}{2} \end{bmatrix}$ (f) $\begin{bmatrix} -1 & & \\ 1 & 1 & \\ & 1 & 1 \end{bmatrix}$

2. For A as in (a), (b), (c) of Problem 1, solve the initial value problem

$$x' = Ax, \qquad x(0) = (k_1, k_2, k_3).$$

3. Let A be as in (e), Problem 1. Find constants a, b, c such that the curve $t \to (a \cos t, b \sin t, ce^{-t/2})$ is a solution to $x' = Ax$ with $x(0) = (1, 0, 3)$.

4. Find two different matrices A, B such that the curve

$$x(t) = (e^t, 2e^{2t}, 4e^{2t})$$

satisfies both the differential equations

$$x' = Ax \qquad \text{and} \qquad x' = Bx.$$

5. Let $A = [a_{ij}]$ be an $n \times n$ diagonal matrix, that is, $a_{ij} = 0$ if $i \neq j$. Show that the differential equation

$$x' = Ax$$

has a unique solution for every initial condition.

6. Let A be an $n \times n$ diagonal matrix. Find conditions on A guaranteeing that

$$\lim_{t \to \infty} x(t) = 0$$

for all solutions to $x' = Ax$.

7. Let $A = [a_{ij}]$ be an $n \times n$ matrix. Denote by $-A$ the matrix $[-a_{ij}]$.
   (a) What is the relation between the vector fields $x \to Ax$ and $x \to (-A)x$?
   (b) What is the geometric relation between solution curves of $x' = Ax$ and of $x' = -Ax$?

8. (a) Let $u(t)$, $v(t)$ be solutions to $x' = Ax$. Show that the curve $w(t) = \alpha u(t) + \beta v(t)$ is a solution for all real numbers α, β.

(b) Let $A = \begin{bmatrix} 1 & \\ & -2 \end{bmatrix}$. Find solutions $u(t)$, $v(t)$ to $x' = Ax$ such that every solution can be expressed in the form $\alpha u(t) + \beta v(t)$ for suitable constants α, β.

Notes

The background needed for a reader of Chapter 1 is a good first year of college calculus. One good source is S. Lang's *Second Course in Calculus* [12, Chapters I, II, and IX]. In this reference the material on derivatives, curves, and vectors in $\mathbf{R}^n$ and matrices is discussed much more thoroughly than in our Section 2.

Chapter 2

Newton's Equation and Kepler's Law

We develop in this chapter the earliest important examples of differential equations, which in fact are connected with the origins of calculus. These equations were used by Newton to derive and unify the three laws of Kepler. These laws were found from the earlier astronomical observations of Tycho Brahe. Here we give a brief derivation of two of Kepler's laws, while at the same time setting forth some general ideas about differential equations.

The equations of Newton, our starting point, have retained importance throughout the history of modern physics and lie at the root of that part of physics called classical mechanics.

The first chapter of this book dealt with linear equations, but Newton's equations are nonlinear in general. In later chapters we shall pursue the subject of nonlinear differential equations somewhat systematically. The examples here provide us with concrete examples of historical and scientific importance. Furthermore, the case we consider most thoroughly here, that of a particle moving in a central force gravitational field, is simple enough so that the differential equations can be solved explicitly using exact, classical methods (just calculus!). This is due to the existence of certain invariant functions called *integrals* (sometimes called "first integrals"; we do not mean the integrals of elementary calculus). Physically, an integral is a conservation law; in the case of Newtonian mechanics the two integrals we find correspond to conservation of energy and angular momentum. Mathematically an integral reduces the number of dimensions.

We shall be working with a particle moving in a *field of force* F. Mathematically F is a *vector field* on the (configuration) space of the particle, which in our case we suppose to be Cartesian three space $\mathbf{R}^3$. Thus F is a map $F\colon \mathbf{R}^3 \to \mathbf{R}^3$ that assigns to a point x in $\mathbf{R}^3$ another point $F(x)$ in $\mathbf{R}^3$. From the mathematical point of view, $F(x)$ is thought of as a vector based at x. From the physical point of view, $F(x)$ is the force exerted on a particle located at x.

The example of a force field we shall be most concerned with is the gravitational field of the sun: $F(x)$ is the force on a particle located at x attracting it to the sun.

We shall go into details of this field in Section 6. Other important examples of force fields are derived from electrical forces, magnetic forces, and so on.

The connection between the physical concept of force field and the mathematical concept of differential equation is *Newton's second law*: $F = ma$. This law asserts that a particle in a force field moves in such a way that the force vector at the location of the particle, at any instant, equals the acceleration vector of the particle times the mass m. If $x(t)$ denotes the position vector of the particle at time t, where $x\colon \mathbf{R} \to \mathbf{R}^3$ is a sufficiently differentiable curve, then the acceleration vector is the second derivative of $x(t)$ with respect to time

$$a(t) = \ddot{x}(t).$$

(We follow tradition and use dots for time derivatives in this chapter.) Newton's second law states

$$F(x(t)) = m\ddot{x}(t).$$

Thus we obtain a second order differential equation:

$$\ddot{x} = \frac{1}{m} F(x).$$

In Newtonian physics it is assumed that m is a positive constant. Newton's law of gravitation is used to derive the exact form of the function $F(x)$. While these equations are the main goal of this chapter, we first discuss simple harmonic motion and then basic background material.

§1. Harmonic Oscillators

We consider a particle of mass m moving in one dimension, its position at time t given by a function $t \to x(t)$, $x\colon \mathbf{R} \to \mathbf{R}$. Suppose the force on the particle at a point $x \in \mathbf{R}$ is given by $-mp^2x$, where p is some real constant. Then according to the laws of physics (compare Section 3) the motion of the particle satisfies

$$\ddot{x} + p^2x = 0. \tag{1}$$

This model is called the harmonic oscillator and (1) is the equation of the harmonic oscillator (in one dimension).

An example of the harmonic oscillator is the simple pendulum moving in a plane, when one makes an approximation of $\sin x$ by x (compare Chapter 9). Another example is the case where the force on the particle is caused by a spring.

It is easy to check that for any constants A, B, the function

$$x(t) = A \cos pt + B \sin pt \tag{2}$$

is a solution of (1), with initial conditions $x(0) = A$, $\dot{x}(0) = pB$. In fact, as is proved

often in calculus courses, (2) is the only solution of (1) satisfying these initial conditions. Later we will show in a systematic way that these facts are true.

Using basic trigonometric identities, (2) may be rewritten in the form

$$x(t) = a\cos(pt + t_0), \tag{3}$$

where $a = (A^2 + B^2)^{1/2}$ is called the amplitude, and $\cos t_0 = A(A^2 + B^2)^{-1/2}$.

In Section 6 we will consider equation (1) where a constant term is added (representing a constant disturbing force):

$$\ddot{x} + p^2x = K. \tag{4}$$

Then, similarly to (1), every solution of (4) has the form

$$x(t) = a\cos(pt + t_0) + \frac{K}{p^2}. \tag{5}$$

The two-dimensional version of the harmonic oscillator concerns a map $x: \mathbf{R} \to \mathbf{R}^2$ and a force $F(x) = -mkx$ (where now, of course, $x = (x_1, x_2) \in \mathbf{R}^2$). Equation (1) now has the same form

$$\ddot{x} + k^2x = 0 \tag{1'}$$

with solutions given by

$$\begin{aligned} x_1(t) &= A\cos kt + B\sin kt, \\ x_2(t) &= C\cos kt + D\sin kt. \end{aligned} \tag{2'}$$

See Problem 1.

Planar motion will be considered more generally and in more detail in later sections. But first we go over some mathematical preliminaries.

§2. Some Calculus Background

A path of a moving particle in $\mathbf{R}^n$ (usually $n \leq 3$) is given by a map $f: I \to \mathbf{R}^n$ where I might be the set $\mathbf{R}$ of all real numbers or an interval (a, b) of all real numbers strictly between a and b. The derivative of f (provided f is differentiable at each point of I) defines a map $f': I \to \mathbf{R}^n$. The map f is called C^1, or *continuously differentiable*, if f' is continuous (that is to say, the corresponding coordinate functions $f_i'(t)$ are continuous, $i = 1, \ldots, n$). If $f': I \to \mathbf{R}^n$ is itself C^1, then f is said to be C^2. Inductively, in this way, one defines a map $f: I \to \mathbf{R}^n$ to be C^r, where $r = 3, 4, 5$, and so on.

The *inner product*, or "dot product," of two vectors, x, y in $\mathbf{R}^n$ is denoted by $\langle x, y\rangle$ and defined by

$$\langle x, y\rangle = \sum_{i=1}^{n} x_iy_i.$$

Thus $\langle x, x\rangle = |x|^2$. If $x, y: I \to \mathbf{R}^n$ are C^1 functions, then a version of the Leibnis product rule for derivatives is

$$\langle x, y\rangle' = \langle x', y\rangle + \langle x, y'\rangle,$$

as can be easily checked using coordinate functions.

We will have occasion to consider functions $f: \mathbf{R}^n \to \mathbf{R}$ (which, for example, could be given by temperature or density). Such a map f is called C^1 if the map $\mathbf{R}^n \to \mathbf{R}$ given by each partial derivative $x \to \partial f/\partial x_i(x)$ is defined and continuous (in Chapter 5 we discuss continuity in more detail). In this case the gradient of f, called *grad* f, is the map $\mathbf{R}^n \to \mathbf{R}^n$ that sends x into $(\partial f/\partial x_1(x), \ldots, \partial f/\partial x_n(x))$. Grad f is an example of a vector field on $\mathbf{R}^n$. (In Chapter 1 we considered only linear vector fields, but grad f may be more general.)

Next, consider the composition of two C^1 maps as follows:

$$I \xrightarrow{f} \mathbf{R}^n \xrightarrow{g} \mathbf{R}.$$

The chain rule can be expressed in this context as

$$\frac{d}{dt}g(f(t)) = \langle \text{grad } g(f(t)), f'(t)\rangle;$$

using the definitions of gradient and inner product, the reader can prove that this is equivalent to

$$\sum_{i=1}^{n} \frac{\partial g}{\partial x_i}(f(t))\,\frac{df_i}{dt}(t).$$

§3. Conservative Force Fields

A vector field $F: \mathbf{R}^3 \to \mathbf{R}^3$ is called a force field if the vector $F(x)$ assigned to the point x is interpreted as a force acting on a particle placed at x.

Many force fields appearing in physics arise in the following way. There is a C^1 function

$$V: \mathbf{R}^3 \to \mathbf{R}$$

such that

$$\begin{aligned} F(x) &= -\left(\frac{\partial V}{\partial x_1}(x), \frac{\partial V}{\partial x_2}(x), \frac{\partial V}{\partial x_3}(x)\right) \\ &= -\text{grad } V(x). \end{aligned}$$

(The negative sign is traditional.) Such a force field is called *conservative*. The function V is called the *potential energy* function. (More properly V should be called a potential energy since adding a constant to it does not change the force field $-\text{grad } V(x)$.) Problem 4 relates potential energy to *work*.

The planar harmonic oscillation of Section 1 corresponds to the force field

$$F: \mathbf{R}^2 \to \mathbf{R}^2, \qquad F(x) = -mkx.$$

This field is conservative, with potential energy

$$V(x) = \tfrac{1}{2}mk \mid x \mid^2$$

as is easily verified.

For any moving particle $x(t)$ of mass m, the *kinetic energy* is defined to be

$$T = \tfrac{1}{2}m \mid \dot{x}(t) \mid^2.$$

Here $\dot{x}(t)$ is interpreted as the *velocity vector* at time t; its length $\mid \dot{x}(t) \mid$ is the *speed* at time t. If we consider the function $x: \mathbf{R} \to \mathbf{R}^3$ as describing a curve in $\mathbf{R}^3$, then $\dot{x}(t)$ is the *tangent vector* to the curve at $x(t)$.

For a particle moving in a conservative force field $F = -\text{grad } V$, the potential energy at x is defined to be $V(x)$. Note that whereas the kinetic energy depends on the velocity, the potential energy is a function of position.

The *total energy* (or sometimes simply *energy*) is

$$E = T + V.$$

This has the following meaning. If $x(t)$ is the trajectory of a particle moving in the conservative force field, then E is a real-valued function of time:

$$E(t) = \tfrac{1}{2} \mid m\dot{x}(t) \mid^2 + V(x(t)).$$

Theorem (Conservation of Energy) *Let $x(t)$ be the trajectory of a particle moving in a conservative force field $F = -\text{grad } V$. Then the total energy E is independent of time.*

Proof. It needs to be shown that $E(x(t))$ is constant in t or that

$$\frac{d}{dt}(T + V) = 0,$$

or equivalently,

$$\frac{d}{dt}\left(\frac{1}{2}m \mid \dot{x}(t) \mid^2 + V(x(t))\right) = 0.$$

It follows from calculus that

$$\frac{d}{dt} \mid \dot{x} \mid^2 = 2\langle \ddot{x}, \dot{x} \rangle$$

(a version of the Leibniz product formula); and also that

$$\frac{d}{dt}(V(\dot{x})) = \langle \text{grad } V(x), \dot{x} \rangle$$

(the chain rule).

These facts reduce the proof to showing that

$$m\langle \ddot{x}, \dot{x} \rangle + \langle \text{grad } V, \dot{x} \rangle = 0$$

or $\langle m\ddot{x} + \text{grad } V, \dot{x} \rangle = 0$. But this is so since Newton's second law is $m\ddot{x} + \text{grad } V(x) = 0$ in this instance.

§4. Central Force Fields

A force field F is called *central* if $F(x)$ points in the direction of the line through x, for every x. In other words, the vector $F(x)$ is always a scalar multiple of x, the coefficient depending on x:

$$F(x) = \lambda(x)x.$$

We often tacitly exclude from consideration a particle at the origin; many central force fields are not defined (or are "infinite") at the origin.

Lemma *Let F be a conservative force field. Then the following statements are equivalent:*

(a) *F is central,*
(b) *$F(x) = f(\mid x \mid)x$,*
(c) *$F(x) = -\text{grad } V(x)$ and $V(x) = g(\mid x \mid)$.*

Proof. Suppose (c) is true. To prove (b) we find, from the chain rule:

$$\frac{\partial V}{\partial x_j} = g'(\mid x \mid)\frac{\partial}{\partial x_j}(x_1^2 + x_2^2 + x_3^2)^{1/2}$$

$$= \frac{g'(\mid x \mid)}{\mid x \mid}x_j;$$

this proves (b) with $f(\mid x \mid) = g'(\mid x \mid)/\mid x \mid$. It is clear that (b) implies (a). To show that (a) implies (c) we must prove that V is constant on each sphere.

$$S_\alpha = \{x \in \mathbf{R}^3 \mid \mid x \mid = \alpha\}, \qquad \alpha > 0.$$

Since any two points in S_α can be connected by a curve in S_α, it suffices to show that V is constant on any curve in S_α. Hence if $J \subset \mathbf{R}$ is an interval and $u: J \to S_\alpha$ is a C^1 map, we must show that the derivative of the composition $V \cdot u$

$$J \xrightarrow{u} S_\alpha \subset \mathbf{R}^3 \xrightarrow{V} \mathbf{R}$$

is identically 0. This derivative is

$$\frac{d}{dt}V(u(t)) = \langle \text{grad } V(u(t)), u'(t) \rangle$$

as in Section 2. Now grad $V(x) = -F(x) = -\lambda(x)x$ since F is central:

$$\frac{d}{dt}V(u(t)) = -\lambda(u(t))\langle u(t), u'(t)\rangle$$

$$= \frac{-\lambda u(t))}{2}\frac{d}{dt}|\,u(t)\,|^2$$

$$= 0$$

because $|\,u(t)\,| \equiv \alpha$.

In Section 5 we shall consider a special conservative central force field obtained from Newton's law of gravitation.

Consider now a central force field, not necessarily conservative.

Suppose at some time t_0, that $P \subset \mathbf{R}^3$ denotes the plane containing the particle, the velocity vector of the particle and the origin. The force vector $F(x)$ for any point x in P also lies in P. This makes it plausible that the particle stays in the plane P for all time. In fact, this is true: a particle moving in a central force field moves in a fixed plane.

The proof depends on the *cross product* (or vector product) $u \times v$ of vectors u, v in $\mathbf{R}^3$. We recall the definition

$$u \times v = (u_2v_3 - u_3v_2,\ u_3v_1 - u_1v_3,\ u_1v_2 - u_2v_1) \in \mathbf{R}^3$$

and that $u \times v = -v \times u = |\,u\,|\,|\,v\,|\,N \sin\theta$, where N is a unit vector perpendicular to u and v, (U, v, N) oriented as the axes ("right-hand rule"), and θ is the angle between u and v.

Then the vector $u \times v = 0$ if and only if one vector is a scalar multiple of the other; if $u \times v \neq 0$, then $u \times v$ is orthogonal to the plane containing u and v. If u and v are functions of t in $\mathbf{R}$, then a version of the Leibniz product rule asserts (as one can check using Cartesian coordinates):

$$\frac{d}{dt}(u \times v) = \dot{u} \times v + u \times \dot{v}.$$

Now let $x(t)$ be the path of a particle moving under the influence of a central force field. We have

$$\frac{d}{dt}(x \times \dot{x}) = \dot{x} \times \dot{x} + x \times \ddot{x}$$

$$= x \times \ddot{x}$$

$$= 0$$

because $\ddot{x}$ is a scalar multiple of x. Therefore $x(t) \times \dot{x}(t)$ is a constant vector y. If $y \neq 0$, this means that x and $\dot{x}$ always lie in the plane orthogonal to y, as asserted. If $y = 0$, then $\dot{x}(t) = g(t)x(t)$ for some scalar function $g(t)$. This means that the velocity vector of the moving particle is always directed along the line through the origin and the particle, as is the force on the particle. This makes it plausible that the particle always moves along the same line through the origin. To prove this let $(x_1(t), x_2(t), x_3(t))$ be the coordinates of $x(t)$. Then we have three differential equations

$$\frac{dx_k}{dt} = g(t)x_k(t), \qquad k = 1, 2, 3.$$

By integration we find

$$x_k(t) = e^{h(t)}x_k(t_0), \qquad h(t) = \int_{t_0}^{t} g(s)\,ds.$$

Therefore $x(t)$ is always a scalar multiple of $x(t_0)$ and so $x(t)$ moves in a fixed line, and hence in a fixed plane, as asserted.

We restrict attention to a conservative central force field in a plane, which we take to be the Cartesian plane $\mathbf{R}^2$. Thus x now denotes a point of $\mathbf{R}^2$, the potential energy V is defined on $\mathbf{R}^2$ and

$$F(x) = -\text{grad}\,V(x) = -\left(\frac{\partial V}{\partial x_1}, \frac{\partial V}{\partial x_2}\right).$$

Introduce polar coordinates (r, θ), with $r = |\,x\,|$.

Define the *angular momentum* of the particle to be

$$h = mr^2\dot{\theta},$$

where $\dot{\theta}$ is the time derivative of the angular coordinate of the particle.

Theorem (Conservation of Angular Momentum) *For a particle moving in a central force field:*

$$\frac{dh}{dt} = 0, \qquad \textit{where} \quad h = mr^2\dot{\theta}.$$

Proof. Let $i = i(t)$ be the unit vector in the direction $x(t)$ so $x = ri$. Let $j = j(t)$ be the unit vector with a 90° angle from i to j. A computation shows that $di/dt = \dot{\theta}j$, $dj/dt = -\dot{\theta}i$ and hence

$$\dot{x} = \dot{r}i + r\dot{\theta}j.$$

Differentiating again yields

$$\ddot{x} = (\ddot{r} - r\dot{\theta}^2)i + \frac{1}{r}\frac{d}{dt}(r^2\dot{\theta})j.$$

If the force is central, however, it has zero component perpendicular to x. Therefore, since $\ddot{x} = m^{-1}F(x)$, the component of $\ddot{x}$ along j must be 0. Hence

$$\frac{d}{dt}(r^2\dot{\theta}) = 0,$$

proving the theorem.

We can now prove one of Kepler's laws. Let $A(t)$ denote the area swept out by the vector $x(t)$ in the time from t_0 to t. In polar coordinates $dA = \frac{1}{2}r^2\,d\theta$. We define the *areal velocity* to be

$$\dot{A} = \tfrac{1}{2}r^2\dot{\theta},$$

the rate at which the position vector sweeps out area. Kepler observed that the line segment joining a planet to the sun sweeps out equal areas in equal times, which we interpret to mean $\dot{A} =$ constant. We have proved more generally that this is true for any particle moving in a conservative central force field; this is a consequence of conservation of angular momentum.

§5. States

We recast the Newtonian formulation of the preceding sections in such a way that the differential equation becomes first order, the states of the system are made explicit, and energy becomes a function on the space of states.

A *state* of a physical system is information characterizing it at a given time. In particular, a state of the physical system of Section 1 is the position and velocity of the particle. The space of states is the Cartesian product $\mathbf{R}^3 \times \mathbf{R}^3$ of pairs (x, v), x, v in $\mathbf{R}^3$; x is the position, v the velocity that a particle might have at a given moment.

We may rewrite Newton's equation

$$m\dot{x} = F(x) \tag{1}$$

as a first order equation in terms of x and v. (The *order* of a differential equation is the order of the highest derivative that occurs explicitly in the equation.) Consider the differential equation

$$\frac{dx}{dt} = v, \qquad m\frac{dv}{dt} = F(x). \tag{1'}$$

A solution to (1') is a curve $t \to (x(t), v(t))$ in the state space $\mathbf{R}^3 \times \mathbf{R}^3$ such that

$$\dot{x}(t) = v(t) \quad \text{and} \quad \dot{v}(t) = m^{-1}F(x(t)) \quad \text{for all } t.$$

It can be seen then that the solutions of (1) and (1') correspond in a natural fashion. Thus if $x(t)$ is a solution of (1), we obtain a solution of (1') by setting $v(t) = \dot{x}(t)$. *The map* $\mathbf{R}^3 \times \mathbf{R}^3 \to \mathbf{R}^3 \times \mathbf{R}^3$ *that sends* (x, v) *into* $(v, m^{-1}F(x))$ *is a vector field on the space of states, and this vector field defines the differential equation,* (1').

A solution $(x(t), v(t))$ to (1') gives the passage of the state of the system in time.

Now we may interpret energy as a function on the state space, $\mathbf{R}^3 \times \mathbf{R}^3 \to \mathbf{R}$, defined by $E(x, v) = \frac{1}{2}m|v|^2 + V(x)$. The statement that "the energy is an integral" then means that the composite function

$$t \to (x(t), v(t)) \to E(x(t), v(t))$$

is constant, or that *on a solution curve in the state space, E is constant.*

We abbreviate $\mathbf{R}^3 \times \mathbf{R}^3$ by $\mathbf{S}$. An *integral* (for (1')) on $\mathbf{S}$ is then any function that is constant on every solution curve of (1'). It was shown in Section 4 that in addition to energy, angular momentum is also an integral for (1'). In the nineteenth century, the idea of solving a differential equation was tied to the construction of a sufficient number of integrals. However, it is realized now that integrals do not exist for differential equations very generally; the problems of differential equations have been considerably freed from the need for integrals.

Finally, we observe that the force field may not be defined on all of $\mathbf{R}^3$, but only on some portion of it, for example, on an open subset $U \subset \mathbf{R}^3$. In this case the path $x(t)$ of the particle is assumed to lie in U. The force and velocity vectors, however, are still allowed to be arbitrary vectors in $\mathbf{R}^3$. The force field is then a vector field on U, denoted by $F: U \to \mathbf{R}^3$. The state space is the Cartesian product $U \times \mathbf{R}^3$, and (1') is a first order equation on $U \times \mathbf{R}^3$.

§6. Elliptical Planetary Orbits

We now pass to consideration of Kepler's first law, that planets have elliptical orbits. For this, a central force is not sufficient. We need the precise form of V as given by the "inverse square law."

We shall show that in polar coordinates (r, θ), an orbit with nonzero angular momentum h is the set of points satisfying

$$r(1 + \epsilon\cos\theta) = l = \text{constant}; \qquad \epsilon = \text{constant},$$

which defines a conic, as can be seen by putting $r\cos\theta = x$, $r^2 = x^2 + y^2$.

Astronomical observations have shown the orbits of planets to be (approximately) ellipses.

Newton's law of gravitation states that a body of mass m_1 exerts a force on a body of mass m_2. The magnitude of the force is gm_1m_2/r^2, where r is the distance between their centers of gravity and g is a constant. The direction of the force on m_2 is from m_2 to m_1.

Thus if m_1 lies at the origin of $\mathbf{R}^3$ and m_2 lies at $x \in \mathbf{R}^3$, the force on m_2 is

$$-gm_1m_2\frac{x}{|x|^3}.$$

The force on m_1 is the negative of this.

We must now face the fact that *both* bodies will move. However, if m_1 is much greater than m_2, its motion will be much less since acceleration is inversely proportional to mass. We therefore make the simplifying assumption that one of the bodies does not move; in the case of planetary motion, of course it is the sun that is assumed at rest. (One might also proceed by taking the center of mass at the origin, without making this simplifying assumption.)

We place the sun at the origin of $\mathbf{R}^3$ and consider the force field corresponding to a planet of given mass m. This field is then

$$F(x) = -C\frac{x}{|x|^3},$$

where C is a constant. We then change the units in which force is measured to obtain the simpler formula

$$F(x) = -\frac{x}{|x|^3}.$$

It is clear this force field is central. Moreover, it is conservative, since

$$\frac{x}{|x|^3} = \text{grad } V,$$

where

$$V = \frac{-1}{|x|}.$$

Observe that $F(x)$ is not defined at 0.

As in the previous section we restrict attention to particles moving in the plane $\mathbf{R}^2$; or, more properly, in $\mathbf{R}^2 - 0$. The force field is the Newtonian gravitational field in $\mathbf{R}^2$, $F(x) = -x/|x|^3$.

Consider a particular solution curve of our differential equation $\ddot{x} = m^{-1}F(x)$. The angular momentum h and energy E are regarded as constants in time since they are the same at all points of the curve. The case $h = 0$ is not so interesting; it corresponds to motion along a straight line toward or away from the sun. Hence we assume $h \neq 0$.

Introduce polar coordinates (r, θ); along the solution curve they become functions of time $(r(t), \theta(t))$. Since $r^2\dot{\theta}$ is constant and not 0, *the sign of $\dot{\theta}$ is constant* along the curve. Thus θ is always increasing or always decreasing with time. Therefore *r is a function of θ along the curve.*

Let $u(t) = 1/r(t)$; then u is also a function of $\theta(t)$. Note that

$$u = -V.$$

We have a convenient formula for kinetic energy T.

Lemma

$$T = \frac{1}{2}\frac{h^2}{m}\left[\left(\frac{du}{d\theta}\right)^2 + u^2\right].$$

Proof. From the formula for $\dot{x}$ in Section 4 and the definition of T we have

$$T = \tfrac{1}{2}m[\dot{r}^2 + (r\dot{\theta})^2].$$

Also,

$$\dot{r} = \frac{-1}{u^2}\frac{du}{d\theta}\dot{\theta} = -\frac{h}{m}\frac{du}{d\theta}$$

by the chain rule and the definitions of u and h; and also

$$r\dot{\theta} = \frac{h}{mr} = \frac{hu}{m}.$$

Substitution in the formula for T proves the lemma.

Now we find a differential equation relating u and θ along the solution curve. Observe that $T = E - V = E + u$. From the lemma we get

$$\left(\frac{du}{d\theta}\right)^2 + u^2 = \frac{2m}{h^2}(E + u). \tag{1}$$

Differentiate both sides by θ, divide by $2\,du/d\theta$, and use $dE/d\theta = 0$ (conservation of energy). We obtain another equation

$$\frac{d^2u}{d\theta^2} + u = \frac{m}{h^2}, \tag{2}$$

where m/h^2 is a constant.

We re-examine the meaning of just what we are doing and of (2). A particular orbit of the planar central force problem is considered, the force being gravitational. Along this orbit, the distance r from the origin (the source of the force) is a function of θ, as is $1/r = u$. We have shown that this function $u = u(\theta)$ satisfies (2), where h is the constant angular momentum and m is the mass.

The solution of (2) (as was seen in Section 1) is

$$u = \frac{m}{h^2} + C\cos(\theta + \theta_0), \tag{3}$$

where C and θ_0 are arbitrary constants.

To obtain a solution to (1), use (3) to compute $du/d\theta$ and $d^2u/d\theta^2$, substitute the resulting expression into (1) and solve for C. The result is

$$C = \pm\frac{1}{h^2}(2mh^2E + m^2)^{1/2}.$$

Putting this into (3) we get

$$u = \frac{m}{h^2}\left[1 \pm \left(1 + 2\frac{Eh^2}{m}\right)^{1/2} \cos(\theta + q)\right],$$

where q is an arbitrary constant. There is no need to consider both signs in front of the radical since $\cos(\theta + q + \pi) = -\cos(\theta + q)$. Moreover, by changing the variable θ to $\theta - q$ we can put any particular solution in the form

$$u = \frac{m}{h^2}\left[1 + \left(1 + 2\frac{Eh^2}{m}\right)^{1/2} \cos\theta\right]. \tag{4}$$

We recall from analytic geometry that the equation of a conic in polar coordinates is

$$u = \frac{1}{l}(1 + \epsilon\cos\theta), \qquad u = \frac{1}{r}. \tag{5}$$

Here l is the *latus rectum* and $\epsilon \geq 0$ is the *eccentricity*. The origin is a focus and the three cases $\epsilon > 1$, $\epsilon = 1$, $\epsilon < 1$ correspond respectively to a hyperbola, parabola, and ellipse. The case $\epsilon = 0$ is a circle.

Since (4) is in the form (5) we have shown that *the orbit of a particle moving under the influence of a Newtonian gravitational force is a conic of eccentricity*

$$\epsilon = \left(1 + \frac{2Eh^2}{m}\right)^{1/2}$$

Clearly, $\epsilon \geq 1$ if and only if $E \geq 0$. Therefore the orbit is a hyperbola, parabola, or ellipse according to whether $E > 0$, $E = 0$, or $E < 0$.

The quantity $u = 1/r$ is always positive. From (4) it follows that

$$\left(1 + \frac{2Eh^2}{m}\right)^{1/2} \cos\theta > -1.$$

But if $\theta = \pm\pi$ radians, $\cos\theta = -1$ and hence

$$\left(1 + \frac{2Eh^2}{m}\right)^{1/2} < 1.$$

This is equivalent to $E < 0$. For some of the planets, including the earth, complete revolutions have been observed; for these planets $\cos\theta = -1$ at least once a year. Therefore their orbits are ellipses. In fact from a few observations of any planet it can be shown that the orbit is in fact an ellipse.

PROBLEMS

1. A particle of mass m moves in the plane $\mathbf{R}^2$ under the influence of an elastic band tying it to the origin. The length of the band is negligible. Hooke's law states that the force on the particle is always directed toward the origin and is proportional to the distance from the origin. Write the force field and verify that it is conservative and central. Write the equation $F = ma$ for this case and solve it. (Compare Section 1.) Verify that for "most" initial conditions the particle moves in an ellipse.
2. Which of the following force fields on $\mathbf{R}^2$ are conservative?
   (a) $F(x, y) = (-x^2, -2y^2)$
   (b) $F(x, y) = (x^2 - y^2, 2xy)$
   (c) $F(x, y) = (x, 0)$
3. Consider the case of a particle in a gravitational field moving directly away from the origin at time $t = 0$. Discuss its motion. Under what initial conditions does it eventually reverse direction?
4. Let $F(x)$ be a force field on $\mathbf{R}^3$. Let x_0, x_1 be points in $\mathbf{R}^3$ and let $y(s)$ be a path in $\mathbf{R}^3$, $s_0 \leq s \leq s_1$, parametrized by arc length s, from x_0 to x_1. The *work* done in moving a particle along this path is defined to be the integral
   $$\int_{s_0}^{s_1} \langle F(y(s)), y'(s)\rangle\, ds,$$
   where $y'(s)$ is the (unit) tangent vector to the path. Prove that the force field is conservative if and only if the work is independent of the path. In fact if $F = -\text{grad } V$, then the work done is $V(x_1) - V(x_0)$.
5. How can we determine whether the orbit of (a) Earth and (b) Pluto is an ellipse, parabola, or hyperbola?
6. Fill in the details of the proof of the theorem in Section 4.
7. Prove the angular momentum h, energy E, and mass m of a planet are related by the inequality
   $$E \geq -\frac{m}{2h^2}.$$

Notes

Lang's *Second Course in Calculus* [12] is a good background reference for the mathematics in this chapter, especially his Chapters 3 and 4. The physics material is covered extensively in a fairly elementary (and perhaps old-fashioned) way in

Principles of Mechanics by Synge and Griffith [23]. One can also find the mechanics discussed in the book on advanced calculus by Loomis and Sternberg [15, Chapter 13].

The unsystematic ad hoc methods used in Section 6 are successful here because of the relative simplicity of the equations. These methods do not extend very far into mechanics. In general, there are not enough "integrals."

The model of planetary motion in this chapter is quite idealized; it ignores the gravitational effect of the other planets.

Chapter 3

Linear Systems with Constant Coefficients and Real Eigenvalues

The purpose of this chapter is to begin the study of the theory of linear operators, which are basic to differential equations. Section 1 is an outline of the necessary facts about vector spaces. Since it is long it is divided into Parts A through F. A reader familiar with some linear algebra should use Section 1 mainly as a reference. In Section 2 we show how to diagonalize an operator having real, distinct eigenvalues. This technique is used in Section 3 to solve the linear, constant coefficient system $x' = Ax$, where A is an operator having real distinct eigenvalues. The last section is an introduction to complex eigenvalues. This subject will be studied further in Chapter 4.

§1. Basic Linear Algebra

We emphasize that for many readers this section should be used only as a reference or a review.

A. Matrices and operators

The setting for most of the differential equations in this book is Cartesian space $\mathbf{R}^n$; this space was defined in Chapter 1, Section 2, as were the operators of addition and scalar multiplication of vectors. The following familiar properties of these

operations are immediate consequences of the definitions:

VS1: $x + y = y + x$,
$x + 0 = x$,
$x + (-x) = 0$,
$(x + y) + z = x + (y + z)$.

Here $x, y, z \in \mathbf{R}^n$, $-x = (-1)x$, and $0 = (0, \ldots, 0) \in \mathbf{R}^n$.

VS2: $(\lambda + \mu)x = \lambda x + \mu x$,
$\lambda(x + y) = \lambda x + \lambda y$,
$1x = x$,
$0x = 0$ (the first 0 in $\mathbf{R}$, the second in $\mathbf{R}^n$).

These operations satisfying VS1 and VS2 define the *vector space* structure on $\mathbf{R}^n$. Frequently, our development relies only on the vector space structure and ignores the Cartesian (that is, coordinate) structure of $\mathbf{R}^n$. To emphasize this idea, we may write E for $\mathbf{R}^n$ and call E a vector space.

The standard coordinates are often ill suited to the differential equation being studied; we may seek new coordinates, as we did in Chapter 1, giving the equation a simpler form. The goal of this and subsequent chapters on algebra is to explain this process. It is very useful to be able to treat vectors (and later, operators) as objects independent of any particular coordinate system.

The reader familiar with linear algebra will recognize VS1 and VS2 as the defining axioms of an abstract vector space. With the additional axiom of finite dimensionality, abstract vector spaces could be used in place of $\mathbf{R}^n$ throughout most of this book.

Let $A = [a_{ij}]$ be some $n \times m$ matrix as in Section 2 of Chapter 1. Thus each a_{ij} is a real number, where (i, j) ranges over all ordered pairs of integers with $1 \leq i \leq n$, $1 \leq j \leq n$. The matrix A can be considered as a map $A: \mathbf{R}^n \to \mathbf{R}^n$ where the ith coordinate of Ax is $\sum_{j=1}^n a_{ij}x_j$, for each $x = (x_1, \ldots, x_n)$ in $\mathbf{R}^n$. It is easy to check that this map satisfies, for $x, y \in \mathbf{R}^n$, $\lambda \in \mathbf{R}$:

L1: $A(x + y) = Ax + Ay$,
L2: $A(\lambda x) = \lambda Ax$.

These are called *linearity properties*. Any map $A: \mathbf{R}^n \to \mathbf{R}^n$ satisfying L1 and L2 is called a *linear map*. Even more generally, a map $A: \mathbf{R}^n \to \mathbf{R}^m$ (perhaps different domain and range) that satisfies L1 and L2 is called *linear*. In the case where the domain and range are the same, A is also called an *operator*. The set of all operators on $\mathbf{R}^n$ is denoted by $L(\mathbf{R}^n)$.

Note that if $e_k \in \mathbf{R}^n$ is the vector

$$e_k = (0, \ldots, 0, 1, 0, \ldots, 0),$$

with a 1 in the kth place, zeros elsewhere, then

$$Ae_k = (a_{1k}, a_{2k}, \ldots, a_{nk}) = \sum_{i=1}^n a_{ik}e_i. \tag{1}$$

Thus the image of e_k is the kth column of the matrix A.

Let M_n be the set of all $n \times n$ matrices. Then from what we have just described, there is a natural map

$$M_n \to L(\mathbf{R}^n) \tag{2}$$

that associates to each matrix the corresponding linear map. There is an inverse process that associates to every operator on $\mathbf{R}^n$ a matrix. In fact let $T: \mathbf{R}^n \to \mathbf{R}^n$ be any operator. Then define a_{ij} = the ith coordinate of Te_j. The matrix $A = [a_{ij}]$ obtained in this way has for its kth column the vector Te_k. Thus

$$Te_k = Ae_k; \qquad k = 1, \ldots, n.$$

It follows that the operator defined by A is exactly T. For let $x \in \mathbf{R}^n$ be any vector, $x = (x_1, \ldots, x_n)$. Then

$$x = x_1e_1 + \cdots + x_ne_n.$$

Hence

$$\begin{aligned} Ax &= A(\textstyle\sum x_ke_k) = \sum x_k(Ae_k) \qquad \text{(by L1 and L2)} \\ &= \textstyle\sum x_k(Te_k) \\ &= T(\textstyle\sum x_ke_k) \\ &= Tx. \end{aligned}$$

In this way we obtain a natural correspondence between operators on $\mathbf{R}^n$ and $n \times n$ matrices.

More generally, to every linear map $\mathbf{R}^n \to \mathbf{R}^m$ corresponds an $m \times n$ matrix, and conversely. In this book we shall usually be concerned with only operators and $n \times n$ matrices.

Let S, T be operators on $\mathbf{R}^n$. The composite map TS, sending the vector x to $T(S(x))$, is again an operator on $\mathbf{R}^n$. If S has the matrix $[a_{ij}] = A$ and T has the matrix $[b_{ij}] = B$, then TS has the matrix $[c_{ij}] = C$, where

$$c_{ij} = \sum_{k=1}^n b_{ik}a_{kj}.$$

To see this we compute the image of e_j under TS:

$$\begin{aligned} (TS)e_j &= B(Ae_j) = B(\textstyle\sum_k a_{kj}e_k) \\ &= \textstyle\sum_k a_{kj}(Be_k) \\ &= \textstyle\sum_k a_{kj}(\sum_i b_{ij}e_i). \end{aligned}$$

Therefore

$$(TS)e_j = \sum_i \left(\sum_k b_{ik}a_{kj}\right)e_i.$$

This formula says that the ith coordinate of $(TS)e_j$ is

$$\sum_k b_{ik}a_{kj}.$$

Since this ith coordinate is c_{ij}, our assertion follows.

We call the matrix C obtained in this way the *product* BA of B and A (in that order).

Since composition of mappings is associative, it follows that $C(BA) = (CB)A$ if A, B, C are $n \times n$ matrices.

The sum $S + T$ of operators $S, T \in L(\mathbf{R}^n)$ is defined to be the operator

$$x \to Sx + Tx.$$

It is easy to see that if A and B are the respective matrices of S and T, then the matrix of $S + T$ is $A + B = [a_{ij} + b_{ij}]$.

Operators and matrices obey the two distributive laws

$$P(Q + R) = PQ + PR; \qquad (Q + R)P = QP + RP.$$

Two special operators are $O: x \to 0$ and $I: x \to x$. We also use O and I to denote the corresponding matrices. All entries of O are $0 \in \mathbf{R}$ while $I = [\delta_{ij}]$ where δ_{ij} is the Kronecker function:

$$\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases}$$

Thus I has ones on the diagonal (from upper left to lower right) and zeros elsewhere.

It is clear that $A + O = O + A = A$, $OA = AO = O$, and $AI = IA = A$, for both operators and matrices.

If T is an operator and λ any real number, a new operator T is defined by

$$(\lambda T)x = \lambda(Tx).$$

If $A = [a_{ij}]$ is the matrix of T, then the matrix of λT is $\lambda A = [\lambda a_{ij}]$, obtained by multiplying each entry in A by λ. It is clear that

$$0T = 0,$$

$$1T = T,$$

and similarly for matrices. Here 0 and 1 are real numbers.

The set $L(\mathbf{R}^n)$ of all operators on $\mathbf{R}^n$, like the set M_n of all $n \times n$ matrices, satisfies the vector space axiom VS1, VS2 with 0 as O and x, y, z as operators (or matrices). If we consider an $n \times n$ matrix as a point in $\mathbf{R}^{n^2}$, the Cartesian space of dimension n^2, then the vector space operations on $L(\mathbf{R}^n)$ and M_n are the usual ones.

An operator T is called *invertible* if there exists an operator S such that $ST = TS = I$. We call S the *inverse* of T and write $S = T^{-1}$, $T = S^{-1}$. If A and B are the matrices corresponding to S and T, then $AB = BA = I$. We also say A is invertible, $B = A^{-1}$, $A = B^{-1}$.

It is not easy to find the inverse of a matrix (supposing it has one) in general; we discuss this further in the appendix. The 2×2 case is quite simple, however. The inverse of

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

is

$$\begin{bmatrix} \frac{d}{D} & -\frac{b}{D} \\ -\frac{c}{D} & \frac{a}{D} \end{bmatrix}; \qquad D = ad - bc,$$

provided the *determinant* $D \neq 0$. If $D = 0$, A is not invertible. (Determinants are considered in Part E.)

B. *Subspaces, bases, and dimension*

Let $E = \mathbf{R}^n$. A nonempty subset $F \subset E$ is called a subspace (more properly, a *linear subspace*) if F is closed under the operations of addition and scalar multiplication in E; that is, for all $x \in F$, $y \in F$, $\lambda \in \mathbf{R}$:

$$x + y \in F, \qquad \lambda x \in F.$$

It follows that with these operations F satisfies VS1 and VS2 of Part A.

If F contains only 0, we write $F = 0$ and call F the *trivial* subspace. If $F \neq E$, we call F a *proper* subspace.

If F_1 and F_2 are subspaces and $F_1 \subset F_2$, we call F_1 a subspace of F_2.

Since a subspace satisfies VS1 and VS2, the concept of a *linear map* $T: F_1 \to F_2$ between subspaces $F_1 \subset \mathbf{R}^n$, $F_2 \subset \mathbf{R}^m$, makes sense: T is a map satisfying L1 and L2 in Part A. In particular, if $m = n$ and $F_1 = F_2$, T is an *operator* on a subspace.

Henceforth we shall use the term *vector space* to mean "subspace of a Cartesian space." An element of a vector space will be called a *vector* (also a point). To distinguish them from vectors, real numbers are called *scalars*.

Two important subspaces are determined by a linear map

$$A: E_1 \to E_2,$$

where E_1 and E_2 are vector spaces. The *kernel of* A is the set

$$\operatorname{Ker} A = \{x \in E_1 \mid Ax = 0\} = A^{-1}(0).$$

The *image of* A is the set

$$\begin{aligned}\operatorname{Im} A &= \{y \in E_2 \mid Ax = y \text{ for some } x \in E_1\} \\ &= A(E_1).\end{aligned}$$

Let F be a vector space. A set $S = \{a_1, \ldots, a_k\}$ of vectors in F is said to *span* F if every vector in F is a linear combination of $a_1, \ldots, a_k$; that is, for every $x \in F$ there are scalars $t_1, \ldots, t_k$ such that

$$x = t_1a_1 + \cdots + t_ka_k.$$

The set S is called *independent* if whenever $t_1, \ldots, t_k$ are scalars such that

$$t_1a_1 + \cdots + t_ka_k = 0,$$

then $t_1 = \cdots = t_k = 0$.

A *basis* of F is an ordered set of vectors in F that is independent and which spans F.

The following basic fact is proved in Appendix I.

Proposition 1 *Every vector space F has a basis, and every basis of F has the same number of elements. If $\{e_1, \ldots, e_k\} \subset F$ is an independent subset that is not a basis, by adjoining to it suitable vectors $e_{k+1}, \ldots, e_m$ one can form a basis $\{e_1, \ldots, e_m\}$.*

The number of elements in a basis of F is called the *dimension* of F, denoted by $\dim F$. If $\{e_1, \ldots, e_m\}$ is a basis of F, then every vector $x \in F$ can be expressed

$$x = \sum_{i=1}^{m} t_ie_i, \qquad t_i \in \mathbf{R},$$

since the e_i span F. Moreover, the numbers $t_1, \ldots, t_m$ are unique. To see this, suppose also that

$$x = \sum_{i=1}^{m} s_ie_i.$$

Then

$$0 = x - x = \sum_i (t_i - s_i)e_i;$$

by independence,

$$t_i - s_i = 0, \qquad i = 1, \ldots, m.$$

These numbers $t_1, \ldots, t_m$ are called the *coordinates of x in the basis* $\{e_1, \ldots, e_m\}$.

The *standard basis* $e_1, \ldots, e_n$ of $\mathbf{R}^n$ is defined by

$$e_i = (0, \ldots, 0, 1, 0, \ldots, 0); \qquad i = 1, \ldots, n,$$

with 1 in the ith place and 0 elsewhere. This is in fact a basis; for $\sum t_ie_i = (t_1, \ldots, t_n)$, so $\{e_1, \ldots, e_n\}$ spans $\mathbf{R}^n$; independence is immediate.

It is easy to check that $\operatorname{Ker} A$ and $\operatorname{Im} A$ are subspaces of E_1 and E_2, respectively.

A simple but important property of $\operatorname{Ker} A$ is this: *A is one-to-one if and only if* $\operatorname{Ker} A = 0$. For suppose A is one-to-one, and $x \in \operatorname{Ker} A$. Then $Ax = 0 = A0$. Hence $x = 0$; therefore 0 is the only element of $\operatorname{Ker} A$. Conversely, suppose $\operatorname{Ker} A = 0$, and $Ax = Ay$. Then $A(x - y) = 0$, so $x - y \in \operatorname{Ker} A$. Then $A(x - y) = 0$, so $x - y \in \operatorname{Ker} A$. Thus $x - y = 0$ so $x = y$.

The kernel of a linear map $\mathbf{R}^n \to \mathbf{R}^m$ is connected with linear equations (algebraic, not differential) as follows. Let $A = [a_{ij}]$ be the $m \times n$ matrix of the map. Then $x = (x_1, \ldots, x_n)$ is in $\operatorname{Ker} A$ if and only if

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= 0, \\ &\vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= 0. \end{aligned}$$

In other words, $(x_1, \ldots, x_n)$ is a solution to the above system of m linear homogeneous equations in n unknowns. In this case $\operatorname{Ker} A$ is called the *solution space* of the system. *"Solving" the system means finding a basis for* $\operatorname{Ker} A$.

If a linear map $T: E \to F$ is both one-to-one and onto, then there is a unique map $S: F \to E$ such that $ST(x) = x$ and $TS(y) = y$ for all $x \in E$, $y \in F$. The map S is also linear. In this case we call T an *isomorphism*, and say that E and F are isomorphic vector spaces.

Proposition 2 *Two vector spaces are isomorphic if and only if they have the same dimension. In particular, every n-dimensional vector space is isomorphic to $\mathbf{R}^n$.*

Proof. Suppose E and F are isomorphic. If $\{e_1, \ldots, e_n\}$ is a basis for E, it is easy to verify that $Te_1, \ldots, Te_n$ span F (since T is onto) and are independent (since T is one-to-one). Therefore E and F have the same dimension, n. Conversely, suppose $\{e_1, \ldots, e_n\}$ and $\{f_1, \ldots, f_n\}$ are bases for E and F, respectively. Define $T: E \to F$ to be the unique linear map such that $Te_i = f_i$, $i = 1, \ldots, n$: if $x = \sum x_ie_i \in E$, then $Tx = \sum x_if_i$. Then T is onto since the f_i span F, and $\operatorname{Ker} T = 0$ since the f_i are independent.

The following important proposition is proved in Appendix I.

Proposition 3 *Let $T: E \to F$ be a linear map. Then*

$$\dim(\operatorname{Im} T) + \dim(\operatorname{Ker} T) = \dim E.$$

In particular, suppose $\dim E = \dim F$. Then the following are equivalent statements:

(a) $\operatorname{Ker} T = 0$,
(b) $\operatorname{Im} T = F$,
(c) *T is an isomorphism.*

C. *Changes of bases and coordinates*

To every basis $\{e_1, \ldots, e_n\}$ of a vector space E we have associated a system of coordinates as follows: to each vector $z \in E$ we assign the unique n-tuple of real numbers $(x_1, \ldots, x_n)$ such that $z = \sum x_i e_i$. If we consider x_i as a function of z, we may define a map

$$\varphi: E \to \mathbf{R}^n, \qquad \varphi(z) = (x_1(z), \ldots, x_n(z)).$$

This is a linear map; it is in fact the unique linear map sending each basis vector e_i of E into the corresponding standard basis vector of $\mathbf{R}^n$, which we denote here by $\bar{e}_i$.

It is easy to see that φ is an isomorphism (see Proposition 2 of Part B). The isomorphism φ sends each vector z into its n-tuple of coordinates in the basis $\{\bar{e}_1, \ldots, \bar{e}_n\}$.

Conversely, let $\varphi: E \to \mathbf{R}^n$ be any isomorphism. If $\{\bar{e}_1, \ldots, \bar{e}_n\}$ is the standard basis of $\mathbf{R}^n$, then define $e_i = \varphi^{-1}(\bar{e}_i)$, $i = 1, \ldots, n$. Then $\{e_1, \ldots, e_n\}$ is a basis of E, and clearly,

$$\varphi(\sum x_i e_i) = (x_1, \ldots, x_n).$$

In this way we arrive at the following definition: A *coordinate system* on a vector space E is an isomorphism $\varphi: E \to \mathbf{R}^n$. (Of course, $n = \dim E$.) The coordinates of $z \in E$ are $(x_1, \ldots, x_n)$, where $\varphi(z) = (x_1, \ldots, x_n)$. Each coordinate x_i is a linear function $x_i: E \to \mathbf{R}$.

We thus have three equivalent concepts: a basis of E, a coordinate system on E, and an isomorphism $E \to \mathbf{R}^n$.

Readers familiar with the theory of dual vector spaces (see Chapter 9) will recognize the coordinate functions x_i as forming the basis of E^* dual to $\{e_1, \ldots, e_n\}$; here E^* is the "dual space" of E, that is, the vector space of linear maps $E \to \mathbf{R}$.

The coordinate functions x_i are the unique linear functions $E \to \mathbf{R}$ such that

$$x_i(e_j) = \delta_{ij}, \qquad i = 1, \ldots, n; \quad j = 1, \ldots, n,$$

where $\delta_{ij} = 0$ if $i \neq j$ and 1 if $i = j$.

Now we investigate the relations between two bases in E and the two corresponding coordinate systems.

Let $\{e_1, \ldots, e_n\}$ be a basis of E and $(x_1, \ldots, x_n)$ the corresponding coordinates. Let $\varphi: E \to \mathbf{R}^n$ be the corresponding isomorphism. Let $\{f_1, \ldots, f_n\}$ be a new basis, with coordinates $(y_1, \ldots, y_n)$. Let $\psi: E \to \mathbf{R}^n$ be the corresponding isomorphism. Each vector f_i is a linear combination of the e_j; hence we define an $n \times n$ matrix:

$$P = [p_{ij}]; \qquad f_i = \sum_j p_{ij} e_j. \tag{3}$$

Each of the new coordinates $y_i: E \to \mathbf{R}$ is a linear map, and so can be expressed in terms of the old coordinates $(x_1, \ldots, x_n)$. In this way another $n \times n$ matrix is

defined:

$$Q = [q_{kl}]; \qquad y_k = \sum_l q_{kl} x_l. \tag{4}$$

In fact, Q is the matrix of the linear operator $\psi\varphi^{-1}: \mathbf{R}^n \to \mathbf{R}^n$.

How are the matrices P and Q related? To answer this we first relate the bases with their corresponding coordinates:

$$x_l(e_j) = \delta_{lj}, \qquad l, j = 1, \ldots, n; \tag{5}$$

$$y_k(f_i) = \delta_{ki}, \qquad k, i = 1, \ldots, n. \tag{6}$$

Substituting (4) and (3) into (6):

$$\delta_{ki} = \sum_l q_{kl} x_l (\sum_j p_{ij} e_j).$$

Since x_l is a linear function, we have

$$\delta_{ki} = \sum_l (\sum_j q_{kl} p_{ij} x_l(e_j))$$

$$= \sum_l (\sum_j q_{kl} p_{ij} \delta_{lj})$$

by (5). Each term of the internal sum on the right is 0 unless $l = j$, in which case it is $q_{kj} p_{ij}$. Thus

$$\delta_{ki} = \sum_j q_{kj} p_{ij}.$$

To interpret this formula, introduce the matrix R which is the *transpose* P^t of P, by

$$R = [r_{ij}], \qquad r_{ij} = p_{ji}.$$

Each row of R is the corresponding column of P. Then,

$$\delta_{ki} = \sum_j q_{kj} r_{ji}$$

tells us that the (k, i)th entry in the matrix QR is δ_{ki}; in other words,

$$I = QR.$$

We finally obtain

$$I = QP^t.$$

Thus

$$Q = (P^t)^{-1} = (P^{-1})^t.$$

The last equality follows from the identities $I^t = I$, and $(AB)^t = B^t A^t$ for any $n \times n$ matrices A, B. Hence

$$I = (PP^{-1})^t = P^t (P^{-1})^t,$$

so $(P^t)^{-1} = (P^{-1})^t$.

We have proved:

Proposition 4 *The matrix expressing new coordinates in terms of the old is the inverse transpose of the matrix expressing the new basis in terms of the old.*

D. Operator, bases, and matrices

In Part A we associated to an operator T on $\mathbf{R}^n$ a matrix $[a_{ij}]$ by the rule

$$Te_j = \sum_i a_{ij}e_i; \qquad i = 1, \ldots, n, \tag{7}$$

where $\{e_1, \ldots, e_n\}$ is the standard basis of $\mathbf{R}^n$. Equivalently, the ith coordinate of Tx, $x = (x_1, \ldots, x_n)$, is

$$\sum_j a_{ij}x_j. \tag{8}$$

It is useful to represent (8) as the product of an $n \times n$ matrix and an $n \times 1$ matrix:

$$\begin{bmatrix} (Tx)_1 \\ \vdots \\ (Tx)_i \\ \vdots \\ (Tx)_n \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{i1} & \cdots & a_{in} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ \\ x_n \end{bmatrix}$$

We carry out exactly the same procedure for an operator $T: E \to E$, where E is any vector space and $\{e_1, \ldots, e_n\}$ is a given basis of E. Namely, (7) defines a matrix $[a_{ij}]$. The coordinates of Tx for the basis $\{e_1, \ldots, e_n\}$ are computed by (8).

It is helpful to use the following rules in constructing the matrix of an operator in a given basis:

The jth *column* of the matrix gives the coordinates of the image of the jth basis vector, as in (7).

The ith *row* of the matrix expresses the ith *coordinate* of the image of x as a linear function of the coordinates of x, as in (8).

If we think of the coordinates as linear functions $x_i: E \to \mathbf{R}$, then (7) is expressed succinctly by

$$x_iT = \sum_j a_{ij}x_j; \qquad i = 1, \ldots, n. \tag{9}$$

This looks very pretty when placed next to (7)! The left side of (9) is the composition

$$E \xrightarrow{T} E \xrightarrow{x_i} R.$$

The right-hand side of (9) is a linear combination of the linear functions $x_1, \ldots, x_n$. The meaning of (9) is that the two linear functions on E, expressed by the left and right sides of (9), are equal.

Now suppose a new system of coordinates $(y_1, \ldots, y_n)$ is introduced in E, corresponding to a new basis $\{f_1, \ldots, f_n\}$. Let B be the matrix of T in the new coordinates. How is B related to A?

The new coordinates are related to the old ones by an invertible matrix $Q = [q_{ij}]$, as explained in Part C. If $z \in E$ is any point, its two sets of coordinates $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ are related by

$$y = Qx; \qquad x = Q^{-1}y.$$

(Here we think of x and y as points in $\mathbf{R}^n$.) The image Tz also has two sets of coordinates, Ax and By, where B is the matrix of T in the new coordinates. Therefore

$$By = QAx.$$

Hence

$$By = QAQ^{-1}y$$

for all $y \in \mathbf{R}^n$. It follows that

$$B = QAQ^{-1}. \tag{10}$$

This is a basic fact. It is worth restating in terms of the matrix P expressing the new basis vectors f_i in terms of the old basis $\{e_1, \ldots, e_n\}$:

$$P = [p_{ij}], \qquad f_i = \sum_j p_{ij}e_j.$$

In Part C we saw that Q is the inverse transpose of P. Therefore

$$B = (P^t)^{-1}AP^t. \tag{11}$$

The matrix P^t can be described as follows: the ith *column* of P^t consists of the coordinates of the new basis vector f_i in the old basis $\{e_1, \ldots, e_n\}$. Observe that in (10) and (11) the inverse signs -1 appear in different places.

Two $n \times n$ matrices B and A related as in (10) by some invertible matrix Q are called *similar*. This is a basic equivalence relation on matrices. Two matrices are similar if and only if they represent the same operator in different bases. Any matrix property that is preserved under similarity is a property of the underlying linear transformation. One of the main goals of linear algebra is to discover criteria for the similarity of matrices.

We also call two operators $S, T \in L(E)$ similar if $T = QSQ^{-1}$ for some invertible operator $Q \in L(E)$. This is equivalent to similarity of their matrices. Similar operators define differential equations that have the same dynamical properties.

E. Determinant, trace, and rank

We recall briefly the main properties of the determinant function

$$\text{Det}: M_n \to \mathbf{R},$$

where M_n is the set of $n \times n$ matrices:

D1: $\mathrm{Det}(AB) = (\mathrm{Det}\, A)(\mathrm{Det}\, B)$,
D2: $\mathrm{Det}\, I = 1$,
D3: $\mathrm{Det}\, A \neq 0$ if and only if A is invertible.

There is a unique function Det having these three properties; it is discussed in more detail in the appendix. For a 1×1 matrix $A = [a]$, $\mathrm{Det}\, A = a$. For a 2×2 matrix $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] = A$,

$$\mathrm{Det}\, A = ad - bc.$$

From D1 and D2 it follows that if A^{-1} exists, then

$$\mathrm{Det}(A^{-1}) = (\mathrm{Det}\, A)^{-1}.$$

From D1 we then obtain

$$\mathrm{Det}(RAR^{-1}) = \mathrm{Det}\, A.$$

In other words, *similar matrices have the same determinant.* We may therefore define the determinant of an operator $T\colon E \to E$ to be the determinant of any matrix representing T.

For $n = 1$, the determinant of $T\colon \mathbf{R}^1 \to \mathbf{R}^1$ is the factor by which T multiplies lengths, except possibly for sign. Similarly, for $\mathbf{R}^2$ and areas, $\mathbf{R}^3$ and volumes.

If A is a triangular matrix ($a_{ij} = 0$ for $i > j$, or $a_{ij} = 0$ for $i < j$), then $\mathrm{Det}\ A = a_{11} \cdots a_{nn}$, the product of the diagonal elements.

From D3 we deduce:

Proposition 5 *Let A be an operator. Then the following statements are equivalent:*

(a) $\mathrm{Det}\, A \neq 0$,
(b) $\mathrm{Ker}\, A = 0$,
(c) *A is one-to-one,*
(d) *A is onto,*
(e) *A is invertible.*

In particular, $\mathrm{Det}\, A = 0$ *if and only if* $Ax = 0$ *for some vector* $x \neq 0$.

Another important similarity invariant is the *trace* of a matrix $A = [a_{ij}]$:

$$\mathrm{Tr}\, A = \sum_i a_{ii},$$

the sum of the diagonal elements. A computation shows that

$$\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$$

and hence

$$\mathrm{Tr}(RAR^{-1}) = \mathrm{Tr}(R^{-1}RA) = \mathrm{Tr}(A).$$

Therefore we can define the trace of an operator to be the trace of any matrix representing it. It is not easy to interpret the trace geometrically.

Note that

$$\mathrm{Tr}(A + B) = \mathrm{Tr}(A) + \mathrm{Tr}(B).$$

The *rank* of an operator is defined to be the dimension of its image. Since every $n \times n$ matrix defines an operator on $\mathbf{R}^n$, we can define the rank of a matrix A to be the rank of the corresponding operator T. *Rank is invariant under similarity.*

The vector space Im T is spanned by the images under T of the standard basis vector, $e_1, \ldots, e_k$. Since Te_j is the n-tuple that is the jth column of A, it follows that *the rank of A equals the maximum number of independent columns of A.*

This gives a practical method for computing the rank of an operator T. Let A be an $n \times n$ matrix representing T in some basis. Denote the jth column of A by c_j, thought of as an n-tuple of numbers, that is, an element of $\mathbf{R}^n$. The rank of T equals the dimension of the subspace of $\mathbf{R}^n$ spanned by $c_1, \ldots, c_n$. This subspace is also spanned by $c_1, \ldots, c_{j-1}, c_j + \lambda c_k, c_{j+1}, \ldots, c_n$; $\lambda \in \mathbf{R}$. Thus we may replace any column c_j of A by $c_j + \lambda c_k$, for any $\lambda \in \mathbf{R}$, $k \neq j$. In addition, the order of the columns can be changed without altering the rank. By repeatedly transforming A in these two ways we can change A to the form

$$B = \begin{bmatrix} D & 0 \\ C & 0 \end{bmatrix},$$

where D is an $r \times r$ diagonal matrix whose diagonal entries are different from zero and C has $n - r$ rows and r columns, and all other entries are 0. It is easy to see that the rank of B, and hence of A, is r.

From Proposition 3 (Part B) it follows that an operator on an n-dimensional vector space is invertible if and only if it has rank n.

F. Direct sum decomposition

Let $E_1, \ldots, E_r$ be subspaces of E. We say E is the *direct sum* of them if every vector x in E can be expressed uniquely:

$$x = x_1 + \cdots + x_r, \qquad x_i \in E_i, \quad i = 1, \ldots, r.$$

This is denoted

$$E = E_1 \oplus \cdots \oplus E_r = \bigoplus_{i=1}^{r} E_i.$$

Let $T\colon E \to E$ and $T_i\colon E_i \to E_i$, $i = 1, \ldots, n$ be operators. We say that T is the *direct sum* of the T_i if $E = E_1 \oplus \cdots \oplus E_n$, each E_i is invariant under T, that is, $T(E_i) \subset E_i$, and $Tx = T_i x$ if $x \in E_i$. We denote the situation by $T = T_1 \oplus \cdots \oplus T_n$. If T_j has the matrix A_j in some basis for each E_j, then by taking

the union of the basis elements of the E_j to obtain a basis for E, T has the matrix

$$A = \operatorname{diag}\{A_1, \ldots, A_n\} = \begin{bmatrix} A_1 & & & \\ & \cdot & & \\ & & \cdot & \\ & & & A_n \end{bmatrix}.$$

This means the matrices A_i are put together corner-to-corner diagonally as indicated, all other entries in A being zero. (We adopt the convention that the blank entries in a matrix are zeros.)

For direct sums of operators there is the useful formula:

$$\operatorname{Det}(T_1 \oplus \cdots \oplus T_n) = (\operatorname{Det} T_1) \cdots (\operatorname{Det} T_n),$$

and the equivalent matrix formula:

$$\operatorname{Det} \operatorname{diag}\{A_1, \ldots, A_n\} = (\operatorname{Det} A_1) \cdots (\operatorname{Det} A_n).$$

Also:

$$\operatorname{Tr}(T_1 \oplus \cdots \oplus T_n) = \operatorname{Tr}(T_1) + \cdots + \operatorname{Tr}(T_n),$$

and

$$\operatorname{Tr} \operatorname{diag}(A_1, \ldots, A_n) = \operatorname{Tr}(A_1) + \cdots + \operatorname{Tr}(A_n).$$

We identify the Cartesian product of $\mathbf{R}^m$ and $\mathbf{R}^n$ with $\mathbf{R}^{m+n}$ in the obvious way. If $E \subset \mathbf{R}^m$ and $F \subset \mathbf{R}^n$ are subspaces, then $E \times F$ is a subspace of $\mathbf{R}^{m+n}$ under this identification. Thus the *Cartesian product* of two vector spaces is a vector space.

§2. Real Eigenvalues

Let T be an operator on a vector space E. A nonzero vector $x \in E$ is called a (real) *eigenvector* if $Tx = \alpha x$ for some real number α. This α is called a *real eigenvalue*; we say x *belongs to* α.

Eigenvalues and eigenvectors are very important. Many problems in physics and other sciences, as well as in mathematics, are equivalent to the problem of finding eigenvectors of an operator. Moreover, eigenvectors can often be used to find an especially simple matrix for an operator.

The condition that α is a real eigenvalue of T means that the kernel of the operator

$$T - \alpha I : E \to E$$

is nontrivial. This kernel is called the *α-eigenspace* of T; it consists of all eigenvectors belonging to α together with the 0 vector.

To find the real eigenvalues of T we must find all real numbers λ such that

$$\operatorname{Det}(T - \lambda I) = 0. \tag{1}$$

(See Part E of the previous section.) To do this let A be a representative of T. Then (1) is equivalent to

$$\operatorname{Det}(A - \lambda I) = 0. \tag{2}$$

We consider λ as an indeterminate (that is, an "unknown number") and compute the left-hand side of (2) (see Appendix I). The result is a polynomial $p(\lambda)$ in λ, called the *characteristic polynomial* of A. Thus the real eigenvalues of T are exactly the real roots of the $p(\lambda)$. Actually, $p(\lambda)$ is independent of the basis, for if B is the matrix of T in another basis, then

$$B = QAQ^{-1}$$

for some invertible $n \times n$ matrix Q (Section 1, Part D). Hence

$$\begin{aligned} \operatorname{Det}(B - \lambda I) &= \operatorname{Det}(QAQ^{-1} - \lambda I) \\ &= \operatorname{Det}(Q(A - \lambda I)Q^{-1}) \\ &= \operatorname{Det}(A - \lambda I) \end{aligned}$$

(Section 1, Part E). We therefore call $p(\lambda)$ the *characteristic polynomial of the operator* T. Note that the degree of $p(\lambda)$ is the dimension of E.

A complex root of the characteristic polynomial is called a *complex eigenvalue* of T. These will be considered in Section 4.

Once a real eigenvalue α has been found, the eigenvectors belonging to α are found by solving the equation

$$(A - \lambda I)x = 0. \tag{3}$$

By (2) there must exist a nonzero solution vector x. The solution space of (3) is exactly the α-eigenspace.

Example. Consider the operator $A = \begin{bmatrix} \frac{5}{4} & \frac{3}{4} \\ -\frac{6}{4} & -\frac{4}{4} \end{bmatrix}$ on $\mathbf{R}^2$, used to describe a differential equation (4) in Chapter 1. The characteristic polynomial is

$$\begin{aligned} \operatorname{Det} \begin{bmatrix} 5 - \lambda & 3 \\ -6 & -4 - \lambda \end{bmatrix} &= \lambda^2 - \lambda - 2 \\ &= (\lambda - 2)(\lambda + 1). \end{aligned}$$

The eigenvalues are therefore 2 and -1. The eigenvectors belonging to 2 are solutions of the equations $(T - 2I)x = 0$, or

$$\begin{bmatrix} 3 & 3 \\ -6 & -6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0,$$

which, in coordinates, is

$$3x_1 + 3x_2 = 0,$$

$$-6x_1 - 6x_2 = 0.$$

The solutions are

$$x_1 = t, \qquad x_2 = -t; \qquad t \in \mathbf{R}.$$

Thus the vector

$$f_1 = (1, -1) \in \mathbf{R}^2$$

is a basis for the eigenspace belonging to the real eigenvalue 2.

The -1 eigenspace comprises solutions of

$$(A + I)x = 0,$$

or

$$\begin{bmatrix} 6 & 3 \\ -6 & -3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0.$$

This matrix equation is equivalent to the pair of scalar equations

$$6x_1 + 3x_2 = 0,$$

$$-6x_1 - 3x_2 = 0.$$

It is clear that $(-1, 2)$ is a basis for the solution space. Therefore the vector $f_2 = (-1, 2) \in \mathbf{R}^2$ is a basis for the (-1)-eigenspace of T.

The two vectors

$$f_1 = (1, -1), \qquad f_2 = (-1, 2)$$

form a new basis $\{f_1, f_2\}$ for $\mathbf{R}^2$. In this basis T has the diagonal matrix

$$\begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix}.$$

Note that any vector $x = (x_1, x_2)$ in $\mathbf{R}^2$ can be written in the form $y_1f_1 + y_2f_2$; then $x = (y_1 - y_2, -y_1 + 2y_2)$ using the definition of the f_i. Therefore (y_1, y_2) are the coordinates of x in the new basis. Thus

$$x = By, \qquad B = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}.$$

This is how the diagonalizing change of coordinates was found in Section 1 of Chapter 1.

Example. Let T have the matrix $\left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$. The characteristic polynomial is

$$\operatorname{Det}\begin{bmatrix} -\lambda & -1 \\ 1 & -\lambda \end{bmatrix} = \lambda^2 + 1.$$

Hence T has no real eigenvalues.

If a real eigenvalue α is known, the general procedure for finding eigenvectors belonging to α are found as follows. Let A be the matrix of T in a basis $\mathcal{B}$. The matrix equation $(A - \alpha I)x = 0$ is equivalent to the system of linear equations

$$\begin{aligned} (a_{11} - \alpha)x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0, \\ a_{21}x_1 + (a_{22} - \alpha)x_2 + \cdots + a_{2n}x_n &= 0, \\ &\vdots \\ a_{n1}x_1 + \cdots + a_{n,n-1}x_{n-1} + (a_{nn} - \alpha)x_n &= 0. \end{aligned}$$

The vanishing of $\operatorname{Det}(A - \alpha I)$ guarantees a nonzero solution $x = (x_1, \ldots, x_n)$. Such a solution is an eigenvector for α, expressed in the basis $\mathcal{B}$.

A very fortunate situation occurs when E has a basis $\{f_1, \ldots, f_n\}$ such that each f_i is an eigenvector of T. For the matrix of T in this basis is just the *diagonal* matrix $D = \operatorname{diag}\{\alpha_1, \ldots, \alpha_n\}$, that is,

$$D = \begin{bmatrix} \alpha_1 & & & \\ & \cdot & & \\ & & \cdot & \\ & & & \alpha_n \end{bmatrix},$$

all other entries being 0. We say T is *diagonalizable.*

It is very easy to compute with D. For example, if $x \in E$ has components $(x_1, \ldots, x_n)$, that is, $x = \sum x_if_i$, then $Tx = (\alpha_1x_1, \ldots, \alpha_nx_n)$. The kth power $D^k = D \cdots D$ (k factors) is just $\operatorname{diag}\{\alpha_1^k, \ldots, \alpha_n^k\}$.

An important criterion for diagonalizability is the following.

Theorem 1 *Let T be an operator for an n-dimensional vector space E. If the characteristic polynomial of T has n distinct real roots, then T can be diagonalized.*

Proof. Let $e_1, \ldots, e_n$ be eigenvectors corresponding to distinct eigenvalues $\alpha_1, \ldots, \alpha_n$. If $e_1, \ldots, e_n$ do not form a basis for E, order them so that $e_1, \ldots, e_m$ is a maximal independent subset, $m < n$. Then $e_n = \sum_{j=1}^{m} t_je_j$; and

$$\begin{aligned} 0 = (T - \alpha_nI)e_n &= \sum_{j=1}^{m} t_j(T - \alpha_nI)e_j \\ &= \sum_{j=1}^{m} t_j(Te_j - \alpha_ne_j) \\ &= \sum_{j=1}^{m} t_j(\alpha_j - \alpha_n)e_j. \end{aligned}$$

Since $e_j, \ldots, e_m$ are independent,

$$t_j(\alpha_j - \alpha_n) = 0, \qquad j = 1, \ldots, m.$$

Since $\alpha_j \neq \alpha_n$ by assumption, each $t_j = 0$. Therefore, $e_n = 0$, contradicting e_n being an eigenvector. Hence $\{e_1, \ldots, e_n\}$ is a basis, so T is diagonalizable.

The following theorem interprets Theorem 1 in the language of matrices.

Theorem 2 *Let A be an $n \times n$ matrix having n distinct real eigenvalues $\lambda_1, \ldots, \lambda_n$. Then there exists an invertible $n \times n$ matrix Q such that*

$$QAQ^{-1} = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}.$$

Proof. Let $\{e_1, \ldots, e_n\}$ be the standard basis in $\mathbf{R}^n$ with corresponding coordinates $(x_1, \ldots, x_n)$. Let T be the operator on $\mathbf{R}^n$ where the matrix in the standard basis is A. Suppose $\{f_1, \ldots, f_n\}$ is a basis of eigenvectors of T, so that $Af_j = \lambda_j f_j$, $j = 1, \ldots, n$. Put $f_j = (f_{j1}, \ldots, f_{jn})$. If Q is the matrix whose jth column is f_j, then QAQ^{-1} is the matrix of T in the basis $\{f_1, \ldots, f_n\}$, as shown in Part D of Section 1. But this matrix is $\operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}$.

We will often use the expression "A has *real distinct eigenvalues*" for the hypothesis of Theorems 1 and 2.

Another useful condition implying diagonalizability is that an operator have a *symmetric* matrix ($a_{ij} = a_{ji}$) in some basis; see Chapter 9.

Let us examine a general operator T on $\mathbf{R}^2$ for diagonalizability. Let the matrix be $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$; the characteristic polynomial $p_T(\lambda)$ is

$$\operatorname{Det}\begin{bmatrix} a-\lambda & b \\ c & d-\lambda \end{bmatrix} = (a-\lambda)(d-\lambda) - bc$$

$$= \lambda^2 - (a+d)\lambda + (ad - bc).$$

Notice that $a + d$ is the trace Tr and $ad - bc$ is the determinant Det. The roots of $p_T(\lambda)$, and hence the eigenvalues of T, are therefore

$$\tfrac{1}{2}[\mathrm{Tr} \pm (\mathrm{Tr}^2 - 4\,\mathrm{Det})^{1/2}].$$

The roots are real and distinct if $\mathrm{Tr}^2 - 4\,\mathrm{Det} > 0$; they are nonreal complex conjugates if $\mathrm{Tr} - 4\,\mathrm{Det} < 0$; and there is only one root, necessarily real, if $\mathrm{Tr} - 4\,\mathrm{Det} = 0$. Therefore T is diagonalizable if $\mathrm{Tr}^2 - 4\,\mathrm{Det} > 0$. The remaining case, $\mathrm{Tr}^2 - 4\,\mathrm{Det} = 0$ is ambiguous. If T is diagonalizable, the diagonal elements are eigenvectors. If p_T has only one root α, then T has a matrix $\left[\begin{smallmatrix} \alpha & 0 \\ 0 & \alpha \end{smallmatrix}\right]$. Hence $T = \alpha I$. But this means *any* matrix for T is diagonal (not just diagonalizable)! Therefore when $\mathrm{Tr}^2 - 4\,\mathrm{Det} = 0$ either every matrix for T, or no matrix for T, is diagonal. The operator represented by $\left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]$ cannot be diagonalized, for example.

§3. Differential Equations with Real, Distinct Eigenvalues

We use the results of Section 2 to prove an important result.

Theorem 1 *Let A be an operator on $\mathbf{R}^n$ having n distinct, real eigenvalues. Then for all $x_0 \in \mathbf{R}^n$, the linear differential equation*

$$x' = Ax; \qquad x(0) = x_0, \tag{1}$$

has a unique solution.

Proof. Theorem 2 of Section 2 implies the existence of an invertible matrix Q such that the matrix QAQ^{-1} is diagonal:

$$QAQ^{-1} = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\} = B,$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A. Introducing the new coordinates $y = Qx$ in $\mathbf{R}^n$, with $x = Q^{-1}y$, we find

$$y' = Qx' = QAx = QA(Q^{-1}y)$$

so

$$y' = By. \tag{2}$$

Since B is diagonal, this means

$$y_i' = \lambda_i y_i; \qquad i = 1, \ldots, n. \tag{2'}$$

Thus (2) is an uncoupled form of (1). We know that (2') has unique solutions for every initial condition $y_i(0)$:

$$y_i(t) = y_i(0)\exp(t\lambda_i).$$

To solve (1), put $y(0) = Qx_0$. If $y(t)$ is the corresponding solution of (2), then the solution of (1) is

$$x(t) = Q^{-1}y(t).$$

More explicitly,

$$x(t) = Q^{-1}(y_1(0)\exp(\lambda_1 t), \ldots, y_n(0)\exp(\lambda_n t)).$$

Differentiation shows that

$$x' = Q^{-1}y' = Q^{-1}By$$

$$= Q^{-1}(QAQ^{-1})y$$

$$= AQ^{-1}y;$$

$$x' = Ax.$$

Moreover,

$$x(0) = Q^{-1}y(0) = Q^{-1}Qx_0 = x_0.$$

Thus $x(t)$ really does solve (1).

To prove that there are no other solutions to (1), we note that $x(t)$ is a solution to (1) if and only if $Qx(t)$ is a solution to

$$y' = By, \qquad y(0) = Qx_0. \tag{3}$$

Hence two different solutions to (1) would lead to two different solutions to (3), which is impossible since B is diagonal. This proves the theorem.

It is important to observe that the proof is constructive; it actually shows how to find solutions in any specific case. For the proof of Theorem 1 of Section 2 shows how to find the diagonalizing coordinate change Q (or Q^{-1}). We review this procedure.

First, find the eigenvalues of A by finding the roots of the characteristic polynomial of A. (This, of course, may be very difficult.) For each eigenvalue λ_i find a corresponding eigenvector f_i by solving the system of linear equations corresponding to the vector equation

$$(A - \lambda_i I)f_i = 0.$$

(This is purely mechanical but may take a long time if n is large.) Write out each eigenvector f_i in coordinates:

$$f_i = (p_{i1}, \ldots, p_{in}),$$

obtaining a matrix $P = [p_{ij}]$. Then the y_i are defined by the equation

$$x_j = \sum_i p_{ij}y_i; \qquad j = 1, \ldots, n, \tag{4}$$

or

$$x = P^t y.$$

Note the order of the subscripts in (4)! The ith *column* of P^t consists of the coordinates of f_i. The matrix Q in the proof is the inverse of P^t. However, for some purposes, it is not necessary to compute Q.

In the new coordinates the original differential equation becomes

$$y' = \lambda_i y_i, \qquad i = 1, \ldots, n, \tag{5}$$

so the general solution is

$$y_i(t) = a_i \exp(t\lambda_i); \qquad i = 1, \ldots, n,$$

where $a_1, \ldots, a_n$ are arbitrary constants, $a_i = y_i(0)$. The general solution to the original equation is found from (4):

$$x_j(t) = \sum_i p_{ij}a_i \exp(t\lambda_i); \qquad j = 1, \ldots, n. \tag{6}$$

This substitution is most easily done by matrix multiplication

$$x(t) = P^t y(t),$$

writing $x(t)$ and $y(t)$ as column vectors,

$$x(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix}, \qquad y(t) = \begin{bmatrix} a_1 \exp(t\lambda_1) \\ \vdots \\ a_n \exp(t\lambda_n) \end{bmatrix}.$$

To find a solution $x(t)$ with a specified initial value

$$x(0) = u = (u_1, \ldots, u_n),$$

one substitutes $t = 0$ in (6), equates the right-hand side to u, and solves the resulting system of linear algebraic equations for the unknowns $(a_1, \ldots, a_n)$:

$$\sum_i p_{ij}a_i = u_j; \qquad j = 1, \ldots, n. \tag{7}$$

This is equivalent to the matrix equation

$$P^t a = u; \qquad a = (a_1, \ldots, a_n).$$

Thus $a = (P^t)^{-1}u$. Another way of saying this is that the initial values $x(0) = u$ corresponds to the initial value $y(0) = (P^t)^{-1}u$ of (5). If one is interested only in a specific vector u, it is easier to solve (7) directly then to invert the matrix P^t.

Here is a simple example. Find the general solution to the system

$$\begin{aligned} x_1' &= x_1, \\ x_2' &= x_1 + 2x_2, \\ x_3' &= x_1 - x_3. \end{aligned} \tag{8}$$

The corresponding matrix is

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

Since A is triangular,

$$\mathrm{Det}(A - \lambda I) = (1 - \lambda)(2 - \lambda)(-1 - \lambda).$$

Hence the eigenvalues are 1, 2, -1. They are real and distinct, so the theorem applies.

The matrix B is

$$\begin{bmatrix} 1 & & \\ & 2 & \\ & & -1 \end{bmatrix}.$$

In the new coordinates the equivalent differential equation is

$$y_1' = y_1,$$
$$y_2' = 2y_2,$$
$$y_3' = -y_3,$$

which has the solution

$$y_1(t) = ae^t,$$
$$y_2(t) = be^{2t},$$
$$y_3(t) = ce^{-t}, \qquad a, b, c \text{ arbitrary constants.}$$

To relate the old and new coordinates we must find three eigenvectors f_1, f_2, f_3 of A belonging respectively to the eigenvalues 1, 2, -1. The second column of A shows that we can take

$$f_2 = (0, 1, 0),$$

and the third column shows that we may take

$$f_3 = (0, 0, 1).$$

To find $f_1 = (v_1, v_2, v_3)$ we must solve the vector equation

$$(A - I)f_1 = 0,$$

or

$$\begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = 0;$$

this leads to the numerical equation

$$v_1 + v_2 = 0,$$
$$v_1 - 2v_3 = 0.$$

Any nonzero solution will do; we take $v_1 = 2$, $v_2 = -2$, $v_3 = 1$. Thus

$$f_1 = (2, -2, 1).$$

The matrix P^t has for its columns the triples f_1, f_2, f_3:

$$P^t = \begin{bmatrix} 2 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

From $x = P^t y$ we have

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} ae^t \\ be^{2t} \\ ce^{-t} \end{bmatrix};$$

hence

$$x_1(t) = 2ae^t,$$
$$x_2(t) = -2ae^t + be^{2t}, \tag{9}$$
$$x_3(t) = ae^t + ce^{-t},$$

where a, b, c are arbitrary constants.

The reader should verify that (9) is indeed a solution to (8).

To solve an initial value problem for (8), with

$$x_i(0) = u_i; \qquad i = 1, 2, 3,$$

we must select a, b, c appropriately.
From (9) we find

$$x_1(0) = 2a,$$
$$x_2(0) = -2a + b,$$
$$x_3(0) = a + c.$$

Thus we must solve the linear system

$$2a = u_1,$$
$$-2a + b = u_2, \tag{10}$$
$$a + c = u_3,$$

for the unknowns a, b, c. This amounts to inverting the matrix of coefficients of the left-hand side of (10), which is exactly the matrix P^t. For particular values of u_1, u_2, u_3, it is easier to solve (10) directly.

This procedure can, of course, be used for more general initial values, $x(t_0) = u$.

The following observation is an immediate consequence of the proof of Theorem 1.

Theorem 2 *Let the $n \times n$ matrix A have n distinct real eigenvalues $\lambda_1, \ldots, \lambda_n$. Then every solution to the differential equation*

$$x' = Ax, \qquad x(0) = u,$$

is of the form

$$x_i(t) = c_{i1} \exp(t\lambda_1) + \cdots + c_{in} \exp(t\lambda_n); \qquad i = 1, \ldots, n,$$

for unique constants $c_{i1}, \ldots, c_{in}$ depending on u.

By using this theorem we get much information about the general character of the solutions directly from the knowledge of the eigenvalues, without explicitly solving the differential equation. For example, if all the eigenvalues are negative, evidently

$$\lim_{t\to\infty} x(t) = 0$$

for every solution $x(t)$, and conversely. This aspect of linear equations will be investigated in later chapters.

Theorem 2 leads to another method of solution of (1). Regard the coefficients c_{ij} as unknowns; set

$$x_i(t) = \sum_j c_{ij} \exp(t\lambda_j); \qquad i = 1, \ldots, n,$$

and substitute it into

$$x' = Ax, \qquad x(0) = u.$$

Then equate coefficients of $\exp(t\lambda_j)$ and solve for the c_{ij}. There results a system of linear algebraic equations for the c_{ij} which can always be satisfied *provided* $\lambda_1, \ldots, \lambda_n$ are real and distinct. This is the method of "undetermined coefficients."

As an example we consider the same system as before,

$$x_1' = x_1,$$
$$x_2' = x_1 + 2x_2,$$
$$x_3' = x_1 - x_3,$$

with the initial condition

$$x(0) = (1, 0, 0).$$

The eigenvalues are $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = -1$. Our solution must be of the form

$$x_1(t) = c_{11}e^t + c_{12}e^{2t} + c_{13}e^{-t};$$
$$x_2(t) = c_{21}e^t + c_{22}e^{2t} + c_{23}e^{-t};$$
$$x_3(t) = c_{31}e^t + c_{32}e^{2t} + c_{33}e^{-t}.$$

Then from $x_1'(t) = x_1$ we obtain

$$c_{11}e^t + 2c_{12}e^{2t} - c_{13}e^{-t} = c_{11}e^t + c_{12}e^{2t} + c_{13}e^{-t}$$

for all values of t. This is possible only if

$$c_{13} = c_{11} = 0.$$

(Differentiate and set $t = 0$.) From $x_2' = x_1 + 2x_2$ we get

$$c_{21}e^t + 2c_{22}e^{2t} - c_{23}e^{-t} = (c_{11} + 2c_{21})e^t + (c_{12}2c_{22})e^{2t} + (c_{13} + 2c_{23})e^{-t}.$$

Therefore

$$c_{21} = c_{11} + 2c_{21},$$
$$2c_{22} = c_{12} + 2c_{22},$$
$$-c_{23} = c_{13} + 2c_{23},$$

which reduces to

$$c_{21} = -c_{11},$$
$$c_{22} = 0.$$

From $x_3' = x_1 - x_3$ we obtain

$$c_{31}e^t + 2c_{32}e^{2t} - c_{33}e^{-t} = (c_{11} - c_{31})e^t + (c_{12} - c_{32})e^{2t} + (c_{13} - c_{33})e^{-t}.$$

Therefore

$$c_{31} = c_{11} - c_{31},$$
$$2c_{32} = c_{12} - c_{32},$$
$$-c_{33} = c_{13} - c_{33},$$

which boils down to

$$c_{31} = \tfrac{1}{2}c_{11},$$
$$c_{32} = 0.$$

Without using the initial condition yet, we have found

$$x_1(t) = c_{11}e^t,$$
$$x_2(t) = -c_{11}e^t + c_{22}e^{2t}$$
$$x_3(t) = \tfrac{1}{2}c_{11}e^t + c_{33}e^{-t},$$

which is equivalent to (9). From $(x_1(0), x_2(0), x_3(0)) = (1, 0, 0)$ we find

$$c_{11} = 1, \qquad c_{22} = 1, \qquad c_{33} = -\tfrac{1}{2}.$$

The solution is therefore

$$x(t) = (e^t, -e^t + e^{2t}, \tfrac{1}{2}e^t - \tfrac{1}{2}e^{2t}).$$

We remark that the conclusion of Theorem 2 is definitely false for some operators with real, repeated eigenvalues. Consider the operator $A = \left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]$, whose only eigenvalue is 1 and the system $x' = Ax$:

(11)
$$x_1' = x_1,$$
$$x_2' = x_1 + x_2.$$

Obviously, $x_1(t) = ae^t$, a = constant, but there is no constant b such that $x_2(t) = be^t$ is a solution to (11). The reader can verify that in fact a solution is

$$x_1(t) = ae^t,$$
$$x_2(t) = e^t(at + b),$$

a and b being arbitrary constants. All solutions have this form; see Problem 3.

PROBLEMS

1. Solve the following initial value problems:

(a) $x' = -x$,
$y' = x + 2y$;
$x(0) = 0, y(0) = 3$.

(b) $x_1' = 2x_1 + x_2$,
$x_2' = x_1 + x_2$;
$x_1(1) = 1, x_2(1) = 1$.

(c) $x' = Ax$;
$x(0) = (3, 0)$;
$A = \begin{bmatrix} 0 & 3 \\ 1 & -2 \end{bmatrix}$.

(d) $x' = Ax$,
$x(0) = (0, -b, b)$,
$A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 2 & -3 \end{bmatrix}$.

2. Find a 2×2 matrix A such that one solution to $x' = Ax$ is

$$x(t) = (e^{2t} - e^{-t}, e^{2t} + 2e^{-t}).$$

3. Show that the only solution to

$$x_1' = x_1,$$
$$x_2' = x_1 + x_2;$$
$$x_1(0) = a,$$
$$x_2(0) = b,$$

is

$$x_1(t) = ae^t,$$
$$x_2(t) = e^t(b + at).$$

(*Hint*: If $(y_1(t), y_2(t))$ is another solution, consider the functions $e^{-t}y_1(t)$, $e^{-t}y_2(t)$.)

4. Let an operator A have real, distinct eigenvalues. What condition on the eigenvalues is equivalent to $\lim_{t\to\infty} |x(t)| = \infty$ for every solution $x(t)$ to $x' = Ax$?

5. Suppose the $n \times n$ matrix A has real, distinct eigenvalues. Let $t \to \phi(t, x_0)$ be the solution to $x' = Ax$ with initial value $\phi(0, x_0) = x_0$.

(a) Show that for each fixed t,

$$\lim_{y_0 \to x_0} \phi(t, y_0) = \phi(t, x_0).$$

This means solutions are continuous in initial conditions. (*Hint*: Suppose A is diagonal.)

(b) Improve (a) by finding constants $A \geq 0$, $k \geq 0$ such that

$$|\phi(t, y_0) - \phi(t, x_0)| \leq Ae^{kt}|y_0 - x_0|.$$

(*Hint*: Theorem 2.)

6. Consider a second order differential equation

(*) $\qquad x'' + bx' + cx = 0; \qquad b$ and c constant.

(a) By examining the equivalent first order system

$$x' = y,$$
$$y' = -cx - by,$$

show that if $b^2 - 4c > 0$, then (*) has a unique solution $x(t)$ for every initial condition of the form

$$x(0) = u, \qquad x'(0) = v.$$

(b) If $b^2 - 4c > 0$, what assumption about b and c ensures that

$$\lim_{t\to\infty} x(t) = 0$$

for every solution $x(t)$?

(c) Sketch the graphs of the three solutions of

$$x'' - 3x' + 2x = 0$$

for the initial conditions

$$x(0) = 1, \qquad x'(0) = -1, 0, 1.$$

7. Let a 2×2 matrix A have real, distinct eigenvalues λ, μ. Suppose an eigenvector of λ is $(1, 0)$ and an eigenvector of μ is $(1, 1)$. Sketch the phase portraits of $x' = Ax$ for the following cases:

(a) $0 < \lambda < \mu$; (b) $0 < \mu < \lambda$; (c) $\lambda < \mu < 0$;
(d) $\lambda < 0 < \mu$; (e) $\lambda = 0; \mu > 0$.

§4. Complex Eigenvalues

A class of operators that have no real eigenvalues are the planar operators $T_{a,b}$: $\mathbf{R}^2 \to \mathbf{R}^2$ represented by matrices of the form $A_{a,b} = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$, $b \neq 0$. The characteristic polynomial is

$$\lambda^2 - 2a\lambda + (a^2 + b^2),$$

where roots are

$$a + ib, \qquad a - ib; \qquad i = \sqrt{-1}.$$

We interpret $T_{a,b}$ geometrically as follows. Introduce the numbers r, θ by

$$r = (a^2 + b^2)^{1/2},$$
$$\theta = \arccos\left(\frac{a}{r}\right), \qquad \cos\theta = \frac{a}{r}.$$

Then: Providing $b > 0$, $T_{a,b}$ *is a counterclockwise rotation through* θ *radians followed by a stretching (or shrinking) of the length of each vector by a factor of* r.

That is, if R_θ denotes rotation through θ radians, then

$$T_{a,b}(x) = rR_\theta(x) = R_\theta(rx).$$

To see this first observe that

$$a = r\cos\theta, \qquad b = r\sin\theta.$$

In the standard basis, the matrix of R_θ is

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix};$$

the matrix of scalar multiplication by r is $rI = \left[\begin{smallmatrix} r & 0 \\ 0 & r \end{smallmatrix}\right]$. The equality

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix} = \begin{bmatrix} r & 0 \\ 0 & r \end{bmatrix}\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

yields our assertion.

There is another algebraic interpretation of $T_{a,b}$. Identify the plane $\mathbf{R}^2$ with the field of *complex numbers* under the identification

$$(x, y) \leftrightarrow x + iy.$$

Then with this identification, *the operator* $T_{a,b}$ *corresponds to multiplication by* $a + ib$:

$$\begin{array}{ccc} (x, y) & \longleftrightarrow & x + iy \\ \text{operate by } T_{a,b} \uparrow & & \uparrow \text{ multiply by } a + ib \\ (ax - by, bx + ay) & \longleftrightarrow & (ax - by) + i(bx + ay) \end{array}$$

Notice also that r is the norm (absolute value) of $a + bi$ and θ is its argument. Readers familiar with complex functions will recall the formula $a + ib = re^{i\theta}$ (see Appendix I).

The geometric interpretation of $T_{a,b}$ makes it easy to compute with. For example, to compute the pth power of $T_{a,b}$:

$$(T_{a,b})^p = (rI)^p(R_\theta)^p = (r^pI)(R_{p\theta})$$

$$= \begin{bmatrix} r^p\cos p\theta & -r^p\sin p\theta \\ r^p\sin p\theta & r^p\cos p\theta \end{bmatrix}.$$

Next, we consider the operator T on $\mathbf{R}^2$ where the matrix is $\left[\begin{smallmatrix} 0 & -2 \\ 1 & 2 \end{smallmatrix}\right]$. The characteristic polynomial is $\lambda^2 - 2\lambda + 2$, where roots are

$$1 + i, \qquad 1 - i.$$

T does not correspond to multiplication by a complex number since its matrix is not of the form $A_{a,b}$. But it is possible to introduce new coordinates in $\mathbf{R}^2$—that is, to find a new basis—giving T a matrix $A_{a,b}$.

Let (x_1, x_2) be the standard coordinates in $\mathbf{R}^2$. Make the substitution

$$x_1 = y_1 + y_2,$$

$$x_2 = -y_1,$$

so that the new coordinates are given by

$$y_1 = -x_2,$$

$$-y_2 = x_1 + x_2.$$

The matrix of T in the y-coordinates is $\left[\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right] = A_{1,1}$. For this matrix $r = \sqrt{2}$, $\theta = \pi/4$. Therefore in the (y_1, y_2)-plane T is rotation through $\pi/4$ followed with stretching by $\sqrt{2}$. In the *original* coordinates (x_1, x_2), T is a kind of "elliptical rotation" followed by the $\sqrt{2}$-stretch. If vectors in $\mathbf{R}^2$ are identified with complex numbers via the y-coordinates—the vector whose y-coordinates are (y_1, y_2) becomes $y_1 + iy_2$—then T corresponds to multiplication by $1 + i$.

This shows that although T is not diagonalizable, coordinates can be introduced in which T has a simple geometrical interpretation: a rotation followed by a uniform stretch. Moreover, the amount of the rotation and stretch can be deduced from the roots of the characteristic polynomial, since $\pi/4 = \arg(1 + i)$, $\sqrt{2} = |1 + i|$.

We shall explain in Chapter 4, Section 3 how the new coordinates were found.

We show now how the complex structure on $\mathbf{R}^2$ (that is, the identification of $\mathbf{R}^2$ with $\mathbf{C}$) may be used to solve a corresponding class of differential equations.

Consider the system

$$\frac{dx}{dt} = ax - by, \tag{1}$$

$$\frac{dy}{dt} = bx + ay.$$

We use complex variables to formally find a solution, check that what we have found solves (1), and postpone the uniqueness proof (but see Problem 5).

Thus replace (x, y) by $x + iy = z$, and $\left[\begin{smallmatrix} a & -b \\ b & a \end{smallmatrix}\right]$ by $a + bi = \mu$. Then (1) becomes

$$z' = \mu z. \tag{2}$$

Following the lead from the beginning of Chapter 1, we write a solution for (2), $z(t) = Ke^{\mu t}$. Let us interpret this in terms of complex and real numbers. Write the complex number K as $u + iv$ and set $z(t) = x(t) + iy(t)$, $e^{\mu t} = e^{at}e^{ibt}$. A standard formula from complex numbers (see Appendix I) says that $e^{ibt} = \cos bt + i\sin bt$. Putting this information together and taking real and imaginary parts we

obtain

$$x(t) = ue^{ta}\cos tb - ve^{ta}\sin tb, \tag{3}$$
$$y(t) = ue^{ta}\sin tb + ve^{ta}\cos tb.$$

The reader who is uneasy about the derivation of (3) can regard the preceding paragraph simply as motivation for the formulas (3); it is easy to verify directly by differentiation that (3) indeed provides a solution to (1). On the other hand, all the steps in the derivation of (3) are justifiable.

We have just seen how introduction of complex variables can be an aid in solving differential equations. Admittedly, this use was in a very special case. However, many systems not in the form (1) can be brought to that form through a change of coordinates (see Problem 5). In Chapter 4 we shall pursue this idea systematically. At present we merely give an example which was treated before in the Kepler problem of Chapter 2.

Consider the system

$$x' = y, \tag{4}$$
$$y' = -b^2x; \qquad b > 0.$$

The corresponding matrix is

$$A = \begin{bmatrix} 0 & 1 \\ -b^2 & 0 \end{bmatrix},$$

whose eigenvalues are $\pm bi$. It is natural to ask whether A can be put in the form

$$B = \begin{bmatrix} 0 & -b \\ b & 0 \end{bmatrix}$$

through a coordinate change. The answer is yes; without explaining how we discovered them (this will be done in Chapter 4), we introduce new coordinates (u, v) by setting $x = v$, $y = bu$. Then

$$u' = \frac{1}{b}y' = -bv,$$
$$v' = x' = bu.$$

We have already solved the system

$$u' = -bv,$$
$$v' = bu;$$

the solution with $(u(0), v(0)) = (u_0, v_0)$ is

$$u(t) = u_0\cos tb - v_0\sin tb,$$
$$v(t) = u_0\sin tb + v_0\cos tb.$$

Therefore the solution to (4) with initial condition

$$(x(0), y(0)) = (x_0, y_0)$$

is

$$x(t) = \frac{y_0}{b}\sin tb + x_0\cos tb,$$
$$y(t) = y_0\cos tb - bx_0\sin tb,$$

as can be verified by differentiation.

We can put this solution in a more perspicuous form as follows. Let $C = [(y_0/b)^2 + x_0^2]^{1/2}$ and write, assuming $C \neq 0$,

$$\frac{y_0}{b} = -uC, \qquad x_0 = vC.$$

Then $u^2 + v^2 = 1$, and

$$x(t) = C[v\cos tb - u\sin tb].$$

Let $t_0 = b^{-1}\arccos v$, so that

$$\cos bt_0 = v, \qquad \sin bt_0 = u.$$

Then $x(t) = C(\cos bt\cos bt_0 - \sin bt\sin bt_0)$, or

$$x(t) = C\cos b(t - t_0); \tag{5}$$

and

$$y(t) = bC\sin b(t - t_0) \tag{6}$$

as the reader can verify; C and t_0 are arbitrary constants.

From (5) and (6) we see that

$$\frac{x^2}{c^2} + \frac{y^2}{(bc)^2} = 1.$$

Thus the solution curve $(x(t), y(t))$ goes round and round an ellipse.

Returning to the system (4), the reader has probably recognized that it is equivalent to the second order equation on $\mathbf{R}$

$$x'' + b^2x = 0, \tag{7}$$

obtained by differentiating the first equation of (4) and then substituting the second. This is the famous equation of "simple harmonic motion," whose general solution is (5).

PROBLEMS

1. Solve the following initial value problems.

(a) $x' = -y$, $y' = x$; $x(0) = 1, y(0) = 1$.

(b) $x_1' = -2x_2$, $x_2' = 2x_1$; $x_1(c) = 0, x_2(0) = 2$.

(c) $x' = y$, $y' = -x$; $x(0) = 1, y(0) = 1$.

(d) $x' = Ax$, $x(0) = (3, -9)$;
$$A = \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}.$$

2. Sketch the phase portraits of each of the differential equations in Problem 1.

3. Let $A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$ and let $x(t)$ be a solution to $x' = Ax$, not identically 0. The curve $x(t)$ is of the following form:
(a) a circle if $a = 0$;
(b) a spiral inward toward (0, 0) if $a < 0$, $b \neq 0$;
(c) a spiral outward away from (0, 0) if $a > 0$, $b \neq 0$.
What effect has the sign of b on the spirals in (b) and (c)? What is the phase portrait if $b = 0$?

4. Sketch the phase portraits of:

(a) $x' = -2x$; $y' = 2z$; $z' = -2y$.

(b) $x' = -x + z$; $y' = 3y$; $z' = -x - z$.

Which solutions tend to 0 as $t \to \infty$?

5. Let A be a 2×2 matrix whose eigenvalues are the complex numbers $\alpha \pm \beta i$, $\beta \neq 0$. Let $B = \begin{bmatrix} a & -\beta \\ \beta & a \end{bmatrix}$. Show there exists an invertible matrix Q with $QAQ^{-1} = B$, as follows:
(a) Show that the determinant of the following 4×4 matrix is 0:
$$\begin{bmatrix} A - \alpha I & -\beta I \\ \beta I & A - \alpha I \end{bmatrix},$$
where $I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.
(b) Show that there exists a 2×2 matrix Q such that $AQ = QB$.
(*Hint*: Write out the above equation in the four entries of $Q = [q_{ij}]$ Show that the resulting system of four-linear homogeneous equations in the four unknowns q_{ij} has the coefficient matrix of part (a).)
(c) Show that Q can be chosen invertible.
Therefore the system $x' = Ax$ has unique solutions for given initial conditions.

6. Let $A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$. The solutions of $x' = Ax$ depend continuously on initial values. (See Problem 5, Section 3.)

7. Solve the initial value problem

$$x' = -4y, \quad y' = x; \quad x(0) = 0, \quad y(0) = -7.$$

Chapter 4

Linear Systems with Constant Coefficients and Complex Eigenvalues

As we saw in the last section of the preceding chapter, complex numbers enter naturally in the study and solution of real ordinary differential equations. In general the study of operators of complex vector spaces *facilitates* the solving of linear differential equations. The first part of this chapter is devoted to the linear algebra of complex vector spaces. Subsequently, methods are developed to study almost all first order linear ordinary differential equations with constant coefficients, including those whose associated operator has distinct, though perhaps nonreal, eigenvalues. The meaning of "almost all" will be made precise in Chapter 7.

§1. Complex Vector Spaces

In order to gain a deeper understanding of linear operators (and hence of linear differential equations) we have to find the geometric significance of complex eigenvalues. This is done by extending an operator T on a (real) vector space E to an operator T_C on a complex vector space E_C. Complex eigenvalues of T are associated with complex eigenvectors of T_C. We first develop complex vector spaces.

The definitions and elementary properties of $\mathbf{R}^n$ and (real) vector spaces go over directly to $\mathbf{C}^n$ and complex vector spaces by systematically replacing the real numbers R with complex numbers **C**. We make this more precise now.

Complex Cartesian space $\mathbf{C}^n$ is the set all n-tuples $z = (z_1, \ldots, z_n)$ of complex numbers (see Appendix I for the definition of complex numbers). We call z in $\mathbf{C}^n$ a complex vector or sometimes a point in $\mathbf{C}^n$. Complex vectors are added exactly like vectors in $\mathbf{R}^n$ (see Chapter 1, Section 2). Also, if λ is a complex number and

$z = (z_1, \ldots, z_n)$ is in $\mathbf{C}^n$, then λz is the vector $(\lambda z_1, \ldots, \lambda z_n)$; this is scalar multiplication. Note that $\mathbf{R}^n$ is contained naturally in $\mathbf{C}^n$ as the set of all $(z_1, \ldots, z_n)$, where each z_i is real.

The axioms VS1, VS2 of Section 1A of Chapter 3 are valid for the operations we have just defined for $\mathbf{C}^n$. They define the *complex vector space* structure on $\mathbf{C}^n$.

As in Section 1B, Chapter 3, a nonempty subset F of $\mathbf{C}^n$ is called a subspace or a (complex) linear subspace if it is closed under the operations of addition and scalar multiplication in $\mathbf{C}^n$. The notions of trivial subspace, proper subspace, subspace of a (complex) subspace are defined as in the real case; the same is true for the concept of linear map $T: F_1 \to F_2$ between subspaces F_1, F_2 of $\mathbf{C}^n$. One replaces real scalars by complex scalars (that is, complex numbers) everywhere. A *complex vector space* will mean a subspace of $\mathbf{C}^n$.

The material on kernels and images of linear maps of complex vector spaces goes over directly from the real case as well as the facts about bases, dimension, coordinates. Propositions 1, 2, and 3 of Section 1B, Chapter 3, are all valid for the complex case. In fact, all the algebraic properties of real vector spaces and their linear maps carry over to complex vector spaces and their linear maps. In particular, the determinant of a complex operator T, or a complex $n \times n$ matrix, is defined (in **C**). It is zero if and only if T has a nontrivial kernel.

Consider now an operator on $\mathbf{C}^n$, or more generally, an operator T on a complex vector space $F \subset \mathbf{C}^n$. Thus $T: F \to F$ is a linear map and we may proceed to study its eigenvalues and eigenvectors as in Section 2 of Chapter 3. An eigenvalue λ of T is a complex number such that $Tv = \lambda v$ has a nonzero solution $v \in F$. The vector v of F is called an *eigenvector* belonging to λ. This is exactly analogous to the real case. The methods for finding real eigenvalues and eigenvectors apply to this complex case.

Given a complex operator T as above, one associates to it a polynomial

$$p(\lambda) = \mathrm{Det}(T - \lambda I)$$

(now with complex coefficients) such that the degree of $p(\lambda)$ is the dimension of F and the roots of p are exactly the eigenvalues of T.

The proof of Theorem 1 of Section 2 in the previous chapter applies to yield:

Theorem *Let $T: F \to F$ be an operator on an n-dimensional complex vector space F. If the characteristic polynomial has distinct roots, then T can be diagonalized.*

This implies that when these roots are distinct, then one may find a basis $\{e_1, \ldots, e_n\}$ of eigenvectors for T so that if $z = \sum_{j=1}^n z_j e_j$ is in F, then $Tz = \sum_{j=1}^n \lambda_j z_j e_j$; e_j is the eigenvector belonging to the (complex) eigenvalues λ_j.

Observe that the above theorem is stronger than the corresponding theorem in the real case. The latter demanded the further substantial condition that the roots of the characteristic polynomial be real.

Say that an operator T on a complex vector space is *semisimple* if it is diagonal-

izable. Thus by the theorem above T is semisimple if its characteristic polynomial has distinct roots (but not conversely as we shall see in Chapter 6).

As we have noted, $\mathbf{R}^n \subset \mathbf{C}^n$. We consider now more generally the relations between vector spaces in $\mathbf{R}^n$ and complex vector spaces in $\mathbf{C}^n$. Let F be a complex subspace of $\mathbf{C}^n$. Then $F_{\mathbf{R}} = F \cap \mathbf{R}^n$ is the set of all n-tuples $(z_1, \ldots, z_n)$ that are in F and are real. Clearly, $F_{\mathbf{R}}$ is closed under the operations of addition as well as scalar multiplication by *real* numbers. Thus $F_{\mathbf{R}}$ is a real vector space (subspace of $\mathbf{R}^n$).

Consider now the converse process. Let $E \subset \mathbf{R}^n$ be a subspace and let $E_{\mathbf{C}}$ be the subset of $\mathbf{C}^n$ obtained by taking all linear combinations of vectors in E, with complex coefficients. Thus

$$E_{\mathbf{C}} = \{z \in \mathbf{C}^n \mid z = \sum_{i=1}^{k} \lambda_i z_i,\ z_i \in E,\ \lambda_i \in \mathbf{C}\}$$

and $E_{\mathbf{C}}$ is a complex subspace of $\mathbf{C}^n$. Note that $(E_{\mathbf{C}})_{\mathbf{R}} = E$. We call $E_{\mathbf{C}}$ the *complexification* of E and $F_{\mathbf{R}}$ the *space of real vectors* in F.

In defining $E_{\mathbf{C}}$, $F_{\mathbf{R}}$ we used the fact that all the spaces considered were subsets of $\mathbf{C}^n$. The essential element of structure here, besides the algebraic structure, is the operation of complex *conjugation*.

Recall if $z = x + iy$ is a complex number, then $\bar{z} = x - iy$. We often write $\bar{z} = \sigma(z)$ so that $\sigma: \mathbf{C} \to \mathbf{C}$ as a map with the property $\sigma^2 = \sigma \circ \sigma = \text{identity}$. The set of fixed points of σ, that is, the set of z such that $\sigma(z) = z$, is precisely the set of real numbers in $\mathbf{C}$.

This operation σ, or conjugation, can be extended immediately to $\mathbf{C}^n$ by defining $\sigma: \mathbf{C}^n \to \mathbf{C}^n$ by conjugating each coordinate. That is,

$$\sigma(z_1, \ldots, z_n) = (\bar{z}_1, \ldots, \bar{z}_n).$$

For this extension, *the set of fixed points is* $\mathbf{R}^n$.

Note also that if F is a complex subspace of $\mathbf{C}^n$, such that $\sigma F = F$, then the set of fixed points of σ on F is precisely $F_{\mathbf{R}}$. This map σ plays a crucial role in the relation between real and complex vector spaces.

Let $F \subset \mathbf{C}^n$ be a σ-invariant linear subspace of $\mathbf{C}^n$. Then it follows that for $v \in F$, $\lambda \in \mathbf{C}$, $\sigma(\lambda v) = \sigma(\lambda)\sigma(v)$ or if we write $\sigma(w) = \bar{w}$ for $w \in F$, $\overline{\lambda v} = \bar{\lambda}\bar{v}$. Thus σ is *not* complex linear. However, $\sigma(v + w) = \sigma(v) + \sigma(w)$.

It follows that for any subspace $F \subset \mathbf{C}^n$,

$$F_{\mathbf{R}} = \{z \in F \mid \sigma(z) = z\}.$$

In terms of σ it is easy to see when a subspace $F \subset \mathbf{C}^n$ can be decomplexified, that is, expressed in the form $F = E_{\mathbf{C}}$ for some subspace $E \subset \mathbf{R}^n$: F can be decomplexified if and only if $\sigma(F) \subset F$. For if $\sigma(F) \subset F$, then $x - iy \in F$ whenever $x + iy \in F$ with $x, y \in \mathbf{R}^n$; so $x \in F$ because

$$x = \tfrac{1}{2}[(x + iy) + (x - iy)].$$

Similarly, $y \in F$. It follows easily that $F = F_{\mathbf{RC}}$, that is, F is the complexification of the space of real vectors in F. The converse is trivial.

Just as every subspace $E \subset \mathbf{R}^n$ has a complexification $E_{\mathbf{C}} \subset \mathbf{C}^n$, every operator $T: E \to E$ has an extension to a complex linear operator

$$T_{\mathbf{C}}: E_{\mathbf{C}} \to E_{\mathbf{C}},$$

called the *complexification* of T. To define $T_{\mathbf{C}}$, $z \in E_{\mathbf{C}}$, let

$$z = \sum \lambda_j x_j; \qquad \lambda_j \in \mathbf{C}, \quad x_j \in E. \tag{1}$$

Then

$$T_{\mathbf{C}} z = \sum \lambda_j T x_j.$$

It is easy to see that this definition does not depend on the choice of the representation (1).

If $\{e_1, \ldots, e_k\} = \mathcal{B}$ is a basis for E, it is also a basis for the complex vector space $E_{\mathbf{C}}$; and the $\mathcal{B}$-matrix for $T_{\mathbf{C}}$ is the same as the $\mathcal{B}$-matrix for T.

In particular, if $T \in L(\mathbf{R}^n)$ is represented by an $n \times n$ matrix A (in the usual way), then $T_{\mathbf{C}} \in L(\mathbf{C}^n)$ is also represented by A.

The question arises as to when an operator $Q: E_{\mathbf{C}} \to E_{\mathbf{C}}$ is the complexification of an operator $T: E \to E$.

Proposition *Let $E \subset \mathbf{R}^n$ be a real vector space and $E_{\mathbf{C}} \subset \mathbf{C}^n$ its complexification. If $Q \in L(E_{\mathbf{C}})$ then $Q = T_{\mathbf{C}}$ for some $T \in L(E)$ if and only if*

$$Q\sigma = \sigma Q,$$

where $\sigma: E_{\mathbf{C}} \to E_{\mathbf{C}}$ is conjugation.

Proof. If $Q = T_{\mathbf{C}}$, we leave it to the reader to prove that $Q\sigma = \sigma Q$. Conversely, assume Q commutes with σ. Then $Q(E) \subset E$; for if $x \in E$, then $\sigma x = x$, hence

$$\sigma Q x = Q \sigma x = Q x$$

so

$$Qx \in \{y \in E_{\mathbf{C}} \mid \sigma y = y\} = E_{\mathbf{CR}} = E.$$

Let $Q \mid E = T \in L(E)$; it is clear from the definition of $T_{\mathbf{C}}$ that $T_{\mathbf{C}} = Q$.

We close this section with a property that will be very important in later chapters. An operator T on a real vector space E is *semisimple* if its complexification $T_{\mathbf{C}}$ is a diagonalizable operator on $E_{\mathbf{C}}$. Then the theorem proved earlier implies that a sufficient (but not necessary) condition for semisimplicity is that the characteristic polynomial should have distinct roots.

PROBLEMS

1. Let $F \subset \mathbf{C}^2$ be the subspace spanned by the vector $(1, i)$.
   (a) Prove that F is not invariant under conjugation and hence is not the complexification of any subspace of $\mathbf{R}^2$.
   (b) Find $F_{\mathbf{R}}$ and $(F_{\mathbf{R}})_{\mathbf{C}}$.

2. Let $E \subset \mathbf{R}^n$ and $F \subset \mathbf{C}^n$ be subspaces. What relations, if any, exist between $\dim E$ and $\dim E_{\mathbf{C}}$? Between $\dim F$ and $\dim F_{\mathbf{R}}$?
3. If $F \subset \mathbf{C}^n$ is any subspace, what relation is there between F and $F_{\mathbf{RC}}$?
4. Let E be a real vector space and $T \in L(E)$. Show that $(\operatorname{Ker} T)_{\mathbf{C}} = \operatorname{Ker}(T_{\mathbf{C}})$, $(\operatorname{Im} T)_{\mathbf{C}} = \operatorname{Im}(T_{\mathbf{C}})$, and $(T^{-1})_{\mathbf{C}} = (T_{\mathbf{C}})^{-1}$ if T is invertible.

§2. Real Operators with Complex Eigenvalues

We move toward understanding the linear differential equation with constant coefficients

$$\frac{dx}{dt} = Tx,$$

where T is an operator on $\mathbf{R}^n$. For this purpose, we study further the eigenvalues and eigenvectors of T. This was done thoroughly in Chapter 3 assuming that all the eigenvalues were distinct and real. Now we drop the hypothesis that the eigenvalues must be real.

Proposition. *If T is an operator on a real vector space E, then the set of its eigenvalues is preserved under complex conjugation. Thus if λ is an eigenvalue so is $\bar{\lambda}$. Consequently, we may write the eigenvalues of T as*

$$\lambda_1, \ldots, \lambda_r, \qquad \textit{all real;}$$

$$\mu_1, \bar{\mu}_1, \ldots, \mu_s, \bar{\mu}_s, \qquad \textit{all nonreal.}$$

Proof. First, observe that the eigenvalues of T coincide with the eigenvalues of its complexification $T_{\mathbf{C}}$ because both T and $T_{\mathbf{C}}$ have the same characteristic polynomial. Let λ be an eigenvalue of $T_{\mathbf{C}}$ and φ a corresponding eigenvector in $E_{\mathbf{C}}$, so $T_{\mathbf{C}}\varphi = \lambda\varphi$. Applying the conjugation operation σ to both sides, we find

$$\sigma(T_{\mathbf{C}}\varphi) = \bar{\lambda}\bar{\varphi}.$$

But, by the proposition of Section 1,

$$\sigma(T_{\mathbf{C}}\varphi) = T_{\mathbf{C}}\sigma(\varphi) = T_{\mathbf{C}}(\bar{\varphi}).$$

Hence

$$T_{\mathbf{C}}\bar{\varphi} = \bar{\lambda}\bar{\varphi}.$$

In other words, $\bar{\lambda}$ is an eigenvalue of $T_{\mathbf{C}}$ with corresponding eigenvector $\bar{\varphi}$. This proves the proposition. (Another proof is based on the fact that the characteristic polynomial of T has real coefficients, so the roots occur in conjugate pairs.)

The basic properties of real operators are contained in the following three theorems.

Theorem 1 *Let $T: E \to E$ be a real operator with distinct eigenvalues listed as in the previous proposition. Then E and T have a direct sum decomposition (see Section 1F of Chapter 3),*

$$E = E_a \oplus E_b, \qquad T = T_a \oplus T_b, \qquad T_a: E_a \to E_a, \qquad T_b: E_b \to E_b,$$

where T_a has real eigenvalues and T_b nonreal eigenvalues.

For the proof we pass to the complexification $T_{\mathbf{C}}$ and apply the theorem of the preceding section together with the above proposition. This yields a basis for $E_{\mathbf{C}}$ $(e_1, \ldots, e_r, f_1, \bar{f}_1, \ldots, f_s, \bar{f}_s)$ of eigenvectors of $T_{\mathbf{C}}$ corresponding to the eigenvalues $(\lambda_1, \ldots, \lambda_r, \mu_1, \bar{\mu}_1, \ldots, \mu_s, \bar{\mu}_s)$.

Now let F_a be the complex subspace of $E_{\mathbf{C}}$ spanned by $\{e_1, \ldots, e_r\}$ and F_b be the subspace spanned by $\{f_1, \bar{f}_2, \ldots, f_s, \bar{f}_s\}$. Thus F_a and F_b are invariant subspaces for $T_{\mathbf{C}}$ on $E_{\mathbf{C}}$ and form a direct sum decomposition for $E_{\mathbf{C}}$,

$$E_{\mathbf{C}} = F_a \oplus F_b.$$

Moreover F_a and F_b are invariant under complex conjugation. Set $E_a = E \cap F_a$ and $E_b = E \cap F_b$; then F_a, F_b are the complexifications of E_a, E_b, and $E = E_a \oplus E_b$. It is easy to see that E_a and E_b have the required properties.

Theorem 1 reduces the study of such T to T_a and T_b. The previous chapter analyzed T_a.

We remark that Theorem 1 provides an "uncoupling" of the differential equation

$$\frac{dx}{dt} = Tx$$

mentioned at the beginning of the section. We may rewrite this equation as a pair of equations

$$\frac{dx_a}{dt} = T_a x_a,$$

$$\frac{dx_b}{dt} = T_b x_b,$$

where T_a, T_b are as above and $x_a \in E_a$, $x_b \in E_b$.

We proceed to the study of the operator T_b.

Theorem 2 *Let $T: E \to E$ be an operator on a real vector space with distinct nonreal eigenvalues $(\mu_1, \bar{\mu}_1, \ldots, \mu_s, \bar{\mu}_s)$. Then there is an invariant direct sum decomposition for E and a corresponding direct sum decomposition for T,*

$$E = E_1 \oplus \cdots \oplus E_s,$$

$$T = T_1 \oplus \cdots \oplus T_s,$$

such that each E_i is two dimensional and $T_i \in L(E_i)$ has eigenvalues μ_i, $\bar{\mu}_i$.

For the proof of Theorem 2, simply let F_i be the complex subspace of E_C spanned by the eigenvectors, f_i, $\bar{f}_i$ corresponding to the eigenvalues μ_i, $\bar{\mu}_i$. Then let E_i be $F_i \cap E$. The rest follows.

Theorems 1 and 2 reduce in principle the study of an operator with distinct eigenvalues to the case of an operator on a real two-dimensional vector space with nonreal eigenvalues.

Theorem 3 *Let T be an operator on a two-dimensional vector space $E \subset \mathbf{R}^n$ with nonreal eigenvalues μ, $\bar{\mu}$, $\mu = a + ib$. Then there is a matrix representation A for T*

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

The study of such a matrix A and the corresponding differential equation on $\mathbf{R}^2$, $dx/dt = Ax$, was the content of Chapter 3, Section 4.

We now give the proof of Theorem 3.

Let $T_C: E_C \to E_C$ be the complexification of T. Since T_C has the same eigenvalues as T, there are eigenvectors φ, $\bar{\varphi}$ in E_C belonging to μ, $\bar{\mu}$, respectively.

Let $\varphi = u + iv$ with $u, v \in \mathbf{R}^n$. Then $\bar{\varphi} = u - iv$. Note that u and v are in E_C, for

$$u = \tfrac{1}{2}(\varphi + \bar{\varphi}), \qquad v = \tfrac{1}{2}(\bar{\varphi} - \varphi).$$

Hence u and v are in $E_C \cap \mathbf{R}^n = E$. Moreover, it is easy to see that u and v are independent (use the independence of φ, $\bar{\varphi}$). Therefore $\{v, u\}$ is a basis for E.

To compute the matrix of T in this basis we start from

$$\begin{aligned} T_C(u + iv) &= (a + bi)(u + iv) \\ &= (-bv + au) + i(av + bu). \end{aligned}$$

Also,

$$T_C(u + iv) = Tu + iTv.$$

Therefore

$$\begin{aligned} Tv &= av + bu, \\ Tu &= -bv + au. \end{aligned}$$

This means that the matrix of T in the basis $\{v, u\}$ is $\begin{bmatrix} a & b \\ -b & a \end{bmatrix}$, completing the proof.

In the course of the proof we have found the following interpretation of a complex eigenvalue of a real operator $T \in L(E)$, $E \subset \mathbf{R}^n$:

Corollary *Let $\varphi \in E_C$ be an eigenvector of T belonging to $a + ib$, $b \neq 0$. If $\varphi = u + iv \in \mathbf{C}^n$, then $\{v, u\}$ is a basis for E giving φ the matrix $\begin{bmatrix} a & b \\ -b & a \end{bmatrix}$.*

Note that u and v can be obtained directly from φ and σ (without reference to $\mathbf{C}^n$) by the formulas in the proof of Theorem 3.

PROBLEM

For each of the following operators T on $\mathbf{R}^3$ find an invariant two-dimensional $E \subset \mathbf{R}^3$ and a basis for E giving $T \mid E$ a matrix of the form $\begin{bmatrix} a & b \\ -b & a \end{bmatrix}$:

(a) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & -2 \\ 0 & 1 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 0 & 0 & 15 \\ 1 & 0 & -17 \\ 0 & 1 & 7 \end{bmatrix}$

§3. Application of Complex Linear Algebra to Differential Equations

Consider the linear differential equation on $\mathbf{R}^n$

$$\frac{dx}{dt} = Tx, \tag{1}$$

where T is an operator on $\mathbf{R}^n$ (or equivalently, an $n \times n$ matrix). Suppose that T has n distinct eigenvalues. Then Theorems 1, 2, and 3 of the previous section apply to uncouple the equation and, after finding the new basis, one can obtain the solution. Letting $E = \mathbf{R}^n$, we first apply Theorem 1 to obtain the following system, equivalent to (1):

$$\frac{dx_a}{dt} = T_a x_a, \tag{2a}$$

$$\frac{dx_b}{dt} = T_b x_b. \tag{2b}$$

Here

$$T = T_a \oplus T_b, \qquad x = (x_a, x_b) \in E_a \oplus E_b = E,$$

T_a has real eigenvalues, and T_b nonreal eigenvalues.

Note that (2a) and (2b) *are equations defined not on* $\mathbf{R}^n$, but *on subspaces* E_a and E_b. But our definitions and discussion of differential equations apply just as well to subspaces of $\mathbf{R}^n$. To find explicit solutions to the original equations, bases for these subspaces must be found. This is done by finding eigenvectors of the complexification of T, as will be explained below.

If we obtain solutions and properties of (2a) and (2b) separately, corresponding information is gained for (2) and (1). Furthermore, (2a) received a complete discussion in Chapter 3, Section 3. Thus in principle it is sufficient to give an analysis of (2b). To this end, Theorem 2 of Section 2 applies to give the following system,

equivalent to (2b):

$$\frac{dy_i}{dt} = T_i y_i, \qquad i = 1, \ldots, s, \tag{3}$$

where $T = T_1 \oplus \cdots \oplus T_s$, $y = (y_1, \ldots, y_s) \in E_b = E_1 \oplus \cdots \oplus E_s$ and each E_i has two dimensions.

Thus (2b) and hence (2), (1) are reduced to the study of the equation

$$\frac{dy_i}{dt} = T_i y_i \qquad \text{on two-dimensional } E_i, \tag{4}$$

where each T_i has nonreal eigenvalues. Finally, Theorem 3 of Section 2 applies to put (4) in the form of the equation analyzed in Section 4 of Chapter 3.

Example 1 Consider the equation

$$x_1' = -2x_2,$$
$$x_2' = x_1 + 2x_2,$$

or

$$x' = Ax, \qquad x = (x_1, x_2), \qquad A = \begin{bmatrix} 0 & -2 \\ 1 & 2 \end{bmatrix}.$$

This is the matrix considered in Chapter 3, Section 4. The eigenvalues of A are $\lambda = 1 + i$, $\bar{\lambda} = 1 - i$.

A complex eigenvector belonging to $1 + i$ is found by solving the equation

$$(A - (1 + i))w = 0$$

for $w \in \mathbf{C}^2$,

$$\begin{bmatrix} -1 - i & -2 \\ 1 & 1 - i \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = 0;$$

or

$$(-1 - i)w_1 - 2w_2 = 0,$$
$$w_1 + (1 - i)w_2 = 0.$$

The first equation is equivalent to the second, as is seen by multiplying the second by $(-1 - i)$. From the second equation we see that the solutions are all (complex) multiples of any nonzero complex vector w such that $w_1 = (-1 + i)w_2$; for example, $w_2 = -i$, $w_1 = 1 + i$. Thus

$$w = (1 + i, -i) = (1, 0) + i(1, -1) = u + iv$$

is a complex eigenvector belonging to $1 + i$.

We choose the new basis $\{v, u\}$ for $\mathbf{R}^2 \subset \mathbf{C}^2$, with $v = (1, -1)$, $u = (1, 0)$.

To find new coordinates y_1, y_2 corresponding to this new basis, note that any x can be written

$$x = x_1(1, 0) + x_2(0, 1) = y_1 v + y_2 u = y_1(1, -1) + y_2(1, 0).$$

Thus

$$\left.\begin{aligned} x_1 &= y_1 + y_2, \\ x_2 &= -y_1; \end{aligned}\right\} \quad \text{or} \quad x = Py, \qquad P = \begin{bmatrix} 1 & 1 \\ -1 & 0 \end{bmatrix}.$$

The new coordinates are given by

$$\left.\begin{aligned} y_1 &= -x_2, \\ y_2 &= x_1 + x_2; \end{aligned}\right\} \quad \text{or} \quad y = P^{-1}x, \qquad P^{-1} = \begin{bmatrix} 0 & -1 \\ 1 & 1 \end{bmatrix}.$$

The matrix of A in the y-coordinates is

$$P^{-1}AP = \begin{bmatrix} 0 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & -2 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = B,$$

or $B = A_{1,1}$ in the notation of Section 4, Chapter 3.

Thus, as we saw in that section, our differential equation

$$\frac{dx}{dt} = Ax$$

on $\mathbf{R}^2$, having the form

$$\frac{dy}{dt} = By$$

in the y-coordinates, can be solved as

$$y_1(t) = ue^t \cos t - ve^t \sin t,$$
$$y_2(t) = ue^t \sin t + ve^t \cos t.$$

The original equation has as its general solution

$$x_1(t) = (u + v)e^t \cos t + (u - v)e^t \sin t,$$
$$x_2(t) = -ue^t \cos t + ve^t \sin t.$$

Example 2 Consider on $\mathbf{R}^3$ the differential equation

$$\frac{dx}{dt} = Ax, \qquad A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & -3 \\ 1 & 3 & 2 \end{bmatrix}.$$

The characteristic equation $\mathrm{Det}(A - tI) = 0$ is $(1 - t)((2 - t)^2 + 9) = 0$. Its solutions, the eigenvalues for A, are $\lambda = 1$, $\mu = 2 + 3i$, $\bar{\mu} = 2 - 3i$. Eigenvectors in $\mathbf{C}^3$ for the complexified operator are found by solving the homogeneous systems

of three linear equations,

$$(A - \lambda)e = 0, \qquad \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -3 \\ 1 & 3 & 1 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} = 0;$$

this yields $e = (-10, 3, 1)$. Likewise

$$(A - \mu)w = 0, \qquad \begin{bmatrix} -1 - 3i & 0 & 0 \\ 0 & -3i & -3 \\ 1 & 3 & -3i \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = 0$$

yields $w = (0, i, 1)$. A third eigenvector is $\bar{w} = (0, -i, 1)$.

We now wish to find the matrix P that gives a change of coordinates $x = Py$, $y = P^{-1}x$ where x is in the original coordinate system on $\mathbf{R}^3$ and y corresponds to the basis of eigenvectors. Proposition 4 of Section 1C, Chapter 3, applies.

Thus

$$P = \begin{bmatrix} -10 & 0 & 0 \\ 3 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

Here the columns of P are (e, v, u) where $w = (0, 0, 1) + i(0, 1, 0) = u + iv$. Then

$$P^{-1} = \begin{bmatrix} -\frac{1}{10} & 0 & 0 \\ \frac{3}{10} & 1 & 0 \\ \frac{1}{10} & 0 & 1 \end{bmatrix}$$

and

$$B = P^{-1}AP = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & -3 \\ 0 & 3 & 2 \end{bmatrix}.$$

Now we have transformed our original equation $x' = Ax$ following the outline given in the beginning of this section to obtain

$$y' = By, \qquad B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & -3 \\ 0 & 3 & 2 \end{bmatrix}, \qquad y = P^{-1}x.$$

This can be solved explicitly for y as in the previous example and from this solution one obtains solutions in terms of the original x-coordinates by $x = Py$.

A related approach to the equation (1) is obtained by directly complexifying it, extending (1) to a differential equation on $\mathbf{C}^n$,

$$(1_C) \qquad \frac{dz}{dt} = T_C z, \qquad z \in \mathbf{C}^n.$$

One can make sense of (1_C) as a differential equation either by making definitions directly for derivatives of curves $\mathbf{R} \to \mathbf{C}^n$ or by considering $\mathbf{C}^n$ as $\mathbf{R}^{2n}$, that is,

$$\mathbf{R}^{2n} \to \mathbf{C}^n,$$

$$(x_1, \ldots, x_n, y_1, \ldots, y_n) = (x, y) \to x + iy = z.$$

Application of the theorem of Section 1 diagonalizes T_C, and one may correspondingly rewrite (1_C) as the set of differential equations,

$$\frac{dz_i}{dt} = \lambda_i z_i, \qquad i = 1, \ldots, r;$$

$$\frac{dz_{r+i}}{dt} = \mu_i z_{r+i}, \qquad i = 1, \ldots, s;$$

$$\frac{dw_i}{dt} = \bar{\mu}_i w_i, \qquad i = 1, \ldots, s.$$

(Sometimes $\bar{z}_{r+i}$ is written in place of w_i.) Here z_i, z_{r+i}, w_i are all in one-dimensional complex vector spaces or can be regarded as complex numbers, and $n = r + 2s$. These complex ordinary differential equations may be solved using properties of complex exponentials, as in Section 4 of the previous chapter, obtaining as the general solution:

$$\begin{aligned} z(t) &= (z_1(t), \ldots, z_{r+s}(t), w_1(t), \ldots, w_s(t)) \\ &= (c_1 \exp(\lambda_1 t), \ldots, c_r \exp(\lambda_r t), \\ &\qquad c_{r+1} \exp(\mu_1 t), \ldots, c_{r+s+1} \exp(\bar{\mu}_1 t), \ldots, c_{r+2s} \exp(\bar{\mu}_s t)). \end{aligned}$$

Now it can be checked that if $z(0) \in \mathbf{R}^n$, then $z(t) \in \mathbf{R}^n$ for all t, using formal properties of complex exponentials. This can be a useful approach to the study of (1).

PROBLEM

Solve $x' = Tx$ where T is the operator in (a) and (b) of Problem 1, Section 2.

Chapter 5

Linear Systems and Exponentials of Operators

The object of this chapter is to solve the linear homogeneous system with constant coefficients

$$x' = Ax, \tag{1}$$

where A is an operator on $\mathbf{R}^n$ (or an $n \times n$ matrix). This is accomplished with *exponentials of operators*.

This method of solution is of great importance, although in this chapter we can compute solutions only for special cases. When combined with the operator theory of Chapter 6, the exponential method yields explicit solutions for every system (1).

For every operator A, another operator e^A, called the *exponential of* A, is defined in Section 4. The function $A \to e^A$ has formal properties similar to those of ordinary exponentials of real numbers; indeed, the latter is a special case of the former. Likewise the function $t \to e^{tA}$ ($t \in \mathbf{R}$) resembles the familiar e^{ta}, where $a \in \mathbf{R}$. In particular, it is shown that the solutions of (1) are exactly the maps $x\colon \mathbf{R} \to \mathbf{R}^n$ given by

$$x(t) = e^{tA}K \qquad (K \in \mathbf{R}^n).$$

Thus we establish existence and uniqueness of solution of (1); "uniqueness" means that there is only one solution $x(t)$ satisfying a given initial condition of the form $x(t_0) = K_0$.

Exponentials of operators are defined in Section 3 by means of an infinite series in the operator space $L(\mathbf{R}^n)$; the series is formally the same as the usual series for e^a. Convergence is established by means of a special norm on $L(\mathbf{R}^n)$, the *uniform norm*. Norms in general are discussed in Section 2, while Section 1 briefly reviews some basic topology in $\mathbf{R}^n$.

Sections 5 and 6 are devoted to two less-central types of differential equations. One is a simple inhomogeneous system and the other a higher order equation of one variable. We do not, however, follow the heavy emphasis on higher order equations

of some texts. In geometry, physics, and other kinds of applied mathematics, one seldom encounters naturally any differential equation of order higher than two. Often even the second order equations are studied with more insight after reducing to a first order system (for example, in Hamilton's approach to mechanics).

§1. Review of Topology in $\mathbf{R}^n$

The inner product ("dot product") of vectors x and y in $\mathbf{R}^n$ is

$$\langle x, y\rangle = x_1y_1 + \cdots + x_ny_n.$$

The Euclidean norm of x is $|x| = \langle x, x\rangle^{1/2} = (x_1^2 + \cdots + x_n^2)^{1/2}$. Basic properties of the inner product are

Symmetry: $\langle x, y\rangle = \langle y, x\rangle$;

Bilinearity: $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$,

$$\langle \alpha x, y\rangle = \alpha\langle x, y\rangle, \quad \alpha \in \mathbf{R};$$

Positive definiteness: $\langle x, x\rangle \geq 0$ and

$$\langle x, x\rangle = 0 \text{ if and only if } x = 0.$$

An important inequality is

Cauchy's inequality: $\langle x, y\rangle \leq |x|\,|y|$.

To see this, first suppose $x = 0$ or $y = 0$; the inequality is obvious. Next, observe that for any λ

$$\langle x + \lambda y, x + \lambda y\rangle \geq 0$$

or

$$\langle x, x\rangle + \lambda^2\langle y, y\rangle + 2\lambda\langle x, y\rangle \geq 0.$$

Writing $-\langle x, y\rangle/\langle y, y\rangle$ for λ yields the inequality.

The basic properties of the norm are:

(1) $|x| \geq 0$ and $|x| = 0$ if and only if $x = 0$;
(2) $|x + y| \leq |x| + |y|$;
(3) $|\alpha x| = |\alpha|\,|x|$;

where $|\alpha|$ is the ordinary absolute value of the scalar α. To prove the triangle inequality (2), it suffices to prove

$$|x + y|^2 \leq |x|^2 + |y|^2 + 2|x|\,|y|.$$

Since

$$\begin{aligned} |x + y|^2 &= \langle x + y, x + y\rangle \\ &= |x|^2 + |y|^2 + 2\langle x, y\rangle, \end{aligned}$$

this follows from Cauchy's inequality.

Geometrically, $|x|$ is the length of the vector x and

$$\langle x, y\rangle = |x|\,|y|\cos\theta,$$

where θ is the angle between x and y.

The *distance* between two points $x, y \in \mathbf{R}^n$ is defined to be $|x - y| = d(x, y)$. It is easy to prove:

(4) $|x - y| \geq 0$ and $|x - y| = 0$ if and only if $x = y$;
(5) $|x - z| \leq |x - y| + |y - z|$.

The last inequality follows from the triangle inequality applied to

$$x - z = (x - y) + (y - z).$$

If $\epsilon > 0$ the *ϵ-neighborhood* of $x \in \mathbf{R}^n$ is

$$B_\epsilon(x) = \{y \in \mathbf{R}^n \mid |y - x| < \epsilon\}.$$

A *neighborhood* of x is any subset of $\mathbf{R}^n$ containing an ϵ-neighborhood of x.

A set $X \subset \mathbf{R}^n$ is *open* if it is a neighborhood of every $x \in X$. Explicitly, X is open if and only if for every $x \in X$ there exists $\epsilon > 0$, depending on x, such that

$$B_\epsilon(x) \subset X.$$

A sequence $\{x_k\} = x_1, x_2, \ldots$ in $\mathbf{R}^n$ *converges to the limit* $y \in \mathbf{R}^n$ if

$$\lim_{k\to\infty} |x_k - y| = 0.$$

Equivalently, every neighborhood of y contains all but a finite number of the points of the sequence. We denote this by $y = \lim_{k\to\infty} x_k$ or $x_k \to y$. If $x_k = (x_{k1}, \ldots, x_{kn})$ and $y = (y_1, \ldots, y_n)$, then $\{x_k\}$ converges to y if and only if $\lim_{k\to\infty} x_{kj} = y_j$, $j = 1, \ldots, n$. A sequence that has a limit is called *convergent*.

A sequence $\{x_k\}$ in $\mathbf{R}^n$ is a *Cauchy* sequence if for every $\epsilon > 0$ there exists an integer k_0 such that

$$|x_j - x_k| < \epsilon \quad \text{if } k \geq k_0 \text{ and } j \geq k_0.$$

The following basic property of $\mathbf{R}^n$ is called metric completeness:

A sequence converges to a limit if and only if it is a Cauchy sequence.

A subset $Y \subset \mathbf{R}^n$ is *closed* if every sequence of points in Y that is convergent has its limit in Y. It is easy to see that this is equivalent to: Y is closed if the complement $\mathbf{R}^n - Y$ is open.

Let $X \subset \mathbf{R}^n$ be any subset. A map $f\colon X \to \mathbf{R}^m$ is *continuous* if it takes convergent sequences to convergent sequences. This means: for every sequence $\{x_k\}$ in X with

$$\lim_{k\to\infty} x_k = y \in X,$$

it is true that

$$\lim_{k\to\infty} f(x_k) = f(y).$$

A subset $X \subset \mathbf{R}^n$ is *bounded* if there exists $a > 0$ such that $X \subset B_a(0)$.

A subset X is *compact* if every sequence in X has a subsequence converging to a point in X. The basic *theorem of Bolzano–Weierstrass* says:

A subset of $\mathbf{R}^n$ is compact if and only if it is both closed and bounded.

Let $K \subset \mathbf{R}^n$ be compact and $f\colon K \to \mathbf{R}^m$ be a continuous map. Then $f(K)$ is compact.

A nonempty compact subset of $\mathbf{R}$ has a maximal element and a minimal element. Combining this with the preceding statement proves the familiar result:

Every continuous map $f\colon K \to \mathbf{R}$, defined on a compact set K, takes on a maximum value and a minimum value.

One may extend the notions of distance, open set, convergent sequence, and other topological ideas to vector subspaces of $\mathbf{R}^n$. For example, if E is a subspace of $\mathbf{R}^n$, the distance function $d\colon \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$ restricts to a function $d_E\colon E \times E \to \mathbf{R}$ that also satisfies (4) and (5). Then ϵ-neighborhoods in E may be defined via d_E and thus open sets of E become defined.

§2. New Norms for Old

It is often convenient to use functions on $\mathbf{R}^n$ that are similar to the Euclidean norm, but not identical to it. We define a *norm* on $\mathbf{R}^n$ to be any function $N\colon \mathbf{R}^n \to \mathbf{R}$ that satisfies the analogues of (1), (2), and (3) of Section 1:

(1) $N(x) \geq 0$ and $N(x) = 0$ if and only if $x = 0$;
(2) $N(x + y) \leq N(x) + N(y)$;
(3) $N(\alpha x) = |\alpha|\,N(x)$.

Here are some other norms on $\mathbf{R}^n$:

$$|x|_{\max} = \max\{|x_1|, \ldots, |x_n|\},$$
$$|x|_{\text{sum}} = |x_1| + \cdots + |x_n|.$$

Let $\mathcal{B} = \{f_1, \ldots, f_n\}$ be a basis for $\mathbf{R}^n$ and define the *Euclidean $\mathcal{B}$-norm*:

$$|x|_{\mathcal{B}} = (t_1^2 + \cdots + t_n^2)^{1/2} \quad \text{if} \quad x = \sum_{j=1}^{n} t_j f_j.$$

In other words, $|x|_{\mathcal{B}}$ is the Euclidean norm of x in $\mathcal{B}$-coordinates $(t_1, \ldots, t_n)$.

The $\mathcal{B}$ *max-norm* of x is

$$|x|_{\mathcal{B},\max} = \max\{|t_1|, \ldots, |t_n|\}.$$

The basic fact about norms is the *equivalence of norms*:

Proposition 1 *Let $N: \mathbf{R}^n \to \mathbf{R}$ be any norm. There exist constants $A > 0$, $B > 0$ such that*

$$A|x| \leq N(x) \leq B|x| \tag{4}$$

for all x, where $|x|$ is the Euclidean norm.

Proof. First, consider the max norm. Clearly,

$$(\max|x_j|)^2 \leq \sum_j x_j^2 \leq n(\max|x_j|)^2;$$

taking square roots we have

$$|x|_{\max} \leq |x| \leq \sqrt{n}\,|x|_{\max}.$$

Thus for the max norm we can take $A = 1/\sqrt{n}$, $B = 1$, or, equivalently,

$$\frac{1}{\sqrt{n}}|x| \leq |x|_{\max} \leq |x|.$$

Now let $N: \mathbf{R}^n \to \mathbf{R}$ be any norm. We show that N *is continuous*. We have

$$N(x) = N(\sum x_j e_j) \leq \sum |x_j| N(e_j),$$

where $e_1, \ldots, e_n$ is the standard basis. If

$$\max\{N(e_1), \ldots, N(e_n)\} = M,$$

then

$$N(x) \leq M \sum |x_j| \leq Mn|x|_{\max} \leq Mn|x|.$$

By the triangle inequality,

$$|N(x) - N(y)| \leq N(x - y) \leq Mn|x - y|.$$

This shows that N is continuous; for suppose $\lim x_k = y$ in $\mathbf{R}^n$:

$$|N(x_k) - N(y)| \leq Mn|x_k - y|,$$

so $\lim N(x_k) = N(y)$ in $\mathbf{R}$.

Since N is continuous, it attains a maximum value B and a minimum value A on the closed bounded set

$$\{x \in \mathbf{R}^n \mid |x| = 1\}.$$

Now let $x \in \mathbf{R}^n$. If $x = 0$, (4) is obvious. If $|x| = \alpha \neq 0$, then

$$N(x) = \alpha N(\alpha^{-1}x).$$

Since $|\alpha^{-1}x| = 1$ we have

$$A \leq N(\alpha^{-1}x) \leq B.$$

Hence

$$A \leq \alpha^{-1}N(x) \leq B,$$

which yields (4), since $\alpha = |x|$.

Let $E \subset \mathbf{R}^n$ be a subspace. We define a *norm on E* to be any function

$$N: E \to \mathbf{R}$$

that satisfies (1), (2), and (3). In particular, every norm on $\mathbf{R}^n$ restricts to a norm on E. In fact, every norm on E is obtained from a norm on $\mathbf{R}^n$ by restriction. To see this, decompose $\mathbf{R}^n$ into a direct sum

$$\mathbf{R}^n = E \oplus F.$$

(For example, let $\{e_1, \ldots, e_n\}$ be a basis for $\mathbf{R}^n$ such that $\{e_1, \ldots, e_m\}$ is a basis for E; then F is the subspace whose basis is $\{e_{m+1}, \ldots, e_n\}$.) Given a norm N on E, define a norm N' on $\mathbf{R}^n$ by

$$N'(x) = N(y) + |z|,$$

where

$$x = y + z,\ y \in E,\ z \in F,$$

and $|z|$ is the Euclidean norm of z. It is easy to verify that N' is a norm on $\mathbf{R}^n$ and $N' \mid E = N$.

From this the equivalence of norms on E follows. For let N be a norm on E. Then we may assume N is restriction to E of a norm on $\mathbf{R}^n$, also denoted by N. There exist $A, B \in \mathbf{R}$ such that (4) holds for all x in $\mathbf{R}^n$, so it holds *a fortiori* for all x in E.

We now define a normed vector space (E, N) to be a vector space E (that is, a subspace of some $\mathbf{R}^n$) together with a particular norm N on E.

We shall frequently use the following corollary of the equivalence of norms:

Proposition 2 *Let (E, N) be any normed vector space. A sequence $\{x_k\}$ in E converges to y if and only if*

$$\lim_{k\to\infty} N(x_k - y) = 0. \tag{5}$$

Proof. Let $A > 0$, $B > 0$ be as in (4). Suppose (5) holds. Then the inequality

$$0 \leq |x_k - y| \leq A^{-1}N(x_k - y)$$

shows that $\lim_{k\to\infty} |x_k - y| = 0$, hence $x_k \to y$. The converse is proved similarly.

Another useful application of the equivalence of norms is:

Proposition 3 *Let (E, N) be a normed vector space. Then the unit ball*

$$D = \{x \in E \mid N(x) \leq 1\}$$

is compact.

Proof. Let B be as in (4). Then D is a bounded subset of $\mathbf{R}^n$, for it is contained in

$$\{x \in \mathbf{R}^n \mid |x| \leq B^{-1}\}.$$

It follows from Proposition 2 that D is closed. Thus D is compact.

The Cauchy convergence criterion (of Section 1) can be rephrased in terms of arbitrary norms:

Proposition 4 *Let (E, N) be a normed vector space. Then a sequence $\{x_k\}$ in E converges to an element in E if and only if:*

(6) *for every $\epsilon > 0$, there exists an integer $n_0 > 0$ such that if $p > n \geq n_0$, then*

$$N(x_p - x_n) < \epsilon.$$

Proof. Suppose $E \subset \mathbf{R}^n$, and consider $\{x_k\}$ as a sequence in $\mathbf{R}^n$. The condition (6) is equivalent to the Cauchy condition by the equivalence of norms. Therefore (6) is equivalent to convergence of the sequence to some $y \in \mathbf{R}^n$. But $y \in E$ because subspaces are closed sets.

A sequence in $\mathbf{R}^n$ (or in a subspace of $\mathbf{R}^n$) is often denoted by an *infinite series* $\sum_{k=0}^{\infty} x_k$. This is merely a suggestive notation for the *sequence of partial sums* $\{s_k\}$, where

$$s_k = x_1 + \cdots + x_k.$$

If $\lim_{k\to\infty} s_k = y$, we write

$$\sum_{k=1}^{\infty} x_k = y$$

and say the *series* $\sum x_k$ *converges* to y. If all the x_k are in a subspace $E \subset \mathbf{R}^n$, then also $y \in E$ because E is a closed set.

A series $\sum x_k$ in a normed vector space (E, N) is *absolutely convergent* if the series of real numbers $\sum_{k=0}^{\infty} N(x_k)$ is convergent. This condition implies that $\sum x_k$ is convergent in E. Moreover, it is independent of the norm on E, as follows easily from equivalence of norms. Therefore it is meaningful to speak of absolute convergence of a series in a vector space E, without reference to a norm.

A useful criterion for absolute convergence is the *comparison test*: a series $\sum x_k$ in a normed vector space (E, N) converges absolutely provided there is a convergent series $\sum a_k$ of nonnegative real numbers a_k such that

$$N(x_k) \leq a_k; \qquad k = 1, 2, \ldots.$$

For

$$0 \leq \sum_{k=n+1}^{p} N(x_k) \leq \sum_{k=n+1}^{p} a_k;$$

hence $\sum_{k=0}^{\infty} N(x_k)$ converges by applying the Cauchy criterion to the partial sum sequences of $\sum N(x_k)$ and $\sum a_k$.

PROBLEMS

1. Prove that the norms described in the beginning of Section 2 actually are norms.
2. $|x|_p$ is a norm on $\mathbf{R}^n$, where

$$|x|_p = \left(\sum_{j=1}^{n} |x_j|^p\right)^{1/p}; \qquad 1 \leq p < \infty.$$

   Sketch the unit balls in $\mathbf{R}^2$ and $\mathbf{R}^3$ under the norm $|x|_p$ for $p = 1, 2, 3$.
3. Find the largest $A > 0$ and smallest $B > 0$ such that

$$A|x| \leq |x|_{\text{sum}} \leq B|x|$$

   for all $x \in \mathbf{R}^n$.
4. Compute the norm of the vector $(1, 1) \in \mathbf{R}^2$ under each of the following norms:
   (a) the Euclidean norm;
   (b) the Euclidean $\mathcal{B}$-norm, where $\mathcal{B}$ is the basis $\{(1, 2), (2, 2)\}$;
   (c) the max norm;
   (d) the $\mathcal{B}$-max norm;
   (e) the norm $|x|_p$ of Problem 2, for all p.
5. An *inner product* on a vector space E is any map $\mathbf{R}^n \times \mathbf{R}^n \to \mathbf{R}$, denoted by $(x, y) \to \langle x, y\rangle$, that is *symmetric*, *bilinear*, and *positive definite* (see Section 1).
   (a) Given any inner product show that the function $\langle x, x\rangle^{1/2}$ is a norm.
   (b) Prove that a norm N on E comes from an inner product as in (a) if and only if it satisfies the "parallelogram law":

$$N(x + y)^2 + N(x - y)^2 = 2(N(x)^2 + N(y)^2).$$

   (c) Let $a_1, \ldots, a_n$ be positive numbers. Find an inner product on $\mathbf{R}^n$ whose corresponding norm is

$$N(x) = \left(\sum a_k x_k^2\right)^{1/2}.$$

(d) Let $\{e_1, \ldots, e_m\}$ be a basis for E. Show that there is a unique inner product on E such that

$$\langle e_i, e_j \rangle = \delta_{ij} \qquad \text{for all} \quad i, j.$$

6. Which of the following formulas define norms on $\mathbf{R}^2$? (Let (x, y) be the coordinates in $\mathbf{R}^2$.)
(a) $(x^2 + xy + y^2)^{1/2}$; (b) $(x^2 - 3xy + y^2)^{1/2}$;
(c) $(|x| + |y|)^2$; (d) $\frac{1}{2}(|x| + |y|) + \frac{3}{2}(x^2 + y^2)^{1/2}$.

7. Let $U \subset \mathbf{R}^n$ be a bounded open set containing 0. Suppose U is *convex*: if $x \in U$ and $y \in U$, then the line segment $\{tx + (1 - t)y \mid 0 \leq t \leq 1\}$ is in U. For each $x \in \mathbf{R}^n$ define

$$\sigma(x) = \text{least upper bound of } \{\lambda \geq 0 \mid \lambda x \in U\}.$$

Then the function

$$N(x) = \frac{1}{\sigma(x)}$$

is a norm on $\mathbf{R}^n$.

8. Let M_n be the vector space of $n \times n$ matrices. Denote the transpose of $A \in M_n$ by A^t. Show that an inner product (see Problem 5) on M_n is defined by the formula

$$\langle A, B \rangle = \mathrm{Tr}(A^t B).$$

Express this inner product in terms of the entries in the matrices A and B.

9. Find the orthogonal complement in M_n (see Problem 8) of the subspace of diagonal matrices.

10. Find a basis for the subspace of M_n of matrices of trace 0. What is the orthogonal complement of this subspace?

§3. Exponentials of Operators

The set $L(\mathbf{R}^n)$ of operators on $\mathbf{R}^n$ is identified with the set M_n of $n \times n$ matrices. This in turn is the same as $\mathbf{R}^{n^2}$ since a matrix is nothing but a list of n^2 numbers. (One chooses an ordering for these numbers.) Therefore $L(\mathbf{R}^n)$ is a vector space under the usual addition and scalar multiplication of operators (or matrices). We may thus speak of norms on $L(\mathbf{R}^n)$, convergence of series of operators, and so on.

A frequently used norm on $L(\mathbf{R}^n)$ is the *uniform norm*. This norm is defined in terms of a given norm on $\mathbf{R}^n = E$, which we shall write as $|x|$. If $T: E \to E$ is an operator, the uniform norm of T is defined to be

$$\|T\| = \max\{|Tx| \mid |x| \leq 1\}.$$

In other words, $\|T\|$ is the maximum value of $|Tx|$ on the *unit ball*

$$D = \{x \in E \mid |x| \leq 1\}.$$

The existence of this maximum value follows from the compactness of D (Section 1, Proposition 3) and the continuity of $T: \mathbf{R}^n \to \mathbf{R}^n$. (This continuity follows immediately from a matrix representation of T.)

The uniform norm on $L(\mathbf{R}^n)$ depends on the norm chosen for $\mathbf{R}^n$. If no norm on $\mathbf{R}^n$ is specified, the standard Euclidean norm is intended.

Lemma 1 *Let $\mathbf{R}^n$ be given a norm $|x|$. The corresponding uniform norm on $L(\mathbf{R}^n)$ has the following properties*:

(a) *If $\|T\| = k$, then $|Tx| \leq k|x|$ for all x in $\mathbf{R}^n$.*
(b) $\|ST\| \leq \|S\| \cdot \|T\|$.
(c) $\|T^m\| \leq \|T\|^m$ *for all* $m = 0, 1, 2, \ldots$.

Proof. (a) If $x = 0$, then $|Tx| = 0 = k|x|$. If $x \neq 0$, then $|x| \neq 0$. Let $y = |x|^{-1}x$, then

$$|y| = \frac{1}{|x|}|x| = 1.$$

Hence

$$k = \|T\| \geq |Ty| = \frac{1}{|x|}|Tx|$$

from which (a) follows.

(b) Let $|x| \leq 1$. Then from (a) we have

$$\begin{aligned} |S(Tx)| &\leq \|S\| \cdot |Tx| \\ &\leq \|S\| \cdot \|T\| \cdot |x| \\ &\leq \|S\| \cdot \|T\|. \end{aligned}$$

Since $\|ST\|$ is the maximum value of $|STx|$, (b) follows.

Finally, (c) is an immediate consequence of (b).

We now define an important series generalizing the usual exponential series. For any operator $T: \mathbf{R}^n \to \mathbf{R}^n$ define

$$\exp(T) = e^T = \sum_{k=0}^{\infty} \frac{T^k}{k!}.$$

(Here $k!$ is k *factorial*, the product of the first k positive integers if $k > 0$, and $0! = 1$ by definition.) This is a series in the vector space $L(\mathbf{R}^n)$.

Theorem *The exponential series $\sum_{k=0}^{\infty} T^k/k!$ is absolutely convergent for every operator T.*

Proof. Let $\| T \| = a \geq 0$ be the uniform norm (for some norm on $\mathbf{R}^n$). Then $\| T^k/k! \| \leq a^k/k!$, by Lemma 1, proved earlier. Now the real series $\sum_{k=0}^{\infty} a^k/k!$ converges to e^a (where e is the base of natural logarithms). Therefore the exponential series for T converges absolutely by the comparison test (Section 2).

We have also proved that

$$\| e^A \| \leq e^{\|A\|}.$$

We shall need the following result.

Lemma 2 *Let $\sum_{j=0}^{\infty} A_j = A$ and $\sum_{k=0}^{\infty} B_k = B$ be absolutely convergent series of operators on $\mathbf{R}^n$. Then $AB = C = \sum_{l=0}^{\infty} C_l$, where $C_l = \sum_{j+k=l} A_j B_k$.*

Proof. Let the nth partial sum of the series $\sum A_j$, $\sum B_k$, $\sum C_l$ be denoted respectively by α_n, β_n, γ_n. Then

$$AB = \lim_{n \to \infty} \alpha_n \beta_n,$$

while

$$C = \lim_{n \to \infty} \gamma_{2n}.$$

If $\gamma_{2n} - \alpha_n \beta_n$ is computed, it is found that it equals

$$\sum{}' A_j B_k + \sum{}'' A_j B_k,$$

where $\sum'$ denotes the sum over terms with indices satisfying

$$j + k \leq 2n, \qquad 0 \leq j \leq n, \qquad n + 1 \leq k \leq 2n,$$

while $\sum''$ is the sum corresponding to

$$j + k \leq 2n, \qquad n + 1 \leq j \leq 2n, \qquad 0 \leq k \leq n.$$

Therefore

$$\| \gamma_{2n} - \alpha_n \beta_n \| \leq \sum{}' \| A_j \| \cdot \| B_k \| + \sum{}'' \| A_j \| \cdot \| B_k \|.$$

Now

$$\sum{}' \| A_j \| \cdot \| B_k \| \leq \Big(\sum_{j=0}^{\infty} \| A_j \|\Big)\Big(\sum_{k=n+1}^{2n} \| B_k \|\Big).$$

This tends to 0 as $n \to \infty$ since $\sum_{j=0}^{\infty} \| A_j \| < \infty$. Similarly, $\sum'' \| A_j \| \cdot \| B_k \| \to 0$ as $n \to \infty$. Therefore $\lim_{n \to \infty} (\gamma_{2n} - \alpha_n \beta_n) = 0$, proving the lemma.

The next result is useful in computing with exponentials.

Proposition *Let P, S, T denote operators on $\mathbf{R}^n$. Then:*

(a) *if $Q = PTP^{-1}$, then $e^Q = Pe^T P^{-1}$;*

(b) *if $ST = TS$, then $e^{S+T} = e^S e^T$;*

(c) *$e^{-S} = (e^S)^{-1}$;*

(d) *if $n = 2$ and $T = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$, then*

$$e^T = e^a \begin{bmatrix} \cos b & -\sin b \\ \sin b & \cos b \end{bmatrix}.$$

The proof of (a) follows from the identities $P(A + B)P^{-1} = PAP^{-1} + PBP^{-1}$ and $(PTP^{-1})^k = PT^kP^{-1}$. Therefore

$$P\left(\sum_{k=0}^{n} \frac{T^k}{k!}\right) P^{-1} = \sum_{k=0}^{n} \frac{(PTP^{-1})^k}{k!}$$

and (a) follows by taking limits. To prove (b), observe that because $ST = TS$ we have by the binomial theorem

$$(S + T)^n = n! \sum_{j+k=n} \frac{S^j}{j!} \frac{T^k}{k!}.$$

Therefore

$$e^{S+T} = \sum_{n=0}^{\infty} \left(\sum_{j+k=n} \frac{S^j}{j!} \frac{T^k}{k!} \right)$$

$$= \left(\sum_{j=0}^{\infty} \frac{S^j}{j!}\right)\left(\sum_{k=0}^{\infty} \frac{T^k}{k!}\right)$$

by Lemma 2, which proves (b). Putting $T = -S$ in (b) gives (c).

The proof of (d) follows from the correspondence

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix} \leftrightarrow a + ib$$

of Chapter 3, which preserves sums, products, and real multiples. It is easy to see that it also preserves limits. Therefore

$$e^T \leftrightarrow e^a e^{ib},$$

where e^{ib} is the complex number $\sum_{k=0}^{\infty} (ib)^k/k!$. Using $i^2 = -1$, we find the real part of e^{ib} to be the sum of the Taylor series (at 0) for $\cos b$; similarly, the imaginary part is $\sin b$. This proves (d).

Observe that (c) implies that e^S is invertible for every operator S. This is analogous to the fact that $e^s \neq 0$ for every real number s.

As an example we compute the exponential of $T = \begin{bmatrix} a & 0 \\ b & a \end{bmatrix}$. We write

$$T = aI + B, \qquad B = \begin{bmatrix} 0 & 0 \\ b & 0 \end{bmatrix}.$$

Note that aI commutes with B. Hence

$$e^T = e^{aI}e^B = e^a e^B.$$

Now $B^2 = 0$; hence $B^k = 0$ for all $k > 1$, and

$$\begin{aligned} e^B &= \sum_{k=0}^{1} \frac{1}{k!} B^k \\ &= I + B. \end{aligned}$$

Thus

$$e^T = e^a(I + B) = e^a \begin{bmatrix} 1 & 0 \\ b & 1 \end{bmatrix}$$

$$= \begin{bmatrix} e^a & 0 \\ e^a b & e^a \end{bmatrix}.$$

We can now compute e^A for any 2×2 matrix A. We will see in Chapter 6 that can find an invertible matrix P such that the matrix

$$B = PAP^{-1}$$

has one of the following forms:

$$(1)\ \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}; \quad (2)\ \begin{bmatrix} a & -b \\ b & a \end{bmatrix}; \quad (3)\ \begin{bmatrix} \lambda & 0 \\ 1 & \lambda \end{bmatrix}.$$

We then compute e^B. For (1),

$$e^B = \begin{bmatrix} e^\lambda & 0 \\ 0 & e^\mu \end{bmatrix}.$$

For (2)

$$e^B = e^a \begin{bmatrix} \cos b & -\sin b \\ \sin b & \cos b \end{bmatrix}$$

as was shown in the proposition above. For (3)

$$e^B = e^\lambda \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

as we have just seen. Therefore e^A can be computed from the formula

$$e^A = e^{P^{-1}BP} = P^{-1}e^B P.$$

There is a very simple relationship between the eigenvectors of T and those of e^T:

If $x \in \mathbf{R}^n$ is an eigenvector of T belonging to the real eigenvalue α of T, then x is also an eigenvector of e^T belonging to e^α.

For, from $Tx = \alpha x$, we obtain

$$\begin{aligned} e^T x &= \lim_{n\to\infty} \left(\sum_{k=0}^{n} \frac{T^k x}{k!} \right) \\ &= \lim_{n\to\infty} \left(\sum_{k=0}^{n} \frac{\alpha^k}{k!} x \right) \\ &= \left(\sum_{k=0}^{\infty} \frac{\alpha^k}{k!} \right) x \\ &= e^\alpha x. \end{aligned}$$

We conclude this section with the observation that all that has been said for exponentials of operators on $\mathbf{R}^n$ also holds for operators on the *complex* vector space $\mathbf{C}^n$. This is because $\mathbf{C}^n$ can be considered as the real vector space $\mathbf{R}^{2n}$ by simply ignoring nonreal scalars; every complex operator is *a fortiori* a real operator. In addition, the preceding statement about eigenvectors is equally valid when complex eigenvalues of an operator on $\mathbf{C}^n$ are considered; the proof is the same.

PROBLEMS

1. Let N be any norm on $L(\mathbf{R}^n)$. Prove that there is a constant K such that
$$N(ST) \leq KN(S)N(T)$$
for all operators S, T. Why must $K \geq 1$?

2. Let $T\colon \mathbf{R}^n \to \mathbf{R}^m$ be a linear transformation. Show that T is *uniformly continuous*: for all $\epsilon > 0$ there exists $\delta > 0$ such that if $|x - y| < \delta$ then
$$|Tx - Ty| < \epsilon.$$

3. Let $T\colon \mathbf{R}^n \to \mathbf{R}^n$ be an operator. Show that
$$\|T\| = \text{least upper bound} \left\{ \frac{|Tx|}{|x|} \,\middle|\, x \neq 0 \right\}.$$

4. Find the uniform norm of each of the following operators on $\mathbf{R}^2$:
(a) $\begin{bmatrix} 3 & 0 \\ 0 & -4 \end{bmatrix}$ (b) $\begin{bmatrix} \frac{1}{2} & 0 \\ 10 & \frac{1}{2} \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$

5. Let
$$T = \begin{bmatrix} \frac{1}{2} & 0 \\ 10 & \frac{1}{2} \end{bmatrix}.$$

(a) Show that

$$\lim_{n\to\infty} \| T^n \|^{1/n} = \tfrac{1}{2}.$$

(b) Show that for every $\epsilon > 0$ there is a basis $\mathcal{B}$ of $\mathbf{R}^2$ for which

$$\| T \|_{\mathcal{B}} < \tfrac{1}{2} + \epsilon,$$

where $\| T \|_{\mathcal{B}}$ is the uniform norm of T corresponding to the Euclidean $\mathcal{B}$-norm on $\mathbf{R}^2$.

(c) For any basis $\mathcal{B}$ of $\mathbf{R}^2$,

$$\| T \|_{\mathcal{B}} > \tfrac{1}{2}.$$

6. (a) Show that

$$\| T \| \cdot \| T^{-1} \| \geq 1$$

for every invertible operator T.

(b) If T has two distinct real eigenvalues, then

$$\| T \| \cdot \| T^{-1} \| > 1.$$

(*Hint*: First consider operators on $\mathbf{R}^2$.)

7. Prove that if T is an operator on $\mathbf{R}^n$ such that $\| T - I \| < 1$, then T is invertible and the series $\sum_{k=0}^{\infty} (I - T)^k$ converges absolutely to T^{-1}. Find an upper bound for $\| T^{-1} \|$.

8. Let $A \in L(\mathbf{R}^n)$ be invertible. Find $\epsilon > 0$ such that if $\| B - A \| < \epsilon$, then B is invertible. (*Hint*: First show $A^{-1}B$ is invertible by applying Problem 7 to $T = A^{-1}B$.)

9. Compute the exponentials of the following matrices ($i = \sqrt{-1}$):

(a) $\begin{bmatrix} 5 & -6 \\ 3 & -4 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}$ (c) $\begin{bmatrix} 2 & -1 \\ 0 & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

(e) $\begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix}$ (f) $\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 1 & 3 \end{bmatrix}$ (g) $\begin{bmatrix} \lambda & 0 & 0 \\ 1 & \lambda & 0 \\ 0 & 1 & \lambda \end{bmatrix}$

(h) $\begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}$ (i) $\begin{bmatrix} 1+i & 0 \\ 2 & 1+i \end{bmatrix}$ (j) $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$

10. For each matrix T in Problem 9 find the eigenvalues of e^T.

11. Find an example of two operators A, B on $\mathbf{R}^2$ such that

$$e^{A+B} \neq e^A e^B.$$

12. If $AB = BA$, then $e^A e^B = e^B e^A$ and $e^A B = B e^A$.

13. Let an operator $A: \mathbf{R}^n \to \mathbf{R}^n$ leave invariant a subspace $E \subset \mathbf{R}^n$ (that is, $Ax \in E$ for all $x \in E$). Show that e^A also leaves E invariant.

14. Show that if $\| T - I \|$ is sufficiently small, then there is an operator S such that $e^S = T$. (*Hint*: Expand $\log(1 + x)$ in a Taylor series.) To what extent is S unique?

15. Show that there is no real 2×2 matrix S such that $e^S = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$.

§4. Homogeneous Linear Systems

Let A be an operator on $\mathbf{R}^n$. In this section we shall express solutions to the equation:

$$x' = Ax \tag{1}$$

in terms of exponentials of operators.

Consider the map $\mathbf{R} \to L(\mathbf{R}^n)$ which to $t \in \mathbf{R}$ assigns the operator e^{tA}. Since $L(\mathbf{R}^n)$ is identified with $\mathbf{R}^{n^2}$, it makes sense to speak of the derivative of this map.

Proposition

$$\frac{d}{dt} e^{tA} = Ae^{tA} = e^{tA}A.$$

In other words, the derivative of the operator-valued function e^{tA} is another operator-valued function Ae^{tA}. This means the composition of e^{tA} with A; the order of composition does not matter. One can think of A and e^{tA} as matrices, in which case Ae^{tA} is their product.

Proof of the proposition.

$$\begin{aligned} \frac{d}{dt} e^{tA} &= \lim_{h\to 0} \frac{e^{(t+h)A} - e^{tA}}{h} \\ &= \lim_{h\to 0} \frac{e^{tA}e^{hA} - e^{tA}}{h} \\ &= e^{tA} \lim_{h\to 0} \left(\frac{e^{hA} - I}{h} \right) \\ &= e^{tA}A; \end{aligned}$$

that the last limit equals A follows from the series definition of e^{hA}. Note that A commutes with each term of the series for e^{tA}, hence with e^{tA}. This proves the proposition.

We can now solve equation (1). We recall from Chapter 1 that the general solution of the scalar equation

$$x' = ax \qquad (a \in \mathbf{R})$$

is

$$x(t) = ke^{ta}; \qquad k = x(0).$$

The same is true where x, a, and k are allowed to be complex numbers (Chapter 3). These results are special cases of the following, which can be considered as the fundamental theorem of linear differential equations with constant coefficients.

Theorem *Let A be an operator on $\mathbf{R}^n$. Then the solution of the initial value problem*

(1′) $$x' = Ax, \qquad x(0) = K \in \mathbf{R}^n,$$

is

(2) $$e^{tA}K,$$

and there are no other solutions.

Proof. The preceding lemma shows that

$$\frac{d}{dt}(e^{tA}K) = \left(\frac{d}{dt}e^{tA}\right)K$$
$$= Ae^{tA}K;$$

since $e^{0A}K = K$, it follows that (2) is a solution of (1′). To see that there are no other solutions, let $x(t)$ be any solution of (1′) and put

$$y(t) = e^{-tA}x(t).$$

Then

$$y'(t) = \left(\frac{d}{dt}e^{-tA}\right)x(t) + e^{-tA}x'(t)$$
$$= -Ae^{-tA}x(t) + e^{-tA}Ax(t)$$
$$= e^{-tA}(-A + A)x(t)$$
$$\equiv 0.$$

Therefore $y(t)$ is a constant. Setting $t = 0$ shows $y(t) = K$. This completes the proof of the theorem.

As an example we compute the general solution of the two-dimensional system

(3) $$x_1' = ax_1,$$
$$x_2' = bx_1 + ax_2,$$

where a, b are constants. In matrix notation this is

$$x' = Ax; \qquad A = \begin{bmatrix} a & 0 \\ b & a \end{bmatrix}; \qquad x = (x_1, x_2).$$

The solution with initial value $K = (K_1, K_2) \in \mathbf{R}^2$ is

$$e^{tA}K.$$

In Section 3 we saw that

$$e^{tA} = e^{ta}\begin{bmatrix} 1 & 0 \\ tb & 1 \end{bmatrix}.$$

Thus

$$e^{tA}K = (e^{ta}K_1, e^{ta}(tbK_1 + K_2)).$$

Thus the solution to (3) satisfying

$$x_1(0) = K_1, \qquad x_2(0) = K_2$$

is

$$x_1(t) = e^{ta}K_1,$$
$$x_2(t) = e^{ta}(tbK_1 + K_2).$$

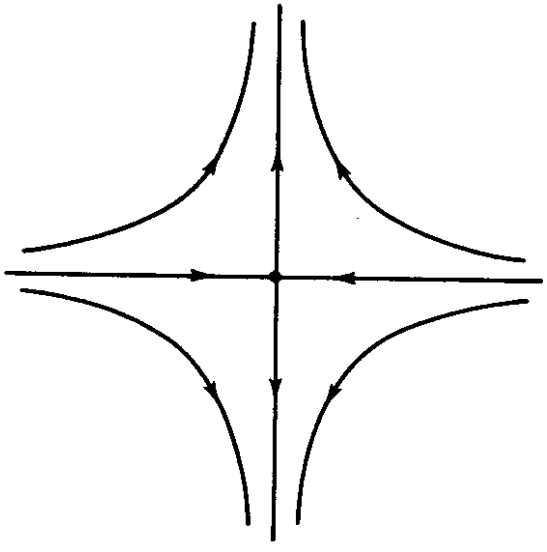


FIG. A. Saddle: $B = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}$, $\lambda < 0 < \mu$.

Since we know how to compute the exponential of any 2×2 matrix (Section 3), we can explicitly solve any two-dimensional system of the form $x' = Ax, A \in L(\mathbf{R}^2)$. Without finding explicit solutions, we can also obtain important qualitative information about the solutions from the eigenvalues of A. We consider the most important special cases.

Case I. *A has real eigenvalues of opposite signs.* In this case the origin (or sometimes the differential equation) is called a *saddle*. As we saw in Chapter 3, after a suitable change of coordinates $x = Py$, the equation becomes

$$y' = By,$$

$$B = PAP^{-1} = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}, \qquad \lambda < 0 < \mu.$$

In the (y_1, y_2) plane the phase portrait looks like Fig. A on p. 91.

Case II. *All eigenvalues have negative real parts.* This important case is called a *sink*. It has the characteristic property that

$$\lim_{t \to \infty} x(t) = 0$$

for every solution $x(t)$. If A is diagonal, this is obvious, for the solutions are

$$y(t) = (c_1 e^{\lambda t}, c_2 e^{\mu t}); \qquad \lambda < 0, \quad \mu < 0.$$

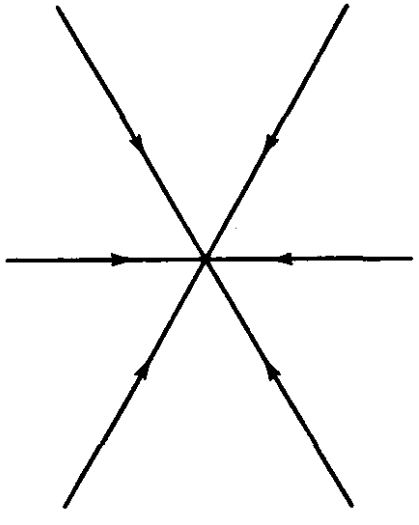


FIG. B. Focus: $B = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}, \lambda < 0.$

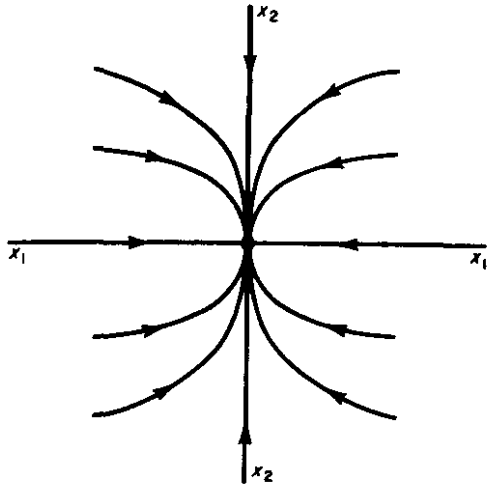


FIG. C. Node: $B = \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix}, \lambda < \mu < 0.$

If A is diagonalizable, the solutions

$$x(t) = Py(t)$$

are of the form with $y(t)$ as above and $P \in L(\mathbf{R}^2)$; clearly, $x(t) \to 0$ as $t \to \infty$.

The phase portrait for these subcases looks like Fig. B if the eigenvalues are equal (a *focus*) and like Fig. C if they are unequal (a *node*).

If the eigenvalues are negative but A is not diagonalizable, there is a change of coordinates $x = Py$ (see Chapter 6) giving the equivalent equation

$$y' = By,$$

where

$$B = P^{-1}AP = \begin{bmatrix} \lambda & 0 \\ 1 & \lambda \end{bmatrix}, \qquad \lambda < 0.$$

We have already solved such an equation; the solutions are

$$y_1(t) = K_1 e^{t\lambda},$$

$$y_2(t) = K_2 e^{t\lambda} + K_1 t e^{t\lambda},$$

which tend to 0 as t tends to ∞. The phase portrait looks like Fig. D (an *improper node*).

If the eigenvalues are $a \pm ib$, $a < 0$ we can change coordinates as in Chapter 4

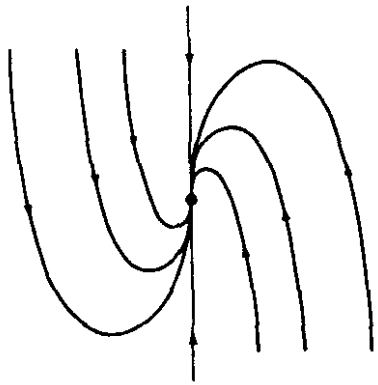


FIG. D. Improper node: $B = \begin{bmatrix} \lambda & 0 \\ 1 & \lambda \end{bmatrix}$, $\lambda < 0$.

to obtain the equivalent system

$$y' = By, \qquad B = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}.$$

From Section 3 we find

$$e^{tB} = e^{ta} \begin{bmatrix} \cos tb & -\sin tb \\ \sin tb & \cos tb \end{bmatrix}.$$

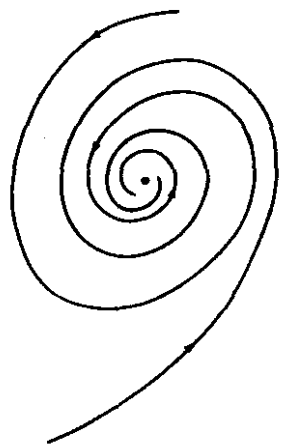


FIG. E. Spiral sink: $B = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$, $b > 0 > a$.

Therefore the general solution is expressed in y-coordinates as

$$y(t) = e^{ta}(K_1 \cos tb - K_2 \sin tb,\ K_2 \cos tb + K_1 \sin tb).$$

Since $|\cos tb| \leq 1$ and $|\sin tb| \leq 1$, and $a < 0$, it follows that

$$\lim_{t \to \infty} y(t) = 0.$$

If $b > 0$, the phase portrait consists of counterclockwise spirals tending to 0 (Fig. E), and clockwise spirals tending to 0 if $b < 0$.

Case III. *All eigenvalues have positive real part.* In this case, called a *source*, we have

$$\lim_{t \to \infty} |x(t)| = \infty \qquad \text{and} \qquad \lim_{t \to -\infty} |x(t)| = 0.$$

A proof similar to that of Case II can be given; the details are left to the reader. The phase portraits are like Figs. B–E with the arrows reversed.

Case IV. *The eigenvalues are pure imaginary.* This is called a *center*. It is characterized by the property that all solutions are *periodic* with the same period. To see this, change coordinates to obtain the equivalent equation

$$y' = By, \qquad B = \begin{bmatrix} 0 & -b \\ b & 0 \end{bmatrix}.$$

We know that

$$e^{tB} = \begin{bmatrix} \cos tb & -\sin tb \\ \sin tb & \cos tb \end{bmatrix}.$$

Therefore if $y(t)$ is any solution,

$$y\left(t + \frac{2\pi}{b}\right) \equiv y(t).$$

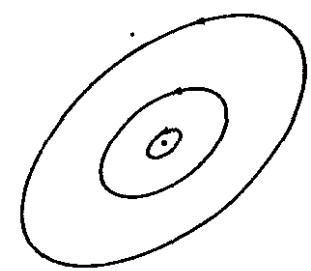


FIG. F. Center: $B = \begin{bmatrix} 0 & -b \\ b & 0 \end{bmatrix}$, $b > 0$.

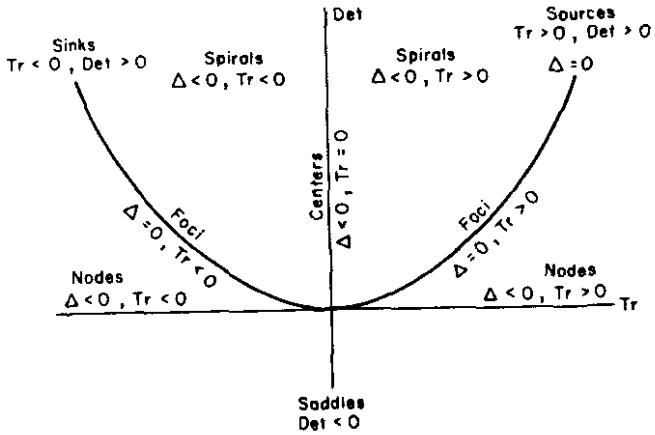


FIG. G

The phase portrait in the y-coordinates consists of concentric circles. In the original x-coordinates the orbits may be ellipses as in Fig. F. (If $b < 0$, the arrows point clockwise.)

Figure G summarizes the geometric information about the phase portrait of $x' = Ax$ that can be deduced from the characteristic polynomial of A. We write this polynomial as

$$\lambda^2 - (\operatorname{Tr} A)\lambda + \operatorname{Det} A.$$

The *discriminant* Δ is defined to be

$$\Delta = (\operatorname{Tr} A)^2 - 4 \operatorname{Det} A.$$

The eigenvalues are

$$\tfrac{1}{2}(\operatorname{Tr} A \pm \sqrt{\Delta}).$$

Thus real eigenvalues correspond to the case $\Delta \geq 0$; the eigenvalues have negative real part when $\operatorname{Tr} A < 0$; and so on.

The geometric interpretation of $x' = Ax$ is as follows (compare Chapter 1). The map $\mathbf{R}^n \to \mathbf{R}^n$ which sends x into Ax is a vector field on $\mathbf{R}^n$. Given a point K of $\mathbf{R}^n$, there is a *unique* curve $t \to e^{tA}K$ which starts at K at time zero, and is a solution of (1). (We interpret t as time.) The tangent vector to this curve at a time t_0 is the vector $Ax(t_0)$ of the vector field at the point of the curve $x(t_0)$.

We may think of points of $\mathbf{R}^n$ flowing simultaneously along these solution curves. The position of a point $x \in \mathbf{R}^n$ at time t is denoted by

$$\phi_t(x) = e^{tA}x.$$

Thus for each $t \in \mathbf{R}$ we have a map

$$\phi_t\colon \mathbf{R}^n \to \mathbf{R}^n \qquad (t \in \mathbf{R})$$

given by

$$\phi_t(x) = e^{tA}x.$$

The collection of maps $\{\phi_t\}_{t\in\mathbf{R}}$ is called the *flow* corresponding to the differential equation (1). This flow has the basic property

$$\phi_{s+t} = \phi_s \circ \phi_t,$$

which is just another way of writing

$$e^{(s+t)A} = e^{sA}e^{tA};$$

this is proved in the proposition in Section 2. The flow is called *linear* because each map $\phi_t\colon \mathbf{R}^n \to \mathbf{R}^n$ is a linear map. In Chapter 8 we shall define more general *nonlinear* flows.

The phase portraits discussed above give a good visualization of the corresponding flows. Imagine points of the plane all moving at once along the curves in the direction of the arrows. (The origin stays put.)

PROBLEMS

1. Find the general solution to each of the following systems:

(a) $\begin{cases} x' = 2x - y \\ y' = 2y \end{cases}$ (b) $\begin{cases} x' = 2x - y \\ y' = x + 2y \end{cases}$

(c) $\begin{cases} x' = y \\ y' = x \end{cases}$ (d) $\begin{cases} x' = -2x \\ y' = x - 2y \\ z' = y - 2z \end{cases}$

(e) $\begin{cases} x' = y + z \\ y' = z \\ z' = 0 \end{cases}$

2. In (a), (b), and (c) of Problem 1, find the solutions satisfying each of the following initial conditions:
(a) $x(0) = 1, y(0) = -2$; (b) $x(0) = 0, y(0) = -2$;
(c) $x(0) = 0, y(0) = 0$.

3. Let $A\colon \mathbf{R}^n \to \mathbf{R}^n$ be an operator that leaves a subspace $E \subset \mathbf{R}^n$ invariant. Let $x\colon \mathbf{R} \to \mathbf{R}^n$ be a solution of $x' = Ax$. If $x(t_0) \in E$ for some $t_0 \in \mathbf{R}$, show that $x(t) \in E$ for all $t \in \mathbf{R}$.

4. Suppose $A \in L(\mathbf{R}^n)$ has a real eigenvalue $\lambda < 0$. Then the equation $x' = Ax$

has at least one nontrivial solution $x(t)$ such that

$$\lim_{t\to\infty} x(t) = 0.$$

5. Let $A \in L(\mathbf{R}^2)$ and suppose $x' = Ax$ has a nontrivial *periodic solution*, $u(t)$: this means $u(t+p) \equiv u(t)$ for some $p > 0$. Prove that every solution is periodic, with the same period p.

6. If $u: \mathbf{R} \to \mathbf{R}^n$ is a nontrivial solution of $x' = Ax$, then

$$\frac{d}{dt}(|u|) = \frac{1}{|u|}\langle u, Au\rangle.$$

7. Supply the details of Case II in the text.

8. Classify and sketch the phase portraits of planar differential equations $x' = Ax$, $A \in L(\mathbf{R}^2)$, where A has zero as an eigenvalue.

9. For each of the following matrices A consider the corresponding differential equation $x' = Ax$. Decide whether the origin is a sink, source, saddle, or none of these. Identify in each case those vectors u such that $\lim_{t\to\infty} x(t) = 0$, where $x(t)$ is the solution with $x(0) = u$:

(a) $\begin{bmatrix} -1 & 0 \\ 2 & -2 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 2 \\ 0 & 2 \end{bmatrix}$ (c) $\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$

(d) $\begin{bmatrix} -1 & 2 \\ -1 & 1 \end{bmatrix}$ (e) $\begin{bmatrix} 1 & -2 \\ -2 & 4 \end{bmatrix}$

10. Which values (if any) of the parameter k in the following matrices makes the origin a sink for the corresponding differential equation $x' = Ax$?

(a) $\begin{bmatrix} a & -k \\ k & 2 \end{bmatrix}$ (b) $\begin{bmatrix} 3 & 0 \\ k & -4 \end{bmatrix}$

(c) $\begin{bmatrix} k^2 & 1 \\ 0 & k \end{bmatrix}$ (d) $\begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ -1 & 0 & k \end{bmatrix}$

11. Let $\phi_t: \mathbf{R}^2 \to \mathbf{R}^2$ be the flow corresponding to the equation $x' = Ax$. (That is, $t \to \phi_t(x)$ is the solution passing through x at $t = 0$.) Fix $\tau > 0$, and show that ϕ_τ is a linear map of $\mathbf{R}^2 \to \mathbf{R}^2$. Then show that ϕ_t preserves area if and only if $\operatorname{Tr} A = 0$, and that in this case the origin is not a sink or a source. (*Hint*: An operator is area-preserving if and only if the determinant is ± 1.)

12. Describe in words the phase portraits of $x' = Ax$ for

(a) $A = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$ (b) $A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$

(c) $A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ (d) $A = \begin{bmatrix} 2 & 0 \\ 1 & 2 \end{bmatrix}$

13. Suppose A is an $n \times n$ matrix with n distinct eigenvalues and the real part of every eigenvalue is less than some negative number α. Show that for every solution to $x' = Ax$, there exists $t_0 > 0$ such that

$$|x(t)| < e^{t\alpha} \quad \text{if} \quad t \geq t_0.$$

14. Let T be an invertible operator on $\mathbf{R}^n$, n odd. Then $x' = Tx$ has a nonperiodic solution.

15. Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ have nonreal eigenvalues. Then $b \neq 0$. The nontrivial solutions curves to $x' = Ax$ are spirals or ellipses that are oriented clockwise if $b > 0$ and counterclockwise if $b < 0$. (*Hint*: Consider the sign of

$$\frac{d}{dt} \arctan(x_2(t)/x_1(t)).)$$

§5. A Nonhomogeneous Equation

We consider a nonhomogeneous nonautonomous linear differential equation

$$x' = Ax + B(t). \tag{1}$$

Here A is an operator on $\mathbf{R}^n$ and $B: \mathbf{R} \to \mathbf{R}^n$ is a continuous map. This equation is called *nonhomogeneous* because of the term $B(t)$ which prevents (1) from being strictly linear; the fact that the right side of (1) depends explicitly on t makes it *nonautonomous*. It is difficult to interpret solutions geometrically.

We look for a solution having the form

$$x(t) = e^{tA}f(t), \tag{2}$$

where $f: \mathbf{R} \to \mathbf{R}^n$ is some differentiable curve. (This method of solution is called "variation of constants," perhaps because if $B(t) \equiv 0$, $f(t)$ is a constant.) Every solution can in fact be written in this form since e^{tA} is invertible.

Differentiation of (2) using the Leibniz rule yields

$$x'(t) = Ae^{tA}f(t) + e^{tA}f'(t).$$

Since x is assumed to be a solution of (2),

$$Ax(t) + B(t) = Ax(t) + e^{tA}f'(t)$$

or

$$f'(t) = e^{-tA}B(t).$$

By integration

$$f(t) = \int_0^t e^{-As}B(s)\,ds + K,$$

so as a candidate for a solution of (1) we have

$$x(t) = e^{At}\left[\int_0^t e^{-As}B(s)\,ds + K\right], \qquad K \in \mathbf{R}^n. \tag{3}$$

Let us examine (3) to see that it indeed makes sense. The integrand in (3) and the previous equation is the vector-valued function $s \to e^{-As}B(s)$ mapping R into $\mathbf{R}^n$. In fact, for any continuous map g of the reals into a vector space $\mathbf{R}^n$, the integral can be defined as an element of $\mathbf{R}^n$. Given a basis of $\mathbf{R}^n$, this integral is a vector whose coordinates are the integrals of the coordinate functions of g.

The integral as a function of its upper limit t is a map from R into $\mathbf{R}^n$. For each t the operator acts on the integral to give an element of $\mathbf{R}^n$. So $t \to x(t)$ is a well-defined map from R into E.

To check that (3) is a solution of (1), we differentiate $x(t)$ in (3):

$$x'(t) = B(t) + Ae^{At}\left[\int_0^t e^{-As}B(s)\,ds + K\right]$$

$$= B(t) + Ax(t).$$

Thus (3) is indeed a solution of (1).

That every solution of (1) must be of the form (3) can be seen as follows. Let $y: \mathbf{R}^n \to E$ be a second solution of (1). Then

$$x' - y' = A(x - y)$$

so that from Section 1

$$x - y = e^{tA}K_0 \qquad \text{for some} \quad K_0 \text{ in } \mathbf{R}^n.$$

This implies that y is of the form (3) (with perhaps a different constant $K \in \mathbf{R}^n$).

We remark that if B in (1) is only defined on some interval, instead of on all of R, then by the above methods, we obtain a solution $x(t)$ defined for t in that same interval.

We obtain further insight into (1) by rewriting the general solution (3) in the form

$$x(t) = u(t) + e^{At}K,$$

$$u(t) = e^{-\alpha t}\int_0^t e^{-As}B(s)\,ds.$$

Note that $u(t)$ is also a solution to (1), while $e^{At}K$ is a solution to the homogeneous equation

$$y' = Ay \tag{4}$$

obtained from (1) by replacing $B(t)$ with 0. In fact, if $v(t)$ is any solution to (1) and $y(t)$ any solution to (4), then clearly $x = v + y$ is another solution to (1). Hence the general solution to (1) is obtained from a particular solution by adding to it the general solution of the corresponding homogeneous equation. In summary

Theorem *Let $u(t)$ be a particular solution of the nonhomogeneous linear differential equation*

$$x' = Ax + B(t). \tag{1}$$

Then every solution of (1) *has the form $u(t) + v(t)$ where $v(t)$ is a solution of the homogeneous equation*

$$x' = Ax. \tag{4'}$$

Conversely, the sum of a solution of (1) *and a solution of* (4') *is a solution of* (1).

If the function $B(t)$ is at all complicated it will probably be impossible to replace the integral in (3) by a simple formula; sometimes, however, this can be done.

Example. Find the general solution to

$$\begin{aligned} x_1' &= -x_2, \\ x_2' &= x_1 + t. \end{aligned} \tag{5}$$

Here

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \qquad B(t) = \begin{bmatrix} 0 \\ t \end{bmatrix}.$$

Hence

$$e^{-As} = \begin{bmatrix} \cos s & \sin s \\ -\sin s & \cos s \end{bmatrix}$$

and the integral in (3) is

$$\int_0^t \begin{bmatrix} \cos s & \sin s \\ -\sin s & \cos s \end{bmatrix}\begin{bmatrix} 0 \\ s \end{bmatrix} ds = \int_0^t \begin{bmatrix} s \sin s \\ s \cos s \end{bmatrix} ds$$

$$= \begin{bmatrix} \sin t - t\cos t \\ \cos t + t \sin t - 1 \end{bmatrix}.$$

To compute (3) we set

$$e^{At} = \begin{bmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{bmatrix}, \qquad K = \begin{bmatrix} K_1 \\ K_2 \end{bmatrix};$$

hence the general solution is

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{bmatrix} \begin{bmatrix} \sin t - t\cos t + K_1 \\ \cos t + t\sin t - 1 + K_2 \end{bmatrix}.$$

Performing the matrix multiplication and simplifying yields

$$x_1(t) = -t + K_1 \cos t + (1 - K_2) \sin t,$$

$$x_2(t) = 1 - (1 - K_2) \cos t + K_1 \sin t.$$

This is the solution whose value at $t = 0$ is

$$x_1(0) = K_1, \qquad x_2(0) = K_2.$$

PROBLEMS

1. Find all solutions to the following equations or systems:
   (a) $x' - 4x - \cos t = 0$; (b) $x' - 4x - t = 0$; (c) $x' = y$, $y' = 2 - x$;
   (d) $x' = y$, $y' = -4x + \sin 2t$; (e) $x' = x + y + z$, $y' = -2y + t$, $z' = 2z + \sin t$.
2. Suppose $T\colon \mathbf{R}^n \to \mathbf{R}^n$ is an invertible linear operator and $c \in E$ is a nonzero constant vector. Show there is a change of coordinates of the form
$$x = Py + b, \qquad b \in \mathbf{R}^n,$$
transforming the nonhomogeneous equation $x = Tx + c$ into homogeneous form $y' = Sy$. Find P, b, and S. (*Hint*: Where is $x' = 0$?)
3. Solve Problem 1(c) using the change of coordinates of Problem 2.

§6. Higher Order Systems

Consider a linear differential equation with constant coefficients which involves a derivative higher than the first; for example,

$$s'' + as' + bs = 0. \tag{1}$$

By introducing new variables we are able to reduce (1) to a first order system of two equations. Let $x_1 = s$ and $x_2 = x_1' = s'$. Then (1) becomes equivalent to the system:

$$\begin{aligned} x_1' &= x_2, \\ x_2' &= -bx_1 - ax_2. \end{aligned} \tag{2}$$

Thus if $x(t) = (x_1(t), x_2(t))$ is a solution of (2), then $s(t) = x_1(t)$ is a solution of (1); if $s(t)$ is a solution of (1), then $x(t) = (s(t), s'(t))$ is a solution of (2).

This procedure of introducing new variables works very generally to reduce higher order equations to first order ones. Thus consider

$$s^{(n)} + a_1 s^{(n-1)} + \cdots + a_{n-1} s' + a_n s = 0. \tag{3}$$

Here s is a real function of t and $s^{(n)}$ is the nth derivative of s, while $a_1, \ldots, a_n$ are constants.

In this case the new variables are $x_1 = s$, $x_2 = x_1', \ldots, x_n = x_{n-1}'$ and the equation (3) is equivalent to the system

$$\begin{aligned} x_1' &= x_2, \\ x_2' &= x_3, \\ &\vdots \\ x_n' &= -a_n x_1 - a_{n-1} x_2 - \cdots - a_1 x_n. \end{aligned} \tag{4}$$

In vector notation (4) has the form $x' = Ax$, where A is the matrix

$$\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & 0 \\ 0 & 0 & & \cdots & 0 \quad 1 \\ -a_n & -a_{n-1} & & \cdots & -a_1 \end{bmatrix}. \tag{4'}$$

Proposition *The characteristic polynomial of* (4') *is*

$$p(\lambda) = \lambda^n + a_1 \lambda^{n-1} + \cdots + a_n.$$

Proof. One uses induction on n. For $n = 2$, this is easily checked. Assume the truth of the proposition for $n - 1$, and let A_{n-1} be the $(n-1) \times (n-1)$ submatrix of A consisting of the last $(n-1)$ rows and last $(n-1)$ columns. Then $\operatorname{Det}(\lambda I - A)$ is easily computed to be $\lambda \operatorname{Det}(\lambda I - A_{n-1}) + a_n$ by expanding along the first column. The induction hypothesis yields the desired characteristic polynomial.

The point of the proposition is that it gives the characteristic polynomial directly from the equation for the higher order differential equation (3).

Let us now return to our first equation

$$s'' + as' + bs = 0. \tag{1}$$

Denote the roots of the polynomial equation $\lambda^2 + a\lambda + b = 0$ by λ_1, λ_2. Suppose at first that these roots are real and distinct. Then (1) reduces to the equation of first order (2); one can find a diagonalizing system of coordinates (y_1, y_2). Every solution of (2) for these coordinates is then $y_1(t) = K_1 \exp(\lambda_1 t)$, $y_2(t) = K_2 \exp(\lambda_2 t)$, with arbitrary constants K_1, K_2. Thus $x_1(t)$ or $s(t)$ is a certain linear combination $s(t) = p_{11}K_1 \exp(\lambda_1 t) + p_{12}K_2 \exp(\lambda_2 t)$. We conclude that if λ_1, λ_2 are real and distinct then every solution of (1) is of the form

$$s(t) = C_1 \exp(\lambda_1 t) + C_2 \exp(\lambda_2 t)$$

for some (real) constants C_1, C_2. These constants can be found if initial values $s(t_0)$, $s'(t_0)$ are given.

Next, suppose that $\lambda_1 = \lambda_2 = \lambda$ and that these eigenvalues are real. In this case the 2×2 matrix in (2) is similar to a matrix of the form

$$\begin{bmatrix} \lambda & 0 \\ \beta & \lambda \end{bmatrix}, \qquad \beta \neq 0,$$

as will be shown in Chapter 6. In the new coordinates the equivalent first-order system is

$$y_1' = \lambda y_1,$$
$$y_2' = \beta y_1 + \lambda y_2.$$

By the methods of Section 4 we find that the general solution to such a system is

$$y_1(t) = K_1 e^{\lambda t},$$
$$y_2(t) = K_1 \beta t e^{\lambda t} + K_2 e^{\lambda t},$$

K_1 and K_2 being arbitrary constants. In the original coordinates the solutions to the equivalent first order system are linear combinations of these. Thus we conclude that if the characteristic polynomial of (1) has only one root $\lambda \in \mathbf{R}$, the solutions have the form

$$s(t) = C_1 e^{\lambda t} + C_2 t e^{\lambda t}.$$

The values of C_1 and C_2 can be determined from initial conditions.

Example. Solve the initial-value problem

$$s'' + 2s' + s = 0, \tag{5}$$
$$s(0) = 1, \qquad s'(0) = 2.$$

The characteristic polynomial is $\lambda^2 + 2\lambda + 1$; the only root is $\lambda = -1$. Therefore the general solution is

$$s(t) = C_1 e^{-t} + C_2 t e^{-t}.$$

We find that

$$s'(t) = (-C_1 + C_2)e^{-t} - C_2 t e^{-t}.$$

From the initial conditions in (5) we get, setting $t = 0$ in the last two formulas

$$C_1 = 1,$$
$$-C_1 + C_2 = 2.$$

Hence $C_2 = 3$ and the solution to (5) is

$$s(t) = e^{-t} + 3te^{-t}.$$

The reader may verify that this actually is a solution to (5)!

The final case to consider is that when λ_1, λ_2 are nonreal complex conjugate numbers. Suppose $\lambda_1 = u + iv$, $\lambda_2 = u - iv$. Then we get a solution (as in Chapter 3):

$$y_1(t) = e^{ut}(K_1 \cos vt - K_2 \sin vt),$$
$$y_2(t) = e^{ut}(K_1 \sin vt + K_2 \cos vt).$$

Thus we obtain $s(t)$ as a linear combination of $y_1(t)$ and $y_2(t)$, so that finally,

$$s(t) = e^{ut}(C_1 \cos vt + C_2 \sin vt)$$

for some constants C_1, C_2.

A special case of the last equation is the "harmonic oscillator":

$$s'' + b^2 s = 0;$$

the eigenvalues are $\pm ib$, and the general solution is

$$C_1 \cos bt + C_2 \sin bt.$$

We summarize what we have found.

Theorem *Let λ_1, λ_2 be the roots of the polynomial $\lambda^2 + a\lambda + b$. Then every solution of the differential equation*

$$s'' + as' + bs = 0 \tag{1}$$

is of the following type:

Case (a). λ_1, λ_2 *are real distinct:* $s(t) = C_1 \exp(\lambda_1 t) + C_2 \exp(\lambda_2 t)$;
Case (b). $\lambda_1 = \lambda_2 = \lambda$ *is real:* $s(t) = C_1 e^{\lambda t} + C_2 t e^{\lambda t}$;
Case (c). $\lambda_1 = \bar{\lambda}_2 = u + iv$, $v \neq 0$: $s(t) = e^{ut}(C_1 \cos vt + C_2 \sin vt)$.

In each case C_1, C_2 are (real) constants determined by initial conditions of the form

$$s(t_0) = \alpha, \qquad s'(t_0) = \beta.$$

The nth order linear equation (3) can also be solved by changing it to an equivalent first order system. First order systems that come from nth order equations

have special properties which enable them to be solved quite easily. To understand the method of solution requires more linear algebra, however. We shall return to higher order equations in the next chapter.

We make a simple but important observation about the linear homogeneous equation (3):

If $s(t)$ and $q(t)$ are solutions to (3), *so is the function $s(t) + q(t)$; if k is any real number, then $ks(t)$ is a solution.*

In other words, the set of all solutions is a vector space. And since n initial conditions determines a solution uniquely (consider the corresponding first order system), the dimension of the vector space of solutions equals the order of the differential equation.

A higher order inhomogeneous linear equation

$$s^{(n)} + a_1 s^{(n-1)} + \cdots + a_n s = b(t) \tag{6}$$

can be solved (in principle) by reducing it to a first order inhomogeneous linear system

$$\dot{x} = Ax + B(t)$$

and applying variation of constants (Section 5). Note that

$$B(t) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ b(t) \end{bmatrix}.$$

As in the case of first order systems, the general solution to (6) can be expressed as the general solution to the corresponding homogeneous equation

$$s^{(n)} + a_1 s^{(n-1)} + \cdots + a_n s = 0$$

plus a particular solution of (3). Consider, for example,

$$s'' + s = t - 1. \tag{7}$$

The general solution of

$$s'' + s = 0$$

is

$$A \cos t + B \sin t; \qquad A, B \in \mathbf{R}.$$

A particular solution to (7) is

$$s(t) = t - 1.$$

Hence the general solution to (7) is

$$A \cos t + B \sin t + t - 1.$$

Finally, we point out that higher order *systems* can be reduced to first order systems. For example, consider the system

$$x'' + x' + 2y' - 3x = 0,$$
$$y'' + 5x' - 4y = 0.$$

Here $x(t)$ and $y(t)$ are unknown real-valued functions of a real variable. Introduce new functions $u = x'$, $v = y'$. The system is equivalent to the four-dimensional first order system

$$x' = u,$$
$$u' = 3x - u - 2v,$$
$$y' = v,$$
$$v' = -5u + 4y.$$

PROBLEMS

1. Which of the following functions satisfy an equation of the form $s'' + as' + bs = 0$?
   (a) te^t (b) $t^2 - t$ (c) $\cos 2t + 3 \sin 2t$
   (d) $\cos 2t + 2 \sin 3t$ (e) $e^{-t} \cos 2t$ (f) $e^t + 4$
   (g) $3t - 9$
2. Find solutions to the following equations having the specified initial values.
   (a) $s'' + 4s = 0$; $s(0) = 1$, $s'(0) = 0$.
   (b) $s'' - 3s' + 2s = 0$; $s(1) = 0$, $s'(1) = -1$.
3. For each of the following equations find a *basis* for the solutions; that is, find two solutions $s_1(t)$, $s_2(t)$ such that every solution has the form $\alpha s_1(t) + \beta s_2(t)$ for suitable constants α, β:
   (a) $s'' + 3s = 0$ (b) $s'' - 3s = 0$
   (c) $s'' - s' - 6s = 0$ (d) $s'' + s' + s = 0$
4. Suppose the roots of the quadratic equation $\lambda^2 + a\lambda + b = 0$ have negative real parts. Prove every solution of the differential equation
   $$s'' + as' + bs = 0$$
   satisfies
   $$\lim_{t \to \infty} s(t) = 0.$$
5. State and prove a generalization of Problem 4 for for nth order differential equations
   $$s^{(n)} + a_1 s^{(n-1)} + \cdots + a_n s = 0,$$
   where the polynomial
   $$\lambda^n + a_1 \lambda^{n-1} + \cdots + a_n$$
   has n distinct roots with negative real parts.

6. Under what conditions on the constants a, b is there a nontrivial solution to $s'' + as + b = 0$ such that the equation

$$s(t) = 0$$

has
(a) no solution;
(b) a positive finite number of solutions;
(c) infinitely many solutions?

7. For each of the following equations sketch the phase portrait of the corresponding first order system. Then sketch the graphs of several solutions $s(t)$ for different initial conditions:
(a) $s'' + s = 0$ (b) $s'' - s = 0$ (c) $s'' + s' + s = 0$
(d) $s'' + 2s' = 0$ (e) $s'' - s' + s$.

8. Which equations $s'' + as' + bs = 0$ have a nontrivial periodic solution? What is the period?

9. Find all solutions to

$$s''' - s'' + 4s' - 4s = 0.$$

10. Find a real-valued function $s(t)$ such that

$$s'' + 4s = \cos 2t,$$

$$s(0) = 0, \qquad s'(0) = 1.$$

11. Find all pairs of functions $x(t)$, $y(t)$ that satisfy the system of differential equations

$$x' = -y,$$

$$y'' = -x - y + y'.$$

12. Let $q(t)$ be a polynomial of degree m. Show that any equation

$$s^{(n)} + a_1 s^{(n-1)} + \cdots + a_n s = q(t)$$

has a solution which is a polynomial of degree $\leq m$.

Notes

A reference to some of the topological background in Section 1 is Bartle's *The Elements of Real Analysis* [2]. Another is Lang's *Analysis I* [11].

Chapter 6

Linear Systems and Canonical Forms of Operators

The aim of this chapter is to achieve deeper insight into the solutions of the differential equation

$$x' = Ax, \qquad A \in L(E), \qquad E = \mathbf{R}^n, \tag{1}$$

by decomposing the operator A into operators of particularly simple kinds. In Sections 1 and 2 we decompose the vector space E into a direct sum

$$E = E_1 \oplus \cdots \oplus E_r$$

and A into a direct sum

$$A = A_1 \oplus \cdots \oplus A_r, \qquad A_k \in L(E_k).$$

Each A_k can be expressed as a sum

$$A_k = S_k + N_k; \qquad S_k, N_k \in L(E_k),$$

with S_k semisimple (that is, its complexification is diagonalizable), and N_k nilpotent (that is, $(N_k)^m = O$ for some m); moreover, S_k and N_k commute. This reduces the series for e^{tA} to a finite sum which is easily computed. Thus solutions to (1) can be found for any A.

Section 3 is devoted to nilpotent operators. The goal is a special, essentially unique matrix representation of a nilpotent operator. This special matrix is applied in Section 4 to the nilpotent part of any operator T to produce special matrices for T called the Jordan form; and for operators on real vector spaces, the real canonical form. These forms make the structure of the operator quite clear.

In Section 5 solutions of the differential equation $x' = Ax$ are studied by means of the real canonical form of A. It is found that all solutions are linear combinations of certain simple functions. Important information about the nature of the solutions can be obtained without explicitly solving the equation.

Section 6 applies the results of Section 5 to the higher order one-dimensional linear homogeneous equation with constant coefficients

$$s^{(n)} + a_1 s^{(n-1)} + \cdots + a_n s = 0. \tag{2}$$

Solutions are easily found if the roots of the characteristic polynomial

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

are known. A different approach to (2), via operators on function spaces, is very briefly discussed in the last section.

The first four sections deal not with differential equations, only linear algebra. This linear algebra, the eigenvector theory of a real operator, is, on one hand, rarely treated in texts, and, on the other hand, important for the study of linear differential equations.

§1. The Primary Decomposition

In this section we state a basic decomposition theorem for operators; the proof is given in Appendix III. It is not necessary to know the proof in order to use the theorem, however.

In the rest of this section T denotes an operator on a vector space E, which may be real or complex; but if E is real it is assumed that all eigenvalues of T are real.

Let the characteristic polynomial of T be given as the product

$$p(t) = \prod_{k=1}^{r} (t - \lambda_k)^{n_k}.$$

Here $\lambda_1, \ldots, \lambda_k$ are the distinct roots of $p(t)$, and the integer $n_k \geq 1$ is the *multiplicity of* λ_k; note that $n_1 + \cdots + n_k = \dim E$.

We recall that the eigenspace of T belonging to λ_k is the subspace

$$\operatorname{Ker}(T - \lambda_k) \subset E$$

(we write λ_k for the operator $\lambda_k I$). Note that T is diagonalizable if and only if E is the direct sum of the eigenspaces (for this means E has a basis of eigenvectors).

We define the *generalized eigenspace of T belonging to* λ_k to be the subspace

$$E(T, \lambda_k) = \operatorname{Ker}(T - \lambda_k)^{n_k} \subset E.$$

Note that this subspace is invariant under T.

The following *primary decomposition theorem* is proved in Appendix. III.

Theorem 1 *Let T be an operator on E, where E is a complex vector space, or else E is real and T has real eigenvalues. Then E is the direct sum of the generalized eigenspaces of T. The dimension of each generalized eigenspace equals the multiplicity of the corresponding eigenvalue.*

Let us see what this decomposition means. Suppose first that there is only one eigenvalue λ, of multiplicity $n = \dim E$. The theorem implies $E = E(T, \lambda)$. Put

$$N = T - \lambda I, \qquad S = \lambda I.$$

Then, clearly, $T = N + S$ and $SN = NS$. Moreover, S is diagonal (in *every* basis) and N is nilpotent, for $E = E(T, \lambda) = \operatorname{Ker} N^n$. We can therefore immediately compute

$$e^T = e^S e^N = e^\lambda \sum_{k=0}^{n-1} \frac{N^k}{k!};$$

there is no difficulty in finding it.

Example 1 Let $T = \begin{bmatrix} 1 & -1 \\ 1 & 3 \end{bmatrix}$. The characteristic polynomial is

$$p(t) = t^2 - 4t + 4 = (t - 2)^2.$$

There is only one eigenvalue, 2, of multiplicity 2. Hence

$$S = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix},$$

$$N = T - S = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}.$$

We know without further computation that N commutes with S and is nilpotent of order 2: $N^2 = 0$. (The reader can verify these statements.) Therefore

$$e^T = e^S e^N = e^2(I + N)$$

$$= e^2 \begin{bmatrix} 0 & -1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 0 & -e^2 \\ e^2 & 2e^2 \end{bmatrix}.$$

More generally,

$$e^{tT} = e^{tS} e^{tN} = e^{2t}(I + tN)$$

$$= e^{2t} \begin{bmatrix} 1 - 2t & -t \\ t & 1 + t \end{bmatrix}.$$

Thus the method applies directly to solving the differential equation $x' = Tx$ (see the previous chapter).

For comparison, try to compute directly the limit of

$$\sum_{k=0}^{\infty} \frac{t^k}{k!} \begin{bmatrix} 1 & -1 \\ 1 & 3 \end{bmatrix}^k.$$

In the general case put

$$T_k = T \mid E(\lambda_k, T).$$

Then $T = T_1 \oplus \cdots \oplus T_r$. Since each T_k has only the one eigenvalue λ_k, we can

apply the previous result. Thus

$$T_k = S_k + N_k; \qquad S_k, N_k \in L(E(\lambda_k, T)),$$

where $S_k = \lambda_k I$ on $E(\lambda_k, T)$, and $N_k = T_k - S_k$ is nilpotent of order n_k. Then

$$T = S + N,$$

where

$$S = S_1 \oplus \cdots \oplus S_r,$$
$$N = N_1 \oplus \cdots \oplus N_r.$$

Clearly, $SN = NS$. Moreover, N *is nilpotent and* S *is diagonalizable.* For if $m = \max(n_1, \ldots, n_r)$, then

$$N^m = (N_1)^m \oplus \cdots \oplus (N_r)^m = 0;$$

and S is diagonalized by a basis for E which is made up of bases for the generalized eigenspaces.

We have proved:

Theorem 2 *Let $T \in L(E)$, where E is complex if T has a nonreal eigenvalue. Then $T = S + N$, where $SN = NS$ and S is diagonalizable and N is nilpotent.*

In Appendix III we shall prove that S *and* N *are uniquely determined by* T.

Using Theorem 2 one can compute the exponential of any operator $T: E \to E$ for which the eigenvalues are known. (Recall we are making the general assumption that if E is real, all the eigenvalues of T must be real.) The method is made clear by the following example.

Example 2 Let $T \in L(\mathbf{R}^3)$ be the operator whose matrix in standard coordinates is

$$T_0 = \begin{bmatrix} -1 & 1 & -2 \\ 0 & -1 & 4 \\ 0 & 0 & 1 \end{bmatrix}.$$

We analyze T_0 with a view toward solving the differential equation

$$x' = T_0 x.$$

The characteristic polynomial of T_0 can be read off from the diagonal because all subdiagonal entries are 0; it is

$$p(t) = (t+1)^2(t-1).$$

The eigenvalues are -1 with multiplicity 2, and 1 with multiplicity 1.

The two-dimensional generalized eigenspace of -1 is spanned by the basis

$$a_1 = (1, 0, 0), \qquad a_2 = (0, 1, 0);$$

this can be read off directly from the first two columns of T_0.

The one-dimensional generalized eigenspace of $+1$ is the solution space of the system of equations

$$(T_0 - 1)x = 0,$$

or

$$\begin{bmatrix} -2 & 1 & -2 \\ 0 & -2 & 4 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0;$$

one can verify that the vector

$$a_3 = (0, 2, 1)$$

is a basis.

Let $\mathcal{B}$ *be the basis* $\{a_1, a_2, a_3\}$ *of* $\mathbf{R}^3$. *Let* $T = S + N$ *be as in Theorem 2. In* $\mathcal{B}$*-coordinates,* S has the matrix

$$S_1 = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix};$$

this follows from the eigenvalues of T being $-1, -1, 1$. Let S_0 be the matrix of S in standard coordinates. Then

$$S_1 = PS_0P^{-1},$$

where P is the inverse transpose of the matrix whose rows are a_1, a_2, a_3. Hence

$$(P^{-1})^t = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix},$$

$$P^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix},$$

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix}.$$

Therefore

$$S_0 = P^{-1}S_1P.$$

Matrix multiplication gives

$$S_0 = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 4 \\ 0 & 0 & 1 \end{bmatrix}.$$

We can now find the matrix N_0 of N in the standard basis for $\mathbf{R}^3$,

$$N_0 = T_0 - S_0$$
$$= \begin{bmatrix} -1 & 1 & -2 \\ 0 & -1 & 4 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 4 \\ 0 & 0 & 1 \end{bmatrix}$$
$$= \begin{bmatrix} 0 & 1 & -2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

We have now computed the matrices of S and N. The reader might verify that $N^2 = O$ and $SN = NS$.

We compute the matrix in standard coordinates of e^S *not* by computing the matrix e^{S_0} directly from the definition, which involves an infinite series, but as follows:

$$\exp(S_0) = \exp(P^{-1}S_1P) = P^{-1}\exp(S_1)P$$
$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} e^{-1} & 0 & 0 \\ 0 & e^{-1} & 0 \\ 0 & 0 & e \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix},$$

which turns out to be

$$\exp(S_0) = \begin{bmatrix} e^{-1} & 0 & 0 \\ 0 & e^{-1} & -2e^{-1} + 2e \\ 0 & 0 & e \end{bmatrix}.$$

It is easy to compute $\exp(N_0)$:

$$\exp(N_0) = I + N_0$$
$$= \begin{bmatrix} 1 & 1 & -2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Finally, we obtain

$$\exp(T_0) = \exp(S_0 + N_0) = \exp(S_0)\exp(N_0),$$

which gives

$$\exp(T_0) = \begin{bmatrix} e^{-1} & e^{-1} & -2e^{-1} \\ 0 & e^{-1} & -2e^{-1} + 2e \\ 0 & 0 & e \end{bmatrix}.$$

It is no more difficult to compute e^{tT_0}, $t \in \mathbf{R}$. Replacing T_0 by tT_0 transforms S_0 to tS_0, N_0 to tN_0, and so on; the point is that the *same* matrix P is used for all values to t. One obtains

$$\exp(tT_0) = \exp(tS_0)\exp(tN_0)$$
$$= \begin{bmatrix} e^{-t} & 0 & 0 \\ 0 & e^{t} & -2e^{-t} + 2e^{t} \\ 0 & 0 & e^{t} \end{bmatrix} \begin{bmatrix} 1 & t & -2t \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$
$$= \begin{bmatrix} e^{-t} & te^{-t} & -2te^{-t} \\ 0 & e^{-t} & -2e^{-t} + 2e^{t} \\ 0 & 0 & e^{t} \end{bmatrix}.$$

The solution of $x' = T_0x$ is given in terms of $\exp(tT_0)$.

The following consequence of the primary decomposition is called the Cayley-Hamilton theorem.

Theorem 3 *Let A be any operator on a real or complex vector space. Let its characteristic polynomial be*

$$p(t) = \sum_{k=0}^{n} a_k t^k.$$

Then $p(A) = 0$, that is,

$$\sum_{k=0}^{n} a_k A^k(x) = 0$$

for all $x \in E$.

Proof. We may assume $E = \mathbf{R}^n$ or $\mathbf{C}^n$; since an operator on $\mathbf{R}^n$ and its complexification have the same characteristic polynomial, there is no loss of generality in assuming E is a complex vector space.

It suffices to show that $P(A)x = 0$ for all x in an arbitrary generalized eigenspace

$E(A, \lambda)$, where $p(\lambda) = 0$. Now

$$E(\lambda, A) = \mathrm{Ker}(A - \lambda)^m,$$

where m is the multiplicity of λ. Since $(t - \lambda)^m$ divides $p(t)$ we can write

$$p(t) = q(t)(t - \lambda)^m.$$

Hence, for $x \in E(\lambda, A)$:

$$\begin{aligned} p(A)x &= q(A)[(A - \lambda)^m x] \\ &= q(A)(0) = 0. \end{aligned}$$

§2. The $S + N$ Decomposition

Let T be an operator on $\mathbf{R}^n$ and $T_C: \mathbf{C}^n \to \mathbf{C}^n$ its complexification. If T_C is diagonalizable, we say T is *semisimple.*

Theorem 1 *For any operator $T \in L(\mathbf{R}^n)$ there are unique operators S, N on $\mathbf{R}^n$ such that $T = S + N$, $SN = NS$, S is semisimple, and N is nilpotent.*

Proof. We have already seen a similar theorem for operators on complex vector spaces; now we apply this result to prove the theorem for operators on $\mathbf{R}^n$. Let $\sigma: \mathbf{C}^n \to \mathbf{C}^n$ be the operator of conjugation (Chapter 4); if $z = x + iy \in \mathbf{C}^n$, where $x, y \in \mathbf{R}^n$, then $\sigma z = x - iy$. An operator Q on $\mathbf{C}^n$ is the complexification of an operator on $\mathbf{R}^n$ if and only $Q\sigma = \sigma Q$.

Given $T \in L(\mathbf{R}^n)$ let $T_C \in L(\mathbf{C}^n)$ be its complexification. By Theorem 2, Section 1, there are *unique* operators S_0, N_0 on $\mathbf{C}^n$ such that

$$T_C = S_0 + N_0,$$

$S_0N_0 = N_0S_0$, S_0 diagonalizable, and N_0 nilpotent. We assert that S_0 and N_0 are complexifications of operators on $\mathbf{R}^n$. This is proved by showing they commute with σ, as follows. Put $S_1 = \sigma S_0 \sigma^{-1}$, $N_1 = \sigma N_0 \sigma^{-1}$. Then

$$T_C = \sigma T_C \sigma^{-1} = S_1 + N_1.$$

It is easy to see that S_1 is diagonalizable, N_1 is nilpotent, and $S_1N_1 = N_1S_1$. Therefore $S_0 = S_1$ and $N_0 = N_1$. This means that S_0 and N_0 commute with σ as asserted.

There are unique operators S, N in $L(\mathbf{R}^n)$ such that

$$S_0 = S_C, \qquad N_0 = N_C.$$

Since the map $A \to A_C$ is one-to-one, it follows that

$$SN = NS$$

for

$$(SN - NS)_C = S_0N_0 - N_0S_0 = O.$$

Similar reasoning shows that N is nilpotent, and also $S + N = T$. The uniqueness of S and N follows from uniqueness of S_0 and N_0. This completes the proof.

Definition S is called the *semisimple part* of T and N the *nilpotent part.*

Let $T = S + N$ as in Theorem 1. Since S_C is diagonalizable, it follows from Chapter 4 that in a suitable basis $\mathcal{B}$ of $\mathbf{R}^n$, described below, S has a matrix of the form

$$L = \begin{bmatrix} \lambda_1 & & & & & & \\ & \ddots & & & & & \\ & & \lambda_r & & & & \\ & & & \begin{bmatrix} a_1 & -b_1 \\ b_1 & a_1 \end{bmatrix} & & & \\ & & & & \ddots & & \\ & & & & & \begin{bmatrix} a_s & -b_s \\ b_s & a_s \end{bmatrix} \end{bmatrix}$$

Here $\lambda_1, \ldots, \lambda_r$ are the real eigenvalues of T, with multiplicity; and the complex numbers

$$a_k + ib_k; \qquad k = 1, \ldots, s,$$

are the complex eigenvalues with positive imaginary part, with multiplicity. Note that T, T_C, S_C, and S have the same eigenvalues.

The exponential of the matrix tL, $t \in \mathbf{R}$ is easy to calculate since

$$\exp \begin{bmatrix} ta & -tb \\ tb & ta \end{bmatrix} = e^{ta} \begin{bmatrix} \cos tb & -\sin tb \\ \sin tb & \cos tb \end{bmatrix}.$$

The basis $\mathcal{B}$ that gives S the matrix L is obtained as follows. The first r vectors in $\mathcal{B}$ are from bases for the generalized eigenspaces of T that belong to real eigenvalues. The remaining $2s$ vectors are the imaginary and real parts of bases of the generalized eigenspaces of T_C that belong to eigenvalues $a + ib$, $b > 0$.

In this way e^{tT} can be computed for any operator T, provided the eigenvalues of T are known.

Example. Let $T \in L(\mathbf{R}^4)$ be the operator whose matrix in standard coordinates

is

$$T_0 = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 2 & 0 & 1 & 0 \end{bmatrix}.$$

In $\mathbf{C}^4$ the generalized i-eigenspace is the solution space of

$$(T_0 - i)^4 z = 0,$$

or of

$$\begin{aligned} -2z_1 - 2iz_2 &= 0, \\ -2iz_1 - 2z_2 &= 0, \\ -2z_1 - 2z_3 + 2iz_4 &= 0, \\ -4iz_1 - 2z_2 - 2iz_3 - 2z_4 &= 0. \end{aligned}$$

These are equivalent to

$$\begin{aligned} z_1 &= iz_2, \\ -z_3 + iz_4 &= iz_2. \end{aligned}$$

As a basis for the solution space we pick the complex vectors

$$u = (i, 1, 0, 1), \qquad v = (i, 1, -i, 0).$$

From these we take imaginary and real parts:

$$\begin{aligned} \mathrm{I}\, u &= (1, 0, 0, 0) = e_1, & \mathrm{I}\, v &= (1, 0, -1, 0) = e_3, \\ \mathrm{R}\, u &= (0, 1, 0, 1) = e_2, & \mathrm{R}\, v &= (0, 1, 0, 0) = e_4. \end{aligned}$$

These four vectors, in order, form a basis $\mathfrak{B}$ of $\mathbf{R}^4$. This basis gives S the matrix

$$S_1 = \begin{bmatrix} 0 & -1 & & \\ 1 & 0 & & \\ & & 0 & -1 \\ & & 1 & 0 \end{bmatrix}.$$

(We know this without further computation.)

The matrix of S in standard coordinates is

$$S_0 = P^{-1} S_1 P,$$

where P^{-1} is the transpose of the matrix of components of $\mathfrak{B}$; thus

$$P^{-1} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

and one finds that

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

Hence

$$S_0 = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & 1 & 0 \end{bmatrix}.$$

The matrix of N in standard coordinates is then

$$\begin{aligned} N_0 &= T_0 - S_0 \\ &= \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 2 & 0 & 1 & 0 \end{bmatrix} - \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \\ &= \begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 0 & -1 & \cdot & \cdot \\ 1 & 0 & \cdot & \cdot \end{bmatrix}, \end{aligned}$$

which indeed is nilpotent of order 2 (where $\cdot$ denotes a zero).

The matrix of e^{tT} in standard coordinates is

$$\begin{aligned} \exp(tT_0) &= \exp(tN_0 + tS_0) = \exp(tN_0)\exp(tS_0) \\ &= (I + tN_0)P \exp(tS_1)P^{-1}. \end{aligned}$$

From

$$\exp(tS_1) = \begin{bmatrix} \cos t & -\sin t & & \\ \sin t & \cos t & & \\ & & \cos t & -\sin t \\ & & \sin t & \cos t \end{bmatrix}$$

the reader can complete the computation.

PROBLEMS

1. For each of the following operators T find bases for the generalized eigenspaces; give the matrices (for the standard basis) of the semisimple and nilpotent parts of T.

(a) $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

(d) $\begin{bmatrix} 0 & 2 & 0 \\ -2 & 0 & 0 \\ 2 & 0 & 6 \end{bmatrix}$ (e) $\begin{bmatrix} 0 & 0 & 8 \\ 0 & 0 & 4 \\ 0 & 1 & -2 \end{bmatrix}$ (f) $\begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 4 & 4 & 4 & 4 \end{bmatrix}$

2. A matrix $[a_{ij}]$ such that $a_{ij} = 0$ for $i \leq j$ is nilpotent.

3. What are the eigenvalues of a nilpotent matrix?

4. For each of the following matrices A, compute e^{tA}, $t \in \mathbf{R}$:

(a) $\begin{bmatrix} 0 & 0 & 5 \\ 2 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & -1 & 1 & 0 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 0 & 0 \\ -1 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 0 & 2 & 0 \\ 0 & 0 & 3 \\ 1 & 0 & 0 \end{bmatrix}$

5. Prove that an operator is nilpotent if all its eigenvalues are zero.

6. The semisimple and nilpotent parts of T commute with A if T commutes with A.

7. If A is nilpotent, what kind of functions are the coordinates of solutions to $x' = Ax$?

8. If N is a nilpotent operator on an n-dimensional vector space then $N^n = 0$.

9. What can be said about AB and $A + B$ if $AB = BA$ and
   (a) A and B are nilpotent?
   (b) A and B are semisimple?
   (c) A is nilpotent and B is semisimple?

10. If A and B are commuting operators, find a formula for the semisimple and nilpotent parts of AB and $A + B$ in terms of the corresponding parts of A and B. Show by example that the formula is not always valid if A and B do not commute.

11. Identify $\mathbf{R}^{n+1}$ with the set P_n of polynomials of degree $\leq n$, via the correspondence

$$(a_n, \ldots, a_0) \leftrightarrow a_n t^n + \cdots + a_1 t + a_0.$$

Let $D\colon P_n \to P_n$ be the differentiation operator. Prove D is nilpotent.

12. Find the matrix of D in the standard basis in Problem 11.

13. A rotation around a line in $\mathbf{R}^3$ and reflection in a plane in $\mathbf{R}^3$ are semisimple operators.

14. Let S be semisimple and N nilpotent. If $SN = NS = O$, then $S = O$ or $N = O$. (*Hint*: Consider generalized eigenspaces of $S + N$.)

15. If $T^2 = T$, then T is diagonalizable. (*Hint*: Do not use any results in this chapter!)

16. Find necessary and sufficient conditions on a, b, c, d in order that the operator $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ be
   (a) diagonalizable; (b) semisimple; (c) nilpotent.

17. Let $F \subset E$ be invariant under $T \in L(E)$. If T is nilpotent, or semisimple, or diagonalizable, so is $T \mid F$.

18. An operator $T \in L(E)$ is semisimple if and only if for every invariant subspace $F \subset E$, there is another invariant subspace $F' \subset E$ such that $E = F \oplus F'$.

19. Suppose T is nilpotent and

$$T^k = \sum_{j=0}^{k-1} a_j T^j, \qquad a_j \in \mathbf{R}.$$

Then $T^k = O$.

20. What values of a, b, c, d make the following operators semisimple?

(a) $\begin{bmatrix} 0 & a \\ -1 & 2 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & -1 \\ 1 & b \end{bmatrix}$ (c) $\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 1 \\ 0 & 0 & c \end{bmatrix}$ (d) $\begin{bmatrix} 0 & d & 0 \\ 1 & 0 & d \\ d & 1 & 0 \end{bmatrix}$

21. What values of a, b, c, d make the operators in Problem 20 nilpotent?

§3. Nilpotent Canonical Forms

In the previous section we saw that any operator T can be decomposed uniquely as

$$T = S + N$$

with S semisimple, N nilpotent, and $SN = NS$. We also found a *canonical form* for S, that is, a type of matrix representing S which is uniquely determined by T, except for the ordering of diagonal blocks. In the complex case, for example,

$$S = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\},$$

where $\lambda_1, \ldots, \lambda_n$ are the roots of the characteristic polynomial of T, listed with their proper multiplicities.

Although we showed how to find *some* matrix representation of N, we did not give any special one. In this section we shall find for any nilpotent operator a matrix that is uniquely determined by the operator (except for order of diagonal blocks). From this we shall obtain a special matrix for any operator, called the Jordan canonical form.

An *elementary nilpotent block* is a matrix of the form

(1)
$$\begin{bmatrix} 0 & & & \\ 1 & \cdot & & \\ & \cdot & \cdot & \\ & & \cdot & \cdot \\ & & & 1 \quad 0 \end{bmatrix}$$

with 1's just below the diagonal and 0's elsewhere. We include the one-by-one matrix $[0]$.

If $N: E \to E$ is an operator represented by such a matrix in a basis $e_1, \ldots, e_n$, then N behaves as follows on the basis elements:

$$\begin{aligned} N(e_1) &= e_2, \\ N(e_2) &= e_3, \\ &\vdots \\ N(e_{n-1}) &= e_n, \\ N(e_n) &= 0. \end{aligned}$$

It is obvious that $N^n(e_k) = 0$, $k = 1, \ldots, n$; hence $N^n = O$. Thus N is nilpotent of order n. Moreover, $N^k \neq O$ if $0 \leq k < n$, since $N^k e_1 = e_{k+1} \neq 0$.

In Appendix III we shall prove

Theorem 1 *Let N be a nilpotent operator on a real or complex vector space E. Then E has a basis giving N a matrix of the form*

$$A = \operatorname{diag}\{A_1, \ldots, A_r\},$$

where A_i is an elementary nilpotent block, and the size of A_k is a nonincreasing function of k. The matrices $A_1, \ldots, A_r$ are uniquely determined by the operator N.

We call the matrix in Theorem 1 the *canonical form* of N.

Let A be an elementary nilpotent matrix. It is evident that the rank of A is $n - 1$; hence

$$\dim \operatorname{Ker} A = 1.$$

This implies the following corollary of Theorem 1.

Theorem 2 *In Theorem 1 the number r of blocks is equal to* dim Ker A.

We define the canonical form of a nilpotent matrix to be the canonical form of the corresponding operator; this is a matrix similar to the original one. Since similar matrices correspond to the same operator, it follows that they have the same canonical form. From this we conclude:

Theorem 3 *Two nilpotent $n \times n$ matrices, or two nilpotent operators on the same vector space, are similar if and only if they have the same canonical form.*

The question arises: given a nilpotent operator, how is its canonical form found? To answer this let us examine a nilpotent matrix which is already in canonical form, say, the 10×10 matrix

$$N = \begin{bmatrix} \begin{bmatrix} 0 & \\ 1 & 0 \\ & 1 \quad 0 \end{bmatrix} & & & & \\ & \begin{bmatrix} 0 & \\ 1 & 0 \end{bmatrix} & & & \\ & & \begin{bmatrix} 0 & \\ 1 & 0 \end{bmatrix} & & \\ & & & [0] & \\ & & & & [0] \\ & & & & & [0] \end{bmatrix}.$$

We consider N as representing a nilpotent operator T on $\mathbf{R}^{10}$. Consider the relations between the following sets of numbers:

$$\delta_k = \dim \operatorname{Ker} T^k, \qquad 1 \leq k \leq 10,$$

and

$$\nu_k = \text{number of elementary nilpotent } k \times k \text{ blocks}, \qquad 1 \leq k \leq 10.$$

Note that $\nu_k = 0$ if $k > 3$. The numbers δ_k depend on the operator, and can be computed from any matrix for T. On the other hand, if we know the ν_k we can immediately write down the matrix N. The problem, then, is to compute the ν_k *in terms of the* δ_k.

Consider

$$\delta_1 = \dim \operatorname{Ker} T.$$

From Theorem 2 we find

$$\delta_1 = \text{total number of blocks} = \nu_1 + \nu_2 + \nu_3.$$

Next, consider $\delta_2 = \dim \operatorname{Ker} T^2$. Each 1×1 block (that is, the blocks $[0]$) contributes one dimension to $\operatorname{Ker} T^2$. Each 2×2 block contributes 2, while the 3×3 block also contributes 2. Thus

$$\delta_2 = \nu_1 + 2\nu_2 + 2\nu_3.$$

For $\delta_3 = \dim \operatorname{Ker} T^3$, we see that the 1×1 blocks each contribute 1; the 2×2 blocks each contribute 2; and the 3×3 block contributes 3. Hence

$$\delta_3 = \nu_1 + 2\nu_2 + 3\nu_3.$$

In this example $N^3 = 0$, hence $\delta_3 = \delta_k$, $k > 3$.

For an arbitrary nilpotent operator T on a vector space of dimension n, let N be the canonical form; define the numbers δ_k and ν_k, $k = 1, \ldots, n$, as before. By the same reasoning we obtain the equations

$$\begin{aligned}
\delta_1 &= \nu_1 + \nu_2 + \cdots + \nu_n, \\
\delta_2 &= \nu_1 + 2(\nu_2 + \cdots + \nu_n), \\
\delta_3 &= \nu_1 + 2\nu_2 + 3(\nu_3 + \cdots + \nu_n), \\
&\vdots \\
\delta_{n-1} &= \nu_1 + 2\nu_2 + \cdots + (n-2)\nu_{n-2} + (n-1)(\nu_{n-1} + \nu_n), \\
\delta_n &= \nu_1 + 2\nu_2 + \cdots + n\nu_n.
\end{aligned}$$

We think of the δ_k as known and solve for the ν_k. Subtracting each equation from

the one below it gives the equivalent system:

$$\begin{aligned}
\delta_1 &= \nu_1 + \cdots + \nu_n, \\
-\delta_1 + \delta_2 &= \nu_2 + \cdots + \nu_n, \\
-\delta_2 + \delta_3 &= \nu_3 + \cdots + \nu_n, \\
&\vdots \\
-\delta_{n-1} + \delta_n &= \nu_n.
\end{aligned}$$

Subtracting the second of these equations from the first gives

$$\nu_1 = 2\delta_1 - \delta_2.$$

Subtracting the $(k+1)$th from the kth gives

$$\nu_k = -\delta_{k-1} + 2\delta_k - \delta_{k+1}, \qquad 1 < k < n;$$

and the last equation gives ν_n. Thus we have proved the following theorem, in which part (b) allows us to compute the canonical form of any nilpotent operator:

Theorem 4 *Let T be a nilpotent operator on an n-dimensional vector space. If ν_k is the number of $k \times k$ blocks in the canonical form of T, and $\delta_k = \dim \operatorname{Ker} T^k$, then the following equations are valid:*

(a) $\delta_m = \sum_{1 \leq k < m} k\nu_k + m \sum_{m \leq j \leq n} \nu_j$; $m = 1, \ldots, n$;

(b) $\nu_1 = 2\delta_1 - \delta_2$,
$\nu_k = -\delta_{k-1} + 2\delta_k - \delta_{k+1}$, $1 < k < n$,
$\nu_n = -\delta_{n-1} + \delta_n$.

Note that the equations in (b) can be subsumed under the single equation

$$\nu_k = -\delta_{k-1} + 2\delta_k - \delta_{k+1},$$

valid for *all* integers $k \geq 1$, if we note that $\delta_0 = 0$ and $\delta_k = \delta_n$ for $k > n$.

There is the more difficult problem of finding a basis that puts a given nilpotent operator in canonical form. An algorithm is implicit in Appendix III. Our point of view, however, is to obtain theoretical information from canonical forms. For example, the equations in the preceding theorem immediately prove that *two nilpotent operators N, M on a vector space E are similar if and only if* $\dim \operatorname{Ker} N^k = \dim \operatorname{Ker} M^k$ *for* $1 \leq k \leq \dim E$.

For computational purposes, the $S + N$ decomposition is usually adequate. On the other hand, the existence and uniqueness of the canonical forms is important for theory.

PROBLEMS

1. Verify that each of the following operators is nilpotent and find its canonical form:

(a) $\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 2 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 0 & 2 & -2 \\ 0 & 0 & 4 \\ 0 & 0 & 0 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 6 & 7 & 0 & 0 \\ 8 & 9 & 0 & 0 \end{bmatrix}$

(d) $\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 3 & 0 \end{bmatrix}$ (e) $\begin{bmatrix} 1 & 1 & 0 & 0 \\ -1 & -1 & 0 & 0 \\ 0 & 1 & 2 & -2 \\ 1 & 0 & 2 & -2 \end{bmatrix}$

2. Let N be a matrix in nilpotent canonical form. Prove N is similar to
   (a) kN for all nonzero $k \in \mathbf{R}$,
   (b) the transpose of N.
3. Let N be an $n \times n$ nilpotent matrix of rank r. If $N^k = 0$, then $k \geq n/(n-r)$.
4. Classify the following operators on $\mathbf{R}^4$ by similarity (missing entries are 0):

(a) $\begin{bmatrix} 0 & 1 & & \\ & 0 & 2 & \\ & & 0 & 3 \\ & & & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 0 & 0 & 2 \\ 0 & & & 0 \\ 0 & & & 0 \\ -2 & 0 & 0 & -2 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 0 & & \\ 4 & 0 & & \\ & & 0 & 4 \\ & & 0 & 0 \end{bmatrix}$

(d) $\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0 & 1 & 0 & \\ -1 & -1 & -1 & 0 \end{bmatrix}$ (e) $\begin{bmatrix} 0 & 0 & 0 & 100 \\ & & & 0 \\ & & & 0 \\ & & & 0 \end{bmatrix}$

§4. Jordan and Real Canonical Forms

In this section canonical forms are constructed for arbitrary operators.

We start with an operator T on E that has only one eigenvalue λ; if λ is nonreal, we suppose E complex. In Section 1 we saw that

$$T = \lambda I + N$$

with N nilpotent. We apply Theorem 1 of Section 3 and give E a basis $\mathcal{B} = \{e_1, \ldots, e_n\}$ that gives N a matrix A in nilpotent canonical form. Hence T has the $\mathcal{B}$-matrix $\lambda I + A$. Since A is composed of diagonal blocks, each of which is an

elementary nilpotent matrix, $\lambda I + A$ has the form

(1) $\begin{bmatrix} \lambda & & & \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & \lambda \end{bmatrix}$ $\begin{bmatrix} \lambda & & \\ 1 & \ddots & \\ & \ddots & \ddots \\ & & 1 \quad \lambda \end{bmatrix}$ $\ddots$ $\begin{bmatrix} \lambda & & \\ 1 & \ddots & \\ & \ddots & \ddots \\ & & 1 \quad \lambda \end{bmatrix}$

(Some of the diagonal blocks may be 1×1 matrices $[\lambda]$.) That is, $\lambda I + A$ has λ's along the diagonal; below the diagonal are 1's and 0's; all other entries are 0. The blocks making up $\lambda I + A$ are called *elementary Jordan matrices*, or *elementary λ-blocks*. A matrix of the form (1) is called a *Jordan matrix belonging to* λ, or briefly, a *Jordan λ-block*.

Consider next an operator $T: E \to E$ whose distinct eigenvalues are $\lambda_1, \ldots, \lambda_m$; as usual E is complex if some eigenvalue is nonreal. Then $E = E_1 \oplus \cdots \oplus E_m$, where E_k is the generalized λ_k-eigenspace, $k = 1, \ldots, m$. We know that $T \mid E_k = \lambda_k I + N_k$ with N_k nilpotent. We give E_k a basis $\mathcal{B}_k$, which gives $T \mid E_k$ a Jordan matrix belonging to λ_k. The basis $\mathcal{B} = \mathcal{B}_1 \cup \cdots \cup \mathcal{B}_k$ of E gives T a matrix of the form

$$C = \operatorname{diag}\{C_1, \ldots, C_m\},$$

where each C_k is a Jordan matrix belonging to λ_k. Thus C is composed of diagonal blocks, each of which is an elementary Jordan matrix C. The matrix C is called the *Jordan form* (or Jordan matrix) of T.

We have constructed a particular Jordan matrix for T, by decomposing E as a direct sum of the generalized eigenspaces of T. But it is easy to see that given any Jordan matrix M representing T, each Jordan λ-block of M represents the restric-

tion of T to the generalized λ-eigenspace. Thus M must be the matrix we constructed, perhaps with the λ-blocks rearranged.

It is easy to prove that similar operators have the same Jordan forms (perhaps with rearranged λ-blocks). For if $PT_0P^{-1} = T_1$, then P maps each generalized λ-eigenspace of T_0 isomorphically onto the generalized λ-eigenspace of T_1; hence the Jordan λ-blocks are the same for T_0 and T_1.

In summary:

Theorem 1 *Let $T \in L(E)$ be an operator; if E is real, assume all eigenvalues of T are real. Then E has a basis giving T a matrix in Jordan form, that is, a matrix made up of diagonal blocks of the form* (1).

Except for the order of these blocks, the matrix is uniquely determined by T. Any operator similar to T has the same Jordan form. The Jordan form can be written $A + B$, where B is a diagonal matrix representing the semisimple part of T while A is a canonical nilpotent matrix which represents the nilpotent part of T; and $AB = BA$.

Note that each elementary λ-block contributes 1 to the dimension of $\operatorname{Ker}(T - \lambda)$. Therefore,

Proposition *In the Jordan form of an operator T, the number of elementary λ-blocks is* $\dim \operatorname{Ker}(T - \lambda)$.

We turn now to an operator T on a real vector space E, allowing T to have nonreal eigenvalues. Let $T_C: E_C \to E_C$ be the complexification of T. Then E_C has a basis $\mathfrak{B}$ putting T_C into Jordan form. This basis $\mathfrak{B}$ is made up of bases for each generalized eigenspace of T_C. We observed in Chapter 4, Section 2, that for a real eigenvalue λ, the generalized eigenspace $E_C(T_C, \lambda)$ is the complexification of a subspace of E, and hence has a basis of vectors in E; the matrix of $T_C \mid E(T_C, \lambda)$ in this basis is thus a real matrix which represents $T \mid E(T, \lambda)$. It is a Jordan λ-block.

Let $\mu = a + ib$, $b > 0$ be a nonreal eigenvalue of T. Let

$$\{x_1 + iy_1, \ldots, x_p + iy_p\}$$

be a basis for $E(\mu, T_C)$, giving $T_C \mid E(\mu, T_C)$ a Jordan matrix belonging to μ. In Section 2 we saw that

$$E(\mu, T_C) \oplus E(\bar{\mu}, T_C)$$

is the complexification of a subspace $E_\mu \subset E$ which is T-invariant; and the vectors

$$\{y_1, x_1, \ldots, y_p, x_p\}$$

are a basis for E. It is easy to see that in this basis, $T \mid E_\mu$ has a matrix composed

of diagonal blocks of the form

(2)
$$\begin{bmatrix} D & & & \\ I_2 & \ddots & & \\ & \ddots & \ddots & \\ & & I_2 & D \end{bmatrix} \quad \text{or} \quad D,$$

where

$$D = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}, \qquad I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Thus $T \mid E_\mu$ has a matrix of the form

(3)
$$\begin{bmatrix} \begin{bmatrix} D & & \\ I_2 & \ddots & \\ & \ddots & D \end{bmatrix} & & & \\ & \begin{bmatrix} D & & \\ I_2 & \ddots & \\ & \ddots & D \end{bmatrix} & & \\ & & \ddots & \\ & & & \begin{bmatrix} D & & \\ I_2 & \ddots & \\ & \ddots & D \end{bmatrix} \end{bmatrix}$$

Combining these bases, we obtain

Theorem 2 *Let $T: E \to E$ be an operator on a real vector space. Then E has a basis giving T a matrix composed of diagonal blocks of the forms* (1) *and* (2). *The diagonal*

elements are the real eigenvalues, with multiplicity. Each block $\left[\begin{smallmatrix} a & -b \\ b & a \end{smallmatrix}\right]$, $b > 0$, *appears as many times as the multiplicity of the eigenvalue* $a + bi$. *Such a matrix is uniquely determined by the similarity class of* T, *except for the order of the blocks.*

Definition The matrix described in the theorem is called the *real* canonical form of T. If T has only real eigenvalues, it is the same as the Jordan form. If T is nilpotent, it is the same as the canonical form discussed earlier for nilpotent operators.

The previous theory applies to T_C to show:

Proposition *In the real canonical form of an operator* T *on a real vector space, the number of blocks of the form*

$$\begin{bmatrix} \lambda & & & \\ 1 & \cdot & & \\ & \cdot & \cdot & \\ & & 1 & \lambda \end{bmatrix}$$

is $\dim \operatorname{Ker}(T - \lambda)$. *The number of blocks of the form* (2) *is* $\dim \operatorname{Ker}(T_C - (a + ib))$.

The real canonical form of an operator T exhibits the eigenvalues as part of a matrix for T. This ties them to T much more directly than their definition as roots of the characteristic polynomial. For example, it is easy to prove:

Theorem 3 *Let* $\lambda_1, \ldots, \lambda_n$ *be the eigenvalues (with multiplicities) of an operator* T. *Then*

(a) $$\operatorname{Tr}(T) = \lambda_1 + \cdots + \lambda_n,$$

(b) $$\operatorname{Det}(T) = \lambda_1 \cdots \lambda_n.$$

Proof. We may replace a real operator by its complexification, without changing its trace, determinant, or eigenvalues. Hence we may assume T operates on a complex vector space. The trace is the sum of the diagonal elements of any matrix for T; looking at the Jordan form proves (a). Since the Jordan form is a triangular matrix, the determinant of T is the product of its diagonal elements. This proves (b).

To compute the canonical form of an operator T we apply Theorem 4 of Section 3 to the nilpotent part of $T - \lambda$ for each real eigenvalue λ, and to $T_C - (a + bi)$ for each complex eigenvalue $a + bi$, $b > 0$. For each real eigenvalue λ define $\nu_k(\lambda)$ =

number of $k \times k$ blocks of the form

$$\begin{bmatrix} \lambda & & & \\ 1 & \cdot & & \\ & \cdot & \cdot & \\ & & 1 & \lambda \end{bmatrix}$$

in the real Jordan form of T; and

$$\delta_k(\lambda) = \dim \operatorname{Ker}(T - \lambda)^k.$$

For each complex eigenvalue $\lambda = a + bi$, $b > 0$, define $\nu_k(\lambda)$ = number of $2k \times 2k$ blocks of the form

$$\begin{bmatrix} D & & & \\ I & \cdot & & \\ & \cdot & \cdot & \\ & & I & D \end{bmatrix}, \qquad D = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}, \qquad I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

in the real Jordan form of T; and

$$\delta_k(\lambda) = \dim \operatorname{Ker}(T_C - \lambda)$$

as a complex vector space. One obtains:

Theorem 4 *Let* T *be an operator on a real* n*-dimensional vector space. Then the real Jordan form of* T *is determined by the following equations:*

$$\nu_k(\lambda) = -\delta_{k-1}(\lambda) + 2\delta_k(\lambda) - \delta_{k+1}(\lambda), \qquad 1 \leq k \leq n,$$

where λ *runs through all real eigenvalues and all complex eigenvalues with positive imaginary part.*

Example. Find the real canonical form of the operator

$$T = \begin{bmatrix} 0 & 0 & 0 & -8 \\ 1 & 0 & 0 & 16 \\ 0 & 1 & 0 & -14 \\ 0 & 0 & 1 & 6 \end{bmatrix}.$$

The characteristic polynomial is

$$(t - (1 + i)(t - (1 - i))(t - 2)^2.$$

The eigenvalues are thus $1 + i$, $1 - i$, 2, 2. Since $1 + i$ has multiplicity 1, there can only be one block $\left[\begin{smallmatrix} 1+i \end{smallmatrix}\right]$. A computation shows

$$\delta_1(2) = 1.$$

This is proved most easily by showing that rank $(T - 2) = 3$. Hence there is only one elementary 2-block. The real canonical form is thus:

$$\begin{bmatrix} 2 & \cdot & \cdot & \cdot \\ 1 & 2 & \cdot & \cdot \\ \cdot & \cdot & 1 & -1 \\ \cdot & \cdot & 1 & 1 \end{bmatrix}.$$

There remains the problem of finding a basis that puts an operator in real canonical form. An algorithm can be derived from the procedure in Appendix III for putting nilpotent operators in canonical form. We shall have no need for it, however.

PROBLEMS

1. Find the Jordan forms of the following operators on $\mathbf{C}^n$:

   (a) $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} i & -1 \\ 1 & i \end{bmatrix}$ (c) $\begin{bmatrix} 1+i & 2 \\ 0 & 1+i \end{bmatrix}$

2. Find the real canonical forms of the operators in Problem 1, Section 2.
3. Find the real canonical forms of operators in Problem 4, Section 2.
4. What are the possible real canonical forms of an operator on $\mathbf{R}^n$ for $n \leq 5$?
5. Let A be a 3×3 real matrix which is not diagonal. If $(A + I)^2 = O$, find the real canonical form of A.
6. Let A be an operator. Suppose $q(\lambda)$ is a polynomial (not identically 0) such that $q(A) = O$. Then the eigenvalues of A are roots of q.
7. Let A, B be commuting operators on $\mathbf{C}^n$ (respectively, $\mathbf{R}^n$). There is a basis putting both of them in Jordan (respectively, real) canonical form.
8. Every $n \times n$ matrix is similar to its transpose.
9. Let A be an operator on $\mathbf{R}^n$. An operator B on $\mathbf{R}^n$ is called a *real logarithm* of A if $e^B = A$. Show that A has a real logarithm if and only if A is an isomorphism and the number of Jordan λ-blocks is even for each negative eigenvalue λ.

10. Show that the number of real logarithms of an operator on $\mathbf{R}^n$ is either 0, 1, or countably infinite.

§5. Canonical Forms and Differential Equations

After a long algebraic digression we return to the differential equation

$$x' = Ax, \qquad A \in L(\mathbf{R}^n). \tag{1}$$

Suppose A is Jordan λ-block, $\lambda \in \mathbf{R}$:

$$\begin{bmatrix} \lambda & & & \\ 1 & \cdot & & \\ & \cdot & \cdot & \\ & & \cdot & \\ & & 1 & \lambda \end{bmatrix}.$$

From the decomposition

$$A = \lambda + N,$$

$$N = \begin{bmatrix} 0 & & & \\ 1 & & & \\ & \cdot & & \\ & & \cdot & \\ & & 1 & 0 \end{bmatrix}$$

we find by the exponential method (Chapter 5) that the solution to (1) with initial value $x(0) = C \in \mathbf{R}^n$ is

$$\begin{aligned} x(t) &= e^{tA}C = e^{t\lambda}e^{tN}C \\ &= \left[e^{t\lambda}\sum_{k=0}^{n-1}\frac{t^kN^k}{k!}\right]C \\ &= e^{t\lambda}\begin{bmatrix} 1 & & & & \\ t & 1 & & & \\ \frac{t^2}{2!} & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & \cdot & \\ \cdot & & \cdot & \cdot & \\ \cdot & & & \cdot & \\ \frac{t^{n-1}}{(n-1)!} & \cdots & \frac{t^2}{2!} & t & 1 \end{bmatrix}\begin{bmatrix} C_1 \\ \cdot \\ \cdot \\ \cdot \\ C_n \end{bmatrix}. \end{aligned}$$

In coordinates,

$$x_j(t) = e^{\lambda t} \sum_{k=0}^{j-1} \frac{t^k}{k!} C_{j-k}. \tag{2}$$

Note that the factorials can be absorbed into the constants.

Suppose instead that $\lambda = a + bi$, $b \neq 0$, and that A is a real λ-block:

$$\begin{bmatrix} D & & & \\ I & \ddots & & \\ & \ddots & \ddots & \\ & & I & D \end{bmatrix}, \qquad D = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}, \qquad I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Let m be the number of blocks D so that $n = 2m$. The solution to (1) can be computed using exponentials. It is easiest to consider the equation

$$z' = Bz, \tag{3}$$

where $z\colon \mathbf{R} \to \mathbf{C}^m$ is an unknown map and B is the complex $m \times m$ matrix

$$\begin{bmatrix} \mu & & & \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & \mu \end{bmatrix}, \qquad \mu = a + ib.$$

We identify $\mathbf{C}^m$ with $\mathbf{R}^{2m}$ by the correspondence

$$(x_1 + iy_1, \ldots, x_m + iy_m) = (x_1, y_1, \ldots, x_m, y_m).$$

The solution to (3) is formally the same as (2) with a change of notation:

$$z_j(t) = e^{\mu t} \sum_{k=0}^{j-1} \frac{t^k}{k!} C_{j-k}; \qquad j = 1, \ldots, m. \tag{4}$$

Put $C_k = L_k + iM_k$, $k = 1, \ldots, m$, and take real and imaginary parts of (4); using the identity

$$e^{t(a+bi)} = e^{at}(\cos bt + i \sin bt)$$

one obtains

$$x_j(t) = e^{at} \sum_{k=0}^{j-1} \frac{t^k}{k!} [L_{j-k} \cos bt - M_{j-k} \sin bt], \tag{5}$$

$$y_j(t) = e^{at} \sum_{k=0}^{j-1} \frac{t^k}{k!} [M_{j-k} \cos bt + L_{j-k} \sin bt];$$

$j = 1, \ldots, m$. This is the solution to (1) with initial conditions

$$x_j(0) = L_j, \qquad y_j(0) = M_j.$$

The reader may verify directly that (5) is a solution to (1).

At this point we are not so much interested in the precise formulas (2) and (5) as in the following observation:

(6) If λ is real, each coordinate $x_j(t)$ of any solution to (1) is a linear combination (with constant coefficients) of the functions

$$e^{\lambda t} t^k, \qquad k = 0, \ldots, n.$$

(7) If $\lambda = a + bi$, $b \neq 0$, and $n = 2m$, then each coordinate $x_j(t)$ of any solution to (1) is a linear combination of the functions

$$e^{at} t^k \cos bt, \qquad e^{at} t^k \sin bt; \qquad 0 \leq k \leq m.$$

Consider now Eq. (1) where A is any real $n \times n$ matrix. By a suitable change or coordinates $x = Py$ we transform A into real canonical form $B = PAP^{-1}$. The equation

$$y' = By \tag{8}$$

is equivalent to (1): every solution $x(t)$ to (1) has the form

$$x(t) = Py(t),$$

where $y(t)$ solves (8).

Equation (8) breaks up into a set of uncoupled equations, each of the form

$$u' = B_r u,$$

where B_r is one of the blocks in the real canonical form B of A. Therefore the coordinates of solutions to (8) are linear coordinates of the function described in (6) and (7), where λ or $a + bi$ is an eigenvalue of B (hence of A). The same therefore is true of the original equation (1).

Theorem 1 *Let $A \in L(\mathbf{R}^n)$ and let $x(t)$ be a solution of $x' = Ax$. Then each coordinate $x_j(t)$ is a linear combination of the functions*

$$t^k e^{at} \cos bt, \qquad t^l e^{at} \sin bt,$$

where $a + bi$ runs through all the eigenvalues of A with $b \geq 0$, and k and l run through all the integers $0, \ldots, n-1$. Moreover, for each $\lambda = a + bi$, k and l are less than the size of the largest λ-block in the real canonical form of A.

Notice that if A has real eigenvalues, then the functions displayed in Theorem 1 include these of the form $t^k e^{at}$.

This result does not tell what the solutions of (1) are, but it tells us what form the solutions take. The following is a typical and very important application of Theorem 1.

Theorem 2 *Suppose every eigenvalue of $A \in L(\mathbf{R}^n)$ has negative real part. Then*

$$\lim_{t\to\infty} x(t) = 0$$

for every solution to $x' = Ax$.

Proof. This is an immediate consequence of Theorem 1, the inequalities

$$|\cos bt| \le 1, \qquad |\sin bt| \le 1,$$

and the fact that

$$\lim_{t\to\infty} t^k e^{at} = 0 \qquad \text{for all } k \quad \text{if } a < 0.$$

The converse to Theorem 2 is easy:

Theorem 3 *If every solution of $x' = Ax$ tends to 0 as $t \to \infty$, then every eigenvalue of A has negative real part.*

Proof. Suppose $\mu = a + ib$ is an eigenvalue with $a \ge 0$. From (5) we obtain a solution (in suitable coordinates)

$$x_1(t) = e^{at}\cos bt,$$
$$y_1(t) = e^{at}\sin bt,$$
$$x_j(t) = y_j(t) = 0, \qquad j \ge 1,$$

which does not tend to zero as $t \to \infty$.

An argument similar to the proof of Theorem 2 shows:

Theorem 4 *If every eigenvalue of $A \in L(\mathbf{R}^n)$ has positive real part, then*

$$\lim_{t\to\infty} |x(t)| = \infty$$

for every solution to $x' = Ax$.

The following corollary of Theorem 1 is useful:

Theorem 5 *If $A \in L(\mathbf{R}^n)$, then the coordinates of every solution to $x' = Ax$ are infinitely differentiable functions (that is, C^m for all m).*

PROBLEMS

1. (a) Suppose that every eigenvalue of $A \in L(\mathbf{R}^n)$ has real part less than $-a < 0$. Prove that there exists a constant $k > 0$ such that if $x(t)$ is a solution to $x' = Ax$, then

$$|x(t)| \le k^{-at}|x(0)|$$

for all $t \ge 0$. Find such a k and a for each of the following operators A:

(b) $\begin{bmatrix} -1 & 1 \\ 0 & -1 \end{bmatrix}$ (c) $\begin{bmatrix} -1 & 100 \\ 0 & -1 \end{bmatrix}$

(d) $\begin{bmatrix} \log\frac{1}{2} & 1 \\ 0 & \log\frac{1}{2} \end{bmatrix}$ (e) $\begin{bmatrix} -1 & -1 \\ 1 & -1 \end{bmatrix}$

2. Let $A \in L(\mathbf{R}^n)$. Suppose all solutions of $x' = Ax$ are periodic with the same period. Then A is semisimple and the characteristic polynomial is a power of $t^2 + a^2$, $a \in \mathbf{R}$.

3. Suppose at least one eigenvalue of $A \in L(\mathbf{R}^n)$ has positive real part. Prove that for any $a \in \mathbf{R}^n$, $\epsilon > 0$ there is a solution $x(t)$ to $x' = Ax$ such that

$$|x(0) - a| < \epsilon \qquad \text{and} \qquad \lim_{t\to\infty} |x(t)| = \infty.$$

4. Let $A \in L(\mathbf{R}^n)$, and suppose all eigenvalues of A have nonpositive real parts.
   (a) If A is semisimple, show that every solution of $x' = Ax$ is bounded (that is, there is a constant M, depending on $x(0)$, such that $|x(t)| \le M$ for all $t \in \mathbf{R}$).
   (b) Show by example that if A is not semisimple, there may exist a solution such that
   $$\lim_{t\to\infty} |x(t)| = \infty.$$

5. For any solution to $x' = Ax$, $A \in L(\mathbf{R}^n)$, show that exactly one of the following alternatives holds:
   (a) $\lim_{t\to\infty} x(t) = 0$ and $\lim_{t\to-\infty} |x(t)| = \infty$;
   (b) $\lim_{t\to\infty} |x(t)| = \infty$ and $\lim_{t\to-\infty} x(t) = 0$;
   (c) there exist constants $M, N > 0$ such that
   $$M < |x(t)| < N$$
   for all $t \in \mathbf{R}$.

6. Let $A \in L(\mathbf{R}^4)$ be semisimple and suppose the eigenvalues of A are $\pm ai$, $\pm bi$; $a > 0, b > 0$.
   (a) If a/b is a rational number, every solution to $x' = Ax$ is periodic.
   (b) If a/b is irrational, there is a nonperiodic solution $x(t)$ such that
   $$M < |x(t)| < N$$
   for suitable constants $M, N > 0$.

§6. Higher Order Linear Equations

Consider the one-dimensional, nth order homogeneous linear differential equation with constant coefficients

$$s^{(n)} + a_1 s^{(n-1)} + \cdots + a_n s = 0. \tag{1}$$

Here $s: \mathbf{R} \to \mathbf{R}$ is an unknown function, $a_1, \ldots, a_n$ are constants, and $s^{(k)}$ means the kth derivative of s.

Proposition 1 (a) *A linear combination of solutions of* (1) *is again a solution.*
(b) *The derivative of a solution of* (1) *is again a solution.*

Proof. By a *linear combination* of functions $f_1, \ldots, f_m$ having a common domain, and whose values are in a common vector space, we mean a function of the form

$$f(x) = c_1 f_1(x) + \cdots + c_m f_m(x),$$

where $c_1, \ldots, c_m$ are *constants*. Thus (a) means that if $s_1(t), \ldots, s_m(t)$ are solutions of (1) and $c_1, \ldots, c_m$ are constants, then $c_1 s_1(t) + \cdots + c_m s_m(t)$ is also a solution; this follows from linearity of derivatives.

Part (b) is immediate by differentiating both sides of (1)—provided we know that a solution is $n + 1$ times differentiable! This is in fact true. To prove it, consider the equivalent linear system

$$\begin{aligned} x_1' &= x_2, \\ &\vdots \\ x_{n-1}' &= x_n, \\ x_n' &= a_n x_1 - a_{n-1} x_2 - \cdots - a_1 x_n. \end{aligned} \tag{2}$$

If s is a solution to (1), then

$$x = (s, s', \ldots, s^{(n-1)})$$

is a solution to (1). From Theorem 4, Section 1 we know that every solution to (2) has derivatives of all orders.

The matrix of coefficients of the linear system (2) is the $n \times n$ matrix

$$\begin{bmatrix} 0 & 1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -a_n & \cdots & & -a_2 & -a_1 \end{bmatrix}. \tag{3}$$

A matrix of this form is called the *companion matrix* of the polynomial

$$p(\lambda) = \lambda^n + a_1 \lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n. \tag{4}$$

In Chapter 5 it was shown that this is the characteristic polynomial of A.

Companion matrices have special properties as operators. The key to solving (1) is the following fact.

Proposition 2 *Let $\lambda \in \mathbf{C}$ be a real or complex eigenvalue of a companion matrix A. Then the real canonical form of A has only one λ-block.*

Proof. We consider A as an operator on $\mathbf{C}^n$. The number of λ blocks is

$$\dim \operatorname{Ker}(A - \lambda),$$

considering $\operatorname{Ker}(A - \lambda)$ as a complex vector space.

The first n columns of $A - \lambda$ form the $(n-1) \times n$ matrix

$$\begin{bmatrix} -\lambda & 1 & & & \\ & -\lambda & \ddots & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ & & & -\lambda & 1 \end{bmatrix}$$

which has rank $n - 1$. Hence $A - \lambda$ has rank n or $n - 1$, but rank n is ruled out since λ is an eigenvalue. Hence $A - \lambda$ has rank $n - 1$, so $\operatorname{Ker}(A - \lambda)$ has dimension 1. This proves Proposition 2.

Definition A *basis* of solutions to (1) is a set of solutions $s_1, \ldots, s_n$ such that every solution is expressible as a linear combination of $s_1, \ldots, s_n$ in one and only one way.

The following theorem is the basic result of this section.

Theorem *The following n functions form a basis for the solutions of* (1):

(a) *the function $t^k e^{\lambda t}$, where λ runs through the distinct real roots of the characteristic polynomial* (4), *and k is a nonnegative integer in the range $0 \le k <$ multiplicity of λ; together with*

(b) *the functions*

$$t^k e^{at} \cos bt \quad \text{and} \quad t^k e^{at} \sin bt,$$

where $a + bi$ runs through the complex roots of (4) *having $b > 0$ and k is a nonnegative integer in the range $0 \le k <$ multiplicity of $a + bi$.*

Proof. We call the functions listed in the proposition *basic* functions. It follows from Theorem 1 of the previous section that every solution is a linear combination of basic functions.

The proof that each basic function is in fact a solution is given in the next section. By Proposition 1 it follows that the solutions to (1) are exactly the linear combinations of basic functions.

It remains to prove that each solution is a *unique* linear combination of basic functions. For this we first note that there are precisely n functions listed in (a) and (b): the number of functions listed equals the sum of the multiplicities of the real roots of $p(\lambda)$, plus twice the sum of the multiplicities of the complex roots with positive imaginary parts. Since nonreal roots come in conjugate pairs, this total is the sum of the multiplicities of all the roots, which is n.

Define a map $\varphi: \mathbf{R}^n \to \mathbf{R}^n$ as follows. Let $f_1, \ldots, f_n$ be an ordering of the basic functions. For each $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbf{R}^n$ let $s_\alpha(t)$ be the solution

$$s_\alpha = \sum_{j=1}^{n} \alpha_j f_j.$$

Define

$$\varphi(\alpha) = (s_\alpha(0), s_\alpha'(0), \ldots, s_\alpha^{(n-1)}(0)) \in \mathbf{R}^n.$$

It is easy to see that φ is a linear map. Moreover, φ is surjective since for each $(a_0, \ldots, a_{n-1}) \in \mathbf{R}^n$ there is some solution s such that

$$s(0) = a_0, \ldots, s^{(n-1)}(0) = a_{n-1}, \tag{5}$$

and $s = s_\alpha$ for some α. It follows that φ is injective.

From this we see at once that every solution s is a unique linear combination of the basic functions, for if $s_\alpha = s_\beta$, then $\varphi(\alpha) = \varphi(\beta)$ and hence $\alpha = \beta$. This completes the proof of the theorem.

Theorem 1 reduces the solution of (1) to elementary linear algebra, provided the roots and multiplicities of the characteristic polynomial are known. For example, consider the equation

$$s^{(4)} + 4s^{(3)} + 5s^{(2)} + 4s' + 4s = 0. \tag{6}$$

The roots of the characteristic polynomial

$$\lambda^4 + 4\lambda^3 + 5\lambda^2 + 4\lambda + 4$$

are

$$-2, \quad -2, \quad i, \quad -i.$$

Therefore the general solution is

$$s(t) = Ae^{-2t} + Bte^{-2t} + C\cos t + D\sin t, \tag{7}$$

where A, B, C, D are arbitrary constants.

To find a solution with given initial conditions, say,

$$\begin{aligned} s(0) &= 0, \\ s'(0) &= -1, \\ s^{(2)}(0) &= -4, \\ s^{(3)}(0) &= 14, \end{aligned} \tag{8}$$

we compute the left-hand side of (8) from (7), to get:

$$\begin{aligned} s(0) &= A + C = 0, \\ s'(0) &= -2A + B + D = -1, \\ s^{(2)}(0) &= 4A - 4B - C = -4, \\ s^{(3)}(0) &= -8A + 12B - D = 14. \end{aligned} \tag{9}$$

The only solution to this system of equations is

$$A = C = 0, \quad B = 1, \quad D = -2.$$

Therefore the solution to (6) and (8) is

$$s(t) = te^{-2t} - 2\sin t.$$

PROBLEMS

1. Find a map $s: \mathbf{R} \to \mathbf{R}$ such that

$$s^{(3)} - s^{(2)} + 4s' - 4s = 0,$$

$$s(0) = 1, \quad s'(0) = -1, \quad s''(0) = 1.$$

2. Consider equation (6) in the text. Find out for which initial conditions $s(0)$, $s'(0)$, $s''(0)$ there is a solution $s(t)$ such that:
   (a) $s(t)$ is periodic; (b) $\lim_{t\to\infty} s(t) = 0$;
   (c) $\lim_{t\to\infty} |s(t)| = \infty$; (d) $|s(t)|$ is bounded for $t \geq 0$;
   (e) $|s(t)|$ is bounded for all $t \in \mathbf{R}$.

3. Find all periodic solutions to

$$s^{(4)} + 2s^{(2)} + s = 0.$$

4. What is the smallest integer $n > 0$ for which there is a differential equation

$$s^{(n)} + a_1 s^{(n-1)} + \cdots + a_n s = 0,$$

having among its solutions the functions

$$\sin 2t, \quad 4te^{2t}, \quad -e^{-t}?$$

Find the constants $a_1, \ldots, a_n \in \mathbf{R}$.

§7. Operators on Function Spaces

We discuss briefly another quite different approach to the equation

$$s^{(n)} + a_1 s^{(n-1)} + \cdots + a_n s = 0; \qquad s: \mathbf{R} \to \mathbf{R}. \tag{1}$$

Let $\mathfrak{F}$ be the set of all infinitely differentiable functions $\mathbf{R} \to \mathbf{R}$ (one could also use maps $\mathbf{R} \to \mathbf{C}$). Under multiplication by constants and addition of functions, $\mathfrak{F}$ satisfies the axioms VS1, VS2 for a vector space (Chapter 3, Section 1, Part A); but $\mathfrak{F}$ is not finite dimensional.

Let $D: \mathfrak{F} \to \mathfrak{F}$ denote the *differentiation operator*; that is,

$$Df = f'.$$

Then D is a linear operator. Some other operators on $\mathfrak{F}$ are:

multiplication of f by a constant λ, which we denote simply by λ; note that $1f = f$ and $0f = 0$;

multiplication of f by the function $i(t) = t$, which we denote by t.

New operators can be built from these by composition, addition, and multiplication by constants. For example,

$$D^2: \mathfrak{F} \to \mathfrak{F}$$

assigns to f the function

$$D(Df) = D(f') = f'';$$

similarly $D^n f = f^{(n)}$, the nth derivative. The operator $D - \lambda$ assigns to f the function $f' - \lambda f$.

More generally, let

$$p(t) = t^n + a_1 t^{n-1} + \cdots + a_n$$

be a polynomial. (There could also be a coefficient a_0 of t^n.) There is defined an operator

$$p(D) = D^n + a_1 D^{n-1} + \cdots + a_{n-1} D + a_n,$$

which assigns to f the function

$$f^{(n)} + a_1 f^{(n-1)} + \cdots + a_{n-1} f' + a_n f.$$

We may then rephrase the problem of solving (1): *find the kernel of the operator* $p(D)$.

This way of looking at things suggests new ways of manipulating higher-order equations. For example, if $p(\lambda)$ factors

$$p(\lambda) = q(\lambda) r(\lambda),$$

then clearly,

$$\operatorname{Ker} r(D) \subset \operatorname{Ker} p(D).$$

Moreover,

$$\operatorname{Ker} q(D) \subset \operatorname{Ker} p(D),$$

since $qr = rq$. In addition, if $f \in \operatorname{Ker} q(D)$ and $g \in \operatorname{Ker} r(D)$, then $f + g \in \operatorname{Ker} p(D)$.

We can now give a proof that if $(t - \lambda)^m$ divides $p(t)$, then $t^k e^{\lambda t} \in \operatorname{Ker} p(D)$, $0 \leq k \leq m - 1$. It suffices to prove

$$(D - \lambda)^{k+1} t^k e^{\lambda t} = 0, \qquad k = 0, 1, \ldots. \tag{2}$$

Note that $D(e^{\lambda t}) = \lambda e^{\lambda t}$, or

$$(D - \lambda) e^{\lambda t} = 0.$$

Next, observe the following relation between operators:

$$Dt - tD = 1$$

(this means $D(tf) - tDf = f$, which follows from the Leibniz formula). Hence also

$$(D - \lambda)t - t(D - \lambda) = 1.$$

It follows easily by induction that

$$(D - \lambda)t^k - t^k(D - \lambda) = kt^{k-1}; \qquad k = 1, 2, \ldots.$$

Therefore

$$\begin{aligned}(D - \lambda)^{k+1}(t^k e^{\lambda t}) &= (D - \lambda)^k (D - \lambda)(t^k e^{\lambda t}) \\ &= (D - \lambda)^k ([t^k (D - \lambda) + kt^{k-1}] e^{\lambda t}) \\ &= k(D - \lambda)^k (t^{k-1} e^{\lambda t}).\end{aligned}$$

Hence (2) is proved by induction on k.

If λ is interpreted as a complex number and $p(D)$ has complex coefficients, the proof goes through without change. If $p(D)$ has real coefficients but $\lambda = a + bi$ is a nonreal root, it follows that both the real and imaginary parts of $t^k e^{\lambda t}$ are annihilated by $p(D)$. This shows that

$$t^k e^{at} \cos bt, \qquad t^k e^{at} \sin bt$$

are in $\operatorname{Ker} p(D)$.

We have proved part of Theorem 1, Section 6 by easy "formal" (but rigorous) methods. The main part—that every solution is a linear combination of basic functions—can be proved by similar means. [See *Linear Algebra* by S. Lang, p. 213 (Reading, Massachusetts: Addison-Wesley, 1966).] Operators on function spaces have many uses for both theoretical and practical work in differential equations.

Chapter 7

Contractions and Generic Properties of Operators

In this chapter we study some important kinds of linear flows e^{tA}, particularly contractions. A (linear) contraction is characterized by the property that every trajectory tends to 0 as $t \to \infty$. Equivalently, the eigenvalues of A have negative real parts. Such flows form the basis for the study of asymptotic stability in Chapter 9. Contractions and their extreme opposites, expansions, are studied in Section 1.

Section 2 is devoted to hyperbolic flows e^{tA}, characterized by the condition that the eigenvalues of A have nonzero real parts. Such a flow is the direct sum of a contraction and an expansion. Thus their qualitative behavior is very simple.

In Section 3 we introduce the notion of a generic property of operators on $\mathbf{R}^n$; this means that the set of operators which have that property contains a dense open subset of $L(\mathbf{R}^n)$. It is shown that "semisimple" is a generic property, and also, "generating hyperbolic flows" is a generic property for operators.

The concept of a generic property of operators is a mathematical way of making precise the idea of "almost all" operators, or of a "typical" operator. This point is discussed in Section 4.

§1. Sinks and Sources

Suppose that a state of some "physical" (or mechanical, biological, economic, etc.) system is determined by the values of n parameters; the space of all states is taken to be an open set $U \subset \mathbf{R}^n$. We suppose that the dynamic behavior of the system is modeled mathematically by the solution curves of a differential equation (or *dynamical system*)

$$x' = f(x), \qquad f: U \to \mathbf{R}^n. \tag{1}$$

We are interested in the long-run behavior of trajectories (that is, solution curves)

of (1). Of especial interest are *equilibrium states*. Such a state $\bar{x} \in U$ is one that does not change with time. Mathematically, this means that the constant map $t \to \bar{x}$ is a solution to (1); equivalently, $f(\bar{x}) = 0$. Hence we define an *equilibrium* of (1) to be a point $\bar{x} \in U$ such that $f(\bar{x}) = 0$.

From a physical point of view only equilibria that are "stable" are of interest. A pendulum balanced upright is in equilibrium, but this is very unlikely to occur; moreover, the slightest disturbance will completely alter the pendulum's behavior. Such an equilibrium is unstable. On the other hand, the downward rest position is stable; if slightly perturbed from it, the pendulum will swing around it and (because of friction) gradually approach it again.

Stability is studied in detail in Chapter 9. Here we restrict attention to linear systems and concentrate on the simplest and most important type of stable equilibrium.

Consider a linear equation

$$x' = Ax, \qquad A \in L(\mathbf{R}^n). \tag{2}$$

The origin $0 \in \mathbf{R}^n$ is called a *sink* if all the eigenvalues of A have negative real parts. We also say the linear flow e^{tA} is a *contraction*.

In Chapter 6, Theorems 2 and 3, Section 5, it was shown that 0 is a sink if and only if every trajectory tends to 0 as $t \to \infty$. (This is called *asymptotic* stability.) From Problem 1, Section 5 of that chapter, it follows that trajectories approach a sink *exponentially*. The following result makes this more precise.

Theorem 1 *Let A be an operator on a vector space E. The following statements are equivalent:*

(a) *The origin is a sink for the dynamical system $x' = Ax$.*

(b) *For any norm in E there are constants $k > 0$, $b > 0$ such that*

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

for all $t \geq 0$, $x \in E$.

(c) *There exists $b > 0$ and a basis $\mathcal{B}$ of E whose corresponding norm satisfies*

$$| e^{tA}x |_{\mathcal{B}} \leq e^{-bt} | x |_{\mathcal{B}}$$

for all $t \geq 0$, $x \in E$.

Proof. Clearly, (c) implies (b) by equivalence of norms; and (b) implies (a) by Theorem 3 of Chapter 6, Section 5. That (a) implies (b) follows easily from Theorem 1 of that section; the details are left to the reader.

It remains to prove that (a) implies (c). For this we use the following purely algebraic fact, whose proof is postponed.

Recall that $\mathrm{R}\,\lambda$ is the real part of λ.

Lemma *Let A be an operator on a real vector space E and suppose*

$$\alpha < \mathrm{R}\,\lambda < \beta \tag{3}$$

for every eigenvalue λ of A. Then E has a basis such that in the corresponding inner product and norm,

$$\alpha \mid x \mid^2 \leq \langle Ax, x \rangle \leq \beta \mid x \mid^2 \tag{4}$$

for all $x \in E$.

Assuming the truth of the lemma, we derive an estimate for solutions of $x' = Ax$. Let $(x_1, \ldots, x_n)$ be coordinates on E corresponding to a basis $\mathfrak{B}$ such that (4) holds, and let

$$x(t) = (x_1(t), \ldots, x_n(t))$$

be a solution to $x' = Ax$. Then for the norm and inner product defined by $\mathfrak{B}$ we have

$$\frac{d}{dt} \mid x \mid = \frac{d}{dt} [\sum (x_j)^2]^{1/2}$$

$$= \frac{\sum x_j x_j'}{[\sum (x_j)^2]^{1/2}}.$$

Hence

$$\frac{d}{dt} \mid x \mid = \frac{\langle x, x' \rangle}{\mid x \mid} = \frac{\langle x, Ax \rangle}{\mid x \mid}.$$

Therefore, from (4), we have

$$\alpha \leq \frac{d/dt \mid x \mid}{\mid x \mid} \leq \beta,$$

or

$$\alpha \leq \frac{d}{dt} \log \mid x \mid \leq \beta.$$

It follows by integration that

$$\alpha t \leq \log \mid x(t) \mid - \log \mid x(0) \mid \leq \beta t;$$

hence

$$\alpha t \leq \frac{\log \mid x(t) \mid}{\log \mid x(0) \mid} \leq \beta t,$$

or

$$e^{\alpha t} \mid x(0) \mid \leq \mid x(t) \mid \leq e^{\beta t} \mid x(0) \mid.$$

Theorem 1 is proved by letting $\beta = -b < 0$ where the eigenvalues of A have real parts less than $-b$.

We now prove the lemma; for simplicity we prove only the second inequality of (4).

Let $c \in \mathbf{R}$ be such that

$$\mathbf{R}\lambda < c < \beta$$

for every eigenvalue λ of A.

Suppose first that A is semisimple. Then $\mathbf{R}^n$ has a direct sum decomposition

$$\mathbf{R}^n = E_1 \oplus \cdots \oplus E_r \oplus F_1 \oplus \cdots \oplus F_s,$$

where each E_j is a one-dimensional subspace spanned by an eigenvector e_j of A corresponding to a real eigenvalue λ_j; and each F_k is a two-dimensional subspace invariant under A, having a basis $\{f_k, g_k\}$ giving $A \mid F_k$ the matrix

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

where $a_k + ib_k$ is an eigenvalue of A. By assumption

$$\lambda_j < c, \qquad a_k < c.$$

Given $\mathbf{R}^n$ the inner product defined by

$$\langle e_j, e_j \rangle = \langle f_k, f_k \rangle = \langle g_k, g_k \rangle = 1,$$

and all other inner products among the e_j, f_k, and g_k being 0. Then a computation shows

$$\langle Ae_j, e_j \rangle = \lambda_j < c, \qquad \langle Af_k, f_k \rangle = a_k < c;$$

it follows easily that

$$\langle Ax, x \rangle \leq c \mid x \mid^2$$

for all $x \in \mathbf{R}^n$, as required.

Now let A be any operator. We first give $\mathbf{R}^n$ a basis so that A has a matrix in real canonical form

$$A = \operatorname{diag}\{A_1, \ldots, A_p\},$$

where each A_j has the form

$$\begin{bmatrix} \alpha_j & & & \\ 1 & \cdot & & \\ & \cdot & \cdot & \\ & & \cdot & \cdot \\ & & & 1 \quad \alpha_j \end{bmatrix} \tag{5}$$

or

$$\begin{bmatrix} D_j & & & \\ I & \cdot & & \\ & \cdot & \cdot & \\ & & \cdot & \cdot \\ & & & I \quad D_j \end{bmatrix}, \qquad D_j = \begin{bmatrix} \alpha_k & -\beta_k \\ \beta_k & \alpha_k \end{bmatrix}, \qquad I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{6}$$

If we give a subspace E_j of E, corresponding to a block A_j a basis satisfying the lemma for A_j, then all these bases together fulfill the lemma for A. Therefore we may assume A is a single block.

For the first kind of block (5), we can write $A = S + N$ where S has the matrix $\alpha_j I$ and N has the matrix

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

Thus the basis vectors $\{e_1, \ldots, e_n\}$ are eigenvectors of S, while

$$\begin{aligned} Ne_1 &= e_2, \\ &\vdots \\ Ne_{n-1} &= e_n, \\ Ne_n &= 0. \end{aligned}$$

Let $\epsilon > 0$ be very small and consider a new basis

$$\mathcal{B}_\epsilon = \left\{e_1, \frac{1}{\epsilon} e_2, \ldots, \frac{1}{\epsilon^{n-1}} e_n\right\} = \{\tilde{e}_1, \ldots, \tilde{e}_n\}.$$

$\mathcal{B}_\epsilon$ is again composed of eigenvectors of S, while now

$$\begin{aligned} N\tilde{e}_1 &= \epsilon\tilde{e}_2, \\ N\tilde{e}_2 &= \epsilon\tilde{e}_3, \\ &\vdots \\ N\tilde{e}_{n-1} &= \epsilon\tilde{e}_n, \\ N\tilde{e}_n &= 0. \end{aligned}$$

Thus the $\mathcal{B}_\epsilon$ matrix of A is

$$\begin{bmatrix} \alpha_j & & & \\ \epsilon & \ddots & & \\ & \ddots & \ddots & \\ & & \epsilon & \alpha_j \end{bmatrix}. \tag{7}$$

Let $\langle x, y\rangle_\epsilon$ denote the inner product corresponding to $\mathcal{B}_\epsilon$. It is clear by considering the matrix (7) that

$$\frac{\langle Ax, x\rangle_\epsilon}{\langle x, x\rangle_\epsilon} \to \frac{\langle Sx, x\rangle}{|x|^2} \quad \text{as} \quad \epsilon \to 0.$$

Therefore if ϵ is sufficiently small, the basis $\mathcal{B}_\epsilon$ satisfies the lemma for a block (5). The case of a block (6) is similar and is left to the reader. This completes the proof of the lemma.

The qualitative behavior of a flow near a sink has a simple geometrical interpretation. Suppose $0 \in \mathbf{R}^n$ is a sink for the linear differential equation $x' = f(x)$. Consider the spheres

$$S_a = \{x \in \mathbf{R}^n \mid |x| = a\}, \qquad a > 0,$$

where $|x|$ is the norm derived from an inner product as in the theorem. Since $|x(t)|$ has negative derivatives, the trajectories all point inside these spheres as in Fig. A.

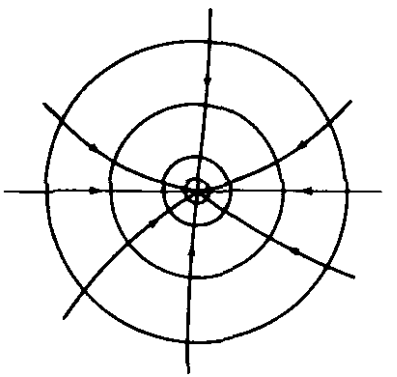


FIG. A

We emphasize that this is true for the spheres in a *special* norm; it may be false for some other norm.

The linear flow e^{tA} that has the extreme opposite character to a contraction is an *expansion*, for which the origin is called a *source*: every eigenvalue of A has positive real part. The following result is the analogue of Theorem 1 for expansions.

Theorem 2 *If $A \in L(E)$, the following are equivalent:*

(a) *The origin is a source for the dynamical system $x' = Ax$;*

(b) *For any norm on E, there are constants $L > 0$, $a > 0$ such that*

$$|e^{tA}x| \geq Le^{at}|x|$$

for all $t \geq 0$, $x \in E$.

(c) *There exists $a > 0$ and a basis $\mathcal{B}$ of E whose corresponding norm satisfies*

$$|e^{tA}x|_{\mathcal{B}} \geq e^{at}|x|$$

for all $t \geq 0$, $x \in E$.

The proof is like that of Theorem 1, using the lemma and the first inequality of (4).

PROBLEMS

1. (a) Show that the operator $A = \begin{bmatrix} -1 & 3 \\ 0 & -2 \end{bmatrix}$ generates a contracting flow e^{tA}
   (b) Sketch the phase portrait of $x' = Ax$ in standard coordinates.
   (c) Show that it is false that $|e^{tA}x| \leq |x|$ for all $t \geq 0$, $x \in \mathbf{R}^2$, where $|x|$ is the Euclidean norm.

2. Let e^{tA} be a contraction in $\mathbf{R}^n$. Show that for $\tau > 0$ sufficiently large, the norm $\|x\|$ on $\mathbf{R}^n$ defined by

$$\|x\| = \int_0^\tau |e^{sA}x|\, ds$$

satisfies, for some $\lambda > 0$,

$$\|e^{tA}x\| \leq e^{-\lambda t}\|x\|.$$

3. (a) If e^{tB} and e^{tA} are both contractions on $\mathbf{R}^n$, and $BA = AB$, then $e^{t(A+B)}$ is a contraction. Similarly for expansions.
   (b) Show that (a) can be false if the assumption that $AB = BA$ is dropped.

4. Consider a mass moving in a straight line under the influence of a spring. Assume there is a retarding frictional force proportional to the velocity but opposite in sign.
   (a) Using Newton's second law, verify that the equation of motion has the form

$$mx'' = ax' + bx; \qquad m > 0, \quad a < 0, \quad b < 0.$$

   (b) Show that the corresponding first order system has a sink at (0, 0).
   (c) What do you conclude about the long-run behavior of this physical system?

5. If e^{tA} is a contraction (expansion), show that $e^{t(-A)}$ is an expansion (respectively, contraction). Therefore a contraction is characterized by every trajectory going to ∞ as $t \to -\infty$; and an expansion, by every trajectory going to 0 as $t \to -\infty$.

§2. Hyperbolic Flows

A type of linear flow e^{tA} that is more general than contractions and expansions is the *hyperbolic flow*: all eigenvalues of A have nonzero real part.

After contractions and expansions, hyperbolic linear flows have the simplest types of phase portraits. Their importance stems from the fact that almost every linear flow is hyperbolic. This will be made precise, and proved, in the next section.

The following theorem says that a hyperbolic flow is the direct sum of a contraction and an expansion.

Theorem *Let e^{tA} be a hyperbolic linear flow, $A \in L(E)$. Then E has a direct sum decomposition*

$$E = E^s \oplus E^u$$

invariant under A, such that the induced flow on E^s is a contraction and the induced flow on E^u is an expansion. This decomposition is unique.

Proof. We give E a basis putting A into real canonical form (Chapter 6). We order this basis so that the canonical form matrix first has blocks corresponding to eigenvalues with negative real parts, followed by blocks corresponding to positive eigenvalues. The first set of blocks represent the restriction of A to a subspace $E^s \subset E$, while the remaining blocks represent the restriction of A to $E^u \subset E$.

Since E^s is invariant under A, it is invariant under e^{tA}. Put $A \mid E^s = A_s$ and $A \mid E^u = A_u$. Then $e^{tA} \mid E^s = e^{tA_s}$. By Theorem 1, Section 1, $e^{tA} \mid E^s$ is a contraction. Similarly, Theorem 2, Section 1 implies that $e^{tA} \mid E^u$ is an expansion.

Thus $A = A_s \oplus A_u$ is the desired decomposition.

To check uniqueness of the decomposition, suppose $F^s \oplus F^u$ is another decomposition of E invariant under the flow such that $e^{tA} \mid F^s$ is a contraction and $e^{tA} \mid F^u$ is an expansion. Let $x \in F^s$. We can write

$$x = y + z, \qquad y \in E^s, \qquad z \in E^u.$$

Since $e^{tA}x \to 0$ as $t \to \infty$, we have $e^{tA}y \to 0$ and $e^{tA}z \to 0$. But

$$|e^{tA}z| \geq e^{ta}|z|, \qquad a > 0,$$

for all $t \geq 0$. Hence $|z| = 0$. This shows that $F^s \subset E^s$. The same argument shows that $E^s \subset F^s$; hence $E^s = F^s$. Similar reasoning about e^{-tA} shows that $E^u = F^u$. This completes the proof.

A hyperbolic flow may be a contraction ($E^u = 0$) or an expansion ($E^s = 0$). If neither E^u nor E^s is 0, the phase portrait may look like Fig. A in the two-dimensional case or like Fig. B in a three-dimensional case.

If, in addition, the eigenvalues of $A \mid E^s$ have nonzero imaginary part, all trajectories will spiral toward E^u (Fig. C).

Other three-dimensional phase portraits are obtained by reversing the arrows in Figs. B and C.

The letters s and u stand for *stable* and *unstable*. E^s and E^u are sometimes called the *stable* and *unstable* subspaces of the hyperbolic flow.

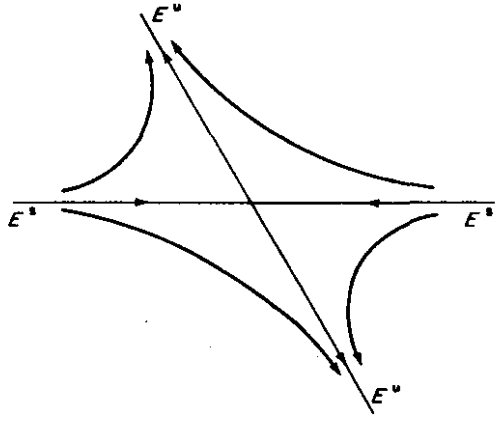


FIG. A

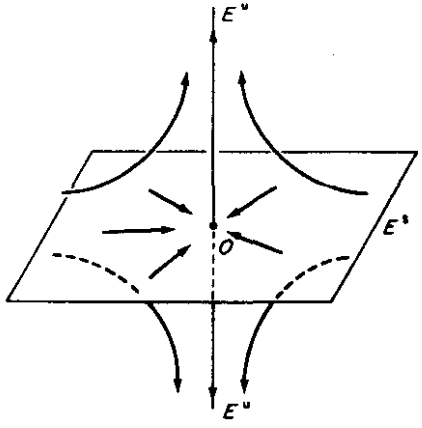


FIG. B

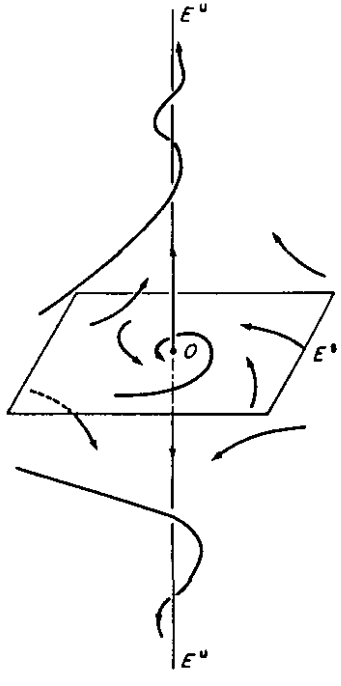


FIG. C

PROBLEMS

1. Let the eigenvalues of $A \in L(\mathbf{R}^3)$ be λ, μ, ν. Notice that e^{tA} is a hyperbolic flow and sketch its phase portrait if:
   (a) $\lambda < \mu < \nu < 0$;
   (b) $\lambda < 0, \mu = a + bi, a < 0, b > 0$;
   (c) $\lambda < 0, \mu = a + bi, a > 0, b > 0$;
   (d) $\lambda < 0 < \mu = \nu$ and A is semisimple;
   (e) $\lambda < \mu < 0 < \nu$.

2. e^{tA} is hyperbolic if and only if for each $x \neq 0$ either

$$|e^{tA}x| \to \infty \quad \text{as} \quad t \to \infty$$

   or

$$|e^{tA}x| \to \infty \quad \text{as} \quad t \to -\infty.$$

3. Show that a hyperbolic flow has no nontrivial periodic solutions.

§3. Generic Properties of Operators

Let F be a normed vector space (Chapter 5). Recall that a set $X \subset F$ is *open* if whenever $x \in X$ there is an open ball about x contained in X; that is, for some $a > 0$ (depending on x) the open ball about x of radius a,

$$\{y \in F \mid |y - x| < a\},$$

is contained in X. From the equivalence of norms it follows that this definition is independent of the norm; any other norm would have the same property (for perhaps a different radius a).

Using geometrical language we say that if x belongs to an open set X, any point sufficiently near to x also belongs to X.

Another kind of subset of F is a *dense* set: $X \subset F$ is dense if every point of F is arbitrarily close to points of X. More precisely, if $x \in F$, then for every $\epsilon > 0$ there exists some $y \in X$ with $|x - y| < \epsilon$. Equivalently, $U \cap X$ is nonempty for every nonempty open set $U \subset F$.

An interesting kind of subset of X is a set $X \subset F$ which is both open and dense. It is characterized by the following properties: every point in the complement of F can be approximated arbitrarily closely by points of X (because X is dense); but no point in X can be approximated arbitrarily closely by points in the complement (because X is open).

Here is a simple example of a dense open set in $\mathbf{R}^2$:

$$X = \{(x, y) \in \mathbf{R}^2 \mid xy \neq 1\}.$$

This, of course, is the *complement* of the hyperbola defined by $xy = 1$. If $x_0y_0 \neq 1$ and $|x - x_0|, |y - y_0|$ are small enough, then $xy \neq 1$; this proves X open. Given any $(x_0, y_0) \in \mathbf{R}^2$, we can find (x, y) as close as we like to (x_0, y_0) with $xy \neq 1$; this proves X dense.

More generally, one can show that the complement of any algebraic curve in $\mathbf{R}^2$ is dense and open.

A dense open set is a very fat set, as the following proposition shows:

Proposition *Let $X_1, \ldots, X_m$ be dense open sets in F. Then*

$$X = X_1 \cap \cdots \cap X_m$$

is also dense and open.

Proof. It can be easily shown generally that the intersection of a finite number of open sets is open, so X is open. To prove X dense let $U \subset F$ be a nonempty open set. Then $U \cap X_1$ is nonempty since X_1 is dense. Because U and X_1 are open, $U \cap X_1$ is open. Since $U \cap X_1$ is open and nonempty, $(U \cap X_1) \cap X_2$ is nonempty because X_2 is dense. Since X_1 is open, $U \cap X_1 \cap X_2$ is open. Thus $(U \cap X_1 \cap X_2) \cap X_3$ is nonempty, and so on. So $U \cap X$ is nonempty, which proves that X is dense in F.

Now consider a subset X of the vector space $L(\mathbf{R}^n)$. It makes sense to call X dense, or open. In trying to prove this for a given X we may use any convenient norm on $L(\mathbf{R}^n)$. One such norm is the $\mathcal{B}$-max norm, where $\mathcal{B}$ is a basis $\mathbf{R}^n$:

$$\| T \|_{\mathcal{B}\text{-max}} = \max\{|a_{ij}| \mid [a_{ij}] = \mathcal{B}\text{-matrix of } T\}.$$

A property $\mathcal{P}$ that refers to operators on $\mathbf{R}^n$ is a *generic property* if the set of operators having property $\mathcal{P}$ contains a dense open set. Thus a property is generic if it is shared by some dense open set of operators (and perhaps other operators as well). Intuitively speaking, a generic property is one which "almost all" operators have.

Theorem 1 *The set $\mathcal{S}_1$ of operators on $\mathbf{R}^n$ that have n distinct eigenvalues is dense and open in $L(\mathbf{R}^n)$.*

Proof. We first prove $\mathcal{S}_1$ dense. Let T be an operator on $\mathbf{R}^n$. Fix a basis $\mathcal{B}$ putting T in real canonical form.

The real canonical form of T can be written as the sum of two matrices

$$T = S + N,$$

where

$$S = \begin{bmatrix} \lambda_1 & & & & & & \\ & \ddots & & & & & \\ & & \lambda_r & & & & \\ & & & D_1 & & & \\ & & & & \ddots & & \\ & & & & & D_s & \end{bmatrix},$$

$$D_k = \begin{bmatrix} a_k & -b_k \\ b_k & a_k \end{bmatrix}, \qquad b_k > 0;\ i = 1, \ldots, s;$$

and

$$N = \begin{bmatrix} 0 & & & & & & \\ 1 & \ddots & & & & & \\ & \ddots & \ddots & & & & \\ & & 1 & 0 & & & \\ & & & I_2 & 0_2 & & \\ & & & & \ddots & \ddots & \\ & & & & & I_2 & 0_2 \end{bmatrix},$$

$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad 0_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

The eigenvalues of T (with multiplicities) are $\lambda_1, \ldots, \lambda_r$, and $a_1 \pm ib_1, \ldots, a_s \pm ib_s$.

Given $\epsilon > 0$, let

$$\lambda_1', \ldots, \lambda_r', a_1', \ldots, a_r'$$

be distinct real numbers such that

$$|\lambda_j' - \lambda_j| < \epsilon \qquad \text{and} \qquad |a_k' - a_k| < \epsilon.$$

Put

$$D_k' = \begin{bmatrix} a_k' & -b_k \\ b_k & a_k' \end{bmatrix}, \qquad S' = \begin{bmatrix} \lambda_1' & & & & & \\ & \ddots & & & & \\ & & \lambda_r' & & & \\ & & & D_1' & & \\ & & & & \ddots & \\ & & & & & D_s' \end{bmatrix},$$

and $T' = S' + N$. Then the ®-max norm of $T - T'$ is less than ϵ, and the eigenvalues of T' are the n distinct numbers

$$\lambda_1', \ldots, \lambda_r', \qquad a_1' \pm ib_1, \ldots, a_s' \pm ib_s.$$

This proves that $\mathcal{S}_1$ is dense.

To prove that $\mathcal{S}_1$ is open we argue by contradiction. If it is not open, then there is a sequence $A_1, A_2, \ldots$ of operators on $\mathbf{R}^n$ that are not in $\mathcal{S}_1$ but which converges to an operator A in $\mathcal{S}_1$. There is an upper bound for the norms of the A_k and hence for their eigenvalues. By assumption each A_k has an eigenvalue λ_k of multiplicity at least two.

Suppose at first that all λ_k are real. Passing to a subsequence we may assume that $\lambda_k \to \lambda \in \mathcal{S}_1$. For each k, there are two independent eigenvectors x_k, y_k for A_k belonging to the eigenvalue λ_k. We may clearly suppose $|x_k| = |y_k| = 1$. Moreover we may assume x_k and y_k orthogonal, otherwise replacing y_k by

$$y_k - \langle y_k, x_k\rangle x_k / |y_k - \langle y_k, x_k\rangle x_k|.$$

Passing again to subsequences we may assume $x_k \to x$ and $y_k \to y$. Then x and y are independent vectors. From the relations $A_k x_k = \lambda_k x_k$ and $A_k y_k = \lambda_k y_k$ we find in the limit that $Ax = \lambda x$ and $Ay = \lambda y$. But this contradicts $A \in \mathcal{S}_1$.

If some of the λ_k are nonreal, the same contradiction is reached by considering the complexifications of the A_k; now x_k and y_k are vectors in $\mathbf{C}^n$. In place of the Euclidean inner product on $\mathbf{R}^n$ we use the *Hermitian inner product* on $\mathbf{C}^n$ defined by $\langle z, w\rangle = \sum_{j=1}^n z_j \bar{w}_j$, and the corresponding norm $|z| = \langle z, z\rangle^{1/2}$. The rest of the argument is formally the same as before.

Note that the operators in $\mathcal{S}_1$ are all semisimple, by Chapter 4. Therefore an immediate consequence of Theorem 1 is

Theorem 2 *Semisimplicity is a generic property in* $L(\mathbf{R}^n)$.

The set of semisimple operators is *not* open. For example, every neighborhood of the semisimple operator $I \in L(\mathbf{R}^2)$ contains a nonsemisimple operator of the form $\left[\begin{smallmatrix} 1 & 0 \\ \epsilon & 1 \end{smallmatrix}\right]$.

We also have

Theorem 3 *The set*

$$\mathcal{S}_2 = \{T \in L(\mathbf{R}^n) \mid e^{tT} \text{ is a hyperbolic flow}\}$$

is open and dense in $L(\mathbf{R}^n)$.

Proof. In the proof of density of $\mathcal{S}_1$ in Theorem 1, we can take the numbers $\lambda_1', \ldots, \lambda_r', a_1', \ldots, a_s'$ (the real parts of eigenvalues of T') to be nonzero; this proves density. Openness is proved by a convergence argument similar to (and easier than) the one given in the proof of Theorem 2.

PROBLEMS

1. Each of the following properties defines a set of real $n \times n$ matrices. Find out which sets are dense, and which are open in the space $L(\mathbf{R}^n)$ of all linear operators on $\mathbf{R}^n$:
   (a) determinant $\neq 0$;
   (b) trace is rational;
   (c) entries are not integers;
   (d) $3 \leq$ determinant < 4;
   (e) $-1 < |\lambda| < 1$ for every eigenvalue λ;
   (f) no real eigenvalues;
   (g) each real eigenvalue has multiplicity one.

2. Which of the following properties of operators on $\mathbf{R}^n$ are generic?
   (a) $|\lambda| \neq 1$ for every eigenvalue λ;
   (b) $n = 2$; some eigenvalue is not real;
   (c) $n = 3$; some eigenvalue is not real;
   (d) no solution of $x' = Ax$ is periodic (except the zero solution);
   (e) there are n distinct eigenvalues, with distinct imaginary parts;
   (f) $Ax \neq x$ and $Ax \neq -x$ for all $x \neq 0$.

3. The set of operators on $\mathbf{R}^n$ that generate contractions is open, but not dense, in $L(\mathbf{R}^n)$. Likewise for expansions.

4. A subset X of a vector space is *residual* if there are dense open sets $A_k \subset E$, $k = 1, 2, \ldots$, such that $\bigcap A_k \subset X$.
   (a) Prove the theorem of Baire: a residual set is dense.
   (b) Show that if X_k is residual, $k = 1, 2, \ldots$, then $\bigcap X_k$ is residual.
   (d) If the set $Q \subset \mathbf{C}$ is residual, show that the set of operators in $\mathbf{R}^n$ whose eigenvalues are in Q, is residual in $L(\mathbf{R}^n)$.

§4. The Significance of Genericity

If an operator $A \in L(\mathbf{R}^n)$ is semisimple, the differential equation $x' = Ax$ breaks down into a number of simple uncoupled equations in one or two dimensions. If the eigenvalues of A have nonzero real parts, the differential equation might be complicated from the analytic point of view, but the geometric structure of the hyperbolic flow e^{tA} is very simple: it is the direct sum of a contraction and an expansion.

In Section 3 we showed that such nice operators A are in a sense *typical*. Precisely, operators that generate hyperbolic flows form a dense open set in $L(\mathbf{R}^n)$; while the set of semisimple operators contains a dense open set. Thus if A generates a hyperbolic flow, so does any operator sufficiently near to A. If A does not, we can approximate A arbitrarily closely by operators that do.

The significance of this for linear differential equations is the following. If there is uncertainty as to the entries in a matrix A, and no reason to assume the contrary, we might as well assume that the flow e^{tA} is hyperbolic. For example, A might be obtained from physical observations; there is a limit to the accuracy of the measuring instruments. Or the differential equation $x' = Ax$ may be used as an abstract model of some general physical (or biological, chemical, etc.) phenomenon; indeed, differential equations are very popular as models. In this case it makes little sense to insist on exact values for the entries in A.

It is often helpful in such situations to assume that A is as simple as possible—until compelled by logic, theory or observation to change that assumption. It is reasonable, then, to ascribe to A any convenient generic property. Thus it is comforting to assume that A is semisimple, for then the operator A (and the flow e^{tA}) are direct sums of very simple, easily analyzed one- and two-dimensional types.

There may, of course, be good reasons for not assuming a particular generic property. If it is suspected that the differential equation $x' = Ax$ has natural symmetry properties, for example, or that the flow must conserve some quantity (for example, energy), then assumption of a generic property could be a mistake.

Chapter 8

Fundamental Theory

This chapter is more difficult than the preceding ones; it is also central to the study of ordinary differential equations. We suggest that the reader browse through the chapter, omitting the proofs until the purpose of the theorems begins to fit into place.

§1. Dynamical Systems and Vector Fields

A dynamical system is a way of describing the passage in time of all points of a given space S. The space S could be thought of, for example, as the space of states of some physical system. Mathematically, S might be a Euclidean space or an open subset of Euclidean space. In the Kepler problem of Chapter 2, S was the set of possible positions and velocities; for the planar gravitational central force problem,

$$S = (\mathbf{R}^2 - 0) \times \mathbf{R}^2 = \{(x, v) \in \mathbf{R}^2 \times \mathbf{R}^2 \mid x \neq 0\}.$$

A *dynamical system on S tells us, for x in S, where x is 1 unit of time later, 2 units* of time later, and so on. We denote these new positions of x by x_1, x_2, respectively. At time zero, x is at x or x_0. One unit before time zero, x was at x_{-1}. If one extrapolates to fill up the real numbers, one obtains the trajectory x_t for all time t. The map $\mathbf{R} \to S$, which sends t into x_t, is a curve in S that represents the life history of x as time runs from $-\infty$ to ∞.

It is assumed that the map from $\mathbf{R} \times S \to S$ defined by $(t, x) \to x_t$ is continuously differentiable or at least continuous and continuously differentiable in t. The map $\phi_t: S \to S$ that takes x into x_t is defined for each t and from our interpretation as states moving in time it is reasonable to expect ϕ_t to have as an inverse ϕ_{-t}. Also, ϕ_0 should be the identity and $\phi_t(\phi_s(x)) = \phi_{t+s}(x)$ is a natural condition (remember $\phi_t(x) = x_t$).

We formalize the above in the following definition:

A *dynamical system* is a C^1 map $\mathbf{R} \times \mathcal{S} \xrightarrow{\phi} \mathcal{S}$ where $\mathcal{S}$ is an open set of Euclidean space and writing $\phi(t, x) = \phi_t(x)$, the map $\phi_t\colon \mathcal{S} \to \mathcal{S}$ satisfies

(a) $\phi_0\colon \mathcal{S} \to \mathcal{S}$ is the identity;
(b) The composition $\phi_t \circ \phi_s = \phi_{t+s}$ for each t, s in $\mathbf{R}$.

Note that the definition implies that the map $\phi_t\colon \mathcal{S} \to \mathcal{S}$ is C^1 for each t and has a C^1 inverse ϕ_{-t} (take $s = -t$ in (b)).

An example of a dynamical system is implicitly and approximately defined by the differential equations in the Newton–Kepler chapter. However, we give a precise example as follows.

Let A be an operator on a vector space E; let $E = \mathcal{S}$ and $\phi\colon \mathbf{R} \times \mathcal{S} \to \mathcal{S}$ be defined by $\phi(t, x) = e^{tA}x$. Thus $\phi_t\colon \mathcal{S} \to \mathcal{S}$ can be represented by $\phi_t = e^{tA}$. Clearly, $\phi_0 = e^0 =$ the identity operator and since $e^{(t+s)A} = e^{tA}e^{sA}$, we have defined a dynamical system on E (see Chapter 5).

This example of a dynamical system is related to the differential equation $dx/dt = Ax$ on E. A dynamical system ϕ_t on $\mathcal{S}$ in general gives rise to a differential equation on $\mathcal{S}$, that is, a vector field on $\mathcal{S}$, $f\colon \mathcal{S} \to E$. Here $\mathcal{S}$ is supposed to be an open set in the vector space E. Given ϕ_t, define f by

$$f(x) = \frac{d}{dt}\phi_t(x)\bigg|_{t=0}; \tag{1}$$

thus for x in $\mathcal{S}$, $f(x)$ is a vector in E which we think of as the tangent vector to the curve $t \to \phi_t(x)$ at $t = 0$. *Thus every dynamical system gives rise to a differential equation.*

We may rewrite this in more conventional terms. If $\phi_t\colon \mathcal{S} \to \mathcal{S}$ is a dynamical system and $x \in \mathcal{S}$, let $x(t) = \phi_t(x)$, and $f\colon \mathcal{S} \to E$ be defined as in (1). Then we may rewrite (1) as

$$x' = f(x). \tag{1'}$$

Thus $x(t)$ or $\phi_t(x)$ is the solution curve of (1′) satisfying the initial condition $x(0) = x$. There is a converse process to the above; given a differential equation one has associated to it, an object that would be a dynamical system if it were defined for all t. This process is the fundamental theory of differential equations and the rest of this chapter is devoted to it.

The equation (1′) we are talking about is called an *autonomous* equation. This means that the function f does not depend on time. One can also consider a C^1 map $f\colon I \times W \to E$ where I is an interval and W is an open set in the vector space. The equation in that case is

$$x' = f(t, x) \tag{2}$$

and is called nonautonomous. The existence and uniqueness theory for (1′) will be developed in this chapter; (2) will be treated in Chapter 15. Our emphasis in this book is completely on the autonomous case.

§2. The Fundamental Theorem

Throughout the rest of this chapter, E will denote a vector space with a norm; $W \subset E$, an open set in E; and $f\colon W \to E$ a continuous map. By a *solution* of the differential equation

$$x' = f(x) \tag{1}$$

we mean a differentiable function

$$u\colon J \to W$$

defined on some interval $J \subset \mathbf{R}$ such that for all $t \in J$

$$u'(t) = f(u(t)).$$

Here J could be an interval of real numbers which is open, closed, or half open, half closed. That is,

$$(a, b) = \{t \in \mathbf{R} \mid a < t < b\},$$

or

$$[a, b] = \{t \in \mathbf{R} \mid a \le t \le b\},$$

or

$$(a, b] = \{t \in \mathbf{R} \mid a < t \le b\},$$

and so on. Also, a or b could be ∞, but intervals like $(a, \infty]$ are not allowed.

Geometrically, u is a curve in E whose tangent vector $u'(t)$ equals $f(u(t))$; we think of this vector as based at $u(t)$. The map $f\colon W \to E$ is a vector field on W. A solution u may be thought of as the path of a particle that moves in E so that at time t, its tangent vector or velocity is given by the value of the vector field at the position of the particle. Imagine a dust particle in a steady wind, for example, or an electron moving through a constant magnetic field. See also Fig. A, where $u(t_0) = x$, $u'(t_0) = f(x)$.

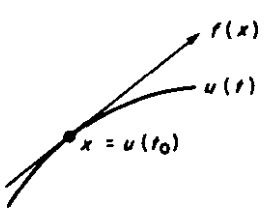


FIG. A

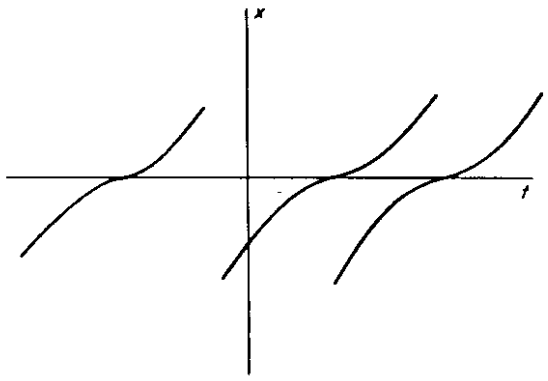


FIG. B

An *initial condition* for a solution $u: J \to W$ is a condition of the form $u(t_0) = x_0$ where $t_0 \in J$, $x_0 \in W$. For simplicity, we usually take $t_0 = 0$.

A differential equation might have several solutions with a given initial condition. For example, consider the equation in $\mathbf{R}$,

$$x' = 3x^{2/3}.$$

Here $W = \mathbf{R} = E$, $f: \mathbf{R} \to \mathbf{R}$ is given by $f(x) = 3x^{2/3}$.

The identically zero function $u_0: \mathbf{R} \to \mathbf{R}$ given by $u_0(t) = 0$ for all t is evidently a solution with initial condition $u(0) = 0$. But so is the function defined by $x(t) = t^3$. The graphs of some solution curves are shown in Fig. B.

Thus it is clear that to ensure unique solutions, extra conditions must be imposed on the function f. That f be continuously differentiable, turns out to be sufficient, as we shall see. Thus the phenomenon of nonuniqueness of solutions with given initial conditions is quite exceptional and rarely arises in practice.

In addition to uniqueness of solutions there is the question of existence. Up to this point, we have been able to compute solutions explicitly. Often, however, this is not possible, and in fact it is not a priori obvious that an arbitrary differential equation has any solutions at all.

We do not give an example of a differential equation without a solution because in fact (1) has a solution for all initial conditions provided f is continuous. We shall not prove this; instead we give an easier proof under hypotheses that also guarantee uniqueness.

The following is the fundamental local theorem of ordinary differential equations. It is called a "local" theorem because it deals with the nature of the vector field $f: W \to E$ near some point x_0 of W.

Theorem 1 *Let $W \subset E$ be an open subset of a normed vector space, $f: W \to E$ a C^1 (continuously differentiable) map, and $x_0 \in W$. Then there is some $a > 0$ and a unique*

solution

$$x: (-a, a) \to W$$

of the differential equation

$$x' = f(x)$$

satisfying the initial condition

$$x(0) = x_0.$$

We will prove Theorem 1 in the next section.

§3. Existence and Uniqueness

A function $f: W \to E$, W an open set of the normed vector space E, is said to be *Lipschitz* on W if there exists a constant K such that

$$|f(y) - f(x)| \leq K |y - x|$$

for all x, y in W. We call K a *Lipschitz constant* for f.

We have assumed a norm for E. In a different norm f will still be Lipschitz because of the equivalence of norms (Chapter 5); the constant K may change, however.

More generally, we call f *locally Lipschitz* if each point of W (the domain of f) has a neighborhood W_0 in W such that the restriction $f \mid W_0$ is Lipschitz. The Lipschitz constant of $f \mid W_0$ may vary with W_0.

Lemma *Let the function $f: W \to E$ be C^1. Then f is locally Lipschitz.*

Before giving the proof we recall the meaning of the derivative $Df(x)$ for $x \in W$. This is a linear operator on E; it assigns to a vector $u \in E$, the vector

$$Df(x)u = \lim_{s \to 0} \frac{1}{s} (f(x + su) - f(x)), \qquad s \in \mathbf{R},$$

which will exist if $Df(x)$ is defined.

In coordinates $(x_1, \ldots, x_n)$ on E, let $f(x) = (f_1(x_1, \ldots, x_n), \ldots, f_n(x_1, \ldots, x_n))$; then $Df(x)$ is represented by the $n \times n$ matrix of partial derivatives

$$(\partial/\partial x_j)(f_i(x_1, \ldots, x_n)).$$

Conversely, if all the partial derivatives exist and are continuous, then f is C^1. For each $x \in W$, there is defined the operator norm $\| Df(x) \|$ of the linear operator $Df(x) \in L(E)$ (see Chapter 5). If $u \in E$, then

$$|Df(x)u| \leq \| Df(x) \| \, |u|. \tag{1}$$

That $f: W \to E$ is C^1 implies that the map $W \to L(E)$ which sends $x \to Df(x)$ is a continuous map (see, for example, the notes at end of this chapter).

Proof of the lemma. Suppose that $f: W \to E$ is C^1 and $x_0 \in W$. Let $b > 0$ be so small that the ball $B_b(x_0)$ is contained in W, where

$$B_b(x_0) = \{x \in W \mid |x - x_0| \leq b\}.$$

Denote by W_0 this ball $B_b(x_0)$. Let K be an upper bound for $\|Df(x)\|$ on W_0; this exists because Df is continuous and W_0 is compact. The set W_0 is *convex*; that is, if $y, z \in W_0$, then the line segment going from y to z is in W_0. (In fact, any compact convex neighborhood of x_0 would work here.) Let y and z be in W_0 and put $u = z - y$. Then $y + su \in W_0$ for $0 \leq s \leq 1$. Let $\phi(s) = f(t, y + su)$; thus $\phi: [0, 1] \to E$ is the composition $[0, 1] \to W_0 \xrightarrow{f} E$ where the first map sends s into $y + su$. By the chain rule

$$\phi'(s) = Df(y + su)u. \tag{2}$$

Therefore

$$f(z) - f(y) = \phi(1) - \phi(0)$$

$$= \int_0^1 \phi'(s)\, ds$$

and, by (2),

$$= \int_0^1 Df(y + su)u\, ds.$$

Hence, by (1),

$$|f(z) - f(y)| \leq \int_0^1 K|u|\, ds = K|z - y|.$$

This proves the lemma.

The following remark is implicit in the proof of the lemma:

If W_0 is convex, and if $\|Df(x)\| \leq K$ for all $x \in W_0$, then K is a Lipschitz constant for $f \mid W_0$.

We proceed now to the proof of the existence part of Theorem 1 of Section 2. Let $x_0 \in W$ and W_0 be as in the proof of the previous lemma. Suppose J is an open interval containing zero and $x: J \to W$ satisfies

$$x'(t) = f(x(t)) \tag{3}$$

and $x(0) = x_0$. Then by integration we have

$$x(t) = x_0 + \int_0^t f(x(s))\, ds. \tag{4}$$

Conversely, if $x: J \to W$ satisfies (4), then $x(0) = x_0$ and x satisfies (3) as is seen by differentiation.

Thus (4) is equivalent to (3) as an equation for $x: J \to W$.

By our choice of W_0, we have a Lipschitz constant K for f on W_0. Furthermore, $|f(x)|$ is bounded on W_0, say, by the constant M. Let $a > 0$ satisfy

$a < \min\{b/M, 1/K\}$, and define $J = [-a, a]$. Recall that b is the radius of the ball W_0. We shall define a sequence of functions $u_0, u_1, \ldots$ from J to W_0. We shall prove they converge uniformly to a function satisfying (4), and later that there are no other solutions of (4). The lemma that is used to obtain the convergence of the $u_k: J \to W_0$ is the following:

Lemma from analysis *Suppose $u_k: J \to E$, $k = 0, 1, 2, \ldots$ is a sequence of continuous functions from a closed interval J to a normed vector space E which satisfy: Given $\epsilon > 0$, there is some $N > 0$ such that for every $p, q > N$*

$$\max_{t \in J} |u_p(t) - u_q(t)| < \epsilon.$$

Then there is a continuous function $u: J \to E$ such that

$$\max_{t \in J} |u_k(t) - u(t)| \to 0 \quad \text{as} \quad k \to \infty.$$

This is called uniform convergence of the functions u_k. This lemma is proved in elementary analysis books and will not be proved here.

The sequence of functions u_k is defined as follows:

Let

$$u_0(t) \equiv x_0.$$

Let

$$u_1(t) = x_0 + \int_0^t f(u_0(s))\, ds.$$

Assuming that $u_k(t)$ has been defined and that

$$|u_k(t) - x_0| \leq b \quad \text{for all} \quad t \in J,$$

let

$$u_{k+1}(t) = x_0 + \int_0^t f(u_k(s))\, ds.$$

This makes sense since $u_k(s) \in W_0$ so the integrand is defined. We show that

$$|u_{k+1}(t) - x_0| \leq b \quad \text{or} \quad u_{k+1}(t) \in W_0 \quad \text{for} \quad t \in J;$$

this will imply that the sequence can be continued to u_{k+2}, u_{k+3}, and so on.

We have

$$|u_{k+1}(t) - x_0| \leq \int_0^t |f(u_k(s))|\, ds$$

$$\leq \int_0^t M\, ds$$

$$\leq Ma < b.$$

Next, we prove that there is a constant $L \geq c$ such that for all $k \geq 0$:

$$|u_{k+1}(t) - u_k(t)| \leq (Ka)^k L.$$

Put $L = \max\{|u_1(t) - u_0(t)|: |t| \leq a\}$. We have

$$|u_2(t) - u_1(t)| = \left| \int_0^t f(u_1(s) - f(u_0(s))\, ds \right|$$

$$\leq \int_0^t K |u_1(s) - u_0(s)|\, ds$$

$$\leq aKL.$$

Assuming by induction that, for some $k \geq 2$, we have already proved

$$|u_k(t) - u_{k-1}(t)| \leq (aK)^{k-1} L, \qquad |t| < a,$$

we have

$$|u_{k+1}(t) - u_k(t)| \leq \int_0^t |f(u_k(s)) - f(u_{k-1}(s))|\, ds$$

$$\leq K \int_0^t |u_k(s) - u_{k-1}(s)|\, ds$$

$$\leq (aK)(aK)^{k-1} L = (aK)^k L.$$

Therefore we see that, putting $aK = \alpha < 1$, for any $r > s > N$

$$|u_r(t) - u_s(t)| \leq \sum_{k=N}^{\infty} |u_{k+1}(t) - u_k(t)|$$

$$\leq \sum_{k=N}^{\infty} \alpha^k L$$

$$\leq \epsilon$$

for any prescribed $\epsilon > 0$ provided N is large enough.

By the lemma from analysis, this shows that the sequence of functions $u_0, u_1, \ldots$ converges uniformly to a continuous function $x: J \to E$. From the identity

$$u_{k+1}(t) = x_0 + \int_0^t f(u_k(s))\, ds,$$

we find by taking limits of both sides that

$$x(t) = x_0 + \lim_{k \to \infty} \int_0^t f(u_k(s))\, ds$$

$$= x_0 + \int_0^t [\lim_{k \to \infty} f(u_k(s))]\, ds$$

(by uniform convergence)

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

(by continuity of f).

Therefore $x: J \to W_0$ satisfies (4) and hence is a solution of (3). In particular, $x: J \to W_0$ is C^1.

This takes care of the existence part of Theorem 1 (of Section 1) and we now prove the uniqueness part.

Let $x, y: J \to W$ be two solutions of (1) satisfying $x(0) = y(0) = x_0$, where we may suppose that J is the closed interval $[-a, a]$. We will show that $x(t) = y(t)$ for all $t \in J$. Let $Q = \max_{t \in J} |x(t) - y(t)|$. This maximum is attained at some point $t_1 \in J$. Then

$$Q = |x(t_1) - y(t_1)| = \left| \int_0^{t_1} x'(s) - y'(s)\, ds \right|$$

$$\leq \int_0^{t_1} |f(x(s)) - f(y(s))|\, ds$$

$$\leq \int_0^{t_1} K |x(s) - y(s)|\, ds$$

$$\leq aKQ.$$

Since $aK < 1$, this is impossible unless $Q = 0$. Thus

$$x(t) \equiv y(t).$$

Another proof of uniqueness follows from the lemma of the next section.

We have proved Theorem 1 of Section 2. Note that in the course of the proof the following was shown: Given any ball $W_0 \subset W$ of radius b about x_0, with $\max_{x \in W_0} |f(x)| \leq M$, where f on W_0 has Lipschitz constant K and $0 < a < \min\{b/M, 1/K\}$, then there is a unique solution $x: (-a, a) \to W$ of (3) such that $x(0) = x_0$.

Some remarks are in order.

Consider the situation in Theorem 1 with a C^1 map $f: W \to E$, W open in E. *Two solution curves of $x' = f(x)$ cannot cross.* This is an immediate consequence of uniqueness but is worth emphasizing geometrically. Suppose $\varphi: J \to W$, $\psi: J_1 \to W$ are two solutions of $x' = f(x)$ such that $\varphi(t_1) = \psi(t_2)$. Then $\varphi(t_1)$ is not a crossing because if we let $\psi_1(t) = \psi(t_2 - t_1 + t)$, then ψ_1 is also a solution. Since $\psi_1(t_1) = \psi(t_2) = \varphi(t_1)$, it follows that ψ_1 and φ agree near t_1 by the uniqueness statement of

Theorem 1. Thus the situation of Fig. A is prevented. Similarly, a solution curve cannot cross itself as in Fig. B.

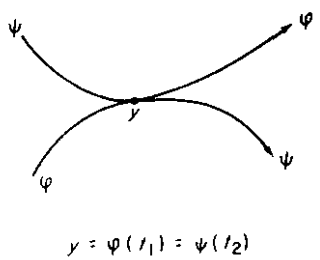


FIG. A

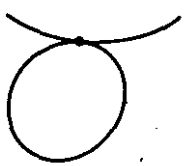


FIG. B

If, in fact, a solution curve $\varphi: J \to W$ of $x' = f(x)$ satisfies $\varphi(t_1) = \varphi(t_1 + w)$ for some t_1 and $w > 0$, then that solution curve must close up as in Fig. C.

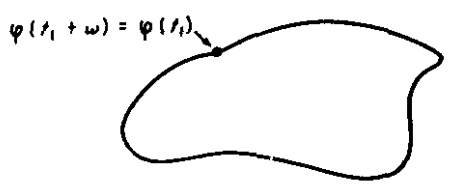


FIG. C

Let us see how the "iteration scheme" used in the proof in this section applies to a very simple differential equation. Consider $W = \mathbf{R}$ and $f(x) = x$, and search for a solution of $x' = x$ in $\mathbf{R}$ (we know already that the solution $x(t)$ satisfying $x(0) = x_0$ is given by $x(t) = x_0e^t$).

Set

$$u_0(t) = x_0,$$

$$u_1(t) = x_0 + \int_0^t x_0\,ds, \qquad u_1(t) = x_0 + tx_0,$$

$$u_2(t) = x_0 + \int_0^t (x_0 + tx_0)\,dt,$$

$$u_2(t) = x_0 + tx_0 + \frac{t^2}{2}x_0,$$

$$u_{k+1}(t) = x_0 + \int_0^t u_k(s)\,ds,$$

and so

$$u_{k+1}(t) = x_0 \sum_{i=0}^{k} \frac{t^i}{i!}.$$

As $k \to \infty$, $u_k(t)$ converges to

$$x_0 \sum_{i=0}^{\infty} \frac{t^i}{i!} = x_0e^t = u(t),$$

which is, of course, the solution of our original equation.

§4. Continuity of Solutions in Initial Conditions

For Theorem 1 of Section 2 to be at all interesting in any physical sense (or even mathematically) it needs to be complemented by the property that the solution $x(t)$ depends continuously on the initial condition $x(0)$. The next theorem gives a precise statement of this property.

Theorem *Let $W \subset E$ be open and suppose $f: W \to E$ has Lipschitz constant K. Let $y(t)$, $z(t)$ be solutions to*

$$x' = f(x) \tag{1}$$

on the closed interval $[t_0, t_1]$. Then, for all $t \in [t_0, t_1]$:

$$|y(t) - z(t)| \leq |y(t_0) - z(t_0)| \exp(K(t - t_0)).$$

The proof depends on a useful inequality (Gronwall's) which we prove first.

Lemma *Let $u: [0, \alpha] \to \mathbf{R}$ be continuous and nonnegative. Suppose $C \geq 0$, $K \geq 0$ are such that*

$$u(t) \leq C + \int_0^t Ku(s)\,ds$$

for all $t \in [0, \alpha]$. Then

$$u(t) \leq Ce^{Kt}$$

for all $t \in [0, \alpha]$.

Proof. First, suppose $C > 0$, let

$$U(t) = C + \int_0^t Ku(s)\,ds > 0;$$

then

$$u(t) \leq U(t).$$

By differentiation of U we find

$$U'(t) = Ku(t);$$

hence

$$\frac{U'(t)}{U(t)} = \frac{Ku(t)}{U(t)} \leq K.$$

Hence

$$\frac{d}{dt}(\log U(t)) \leq K$$

so

$$\log U(t) \leq \log U(0) + Kt$$

by integration. Since $U(0) = C$, we have by exponentiation

$$U(t) \leq Ce^{Kt},$$

and so

$$u(t) \leq Ce^{Kt}.$$

If $C = 0$, then apply the above argument for a sequence of positive c_i that tend to 0 as $i \to \infty$. This proves the lemma.

We turn to the proof of the theorem.

Define

$$v(t) = |y(t) - z(t)|.$$

Since

$$y(t) - z(t) = y(t_0) - z(t_0) + \int_{t_0}^t [f(y(s)) - f(z(s))]ds,$$

we have

$$v(t) \leq v(t_0) + \int_{t_0}^t Kv(s)\,ds.$$

Now apply the lemma to the function $u(t) = v(t_0 + t)$ to get

$$v(t) \leq v(t_0)\exp(K(t - t_0)),$$

which is just the conclusion of the theorem.

§5. On Extending Solutions

Lemma *Let a C^1 map $f: W \to E$ be given. Suppose two solutions $u(t)$, $v(t)$ of $x' = f(x)$ are defined on the same open interval J containing t_0 and satisfy $u(t_0) = v(t_0)$. Then $u(t) = v(t)$ for all $t \in J$.*

We know from Theorem 1 of Section 2 that $u(t) = v(t)$ in some open interval around t_0. The union of all such open intervals is the largest open interval J^* in J around t_0 on which $u = v$. But J^* must equal J. For, if not, J^* has an end point $t_1 \in J$; we suppose t_1 is the right-hand end point, the other case being similar. By continuity, $u(t_1) = v(t_1)$. But, by Theorem 1 of Section 2, $u = v$ in some J', an interval around t_1. Then $u = v$ in $J^* \cup J'$ which is larger than J^*. This contradiction proves the lemma.

There is no guarantee that a solution $x(t)$ to a differential equation can be defined for all t. For example, the equation in $\mathbf{R}$,

$$x' = 1 + x^2,$$

has as solutions the functions

$$x = \tan(t - c), \qquad c = \text{constant}.$$

Such a function cannot be extended over an interval larger than

$$c - \frac{\pi}{2} < t < c + \frac{\pi}{2}$$

since $x(t) \to \pm\infty$ as $t \to c \pm \pi/2$.

Now consider a general equation (1) $x' = f(x)$ where the C^1 function f is defined on an open set $W \subset E$. For *each $x_0 \in W$ there is a maximum open interval (α, β) containing 0 on which there is a solution $x(t)$* with $x(0) = x_0$. There is some such interval by Theorem 1 of Section 2; let (α, β) be the union of *all* open intervals containing 0 on which there is a solution with $x(0) = x_0$. (Possibly, $\alpha = -\infty$ or $\beta = +\infty$, or both.) By the lemma the solutions on any two intervals in the union agree on the intersections of the two intervals. Hence there is a solution on all of (α, β).

Next, we investigate what happens to a solution as the limits of its domain are approached. We state the result only for the right-hand limit; the other case is similar.

Theorem *Let $W \subset E$ be open, let $f: W \to E$ be a C^1 map. Let $y(t)$ be a solution on a maximal open interval $J = (\alpha, \beta) \subset \mathbf{R}$ with $\beta < \infty$. Then given any compact set $K \subset W$, there is some $t \in (\alpha, \beta)$ with $y(t) \notin K$.*

This theorem says that if a solution $y(t)$ cannot be extended to a larger interval, then it leaves any compact set. This implies that *as $t \to \beta$ either $y(t)$ tends to the boundary of W or $|y(t)|$ tends to ∞* (or both).

Proof of the theorem. Suppose $y(t) \in K$ for all $t \in (\alpha, \beta)$. Since f is continuous, there exists $M > 0$ such that $|f(x)| \leq M$ if $x \in K$.

Let $\gamma \in (\alpha, \beta)$. Now we prove that y extends to a continuous map $[\gamma, \beta] \to E$. By a lemma from analysis it suffices to prove $y: J \to E$ uniformly continuous. For $t_0 < t_1$ in J we have

$$|y(t_0) - y(t_1)| = \left| \int_{t_0}^{t_1} y'(s)\, ds \right|$$

$$\leq \int_{t_0}^{t_1} |f(y(s))|\, ds \leq (t_1 - t_0) M.$$

Now the extended curve $y: [\alpha, \beta] \to E$ is differentiable at β. For

$$y(\beta) = y(\gamma) + \lim_{t \to \beta} \int_{\gamma}^{t} y'(s)\, ds$$

$$= y(\gamma) + \lim_{t \to \beta} \int_{\gamma}^{t} f(y(s))\, ds$$

$$= y(\gamma) + \int_{\gamma}^{\beta} f(y(s))\, ds;$$

hence

$$y(t) = y(\gamma) + \int_{\gamma}^{t} f(y(s))\, ds$$

for all t between γ and β. Hence y is differentiable at β, and in fact $y'(\beta) = f(y(\beta))$. Therefore y is a solution on $[\gamma, \beta]$. Since there is a solution on an interval $[\beta, \delta)$, $\delta > \beta$, we can extend y to the interval (α, δ). Hence (α, β) could not be a maximal domain of a solution. This completes the proof of the theorem.

The following important fact follows immediately from the theorem.

Proposition *Let A be a compact subset of the open set $W \subset E$ and let $f: W \to E$ be C^1. Let $y_0 \in A$ and suppose it is known that every solution curve of the form*

$$y: [0, \beta] \to W, \qquad y(0) = y_0,$$

lies entirely in A. Then there is a solution

$$y: [0, \infty) \to W, \qquad y(0) = y_0, \qquad \text{and} \qquad y(t) \in A$$

for all $t \geq 0$.

Proof. Let $[0, \beta)$ be the maximal half-open interval on which there is a solution y as above. Then $y([0, \beta)) \subset A$, and so β cannot be finite by the theorem.

§6. Global Solutions

We give here a stronger theorem on the continuity of solutions in terms of initial conditions.

In the theorem of Section 4 we assumed that both solutions were defined on the same interval. In the next theorem it is not necessary to assume this. The theorem shows that solutions starting at nearby points will be defined on the same closed interval and remain near to each other in this interval.

Theorem *Let $f(x)$ be C^1. Let $y(t)$ be a solution to $x' = f(x)$ defined on the closed interval $[t_0, t_1]$, with $y(t_0) = y_0$. There is a neighborhood $U \subset E$ of y_0 and a constant K such that if $z_0 \in U$, then there is a unique solution $z(t)$ also defined on $[t_0, t_1]$ with $z(t_0) = z_0$; and z satisfies*

$$|y(t) - z(t)| \leq K |y_0 - z_0| \exp(K(t - t_0))$$

for all $t \in [t_0, t_1]$.

For the proof we will use the following lemma.

Lemma *If $f: W \to E$ is locally Lipschitz and $A \subset W$ is a compact (closed and bounded) set, then $f \mid A$ is Lipschitz.*

Proof. Suppose not. Then for every $K > 0$, no matter how large, we can find x and y in A with

$$|f(x) - f(y)| > K |x - y|.$$

In particular, we can find x_n, y_n such that

$$|f(x_n) - f(y_n)| \geq n |x_n - y_n| \qquad \text{for} \quad n = 1, 2, \ldots. \tag{1}$$

Since A is compact, we can choose convergent subsequences of the x_n and y_n. Relabeling, we may assume $x_n \to x^*$ and $y_n \to y^*$ with x^* and y^* in A. We observe that $x^* = y^*$, since we have, for all n,

$$|x^* - y^*| = \lim_{n \to \infty} |x_n - y_n| \leq n^{-1} |f(x_n) - f(y_n)| \leq n^{-1} 2M,$$

where M is the maximum value of f on A. There is a neighborhood W_0 of x^* for which $f \mid W_0$ has a Lipschitz constant K in x. There is an n_0 such that $x_n \in W_0$ if $n \geq n_0$. Therefore, for $n \geq n_0$:

$$|f(x_n) - f(y_n)| \leq K |x_n - y_n|,$$

which contradicts (1) for $n > K$. This proves the lemma.

The proof of the theorem now goes as follows.

By compactness of $[t_0, t_1]$, there exists $\epsilon > 0$ such that $x \in W$ if $|x - y(t)| \leq \epsilon$. The set of all such points is a compact subset A of W. The C^1 map f is locally Lipschitz (Section 3). By the lemma, it follows that $f \mid A$ has a Lipschitz constant k.

Let $\delta > 0$ be so small that $\delta \leq \epsilon$ and $\delta \exp(k \mid t_1 - t_0 \mid) \leq \epsilon$. We assert that if $|z_0 - y_0| < \delta$, then there is a unique solution through z_0 defined on all of $[t_0, t_1]$. First of all, $z_0 \in W$ since $|z_0 - y(t_0)| < \epsilon$, so there is a solution $z(t)$ through z_0 on a maximal interval $[t_0, \beta)$. We prove $\beta > t_1$. For suppose $\beta \leq t_1$. Then by the exponential estimate in Section 4, for all $t \in [t_0, \beta)$, we have

$$\begin{aligned} |z(t) - y(t)| &\leq |z_0 - y_0| \exp(k \mid t - t_0 \mid) \\ &\leq \delta \exp(k \mid t - t_0 \mid) \\ &\leq \epsilon. \end{aligned}$$

Thus $z(t)$ lies in the compact set A; by the theorem of Section 5, $[t_0, \beta)$ could not be a *maximal* solution domain. Therefore $z(t)$ is defined on $[t_0, t_1]$. The exponential estimate follows from Section 4, and the uniqueness from the lemma of Section 5.

We interpret the theorem in another way. Given $f(x)$ as in the theorem and a solution $y(t)$ defined on $[t_0, t_1]$, we see that for all z_0 sufficiently close to $y_0 = y(t_0)$, there is a unique solution on $[t_0, t_1]$ starting at z_0 at time zero. Let us note this solution by $t \to u(t, z_0)$; thus $u(0, z_0) = z_0$, and $u(t, y_0) = y(t)$.

Then the theorem implies:

$$\lim_{z_0 \to y_0} u(t, z_0) = u(t, y_0),$$

uniformly on $[t_0, t_1]$. In other words, *the solution through z_0 depends continuously on z_0.*

§7. The Flow of a Differential Equation

In this section we consider an equation

$$x' = f(x) \tag{1}$$

defined by a C^1 function $f: W \to E$, $W \subset E$ open.

For each $y \in W$ there is a unique solution $\phi(t)$ with $\phi(0) = y$ defined on a maximal open interval $J(y) \subset \mathbf{R}$. To indicate the dependence of $\phi(t)$ on y, we write

$$\phi(t) = \phi(t, y).$$

Thus $\phi(0, y) = y$.

Let $\Omega \subset \mathbf{R} \times W$ be the following set:

$$\Omega = \{(t, y) \in \mathbf{R} \times W \mid t \in J(y)\}.$$

The map $(t, y) \to \phi(t, y)$ is then a function

$$\phi: \Omega \to W.$$

We call ϕ the *flow* of equation (1).

We shall often write

$$\phi(t, x) = \phi_t(x).$$

Example. Let $f(x) = Ax$, $A \in L(E)$. Then $\phi_t(x) = e^{tA}x$.

Theorem 1 *The map ϕ has the following property:*

$$\phi_{s+t}(x) = \phi_s(\phi_t(x)) \tag{2}$$

in the sense that if one side of (2) is defined, so is the other, and they are equal.

Proof. First, suppose s and t are positive and $\phi_s(\phi_t(x))$ is defined. This means $t \in J(x)$ and $s \in J(\phi_t(x))$. Suppose $J(x) = (\alpha, \beta)$. Then $\alpha < t < \beta$; we shall show $\beta > s + t$. Define

$$y: (\alpha, s + t] \to W$$

by

$$y(r) = \begin{cases} \phi(r, x) & \text{if } \alpha < r \leq t; \\ \phi(r - t, \phi_t(x)) & \text{if } t \leq r \leq t + s. \end{cases}$$

Then y is a solution and $y(0) = x$. Hence $s + t \in J(x)$. Moreover,

$$\phi_{s+t}(x) = y(s + t) = \phi_s(\phi_t(x)).$$

The rest of the proof of Theorem 1 uses the same ideas and is left to the reader.

Theorem 2 *Ω is an open set in $\mathbf{R} \times W$ and $\phi: \Omega \to W$ is a continuous map.*

Proof. To prove Ω open, let $(t_0, x_0) \in \Omega$. We suppose $t_0 \geq 0$, the other case being similar. Then the solution curve $t \to \phi(t, x_0)$ is defined on $[0, t_0]$, and hence on an interval $[-\epsilon, t_0 + \epsilon]$, $\epsilon > 0$. By the theorem of Section 6, there is a neighborhood $U \subset W$ of x_0 such that the solution $t \to \phi(t, x)$ is defined on $[-\epsilon, t_0 + \epsilon]$ for all x in U. Thus $(-\epsilon, t_0 + \epsilon) \times U \subset \Omega$, which proves Ω open.

To prove $\phi: \Omega \to W$ continuous at (t_0, x_0), let U and ϵ be as above. We may suppose that U has compact closure $\bar{U} \subset W$. Since f is locally Lipschitz and the set $A = \phi([-\epsilon, t_0 + \epsilon] \times \bar{U})$ is compact, there is a Lipschitz constant K for $f \mid A$. Let $M = \max\{|f(x)| : x \in A\}$. Let $\delta > 0$ satisfy $\delta < \epsilon$, and if $|x_1 - x_0| < \delta$, then $x_1 \in U$. Suppose

$$|t_1 - t_0| < \delta, \qquad |x_1 - x_0| < \delta.$$

Then

$$|\phi(t_1, x_1) - \phi(t_0, x_0)| \leq |\phi(t_1, x_1) - \phi(t_1, x_0)| + |\phi(t_1, x_0) - \phi(t_0, x_0)|.$$

The second term on the right goes to 0 with δ because the solution through x_0 is continuous (even differentiable) in t. The first term on the right, by the estimate in Section 6, is bounded by $\delta e^{K\delta}$ which also goes to 0 with δ. This proves Theorem 2.

In Chapter 16 we shall show that in fact ϕ is C^1.

Now suppose $(t, x_0) \in \Omega$; then x_0 has a neighborhood $U \subset W$ with $t \times U \subset \Omega$, since we know Ω is open in $\mathbf{R} \times W$. The function $x \to \phi_t(x)$ defines a map

$$\phi_t: U \to W.$$

Theorem 3 *The map ϕ_t sends U onto an open set V and ϕ_{-t} is defined on V and sends V onto U. The composition $\phi_{-t}\phi_t$ is the identity map of U; the composition $\phi_t\phi_{-t}$ is the identity map of V.*

Proof. If $y = \phi_t(x)$, then $t \in J(x)$. It is easy to see that then $-t \in J(y)$, for the function

$$s \to \phi_{s-t}(y)$$

is a solution on $[-t, 0]$ sending 0 to y. Thus ϕ_{-t} is defined on $\phi_t(U) = V$; the statement about compositions is obvious. It remains to prove V is open. Let $V^* \supset V$ be the maximal subset of W on which ϕ_{-t} is defined. V^* is open because Ω is open, and $\phi_{-t}: V^* \to W$ is continuous because ϕ is continuous. Therefore the inverse image of the open set U under ϕ_{-t} is open. But this inverse image is exactly V.

We summarize the results of this section:

Corresponding to the autonomous equation $x' = f(x)$, with locally Lipschitz $f: W \to E$, there is a map $\phi: \Omega \to W$ where $(t, x) \in \Omega$ if and only if there is a solution on $[0, t]$ (or $[t, 0]$ if $t < 0$) sending 0 to x. The set Ω is open. ϕ is defined by letting $t \to \phi_t(x) = \phi(t, x)$ be the maximal solution curve taking 0 to x. There is an open set $U_t \subset W$ on which the map $\phi_t: U_t \to W$ is defined. The maps ϕ_t satisfy $\phi_s\phi_t(x) = \phi_{s+t}(x)$ as in Theorem 1. Each map ϕ_t is a homeomorphism; that is, ϕ_t is one-to-one and has a continuous inverse; the inverse is ϕ_{-t}.

If

$$f(x) = Ax, \qquad A \in L(E),$$

then

$$\phi_t(x) = e^{tA}x.$$

In this case $\Omega = \mathbf{R} \times E$ and each ϕ_t is defined in all of E.

PROBLEMS

1. Write out the first few terms of the Picard iteration scheme (Section 3) for each of the following initial value problems. Where possible, use any method to find explicit solutions. Discuss the domain of the solution.
   (a) $x' = x + 2$; $x(0) = 2$.
   (b) $x' = x^{4/3}$; $x(0) = 0$.
   (c) $x' = x^{4/3}$; $x(0) = 1$.
   (d) $x' = \sin x$; $x(0) = 0$.
   (e) $x' = 1/2x$; $x(1) = 1$.

2. Let A be an $n \times n$ matrix. Show that the Picard method for solving $x' = Ax$, $x(0) = u$ gives the solution $e^{tA}u$.

3. Derive the Taylor series for $\sin t$ by applying the Picard method to the first order system corresponding to the second order initial value problem

$$x'' = -x; \qquad x(0) = 0, \qquad x'(0) = 1.$$

4. For each of the following functions, find a Lipschitz constant on the region indicated, or prove there is none:
   (a) $f(x) = |x|$, $-\infty < x < \infty$.
   (b) $f(x) = x^{1/3}$, $-1 \leq x \leq 1$.
   (c) $f(x) = 1/x$, $1 \leq x \leq \infty$.
   (d) $f(x, y) = (x + 2y, -y)$, $(x, y) \in \mathbf{R}^2$.
   (e) $f(x, y) = \dfrac{xy}{1 + x^2 + y^2}$, $x^2 + y^2 \leq 4$.

5. Consider the differential equation

$$x' = x^{2/3}.$$

   (a) There are infinitely many solutions satisfying $x(0) = 0$ on every interval $[0, \beta]$.
   (b) For what values of α are there infinitely many solutions on $[0, \alpha]$ satisfying $x(0) = -1$?

6. Let $f: E \to E$ be continuous; suppose $f(x) \leq M$. For each $n = 1, 2, \ldots$, let $x_n: [0, 1] \to E$ be a solution to $x' = f(x)$. If $x_n(0)$ converges, show that a subsequence of $\{x_n\}$ converges uniformly to a solution. (*Hint*: Look up Ascoli's theorem in a book on analysis.)

7. Use Problem 6 to show that continuity of solutions in initial conditions follows from uniqueness and existence of solutions.

8. Prove the following general fact (see also Section 4): if $C \geq 0$ and $u, v: [0, \beta] \to \mathbf{R}$ are continuous and nonnegative, and

$$u(t) \leq C + \int_0^t u(s)v(s)\,ds \qquad \text{for all} \quad t \in [0, \beta],$$

then

$$u(t) \leq Ce^{V(t)}, \qquad V(t) = \int_0^t v(s)\,ds.$$

9. Define $f: \mathbf{R} \to \mathbf{R}$ by

$$f(x) = 1 \quad \text{if} \quad x \leq 1; \qquad f(x) = 2 \quad \text{if} \quad x > 1.$$

There is no solution to $x' = f(x)$ on any open interval around $t = 1$.

10. Let $g: \mathbf{R} \to \mathbf{R}$ be Lipschitz and $f: \mathbf{R} \to \mathbf{R}$ continuous. Show that the system

$$x' = g(x),$$
$$y' = f(x)y,$$

has at most one solution on any interval, for a given initial value. (*Hint*: Use Gronwall's inequality.)

Notes

Our treatment of calculus tends to be from the modern point of view. The derivative is viewed as a linear transformation.

Suppose that U is an open set of a vector space E and that $g: U \to F$ is some map, F a second vector space. What is the derivative of g at $x_0 \in U_0$? We say that this derivative exists and is denoted by $Dg(x_0) \in L(E, F)$ if

$$\lim_{\substack{|u| \to 0 \\ u \in E \\ u \neq 0}} \frac{|\, g(x_0 + u) - g(x_0) - Dg(x_0)u \,|}{|\, u \,|} = 0.$$

Then, if, for each $x \in U$, the derivative $Dg(x)$ exists, this derivative defines a map

$$U \to L(E, F), \qquad x \to Dg(x).$$

If this map is continuous, then g is said to be C^1. If this map is C^1 itself, then g is said to be C^2.

Now suppose F, G, H are three vector spaces and u, v are open sets of F, G, respectively. Consider C^1 maps f, g,

$$U \xrightarrow{f} V \xrightarrow{g} H.$$

The chain rule of calculus can be stated as: the derivative of the composition is the composition of the derivatives. In other words, if $x \in U$, then

$$D(gf)(x) = Dg(f(x)) \cdot Df(x) \in L(F, H).$$

Consider the case where $F = \mathbf{R}$ and U is an interval; writing $t \in U$, $f'(t) = Df(t)$, the chain rule reads

$$(gf)'(t) = Dg(f(t))(f'(t)).$$

In case H also equals $\mathbf{R}$, the formula becomes

$$(gf)'(t) = \langle \operatorname{grad} g(f(t)), f'(t) \rangle.$$

For more details on this and a further development of calculus along these lines, see S. Lang's *Second Course in Calculus* [12]. S. Lang's *Analysis I* [11] also covers these questions as well as the lemma from analysis used in Section 3 and the uniform continuity statement used in the proof of the theorem of Section 5.

Chapter 9

Stability of Equilibria

In this chapter we introduce the important idea of *stability of an equilibrium point* of a dynamical system. In later chapters other kinds of stability will be discussed, such as stability of periodic solutions and structural stability.

An equilibrium $\bar{x}$ is *stable* if all nearby solutions stay nearby. It is *asymptotically stable* if all nearby solutions not only stay nearby, but also tend to $\bar{x}$. Of course, precise definitions are required; these are given in Section 2. In Section 1 a special kind of asymptotically stable equilibrium is studied first: the *sink*. This is characterized by *exponential* approach to $\bar{x}$ of all nearby solutions. In Chapter 7 the special case of linear sinks was considered. Sinks are useful because they can be detected by the eigenvalues of the linear part of the system (that is, the derivative of the vector field at $\bar{x}$).

In Section 3 the famous stability theorems of Liapunov are proved. This section also contains a rather refined theorem (Theorem 2) which is not essential for the rest of the book, except in Chapter 10.

Sections 4 and 5 treat the important special case of gradient flows. These have special properties that make their analysis fairly simple; moreover, they are of frequent occurrence.

§1. Nonlinear Sinks

Consider a differential equation

$$x' = f(x); \qquad f: W \to \mathbf{R}^n; \qquad W \subset \mathbf{R}^n \text{ open.} \tag{1}$$

We suppose f is C^1. A point $\bar{x} \in W$ is called an *equilibrium* point of (1) if $f(\bar{x}) = 0$. Clearly, the constant function $x(t) \equiv \bar{x}$ is a solution of (1). By uniqueness of solutions, no other solution curve can pass through $\bar{x}$. If W is the state space of some physical (or biological, economic, or the like) system described by (1), then $\bar{x}$ is an "equilibrium state": if the system is at $\bar{x}$ it always will be (and always was) at $\bar{x}$.

Let $\phi: \Omega \to W$ be the flow associated with (1); $\Omega \subset \mathbf{R} \times W$ is an open set, and for each $x \in W$ the map $t \to \phi(t, x) = \phi_t(x)$ is the solution passing through x when $t = 0$; it is defined for t in some open interval. If $\bar{x}$ is an equilibrium, then $\phi_t(\bar{x}) = \bar{x}$ for all $t \in \mathbf{R}$. For this reason, $\bar{x}$ is also called a *stationary point*, or *fixed point*, of the flow. Another name for $\bar{x}$ is a *zero* or *singular point* of the vector field f.

Suppose f is linear: $W = \mathbf{R}^n$ and $f(x) = Ax$ where A is a linear operator on $\mathbf{R}^n$. Then the origin $0 \in \mathbf{R}^n$ is an equilibrium of (1). In Chapter 7 we saw that when $\lambda < 0$ is greater than the real parts of the eigenvalues of A, then solutions $\phi_t(x)$ approach 0 exponentially:

$$|\varphi_t(x)| \leq Ce^{\lambda t}$$

for some $C > 0$.

Now suppose f is a C^1 vector field (not necessarily linear) with equilibrium point $0 \in \mathbf{R}^n$. We think of the derivative $Df(0) = A$ of f at 0 as a linear vector field which approximates f near 0. We call it the *linear part* of f at 0. If all eigenvalues of $Df(0)$ have negative real parts, we call 0 a *sink*. More generally, an equilibrium $\bar{x}$ of (1) is a sink if all eigenvalues of $Df(\bar{x})$ have negative real parts.

The following theorem says that a nonlinear sink $\bar{x}$ behaves locally like a linear sink: nearby solutions approach $\bar{x}$ exponentially.

Theorem *Let $\bar{x} \in W$ be a sink of equation (1). Suppose every eigenvalue of $Df(\bar{x})$ has real part less than $-c$, $c > 0$. Then there is a neighborhood $U \subset W$ of $\bar{x}$ such that*

(a) *$\phi_t(x)$ is defined and in U for all $x \in U$, $t > 0$.*

(b) *There is a Euclidean norm on $\mathbf{R}^n$ such that*

$$|\phi_t(x) - \bar{x}| \leq e^{-ct}|x - \bar{x}|$$

for all $x \in U$, $t \geq 0$.

(c) *For any norm on $\mathbf{R}^n$, there is a constant $B > 0$ such that*

$$|\phi_t(x) - \bar{x}| \leq Be^{-ct}|x - \bar{x}|$$

for all $x \in U$, $t \geq 0$.

In particular, $\phi_t(x) \to \bar{x}$ as $t \to \infty$ for all $x \in U$.

Proof. For convenience we assume $\bar{x} = 0$. (If not, give $\mathbf{R}^n$ new coordinates $y = x - \bar{x}$; in y-coordinates f has an equilibrium at 0; etc.)

Put $A = Df(0)$. Choose $b > 0$ so that the real parts of eigenvalues of A are less than $-b < -c$. The lemma in Chapter 7, Section 1 shows that $\mathbf{R}^n$ has a basis $\mathcal{B}$ whose corresponding norm and inner product satisfy

$$\langle Ax, x \rangle \leq -b|x|^2$$

for all $x \in \mathbf{R}^n$.

Since $A = Df(0)$ and $f(0) = 0$, by the definition of derivative,

$$\lim_{x \to 0} \frac{|f(x) - Ax|}{|x|} = 0.$$

Therefore by Cauchy's inequality,

$$\lim_{x \to 0} \frac{\langle f(x) - Ax, x \rangle}{|x|^2} = 0.$$

It follows that there exists $\delta > 0$ so small that if $|x| \leq \delta$, then $x \in W$ and

$$\langle f(x), x \rangle \leq -c|x|^2.$$

Put $U = \{x \in \mathbf{R}^n \mid |x| \leq \delta\}$. Let $x(t)$, $0 \leq t \leq t_0$, be a solution curve in U, $x(t) \neq 0$. Then

$$\frac{d}{dt}|x| = \frac{1}{|x|}\langle x', x \rangle.$$

Hence, since $x' = f(x)$:

$$\frac{d}{dt}|x| \leq -c|x|. \tag{2}$$

This shows, first, that $|x(t)|$ is decreasing; hence $|x(t)| \in U$ for all $t \in [0, t_0]$. Since U is compact, it follows from Section 5, Chapter 8 that the trajectory $x(t)$ is defined and in U for all $t \geq 0$. Secondly, (2) implies that

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

for all $t \geq 0$. Thus (a) and (b) are proved and (c) follows from equivalence of norms.

The phase portrait at a nonlinear sink $\bar{x}$ looks like that of the linear part of the vector field: in a suitable norm the trajectories point inside all sufficiently small spheres about $\bar{x}$ (Fig. A).

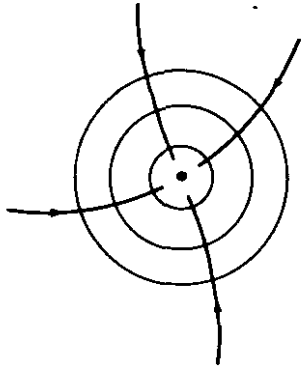


FIG. A. Nonlinear sink.

Remember that the spheres are not necessarily "round" spheres; they are spheres in a special norm. In standard coordinates they may be ellipsoids.

A simple physical example of a nonlinear sink is given by a pendulum moving in a vertical plane (Fig. B). We assume a constant downward gravitational force equal to the mass m of the bob; we neglect the mass of the rod supporting the bob. We assume there is a frictional (or viscous) force resisting the motion, proportional to the speed of the bob.

Let l be the (constant) length of the rod. The bob of the pendulum moves along a circle of radius l. If $\theta(t)$ is the counterclockwise angle from the vertical to the rod at time t, then the angular velocity of the bob is $d\theta/dt$ and the velocity is $l\, d\theta/dt$. Therefore the frictional force is $-kl\, d\theta/dt$, k a nonnegative constant; this force is tangent to the circle.

The downward gravitational force m has component $-m \sin \theta(t)$ tangent to the circle; this is the force on the bob that produces motion. Therefore the total force tangent to the circle at time t is

$$F = -\left(kl\frac{d\theta}{dt} + m \sin \theta\right).$$

The acceleration of the bob tangent to the circle is

$$a = l\frac{d^2\theta}{dt^2};$$

hence, from Newton's law $a = F/m$, we have

$$l\theta'' = -\frac{kl}{m}\theta' - \sin \theta$$

or

$$\theta'' = -\frac{k}{m}\theta' - \frac{1}{l}\sin \theta.$$

Introducing a new variable

$$\omega = \theta'$$

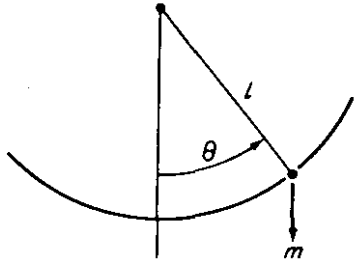


FIG. B. Pendulum.

(interpreted as angular velocity), we obtain the equivalent first order system

$$\theta' = \omega, \tag{3}$$
$$\omega' = -\frac{1}{l}\sin\theta - \frac{k}{m}\omega.$$

This nonlinear, autonomous equation in $\mathbf{R}^2$ has equilibria at the points

$$(\theta, \omega) = (n\pi, 0); \qquad n = 0, \pm 1, \pm 2, \ldots.$$

We concentrate on the equilibrium $(0, 0)$.

The vector field defining (3) is

$$f(\theta, \omega) = \left(\omega, -\frac{1}{l}\sin\theta - \frac{k}{m}\omega\right).$$

Its derivative at (θ, ω) is

$$Df(\theta, \omega) = \begin{bmatrix} 0 & 1 \\ -\frac{1}{l}\cos\theta & -\frac{k}{m} \end{bmatrix}.$$

Hence

$$Df(0, 0) = \begin{bmatrix} 0 & 1 \\ -\frac{1}{l} & -\frac{k}{m} \end{bmatrix}$$

with eigenvalues

$$\frac{1}{2}\left\{-\frac{k}{m} \pm \left[\left(\frac{k}{m}\right)^2 - \frac{4}{l}\right]^{1/2}\right\}.$$

The real part $-k/2m$ is negative as long as the coefficient of friction k is positive and the mass is positive. Therefore the equilibrium $\theta = \omega = 0$ is a sink. We conclude: for all sufficiently small initial angles and velocities, the pendulum tends toward the equilibrium position $(0, 0)$.

This, of course, is not surprising. In fact, from experience it seems obvious that from any initial position and velocity the pendulum will tend toward the downward equilibrium state, except for a few starting states which tend toward the vertically balanced position. To verify this physical conclusion mathematically takes more work, however. We return to this question in Section 3.

Before leaving the pendulum we point out a paradox: *the pendulum cannot come to rest.* That is, once it is in motion—not in equilibrium—it cannot reach an equilibrium state, but only approach one arbitrarily closely. This follows from uniqueness of solutions of differential equations! Of course, one knows that pendulums actually do come to rest. One can argue that the pendulum is not "really" at rest,

but its motion is too small to observe. A better explanation is that the mathematical model (3) of its motion is only an approximation to reality.

PROBLEMS

1. (a) State and prove a converse to the theorem of Section 1.
   (b) Define "sources" for nonlinear vector fields and prove an interesting theorem about them.
2. Show by example that if f is a nonlinear C^1 vector field and $f(0) = 0$, it is possible that $\lim_{t\to\infty} x(t) = 0$ for all solutions to $x' = f(x)$, without the eigenvalues of $Df(0)$ having negative real parts.
3. Assume f is a C^1 vector field on $\mathbf{R}^n$ and $f(0) = 0$. Suppose some eigenvalue of $Df(0)$ has positive real part. Show that in every neighborhood of 0 there is a solution $x(t)$ for which $|\, x(t)\,|$ is increasing on some interval $[0, t_0]$, $t_0 > 0$.
4. If $\bar{x}$ is a sink of a dynamical system, it has a neighborhood containing no other equilibrium.

§2. Stability

The study of equilibria plays a central role in ordinary differential equations and their applications. An equilibrium point, however, must satisfy a certain stability criterion in order to be very significant physically. (Here, as in several other places in this book, we use the word *physical* in a broad sense; thus, in some contexts, *physical* could be replaced by *biological*, *chemical*, or even *ecological*.)

The notion of stability most often considered is that usually attributed to Liapunov. An equilibrium is *stable* if nearby solutions stay nearby for all future time. Since in applications of dynamical systems one cannot pinpoint a state exactly, but only approximately, an equilibrium must be stable to be physically meaningful.

The mathematical definition is:

Definition 1 Suppose $\bar{x} \in W$ is an equilibrium of the differential equation

$$x' = f(x), \tag{1}$$

where $f\colon W \to E$ is a C^1 map from an open set W of the vector space E into E. Then $\bar{x}$ is a *stable* equilibrium if for every neighborhood U of $\bar{x}$ in W there is a neighborhood U_1 of $\bar{x}$ in U such that every solution $x(t)$ with $x(0)$ in U_1 is defined and in U for all $t > 0$. (See Fig. A.)

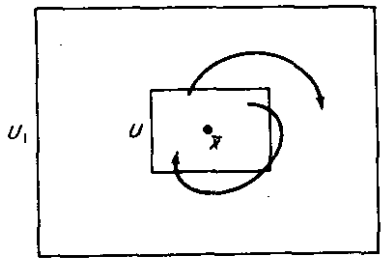


FIG. A. Stability.

Definition 2 If U_1 can be chosen so that in addition to the properties described in Definition 1, $\lim_{t\to\infty} x(t) = \bar{x}$, then $\bar{x}$ is *asymptotically stable*. (See Fig. B.)

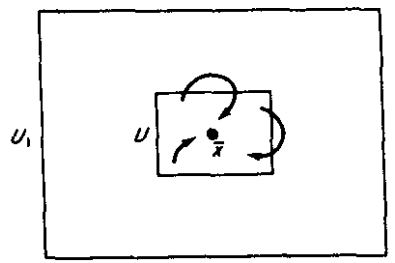


FIG. B. Asymptotic stability.

Definition 3 An equilibrium $\bar{x}$ that is not stable is called *unstable*. This means there is a neighborhood U of $\bar{x}$ such that for *every* neighborhood U_1 of $\bar{x}$ in U, there is at least one solution $x(t)$ starting at $x(0) \in U_1$, which does not lie entirely in U. (See Fig. C.)

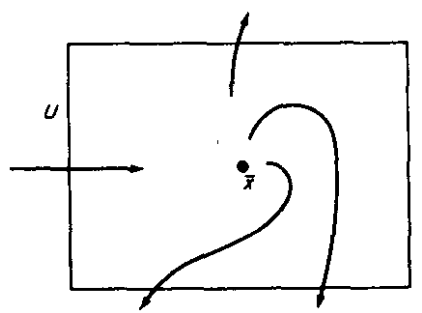


FIG. C. Instability.

A sink is asymptotically stable and therefore stable. An example of an equilibrium that is stable but not asymptotically stable is the origin in $\mathbf{R}^2$ for a linear

equation

$$x' = Ax, \tag{2}$$

where A has pure imaginary eigenvalues. The orbits are all ellipses (Fig. D).

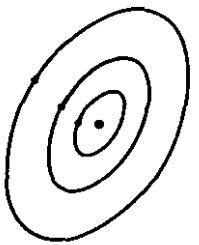


FIG. D. Stable, but not asymptotically stable.

The importance of this example in application is limited (despite the famed harmonic oscillator) because the slightest nonlinear perturbation will destroy its character. Even a small linear perturbation can make it into a sink or a source since "hyperbolicity" is a generic property for linear flows (see Chapter 7).

A source is an example of an unstable equilibrium.

To complement the main theorem of Section 2 we have the following instability theorem. The proof is not essential to the rest of the book.

Theorem *Let $W \subset E$ be open and $f: W \to E$ continuously differentiable. Suppose $f(\bar{x}) = 0$ and $\bar{x}$ is a stable equilibrium point of the equation*

$$x' = f(x).$$

Then no eigenvalue of $Df(\bar{x})$ has positive real part.

We say that an equilibrium $\bar{x}$ is *hyperbolic* if the derivative $Df(\bar{x})$ has no eigenvalue with real part zero.

Corollary *A hyperbolic equilibrium point is either unstable or asymptotically stable.*

Proof of the theorem. Suppose some eigenvalue has positive real part; we shall prove $\bar{x}$ is not stable. We may assume $\bar{x} = 0$, replacing $f(x)$ by $f(x - \bar{x})$ otherwise. By the canonical form theorem (Chapter 2), E has a splitting $E_1 \oplus E_2$ invariant under $Df(0)$, such that eigenvalues of $A = Df(0)\,|\,E_1$ all have positive real part, while those of $B = Df(0)\,|\,E_2$ all have negative or 0 real part.

Let $a > 0$ be such that every eigenvalue of A has real part $> a$. Then there is a Euclidean norm on E_1 such that

$$\langle Ax, x\rangle \geq a\,|\,x\,|^2, \qquad \text{all} \quad x \in E_1. \tag{3}$$

Similarly, for any $b > 0$ there exists a Euclidean norm on E_2 such that

$$\langle By, y\rangle < b \mid y \mid^2, \qquad \text{all} \quad y \in E_2. \tag{4}$$

We choose b so that

$$0 < b < a.$$

We take the inner product on $E = E_1 \oplus E_2$ to be the direct sum of these inner products on E_1 and E_2; we also use the norms associated to these inner products on E_1, E_2, E. If $z = (x, y) \in E_1 \oplus E_2$, then $\mid z \mid = (\mid x \mid^2 + \mid y \mid^2)^{1/2}$.

We shall use the Taylor expansion of f around 0:

$$f(x, y) = (Ax + R(x, y), By + S(x, y)) = (f_1(x, y), f_2(x, y))$$

with

$$(x, y) = z; \qquad (R(x, y), S(x, y)) = Q(z).$$

Thus, given any $\epsilon > 0$, there exists $\delta > 0$ such that if $U = B_\delta(0)$ (the ball of radius δ about 0),

$$\mid Q(z) \mid \leq \epsilon \mid z \mid \qquad \text{for} \quad z \in U. \tag{5}$$

We define the cone $C = \{(x, y) \in E_1 \oplus E_2 \mid \mid x \mid \geq \mid y \mid\}$.

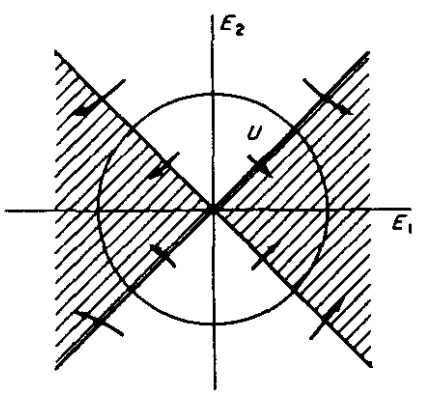


FIG. E. The cone C is shaded.

Lemma *There exists $\delta > 0$ such that if U is the closed ball $B_\delta(0) \subset W$, then for all $z = (x, y) \in C \cap U$,*

(a) $\langle x, f_1(x, y)\rangle - \langle y, f_2(x, y)\rangle > 0$ *if* $x \neq 0$, *and*
(b) *there exists* $\alpha > 0$ *with* $\langle f(z), z\rangle \geq \alpha \mid z \mid^2$.

This lemma yields our instability theorem as follows. We interpret first condition (a). Let $g: E_1 \times E_2 \to \mathbf{R}$ be defined by $g(x, y) = \frac{1}{2}(\mid x \mid^2 - \mid y \mid^2)$. Then g is C^1, $g^{-1}[0, \infty) = C$, and $g^{-1}(0)$ is the boundary of C.

Furthermore, if $(x, y) = z \in U$, then $Dg(z)(f(z)) = Dg(x, y)(f_1(x, y), f_2(x, y)) = \langle x, f_1(x, y)\rangle - \langle y_1, f_2(x, y)\rangle$ which will be positive if $z \in g^{-1}(0)$ by (a). This implies that on a solution $z(t)$ in U passing through the boundary of C, g is *increasing* since by the chain rule, $(d/dt)(g(z(t)) = Dg(z(t))f(z(t))$. Therefore *no solution which starts in C can leave C before it leaves U.* Figure E gives the idea.

Geometrically (b) implies that each vector $f(z)$ at $z \in C$ points outward from the sphere about 0 passing through z. See Fig. F.

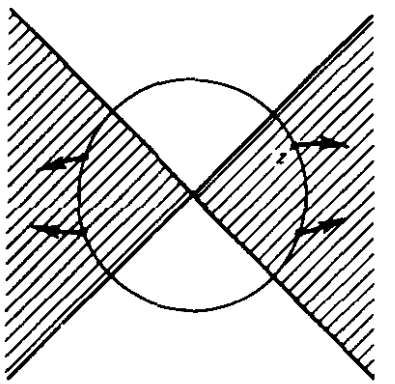


FIG. F

Condition (b) has the following quantitative implication. If $z = z(t)$ is a solution curve in $C \cap U$, then

$$\langle f(z), z\rangle = \langle z', z\rangle = \frac{1}{2}\frac{d}{dt} \mid z \mid^2,$$

so (b) implies

$$\frac{1}{2}\frac{d}{dt} \mid z \mid^2 \geq \alpha \mid z \mid^2$$

or

$$\frac{d/dt \mid z \mid^2}{\mid z \mid^2} \geq 2\alpha,$$

$$\frac{d}{dt} \log \mid z \mid^2 \geq 2\alpha,$$

$$\log \mid z(t) \mid^2 \geq 2\alpha t + \log \mid z(0) \mid^2,$$

$$\mid z(t) \mid^2 \geq e^{2\alpha t} \mid z(0) \mid^2;$$

thus

$$\mid z(t) \mid \geq e^{\alpha t} \mid z(0) \mid.$$

Thus each nontrivial solution $z(t)$ starting in $C \cap U$ moves away from 0 at an exponential rate as long as it is defined and in $C \cap U$.

If $y(t)$ is not defined for all $t \geq 0$, then, by Chapter 8, Section 5, it must leave the compact set $C \cap U$; as we have seen above, it must therefore leave U. On the *other hand, if $y(t)$ is* defined for all t, it must also leave U since U is the ball of radius δ and $e^{\alpha t} \mid z(0) \mid > \delta$ for large t. Therefore there are solutions starting arbitrarily close to 0 and leaving U. Thus (assuming the truth of the lemma), the vector field f does not have 0 as a point of stable equilibrium.

We now give the proof of the lemma. First, part (b): if $(x, y) = z \in C \cap U$,

$$\langle f(z), z \rangle = \langle Ax, x \rangle + \langle By, y \rangle + \langle Q(z), z \rangle,$$

so, by (3), (4), (5):

$$\langle f(z), z \rangle \geq a \mid x \mid^2 - b \mid y \mid^2 - \epsilon \mid z \mid^2.$$

In C, $\mid x \mid \geq \mid y \mid$ and $\mid x \mid^2 \geq \frac{1}{2}(\mid x \mid^2 + \mid y \mid^2) \geq \frac{1}{2} \mid z \mid^2$. Thus $\langle f(z), z \rangle \geq (a/2 - b/2 - \epsilon) \mid z \mid^2$. We choose $\epsilon > 0$ and then $\delta > 0$ so that $\alpha = a/2 - b/2 - \epsilon > 0$. This proves (b).

To check (a), note that the left-hand side of (a) *is*

$$\langle Ax, x \rangle - \langle By, y \rangle + \langle x, R(x, y) \rangle - \langle y, S(x, y) \rangle,$$

but

$$\mid \langle x, R(x, y) \rangle - \langle y, S(x, y) \rangle \mid \leq 2 \mid \langle z, Q(z) \rangle \mid.$$

We may proceed just as in the previous part; finally, $\delta > 0$ is chosen so that $a/2 - b/2 - 2\epsilon > 0$. This yields the proposition.

In Chapter 7 we introduced hyperbolic linear flows. The nonlinear analogue *is* a hyperbolic equilibrium point $\bar{x}$ of a dynamical system $x' = f(x)$; and to repeat, this means that the eigenvalues of $Df(\bar{x})$ have nonzero real parts. If these real parts are all negative, $\bar{x}$ is, of course, a sink; if they are all positive, $\bar{x}$ is called a *source*. If both signs occur, $\bar{x}$ is a *saddle point*. From the preceding theorem we see that a *saddle point is unstable*.

If $\bar{x}$ is an asymptotic equilibrium of a dynamical system, by definition there is a neighborhood N of $\bar{x}$ such that any solution curve starting in N tends toward $\bar{x}$. The union of all solution curves that tend toward x (as $t \to \infty$) *is* called the *basin* of $\bar{x}$, denoted by $B(\bar{x})$.

It is clear that any solution curve which meets N is in $B(\bar{x})$; and, conversely, any solution curve in $B(\bar{x})$ must meet N. It follows that $B(\bar{x})$ *is an open set*; for, by continuity of the flow, *if the trajectory of x meets N*, the trajectory of any nearby point also meets N.

Notice that $B(\bar{x})$ and $B(\bar{y})$ are disjoint if $\bar{x}$ and $\bar{y}$ are different asymptotically stable equilibria. For if a trajectory tends toward $\bar{x}$, it cannot also tend toward $\bar{y}$.

If a dynamical system represents a physical system, one can practically identify the states in $B(\bar{x})$ with $\bar{x}$. For every state in $B(\bar{x})$ will, after a period of transition, stay so close to $\bar{x}$ as to be indistinguishable from it. For some frequently occurring

types of dynamical systems (the gradient systems of Section 4), almost every state is in the basin of some sink; other states are "improbable" (they constitute a set of measure 0). For such a system, the sinks represent the different types of long term behavior.

It is often a matter of practical importance to determine the basin of a sink $\bar{x}$. For example, suppose $\bar{x}$ represents some desired equilibrium state of a physical system. The extent of the basin tells us how large a perturbation from equilibrium we can allow and still be sure that the system will return to equilibrium.

We conclude this section by remarking that James Clerk Maxwell applied stability theory to the study of the rings of the planet Saturn. He decided that they must be composed of many small separate bodies, rather than being solid or fluid, for only in the former case are there *stable* solutions of the equations of motion. He discovered that while solid or fluid rings were mathematically possible, the slightest perturbation would destroy their configuration.

PROBLEMS

1. (a) Let $\bar{x}$ be a stable equilibrium of a dynamical system corresponding to a C^1 vector field on an open set $W \subset E$. Show that for every neighborhood U of $\bar{x}$ in W, there is a neighborhood U' of $\bar{x}$ in U such that every solution curve $x(t)$ with $x(0) \in U'$ is defined and in U' for all $t > 0$.
   (b) If $\bar{x}$ is asymptotically stable, the neighborhood U' in (a) can be chosen to have the additional property that $\lim_{t \to \infty} x(t) = \bar{x}$ if $x(0) \in U'$.
   (*Hint*: Consider the set of all points of U whose trajectories for $t \geq 0$ enter the set U_1 in Definition 1 or 2.)

2. For which of the following linear operators A on $\mathbf{R}^n$ is $0 \in \mathbf{R}^n$ a stable equilibrium of $x' = Ax$?

   (a) $A = O$ (b) $\begin{bmatrix} 0 & -1 & & \\ 1 & 0 & & \\ & & 0 & 1 \\ & & -1 & 0 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}$

   (d) $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ (e) $\begin{bmatrix} 1 & 2 \\ 2 & -2 \end{bmatrix}$

3. Let A be a linear operator on $\mathbf{R}^n$ all of whose eigenvalues have real part 0. Then $0 \in \mathbf{R}^n$ is a stable equilibrium of $x' = Ax$ if and only if A is semisimple; and 0 is never asymptotically stable.

4. Show that the dynamical system in $\mathbf{R}^2$, where equations in polar coordinates are

$$\theta' = 1, \quad r' = \begin{cases} r^2 \sin(1/r), & r > 0, \\ 0, & r = 0, \end{cases}$$

has a stable equilibrium at the origin. (*Hint*: Every neighborhood of the origin contains a solution curve encircling the origin.)

5. Let $f: \mathbf{R}^n \to \mathbf{R}^n$ be C^1 and suppose $f(0) = 0$. If some eigenvalue of $Df(0)$ has positive real part, there is a nonzero solution $x(t)$, $-\infty < t \leq 0$, to $x' = f(x)$, such that $\lim_{t \to -\infty} x(t) = 0$. (*Hint*: Use the instability theorem of Section 3 to find a sequence of solutions $x_n(t)$, $t_n \leq t \leq 0$, in $B_\delta(0)$ with $|x_n(0)| = \delta$ and $\lim_{n \to \infty} x_n(t_n) = 0$.)

6. Let $g: \mathbf{R}^n \to \mathbf{R}^n$ be C^1 and suppose $f(0) = 0$. If some eigenvalue of $Dg(0)$ has negative real part, there is a solution $g(t)$, $0 \leq t < \infty$, to $x' = g(x)$, such that $\lim_{t \to \infty} g(t) = 0$. (*Hint*: Compare previous problem.)

§3. Liapunov Functions

In Section 2 we defined stability and asymptotic stability of an equilibrium $\bar{x}$ of a dynamical system

$$x' = f(x), \tag{1}$$

where $f: W \to \mathbf{R}^n$ is a C^1 map on an open set $W \subset \mathbf{R}^n$. If $\bar{x}$ is a sink, stability can be detected by examining the eigenvalues of the linear part $Df(\bar{x})$. Other than that, however, as yet we have no way of determining stability except by actually finding all solutions to (1), which may be difficult if not impossible.

The Russian mathematician and engineer A. M. Liapunov, in his 1892 doctoral thesis, found a very useful criterion for stability. It is a generalization of the idea that for a sink there is a norm on $\mathbf{R}^n$ such that $|x(t) - \bar{x}|$ decreases for solutions $x(t)$ near $\bar{x}$. Liapunov showed that certain other functions could be used instead of the norm to guarantee stability.

Let $V: U \to R$ be a differentiable function defined in a neighborhood $U \subset W$ of $\bar{x}$. We denote by $\dot{V}: U \to R$ the function defined by

$$\dot{V}(x) = DV(x)(f(x)).$$

Here the right-hand side is simply the operator $DV(x)$ applied to the vector $f(x)$. Then if $\phi_t(x)$ is the solution to (1) passing through x when $t = 0$,

$$\dot{V}(x) = \frac{d}{dt} V(\phi_t x) \Big|_{t=0}$$

by the chain rule. Consequently, if $\dot{V}(x)$ is negative, then V decreases along the solution of (1) through x.

We can now state Liapunov's stability theorem:

Theorem 1 *Let $\bar{x} \in W$ be an equilibrium for (1). Let $V: U \to \mathbf{R}$ be a continuous function defined on a neighborhood $U \subset W$ of x, differentiable on $U - \bar{x}$, such that*

(a) *$V(\bar{x}) = 0$ and $V(x) > 0$ if $x \neq \bar{x}$;*
(b) *$\dot{V} \leq 0$ in $U - \bar{x}$.*

Then $\bar{x}$ is stable. Furthermore, if also

(c) *$\dot{V} < 0$ in $U - \bar{x}$,*

then $\bar{x}$ is asymptotically stable.

A function V satisfying (a) and (b) is called a *Liapunov function* for $\bar{x}$. If (c) also holds, we call V a *strict* Liapunov function. The only equilibrium is the origin $x = y = 0$.

We emphasize that Liapunov's theorem can be applied without solving the differential equation. On the other hand, there is no cut-and-dried method of finding Liapunov functions; it is a matter of ingenuity and trial and error in each case. Sometimes there are natural functions to try. In the case of mechanical or electrical systems, energy is often a Liapunov function.

Example 1 Consider the dynamical system on $\mathbf{R}^3$ described by the system of differential equations

$$\begin{aligned} x' &= 2y(z-1), \\ y' &= -x(z-1), \\ z' &= -z^3. \end{aligned}$$

The z-axis $(= \{(x, y, z) \mid x = y = 0\})$ consists entirely of equilibrium points. Let us investigate the origin for stability.

The linear part of the system at $(0, 0, 0)$ is the matrix

$$\begin{bmatrix} 0 & -2 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

There are two imaginary eigenvalues and one zero eigenvalue. All we can conclude from this is that the origin is not a sink.

Let us look for a Liapunov function for $(0, 0, 0)$ of the form $V(x, y, z) = ax^2 + by^2 + cz^2$, with $a, b, c > 0$. For such a V,

$$\dot{V} = 2(axx' + byy' + czz');$$

so

$$\tfrac{1}{2}\dot{V} = 2axy(z-1) - bxy(z-1) - cz^4.$$

We want $V \leq 0$; this can be accomplished by setting $c = 1$ and $2a = b$. We conclude that $x^2 + 2y^2 + z^2$ is a Liapunov function; therefore the origin is a stable equilibrium. Moreover, the origin is asymptotically stable, since our Liapunov function V is clearly *strict*, that is, it satisfies (c) of p. 193.

Example 2 Consider a (constant) mass m moving under the influence of a conservative force field $-\text{grad } \Phi(x)$ defined by a potential function Φ: $W_0 \to \mathbf{R}$ on an open set $W_0 \subset \mathbf{R}^3$. (See Chapter 2.) The corresponding dynamical system on the state space $W = W_0 \times \mathbf{R}^3 \subset \mathbf{R}^3 \times \mathbf{R}^3$ is, for $(x, v) \in W_0 \times \mathbf{R}^3$:

$$\frac{dx}{dt} = v,$$

$$\frac{dv}{dt} = -\text{grad } \Phi(x).$$

Let $(\bar{x}, \bar{v}) \in W_0 \times \mathbf{R}^3$ be an equilibrium point. Then $\bar{v} = 0$ and grad $\Phi(\bar{x}) = 0$. To investigate stability at $(\bar{x}, 0)$, we try to use the total energy

$$E(x, v) = \tfrac{1}{2}m\,|v|^2 + m\Phi(x)$$

to construct a Liapunov function. Since a Liapunov function must vanish at $(\bar{x}, 0)$, we subtract from $E(x, v)$ the energy of the state $(\bar{x}, 0)$, which is $\Phi(\bar{x})$, and define V: $W_0 \times \mathbf{R}^3 \to \mathbf{R}$ by

$$\begin{aligned} V(x, v) &= E(x, v) - E(\bar{x}, 0) \\ &= \tfrac{1}{2}m\,|v|^2 + m\Phi(x) - m\Phi(\bar{x}). \end{aligned}$$

By conservation of energy, $\dot{V} \equiv 0$. Since $\frac{1}{2}mv^2 \geq 0$, we assume $\Phi(x) > \Phi(\bar{x})$ for x near $\bar{x}$, $x \neq \bar{x}$, in order to make V a Liapunov function. Therefore we have proved the well-known theorem of Lagrange: *an equilibrium $(\bar{x}, 0)$ of a conservative force field is stable if the potential energy has a local absolute minimum at $\bar{x}$.*

Proof of Liapunov's theorem. Let $\delta > 0$ be so small that the closed ball $B_\delta(\bar{x})$ around $\bar{x}$ of radius δ lies entirely in U. Let α be the minimum value of V on the boundary of $B_\delta(\bar{x})$, that is, on the sphere $S_\delta(\bar{x})$ of radius δ and center $\bar{x}$. Then $\alpha > 0$ by (a). Let $U_1 = \{x \in B_\delta(\bar{x}) \mid V(x) < \alpha\}$. Then no solution starting in U_1 can meet $S_\delta(\bar{x})$ since V is nonincreasing on solution curves. Hence every solution starting in U_1 never leaves $B_\delta(\bar{x})$. This proves $\bar{x}$ is stable. Now assume (c) holds as well, so that V is strictly decreasing on orbits in $U - \bar{x}$. Let $x(t)$ be a solution starting in $U_1 - \bar{x}$ and suppose $x(t_n) \to z_0 \in B_\delta(\bar{x})$ for some sequence $t_n \to \infty$; such a sequence exists by compactness of $B_\delta(\bar{x})$. We assert $z_0 = \bar{x}$. To see this, observe that $V(x(t)) > V(z_0)$ for all $t \geq 0$ since $V(x(t))$ decreases and $V(x(t_n)) \to V(z_0)$ by continuity of V. If $z_0 \neq \bar{x}$, let $z(t)$ be the solution starting at z_0. For any $s > 0$, we have $V(z(s)) < V(z_0)$. Hence for any solution $y(s)$ starting

sufficiently near z_0 we have

$$V(y(s)) < V(z_0);$$

putting $y(0) = x(t_n)$ for sufficiently large n yields the contradiction

$$V(x(t_n + s)) < V(z_0).$$

Therefore $z_0 = \bar{x}$. This proves that $\bar{x}$ is the only possible limit point of the set $\{x(t) \mid t \geq 0\}$. This completes the proof of Liapunov's theorem.

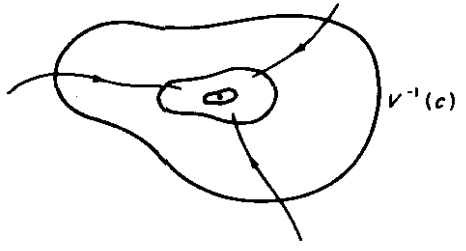


FIG. A. Level surfaces of a Liapunov function.

Figure A makes the theorem intuitively obvious. The condition $\dot{V} \leq 0$ means that when a trajectory crosses a "level surface" $V^{-1}(c)$, it moves inside the set where $V \leq c$ and can never come out again. Unfortunately, it is difficult to justify the diagram; why should the sets $V^{-1}(c)$ shrink down to $\bar{x}$? Of course, in many cases, Fig. A is indeed correct; for example, if V is a positive definite quadratic form, such as $x^2 + 2y^2$. But what if the level surfaces look like Fig. B? It is hard to imagine such a V that fulfills all the requirements of a Liapunov function; but rather than trying to rule out that possibility, it is simpler to give the analytic proof as above.

Liapunov functions not only detect stable equilibria; they can be used to estimate the extent of the basin of an asymptotically stable equilibrium, as the following theorem shows. In order to state it, we make two definitions. A set P is *positively invariant* for a dynamical system if for each x in P, $\phi_t(x)$ is defined and in P for all $t \geq 0$ (where ϕ denotes the flow of the system). An *entire orbit* of the system is a

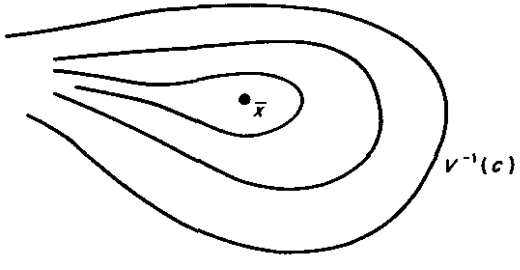


FIG. B. Level surfaces of a Liapunov function?

set of the form

$$\{\phi_t(x) \mid t \in \mathbf{R}\},$$

where $\phi_t(x)$ is defined for all $t \in \mathbf{R}$.

Theorem 2 *Let $\bar{x} \in W$ be an equilibrium of the dynamical system (1) and let $V: U \to \mathbf{R}$ be a Liapunov function for $\bar{x}$, U a neighborhood of $\bar{x}$. Let $P \subset U$ be a neighborhood of $\bar{x}$ which is closed in W. Suppose that P is positively invariant, and that there is no entire orbit in $P - \bar{x}$ on which V is constant. Then $\bar{x}$ is asymptotically stable, and $P \subset B(\bar{x})$.*

Before proving Theorem 2 we apply it to the equilibrium $\bar{x} = (0, 0)$ of the pendulum discussed in Section 1. For a Liapunov function we try the total energy E, which we expect to decrease along trajectories because of friction. Now

$$E = \text{kinetic energy} + \text{potential energy};$$

$$\text{kinetic energy} = \tfrac{1}{2}mv^2 = \tfrac{1}{2}m(l\dot{\theta})^2 = \tfrac{1}{2}ml^2\omega^2.$$

For potential energy we take mass times height above the lowest point of the circle:

$$\text{potential energy} = m(l - l\cos\theta).$$

Thus

$$E = \tfrac{1}{2}ml^2\omega^2 + ml(1 - \cos\theta) = ml(\tfrac{1}{2}l\omega^2 + 1 - \cos\theta).$$

Then

$$\dot{E} = ml(l\omega' + \theta' \sin\theta);$$

using (3) of Section 1 this simplifies to

$$\dot{E} = -kl^2\omega^2.$$

Thus $\dot{E} \leq 0$ and $E(0, 0) = 0$, so that E is indeed a Liapunov function.

To estimate the basin of $(0, 0)$, fix a number c, $0 < c < 2ml$, and define

$$P_c = \{(\theta, \omega) \mid E(\theta, \omega) \leq c \quad \text{and} \quad |\theta| < \pi\}.$$

Clearly, $(0, 0) \in P_c$. We shall prove $P_c \subset B(0, 0)$.

P_c is positively invariant. For suppose

$$(\theta(t), \omega(t)), \qquad 0 \leq t \leq \alpha, \quad \alpha > 0$$

is a trajectory with $(\theta(0), \omega(0)) \in P_c$. To see that $(\theta(\alpha), \omega(\alpha)) \in P_c$, observe that $E(\theta(\alpha), \omega(\alpha)) \leq c$ since $\dot{E} \leq 0$. If $|\theta(\alpha)| \geq \pi$, there must exist a smallest

$t_0 \in [0, \alpha]$ such that $\theta(t_0) = \pm\pi$. Then

$$E(\theta(t_0), \omega(t_0)) = E(\pm\pi, \omega(t_0)) = ml[\tfrac{1}{2}l\omega(t_0)^2 + 2] \geq 2ml.$$

But

$$E(\theta(t_0)), \omega(t_0)) \leq c < 2ml.$$

This contradiction shows that $\theta(\alpha) < \pi$, and so P_c is positively invariant.

We assert that P_c fulfills the second condition of Theorem 2. For suppose E is constant on a trajectory. Then, along that trajectory, $\dot{E} = 0$ and so $\omega = 0$. Hence, from (3) of Section 1, $\theta' = 0$ so θ is constant on the orbit and also $\sin\theta = 0$. Since $|\theta| < \pi$, it follows that $\theta = 0$. Thus the only entire orbit in P_c on which E is constant is the equilibrium orbit $(0, 0)$.

Finally, P_c is a closed set. For if (θ_0, ω_0) is a limit point of P_c, then $|\theta_0| \leq \pi$, and $E(\theta_0, \omega_0) \leq c$ by continuity of E. But $|\theta_0| = \pi$ implies $E(\theta_0, \omega_0) > c$. Hence $|\theta_0| < \pi$ and so $(\theta_0, \omega_0) \in P_c$.

From Theorem 2 we conclude that each $P_c \subset B(0, 0)$; hence the set

$$P = \bigcup\{P_c \mid 0 < c < 2ml\}$$

is contained in $B(0, 0)$. Note that

$$P = \{(\theta, \omega) \mid E(\theta, \omega) < 2ml \quad \text{and} \quad |\theta| < \pi\}.$$

This result is quite natural on physical grounds. For $2ml$ is the total energy of the state $(\pi, 0)$ where the bob of the pendulum is balanced above the pivot. Thus if the pendulum is not pointing straight up, *and* the total energy is less than the total energy of the balanced upward state, then the pendulum will gradually approach the state $(0, 0)$.

There will be other states in the basin of $(0, 0)$ that are not in the set P. Consider a state (π, u), where u *is very small but not zero*. Then $(\pi, u) \notin P$, but the pendulum moves immediately into a state in P, and therefore approaches $(0, 0)$. Hence $(\pi, u) \in B(0, 0)$. See Exercises 5 and 6 for other examples.

Proof of Theorem 2. Imagine a trajectory $x(t)$, $0 \leq t < \infty$, in the positively invariant set P. Suppose $x(t)$ does *not* tend to $\bar{x}$ as $t \to \infty$. Then there must be a point $a \neq \bar{x}$ in P and a sequence $t_n \to \infty$ such that

$$\lim_{n\to\infty} x(t_n) = a.$$

If $\alpha = V(a)$, then α is the greatest lower bound of $\{V(x(t)) \mid t \geq 0\}$; this follows from continuity of V and the fact that V decreases along trajectories.

Let L be the set of all such points a in W:

$$L = \{a \in W \mid \text{there exist } t_n \to \infty \text{ with } x(t_n) \to a\},$$

where $x(t)$ is the trajectory postulated above. Since every point of L is a limit of points in P, and P is closed in W, it follows that $L \subset P$. Moreover, if $a \in L$, then the entire orbit of a is in L; that is, $\phi_t(a)$ is defined and in L for all $t \in \mathbf{R}$. For $\phi_t(a)$ is defined for all $t \geq 0$ since P is positively invariant. On the other hand, each point $\phi_t(x(t_n))$ is defined for all t in the interval $[-t_n, 0]$; since $x(t_n) \to a$ and we may assume $t_1 < t_2 < \cdots$, it follows from Chapter 8 that $\phi_t(a)$ is defined for all $t \in [-t_n, 0]$, $n = 1, 2, \ldots$. Since $-t_n \to -\infty$, $\phi_t(a)$ is defined for all $t \leq 0$. To see that $\phi_s(a) \in L$, for any particular $s \in \mathbf{R}$, note that if $x(t_n) \to a$, then $x(t_n + s) \to \phi_s(a)$.

We reach a contradiction, for $V(a) = a$ for all $a \in L$; hence V is constant on an entire orbit in P. This is impossible; hence $\lim_{t\to\infty} x(t) = \bar{x}$ for all trajectories in P. This proves that $\bar{x}$ is asymptotically stable, and also that $P \subset B(\bar{x})$. This completes the proof of Theorem 2.

The set L defined above is called the set of *ω-limit points*, or the *ω-limit set*, of the trajectory $x(t)$ (or of any point on the trajectory). Similarly, we define the set of *α-limit points*, or the *α-limit set*, of a trajectory $y(t)$ to be the set of all points b such that $\lim_{n\to\infty} y(t_n) = b$ for some sequence $t_n \to -\infty$. (The reason, such as it is, for this terminology is that α is the first letter and ω the last letter of the Greek alphabet.) We will make extensive use of these concepts in Chapter 11.

A set A in the domain W of a dynamical system is *invariant* if for every $x \in A$, $\phi_t(x)$ is defined and in A for all $t \in \mathbf{R}$. The following facts, essentially proved in the proof of Theorem 2, will be used in Chapter 11.

Proposition *The α-limit set and the ω-limit set of a trajectory which is defined for all $t \in \mathbf{R}$ are closed invariant sets.*

PROBLEMS

1. Find a strict Liapunov function for the equilibrium (0, 0) of
$$x' = -2x - y^3.$$
$$y' = -y - x^3.$$
Find $\delta > 0$ as large as you can such that the open disk of radius δ and center (0, 0) is contained in the basin of (0, 0).

2. Discuss the stability and basins of the equilibria of Example 1 in the text.

3. A particle moves on the straight line $\mathbf{R}$ under the influence of a Newtonian force depending only upon the position of the particle. If the force is always directed toward $0 \in \mathbf{R}$, and vanishes at 0, then 0 is a stable equilibrium. (*Hint*: The total energy E is a Liapunov function for the corresponding first order system
$$x' = y,$$
$$y' = -g(x);$$
E is kinetic energy plus potential energy, and the potential energy at $x \in \mathbf{R}$ is the work required to move the mass from 0 to x.)

4. In Problem 3 suppose also that there is a frictional force opposing the motion, of the form $-f(x)v$, $f(x) \geq 0$, where v is the velocity, and x the position of the particle. If $f^{-1}(0) = 0$, then (0, 0) is asymptotically stable, and in fact *every* trajectory tends toward (0, 0).

5. Sketch the phase portraits of
   (a) the pendulum with friction (see also Problem 6);
   (b) the pendulum without friction.

6. (a) For the frictional pendulum, show that for every integer n and every angle θ_0 there is an initial state (θ_0, ω_0) whose trajectory tends toward (0, 0), and which travels n times, but not $n + 1$ times, around the circle.
   (b) Discuss the set of trajectories tending toward the equilibrium $(\pi, 0)$.

7. Prove the following instability theorem: Let V be a C^1 real-valued function defined on a neighborhood U of an equilibrium $\bar{x}$ of a dynamical system. Suppose $V(\bar{x}) = 0$ and $\dot{V} > 0$ in $U - \bar{x}$. If $V(x_n) > 0$ for some sequence $x_n \to \bar{x}$, then $\bar{x}$ is unstable.

8. Let V be a strict Liapunov function for an equilibrium $\bar{x}$ of a dynamical system. Let $c > 0$ be such that $V^{-1}[0, c]$ is compact and contains no other equilibrium. Then $V^{-1}[0, c] \subset B(\bar{x})$.

§4. Gradient Systems

A *gradient system* on an open set $W \subset \mathbf{R}^n$ is a dynamical system of the form

$$x' = -\operatorname{grad} V(x), \tag{1}$$

where

$$V: U \to \mathbf{R}$$

is a C^2 function, and

$$\operatorname{grad} V = \left(\frac{\partial V}{\partial x_1}, \ldots, \frac{\partial V}{\partial x_n}\right)$$

is the gradient vector field

$$\operatorname{grad} V: U \to \mathbf{R}^n$$

of V. (The negative sign in (1) is traditional. Note that $-\text{grad}\, V(x) = \text{grad}(-V(x))$.)

Gradient systems have special properties that make their flows rather simple. The following equality is fundamental:

$$DV(x)y = \langle \text{grad}\, V(x), y \rangle. \tag{2}$$

This says that the derivative of V at x (which is a linear map $\mathbf{R}^n \to \mathbf{R}$), evaluated on $y \in \mathbf{R}^n$, gives the inner product of the vectors grad $V(x)$ and y. To prove (2), we observe that

$$DV(x)y = \sum_{j=1}^{n} \frac{\partial V}{\partial x_j}(x) y_j,$$

which is exactly the inner product of grad $V(x)$ and $y = (y_1, \ldots, y_n)$.

Let $\dot{V}: U \to \mathbf{R}^n$ be the derivative of V along trajectories of (1); that is,

$$\dot{V}(x) = \frac{d}{dt} V(x(t)) \Big|_{t=0}.$$

Theorem 1 *$\dot{V}(x) \leq 0$ for all $x \in U$; and $\dot{V}(x) = 0$ if and only if x is an equilibrium of (1).*

Proof. By the chain rule

$$\begin{aligned} \dot{V}(x) &= DV(x)x' \\ &= \langle \text{grad}\, V(x), -\text{grad}\, V(x) \rangle \end{aligned}$$

by (2); hence

$$\dot{V}(x) = -|\text{grad}\, V(x)|^2.$$

This proves the theorem.

Corollary *Let $\bar{x}$ be an isolated minimum of V. Then $\bar{x}$ is an asymptotically stable equilibrium of the gradient system $x' = -\text{grad}\, V(x)$.*

Proof. It is easy to verify that the function $x \to V(x) - V(\bar{x})$ is a strict Liapunov function for $\bar{x}$, in some neighborhood of $\bar{x}$.

To understand a gradient flow geometrically one looks at the *level surfaces* of the function $V: U \to \mathbf{R}$. These are the subsets $V^{-1}(c)$, $c \in \mathbf{R}$. If $u \in V^{-1}(c)$ is a *regular point*, that is, grad $V(x) \neq 0$, then $V^{-1}(c)$ looks like a "surface" of dimension $n - 1$ near x. To see this, assume (by renumbering the coordinates) that $\partial V/\partial x_n(u) \neq 0$. Using the implicit function theorem, we find a function g: $\mathbf{R}^{n-1} \to \mathbf{R}$ such that for x near u we have identically

$$V(x_1, \ldots, x_{n-1}, g(x_1, \ldots, x_{n-1})) = c;$$

hence near u, $V^{-1}(c)$ looks like the graph of the function g.

The tangent plane to this graph is exactly the kernel of $DV(u)$. But, by (2), this kernel is the $(n-1)$-dimensional subspace of vectors perpendicular to grad $V(u)$ (translated parallelly to u). Therefore we have shown:

Theorem 2 *At regular points, the vector field $-\text{grad}\, V(x)$ is perpendicular to the level surfaces of V.*

Note by (2) that the nonregular or *critical* points of V are precisely the equilibrium points of the system (1).

Since the trajectories of the gradient system (1) are tangent to $-\text{grad}\, V(x)$, we have the following geometric description of the flow of a gradient system:

Theorem 3 *Let*

$$x' = -\text{grad}\, V(x)$$

be a gradient system. At regular points the trajectories cross level surfaces orthogonally. Nonregular points are equilibria of the system. Isolated minima are asymptotically stable.

Example. Let $V: \mathbf{R}^2 \to \mathbf{R}$ be the function $V(x, y) = x^2(x-1)^2 + y^2$. Then we have, putting $z = (x, y)$:

$$f(z) = -\text{grad}\, V(z) = \left(\frac{-\partial V}{\partial x}, \frac{-\partial V}{\partial y}\right) = (-2x(x-1)(2x-1), -2y)$$

or

$$\frac{dx}{dt} = -2x(x-1)(2x-1),$$

$$\frac{dy}{dt} = -2y.$$

The study of this differential equation starts with the equilibria. These are found by setting the right-hand sides equal to 0, or $-2x(x-1)(2x-1) = 0$, $-2y = 0$.

We obtain precisely three equilibria: $z_{\text{I}} = (0, 0)$, $z_{\text{II}} = (\frac{1}{2}, 0)$, $z_{\text{III}} = (1, 0)$. To check their stability properties, we compute the derivative $Df(z)$ which in coordinates is

$$\begin{bmatrix} \dfrac{d}{dx}(-2x(x-1)(2x-1)) & 0 \\ 0 & \dfrac{d}{dy}(-2y) \end{bmatrix}$$

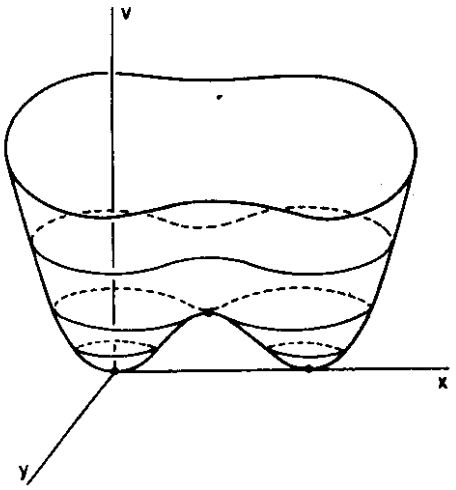


FIG. A. Graph of $V = x^2(x-1)^2 + y^2$.

or

$$Df(z) = \begin{bmatrix} -2(6x^2 - 6x + 1) & 0 \\ 0 & -2 \end{bmatrix}.$$

Evaluating this at the three equilibria gives:

$$Df(z_1) = \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}, \quad Df(z_{II}) = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}, \quad Df(z_{III}) = \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}.$$

We conclude from the main result on nonlinear sinks that z_I, z_{III} are sinks while z_{II} is a saddle. By the theorem of Section 2, z_{II} is not a stable equilibrium.

The graph of V looks like that in Fig. A. The curves on the graph represent intersections with horizontal planes. The level "surfaces" (curves, in this case) look like those in Fig. B. Level curves of $V(x, y) = x^2(x-1)^2 + y^2$ and the phase

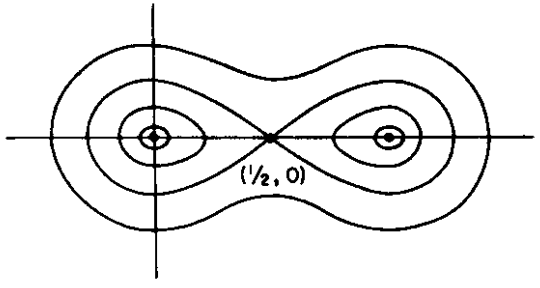


FIG. B. Level curves of $V(x, y)$.

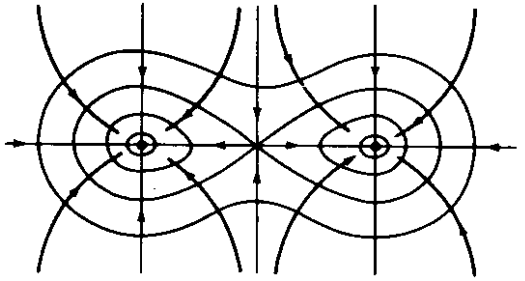


FIG. C. Level curves of $V(x, y)$ and gradient lines of $(x', y') = -\text{grad } V(x, y)$.

portrait of $(x', y') = -\text{grad } V(x, y)$, superimposed on Fig. B, look like Fig. C. The level curve shaped like a reclining figure eight is $V^{-1}(\frac{1}{16})$.

More information about a gradient flow is given by:

Theorem 4 *Let z be an α-limit point or an ω-limit point (Section 3) of a trajectory of a gradient flow. Then z is an equilibrium.*

Proof. Suppose z is an ω-limit point. As in the proof of Theorem 2, Section 3, one shows that V is constant along the trajectory of z. Thus $\dot{V}(z) = 0$; by Theorem 1, z is an equilibrium. The case of α-limit points is similar. In fact, an α-limit point z of $x' = -\text{grad } V(x)$ is an ω-limit point of $x' = \text{grad } V(x)$, whence $\text{grad } V(z) = 0$.

In the case of isolated equilibria this result implies that an orbit must either run off to infinity or else tend to an equilibrium. In the example above we see that the sets

$$V^{-1}([-c, c]), \qquad c \in \mathbf{R},$$

are compact and positively invariant under the gradient flow. Therefore each trajectory entering such a set is defined for all $t \geq 0$, and tends to one of the three equilibria $(0, 0)$, $(1, 0)$, or $(\frac{1}{2}, 0)$. And the trajectory of every point *does* enter such a set, since the trajectory through (x, y) enters the set

$$V^{-1}([-c, c]), \qquad c = V(x, y).$$

The geometrical analysis of this flow is completed by observing that the line $x = \frac{1}{2}$ is made up of the equilibrium $(\frac{1}{2}, 0)$ and two trajectories which approach it, while no other trajectory tends to $(\frac{1}{2}, 0)$. This is because the derivative with respect to t of $|x - \frac{1}{2}|$ is positive if $0 < x < \frac{1}{2}$ or $\frac{1}{2} < x < 1$, as a computation shows.

We have shown: trajectories to the left of the line $x = \frac{1}{2}$ tend toward $(0, 0)$ (as $t \to +\infty$); and trajectories to the right tend toward $(1, 0)$. Trajectories on the line $x = \frac{1}{2}$ tend toward $(\frac{1}{2}, 0)$. This gives a description of the basins of the equilibria $(0, 0)$ and $(1, 0)$. They are the two half planes

$$B(0, 0) = \{(x, y) \in \mathbf{R}^2 \mid x < \tfrac{1}{2}\},$$

$$B(1, 0) = \{(x, y) \in \mathbf{R}^2 \mid x > \tfrac{1}{2}\}.$$

PROBLEMS

1. For each of the following functions $V(u)$, sketch the phase portrait of the gradient flow $u' = -\text{grad } V(u)$. Identify the equilibria and classify them as to stability or instability. Sketch the level surfaces of V on the same diagram.
   (a) $x^3 + 2y^3$ (b) $x^3 - y^3 - 2x + 4y + 5$
   (c) $y \sin x$ (d) $2x^2 - 2xy + 5y^2 + 4x + 4y + 4$
   (e) $x^2 + y^2 - z$ (f) $x^2(x-1) + y^2(y-2) + z^2$

2. Suppose a dynamical system is given. A trajectory $x(t)$, $0 \le t < \infty$, is called *recurrent* if $x(t_n) \to x(0)$ for some sequence $t_n \to \infty$. Prove that a gradient dynamical system has no nonconstant recurrent trajectories.

3. Let $V: E \to \mathbf{R}$ be C^2 and suppose $V^{-1}(-\infty, c]$ is compact for every $c \in \mathbf{R}$. Suppose also $DV(x) \neq 0$ except for a finite number of points $p_1, \ldots, p_r$. Prove:
   (a) Every solution $x(t)$ of $x' = -\text{grad } V(x)$ is defined for all $t \ge 0$;
   (b) $\lim_{t\to\infty} x(t)$ exists and equals one of the equilibrium points $p_1, \ldots, p_r$, for every solution $x(t)$.

§5. Gradients and Inner Products

Here we treat the gradient of a real-valued function V on a vector space E equipped with an inner product $\langle\ ,\ \rangle$. Even if E is $\mathbf{R}^n$, the inner product might not be the standard one. Even if it is, the new definition, while equivalent to the old, has the advantage of being *coordinate free*. As an application we study further the equilibria of a gradient flow.

We define the *dual* of a (real) vector space E to be the vector space

$$E^* = L(E, \mathbf{R})$$

of all linear maps $E \to \mathbf{R}$.

Theorem 1 *E^* is isomorphic to E and thus has the same dimension.*

Proof. Let $\{e_1, \ldots, e_n\}$ be a basis for E and $\langle\ ,\ \rangle$ the induced inner product. Then define $u: E \to E^*$ by $x \to u_x$ where $u_x(y) = \langle x, y\rangle$. Clearly, u is a linear map. Also, $u_x \neq 0$ if $x \neq 0$ since $u_x(x) = \langle x, x\rangle \neq 0$. It remains to show that u is surjective. Let $v \in E^*$ and $v(e_i) = l_i$. Define $x = \sum l_i e_i$, so $u_x(e_k) = \langle e_k, \sum l_i e_i\rangle = l_k$ and $u_x = v$. This proves the theorem.

Since E and E^* have the same dimension, say n, E^* has a basis of n elements. If $\{e_1, \ldots, e_n\} = \mathcal{B}$ is a basis for E, they determine a basis $\{e_1^*, \ldots, e_n^*\} = \mathcal{B}^*$ for E^* by defining

$$e_j^*: E \to \mathbf{R},$$
$$e_j^*\left(\sum_i l_i e_i\right) = l_j;$$

for $i, j = 1, \ldots, n$. Thus e_j^* is characterized by

$$e_j^*(e_i) = \delta_{ij}.$$

$\mathcal{B}^*$ is called the basis *dual* to $\mathcal{B}$.

Now suppose E is given an arbitrary inner product $\langle\ ,\ \rangle$. We define an associated map $\Phi: E \to E^*$ (as in Theorem 1) by $\Phi(x)(y) = \langle x, y\rangle$. Clearly, Φ is an isomorphism by Theorem 1, since its kernel is 0.

Next, let $V: W \to \mathbf{R}$ be a continuously differentiable map defined on an open set $W \subset E$. The derivative of V is a continuous map

$$DV: W \to L(E, \mathbf{R}) = E^*.$$

A map $W \to E^*$ is called a 1-*form* on W. An ordinary differential equation is the same as a vector field on W, that is, a map $W \to E$. We use $\Phi^{-1}: E^* \to E$ to convert the 1-form $DV: W \to E^*$ into a vector field grad $V: W \to E$:

Definition $\text{grad } V(x) = \Phi^{-1}(DV(x))$, $x \in W$.

From the definition of Φ we obtain the equivalent formulation

$$DV(x)y = \langle \text{grad } V(x), y\rangle \quad \text{for all} \quad y \in E. \tag{1}$$

The reader can verify that if $E = \mathbf{R}^n$ with the usual inner product, then this definition of grad $V(x)$ is the same as

$$\left(\frac{\partial V}{\partial x_1}(x), \ldots, \frac{\partial V}{\partial x_n}(x)\right).$$

We now prove some results of the preceding section concerning the differential equation

$$x' = -\text{grad } V(x), \tag{2}$$

using our new definition of grad V.

Theorem 2 *Let $V: W \to \mathbf{R}$ be a C^2 function (that is, $DV: W \to E^*$ is C^1; or V has continuous second partial derivatives) on an open set W in a vector space E with an inner product.*

(a) *$\bar{x}$ is an equilibrium point of the differential equation (2) if and only if $DV(\bar{x}) = 0$.*

(b) *If $x(t)$ is a solution of (2), then*

$$\frac{d}{dt} V(x(t)) = -|\operatorname{grad} V(x(t))|^2.$$

(c) *If $x(t)$ is not constant, then $V(x(t))$ is a decreasing function of t.*

Proof. Since V is C^2, the right side of (2) is a C^1 function of x; therefore the basic uniqueness and existence theory of Chapter 8 applies to (2).

By the definitions $-\operatorname{grad} V(\bar{x}) = 0$ if and only if $DV(\bar{x}) = 0$, since $\Phi: E \to E^*$ is a linear isomorphism; this proves (a). To prove (b) we use the chain rule:

$$\frac{d}{dt} V(x(t)) = DV(x(t))x'(t)$$

$$= DV(x(t))(-\operatorname{grad} V(x(t)));$$

by (1) this equals

$$\langle \operatorname{grad} V(x(t)), -\operatorname{grad} V(x(t)) \rangle = -|\operatorname{grad} V(x(t))|^2.$$

If $x(t)$ is not constant, then by (a), $\operatorname{grad} V(x(t)) \neq 0$; so (b) implies

$$\frac{d}{dt} V(x(t)) < 0.$$

This proves (c).

The dual vector space is also used to study linear operators. We define the *adjoint* of an operator

$$T: E \to E$$

(where E has some fixed inner product) to be the operator

$$T^*: E \to E$$

defined by the equality

$$\langle Tx, y \rangle = \langle x, T^*y \rangle$$

for all x, y in E. To make sense of this, first keep y fixed and note that the map $x \to \langle Tx, y \rangle$ is a linear map $E \to \mathbf{R}$; hence it defines an element $\lambda(y) \in E^*$. We define

$$T^*y = \Phi^{-1}\lambda(y),$$

where

$$\Phi: E \to E^*$$

is the isomorphism defined earlier. It is easy to see that T^* is linear.

If $\mathcal{B}$ is an *orthonormal* basis for E, that is, $\mathcal{B} = \{e_1, \ldots, e_n\}$ and

$$\langle e_i, e_j \rangle = \delta_{ij},$$

then the $\mathcal{B}$-matrix of T^* turns out to be the transpose of the $\mathcal{B}$-matrix for T, as is easily verified.

An operator $T \in L(E)$ is *self-adjoint* if $T^* = T$, that is,

$$\langle Tx, y \rangle = \langle x, Ty \rangle, \qquad \text{for all } x, y \in E.$$

In an orthonormal basis this means the matrix $[a_{ij}]$ of T is *symmetric*, that is, $a_{ij} = a_{ji}$.

Theorem 3 *Let E be a real vector space with an inner product and let T be a self-adjoint operator on E. Then the eigenvalues of T are real.*

Proof. Let $E_{\mathbf{C}}$ be the complexification of E. We extend $\langle\ ,\ \rangle$ to a function $E_{\mathbf{C}} \times E_{\mathbf{C}} \to \mathbf{C}$ as follows. If $x + iy$ and $u + iv$ are in $E_{\mathbf{C}}$, define

$$\langle x + iy, u + iv \rangle = \langle x, u \rangle + i(\langle y, u \rangle - \langle x, v \rangle) + \langle y, v \rangle.$$

It is easy to verify the following for all $a, b \in E_{\mathbf{C}}$, $\lambda \in \mathbf{C}$:

$$\langle a, a \rangle > 0 \quad \text{if } a \neq 0, \tag{3}$$

$$\lambda\langle a, b \rangle = \langle \lambda a, b \rangle = \langle a, \bar{\lambda} b \rangle, \tag{4}$$

where $^-$ denotes the complex conjugate.

Let $T_{\mathbf{C}}: E_{\mathbf{C}} \to E_{\mathbf{C}}$ be the complexification of T; thus $T_{\mathbf{C}}(x + iy) = Tx + i(Ty)$. Let $(T^*)_{\mathbf{C}}$ be the complexification of T^*. It is easy to verify that

$$\langle T_{\mathbf{C}}a, b \rangle = \langle a, (T^*)_{\mathbf{C}}b \rangle. \tag{5}$$

(This is true even if T is not self-adjoint.)

Suppose $\lambda \in \mathbf{C}$ is an eigenvalue for T and $a \in E_{\mathbf{C}}$ an eigenvector for λ; then $T_{\mathbf{C}}a = \lambda a$.

By (5)

$$\langle T_{\mathbf{C}}a, a \rangle = \langle a, (T^*)_{\mathbf{C}}a \rangle$$

$$= \langle a, T_{\mathbf{C}}a \rangle.$$

since $T^* = T$. Hence

$$\langle \lambda a, a \rangle = \langle a, \lambda a \rangle.$$

But, by (4),

$$\lambda\langle a, a \rangle = \langle \lambda a, a \rangle,$$

while

$$\bar{\lambda}\langle a, a \rangle = \langle a, \lambda a \rangle;$$

so, by (3), $\lambda = \bar{\lambda}$ and λ is real.

Corollary *A symmetric real $n \times n$ matrix has real eigenvalues.*

Consider again a gradient vector field

$$F(x) = -\operatorname{grad} V(x).$$

For simplicity we assume the vector space is $\mathbf{R}^n$, equipped with the usual inner product. Let $\bar{x}$ be an equilibrium of the system

$$x' = -\text{grad } V(x).$$

The operator

$$DF(\bar{x})$$

has the matrix

$$-\left[\frac{\partial^2 V}{\partial x_i \partial x_j}(\bar{x})\right]$$

in the standard basis. Since this matrix is symmetric, we conclude:

Theorem 4 *At an equilibrium of a gradient system, the eigenvalues are real.*

This theorem is also true for gradients defined by arbitrary inner products.

For example, a gradient system in the plane cannot have spirals or centers at equilibria. In fact, neither can it have improper nodes because of:

Theorem 5 *Let E be a real vector space with an inner product. Then any self-adjoint operator on E can be diagonalized.*

Proof. Let $T: E \to E$ be self-adjoint. Since the eigenvalues of T are real, there is a nonzero vector $e_1 \in E$ such that $Te_1 = \lambda_1 e_1$, $\lambda_1 \in \mathbf{R}$. Let

$$E_1 = \{x \in E \mid \langle x, e_1 \rangle = 0\},$$

the orthogonal complement of e_1. If $x \in E_1$, then $Tx \in E_1$, for

$$\langle Tx, e_1 \rangle = \langle x, Te_1 \rangle = \langle x, \lambda e_1 \rangle = \lambda \langle x, e_1 \rangle = 0.$$

Hence T leaves E_1 invariant. Give E_1 the same inner product as E; then the operator

$$T_1 = T \mid E_1 \in L(E_1)$$

is self-adjoint. In the same way we find a nonzero vector $e_2 \in E_1$ such that

$$Te_2 = \lambda_2 e_2; \qquad \lambda_2 \in \mathbf{R}.$$

Note that e_1 and e_2 are independent, since $\langle e_1, e_2 \rangle = 0$. Continuing in this way, we find a maximal independent set $\mathcal{B} = \{e_1, \ldots, e_n\}$ of eigenvectors of T. These must span E, otherwise we could enlarge the set by looking at the restriction of T to the subspace orthogonal to $e_1, \ldots, e_n$. In this basis $\mathcal{B}$, T is diagonal.

We have actually proved more. Note that $e_1, \ldots, e_n$ are mutually orthogonal; and we can take them to have norm 1. Therefore a *self-adjoint operator (or a symmetric matrix) can be diagonalized by an orthonormal basis.*

For gradient systems we have proved:

Theorem 6 *At an equilibrium of a gradient flow the linear part of the vector field is diagonalizable by an orthonormal basis.*

PROBLEMS

1. Find an orthonormal diagonalizing basis for each of the following operators:

   (a) $\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 0 & -2 \\ -2 & 0 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 2 \\ 1 & 2 & -1 \end{bmatrix}$

2. Let A be a self-adjoint operator. If x and y are eigenvectors belonging to different eigenvalues then $\langle x, y \rangle = 0$.

3. Show that for each operator A of Problem 1, the vector field $x \to Ax$ is the gradient of some function.

4. If A is a symmetric operator, show that the vector field $x \to Ax$ is the gradient of some function.

Notes

A statement and proof of the implicit function theorem used in Section 4, is given in Appendix 4. See P. Halmos' *Finite Dimensional Vector Spaces* [8] for a more extended treatment of self-adjoint linear operators. One can find more on Liapunov theory in LaSalle and Lefschetz's *Stability by Liapunov's Direct Method with Applications* [14]. Pontryagin's text [10] on ordinary differential equations is recommended; in particular, he has an interesting application of Liapunov theory to the study of the governor of a steam engine.

Chapter 10

Differential Equations for Electrical Circuits

First a simple but very basic circuit example is described and the differential equations governing the circuit are derived. Our derivation is done in such a way that the ideas extend to general circuit equations. That is why we are so careful to make the maps explicit and to describe precisely the sets of states obeying physical laws. This is in contrast to the more typical ad hoc approach to nonlinear circuit theory.

The equations for this example are analyzed from the purely mathematical point of view in the next three sections; these are the classical equations of Lienard and Van der Pol. In particular Van der Pol's equation could perhaps be regarded as the fundamental example of a nonlinear ordinary differential equation. It possesses an oscillation or periodic solution that is a periodic attractor. Every nontrivial solution tends to this periodic solution; no linear flow can have this property. On the other hand, for a periodic solution to be viable in applied mathematics, this or some related stability property must be satisfied.

The construction of the phase portrait of Van der Pol in Section 3 involves some nontrivial mathematical arguments and many readers may wish to skip or postpone this part of the book. On the other hand, the methods have some wider use in studying phase portraits.

Asymptotically stable equilibria connote death in a system, while attracting oscillators connote life. We give an example in Section 4 of a continuous transition from one to the other.

In Section 5 we give an introduction to the mathematical foundations of electrical circuit theory, especially oriented toward the analysis of nonlinear circuits.

§1. An *RLC* Circuit

We give an example of an electrical circuit and derive from it a differential equation that shows how the state of the circuit varies in time. The differential equation is analyzed in the following section. Later we shall describe in greater generality elements of the mathematical theory of electrical circuits.

Our discussion of the example here is done in a way that extends to the more general case.

The circuit of our example is the simple but fundamental series *RLC* circuit in Fig. A. We will try to communicate what this means, especially in mathematical terms. The circuit has three *branches*, one resistor marked by R, one inductor marked by L, and one capacitor marked by C. One can think of a branch as being a certain electrical device with two terminals. In the circuit, branch R has terminals α, β for example and these terminals are wired together to form the points or *nodes* α, β, γ.

The electrical devices we consider in this book are of the three types: resistors, inductors, and capacitors, which we will characterize mathematically shortly.

In the circuit one has flowing through each branch a current which is measured by a real number. More precisely the currents in the circuit are given by the three numbers i_R, i_L, i_C; i_R measures the current through the resistor, and so on. Current in a branch is analogous to water flowing in a pipe; the corresponding measure for water would be the amount flowing in unit time, or better, the rate at which water passes by a fixed point in the pipe. The arrows in the diagram that orient the branches tell us which way the current (read water!) is flowing; if for example i_R is positive, then according to the arrow current flows through the resistor from β to α (the choice of the arrows is made once and for all at the start).

The state of the currents at a given time in the circuit is thus represented by a point $i = (i_R, i_L, i_C) \in \mathbf{R}^3$. But *Kirchhoff's current law* (KCL) says that in reality there is a strong restriction on what i can occur. KCL asserts that the total current

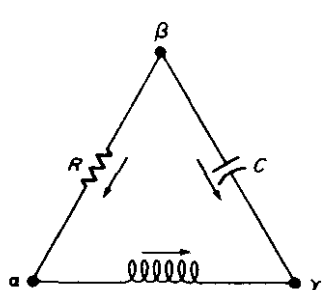


FIG. A

flowing into a node is equal to the total current flowing out of that node. (Think of the water analogy to make this plausible.) For our circuit this is equivalent to

$$\text{KCL:} \qquad i_R = i_L = -i_C.$$

This defines a one-dimensional subspace K_1 of $\mathbf{R}^3$ of *physical current states*. Our choice of orientation of the capacitor branch may seem unnatural. In fact the orientations are arbitrary; in the example they were chosen so that the equations eventually obtained relate most directly to the history of the subject.

The state of the circuit is characterized by the current i together with the voltage (or better, voltage drop) across each branch. These voltages are denoted by v_R, v_L, v_C for the resistor branch, inductor branch, and capacitor branch, respectively. In the water analogy one thinks of the voltage drop as the difference in pressures at the two ends of a pipe. To measure voltage one places a voltmeter (imagine a water pressure meter) at each of the nodes α, β, γ which reads $V(\alpha)$ at α, and so on. Then v_R is the difference in the reading at α and β

$$V(\beta) - V(\alpha) = v_R.$$

The orientation or arrow tells us that $v_R = V(\beta) - V(\alpha)$ rather than $V(\alpha) - V(\beta)$.

An *unrestricted voltage state* of the circuit is then a point $v = (v_R, v_L, v_C)$ in $\mathbf{R}^3$. Again a Kirchhoff law puts a physical restriction on v:

$$\text{KVL:} \qquad v_R + v_L - v_C = 0.$$

This defines a two-dimensional linear subspace K_2 of $\mathbf{R}^3$. From our explanation of the v_R, v_L, v_C in terms of voltmeters, KVL is clear; that is,

$$v_R + v_L - v_C = (V(\beta) - V(\alpha)) + (V(\alpha) - V(\gamma)) - (V(\beta) - V(\gamma)) = 0.$$

In a general circuit, one version of KVL asserts that the voltages can be derived from a "voltage potential" function V on the nodes as above.

We summarize that in the product space, $\mathbf{R}^3 \times \mathbf{R}^3 = \mathcal{S}$, those states (i, v) satisfying Kirchhoff's laws form a three-dimensional subspace K of the form $K = K_1 \times K_2 \subset \mathbf{R}^3 \times \mathbf{R}^3$.

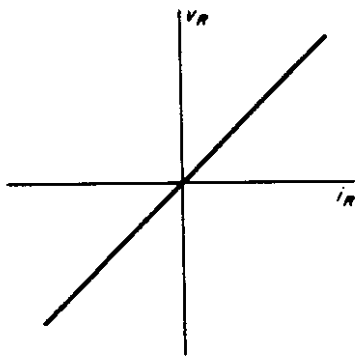


FIG. B

Next, we give a mathematical definition of the three kinds of electrical devices of the circuit.

First consider the resistor element. A resistor in the R branch imposes a "functional relationship" on i_R, v_R. We take in our example this relationship to be defined by a C^1 real function f of a real variable, so that $v_R = f(i_R)$. If R denotes a conventional linear resistor, then f is linear and $v_R = f(i_R)$ is a statement of Ohm's law. The graph of f in the (i_R, v_R) plane is called the *characteristic* of the resistor. A couple of examples of characteristics are given in Figs. B and C. (A characteristic like that in Fig. C occurs in the "tunnel diode.")

A *physical state* $(i, v) \in \mathbf{R}^3 \times \mathbf{R}^3 = \mathcal{S}$ will be one which satisfies KCL and KVL or $(i, v) \in K$ and also $f(i_r) = v_r$. These conditions define a subset $\Sigma \subset K \subset \mathcal{S}$ Thus the *set of physical states* Σ is that set of points $(i_R, i_L, i_C, v_R, v_L, v_C)$ in $\mathbf{R}^3 \times \mathbf{R}^3$ satisfying:

$$\begin{aligned} &i_R = i_L = -i_C && \text{(KCL)}, \\ &v_R + v_L - v_C = 0 && \text{(KVL)}, \\ &f(i_R) = v_R && \text{(generalized Ohm's law)}. \end{aligned}$$

Next we concern ourselves with the passage in time of a state; this defines a curve in the state space $\mathcal{S}$:

$$t \to (i(t), v(t)) = (i_R(t), i_L(t), i_C(t), v_R(t), v_L(t), v_C(t)).$$

The inductor (which one may think of as a coil; it is hard to find a water analogy) specifies that

$$L \frac{di_L(t)}{dt} = v_L(t) \qquad \text{(Faraday's law)},$$

where L is a positive constant called the inductance.

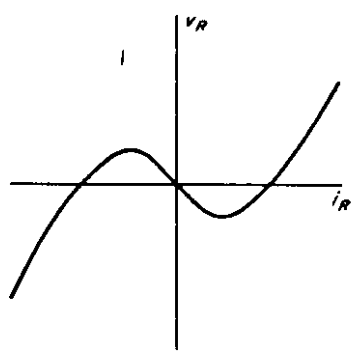


FIG. C

On the other hand, the capacitor (which may be thought of as two metal plates separated by some insulator; in the water model it is a tank) imposes the condition

$$C\frac{dv_C(t)}{dt} = i_C(t),$$

where C is a positive constant called the capacitance.

We summarize our development so far: a state of our circuit is given by the six numbers $(i_R, i_L, i_C, v_R, v_L, v_C)$, that is, an element of $\mathbf{R}^3 \times \mathbf{R}^3$. These numbers are subject to three restrictions: Kirchhoff's current law, Kirchhoff's voltage law, and the resistor characteristic or "generalized Ohm's law." Therefore the space of physical states is a certain subset $\Sigma \subset \mathbf{R}^3 \times \mathbf{R}^3$. The way a state changes in time is determined by two differential equations.

Next, we simplify the state space Σ by observing that i_L and v_C determine the other four coordinates, since $i_R = i_L$ and $i_C = -i_L$ by KCL, $v_R = f(i_R) = f(i_L)$ by the generalized Ohm's law, and $v_L = v_C - v_R = v_C - f(i_L)$ by KVL. Therefore we can use $\mathbf{R}^2$ as the state space, interpreting the coordinates as (i_L, v_C). Formally, we define a map $\pi\colon \mathbf{R}^3 \times \mathbf{R}^3 \to R^3$, sending $(i, v) \in \mathbf{R}^3 \times \mathbf{R}^3$ to (i_L, v_C). Then we set $\pi_0 = \pi \mid \Sigma$, the restriction of π to Σ; this map $\pi_0\colon \Sigma \to \mathbf{R}^2$ is one-to-one and onto; its inverse is given by the map $\varphi\colon \mathbf{R}^2 \to \Sigma$,

$$\varphi(i_L, v_C) = (i_L, i_L, -i_L, f(i_L), v_C - f(i_L), v_C).$$

It is easy to check that $\varphi(i_L, v_C)$ satisfies KCL, KVL, and the generalized Ohm's law, so φ does map $\mathbf{R}^2$ into Σ; it is also easy to see that π_0 and φ are inverse to each other.

We therefore adopt $\mathbf{R}^2$ as our state space. The differential equations governing the change of state must be rewritten in terms of our new coordinates (i_L, v_C):

$$L\frac{di_L}{dt} = v_L \equiv v_C - f(i_L),$$

$$C\frac{dv_C}{dt} = i_C \equiv -i_L.$$

For simplicity, since this is only an example, we make $L = 1$, $C = 1$.

If we write $x = i_L$, $y = v_C$, we have as differential equations on the (x, y) Cartesian space:

$$\frac{dx}{dt} = y - f(x),$$

$$\frac{dy}{dt} = -x.$$

These equations are analyzed in the following section.

PROBLEMS

1. Find the differential equations for the network in Fig. D, where the resistor is voltage controlled, that is, the resistor characteristic is the graph of a C^1 function $g\colon \mathbf{R} \to \mathbf{R}$, $g(v_R) = i_R$.

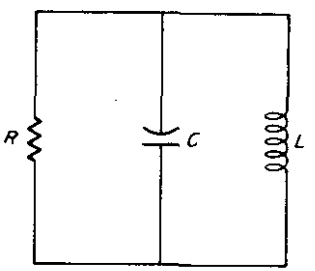


FIG. D

2. Show that the LC circuit consisting of one inductor and one capacitor wired in a closed loop oscillates.

§2. Analysis of the Circuit Equations

Here we begin a study of the phase portrait of the planar differential equation derived from the circuit of the previous section, namely:

$$\frac{dx}{dt} = y - f(x), \tag{1}$$

$$\frac{dy}{dt} = -x.$$

This is one form of *Lienard's equation*. If $f(x) = x^3 - x$, then (1) is a form of *Van der Pol's equation*.

First consider the most simple case of linear f (or ordinary resistor of Section 1). Let $f(x) = Kx$, $K > 0$. Then (1) takes the form

$$z' = Az, \qquad A = \begin{bmatrix} -K & 1 \\ -1 & 0 \end{bmatrix}, \qquad z = (x, y).$$

The eigenvalues of A are given by $\lambda = \frac{1}{2}[-K \pm (K^2 - 4)^{1/2}]$. Since λ always has negative real part, the zero state $(0, 0)$ is an asymptotically stable equilibrium,

in fact a sink. Every state tends to zero; physically this is the dissipative effect of the resistor. Furthermore, one can see that (0, 0) will be a spiral sink precisely when $K < 2$.

Next we consider the equilibria of (1) for a general C^1 function f.

There is in fact a unique equilibrium $\bar{z}$ of (1) obtained by setting

$$y - f(x) = 0,$$

$$-x = 0,$$

or

$$\bar{z} = (0, f(0)).$$

The matrix of first partial derivatives of (1) at $\bar{z}$ is

$$\begin{bmatrix} -f'(0) & 1 \\ -1 & 0 \end{bmatrix}$$

whose eigenvalues are given by

$$\lambda = \tfrac{1}{2}[-f'(0) \pm (f'(0)^2 - 4)^{1/2}].$$

We conclude that this equilibrium satisfies:

$$\bar{z} \quad \text{is a sink if} \quad f'(0) > 0,$$

and

$$\bar{z} \quad \text{is a source if} \quad f'(0) < 0$$

(see Chapter 9).

In particular for Van der Pol's equation ($f(x) = x^3 - x$) the unique equilibrium is a source.

To analyze (1) further we define a function $W: \mathbf{R}^2 \to \mathbf{R}^2$ by $W(x, y) = \frac{1}{2}(x^2 + y^2)$; thus W is half of the norm squared. The following proposition is simple but important in the study of (1).

Proposition *Let $z(t) = (x(t), y(t))$ be a solution curve of Lienard's equation* (1). *Then*

$$\frac{d}{dt} W(z(t)) = -x(t)f(x(t)).$$

Proof. Apply the chain rule to the composition

$$J \xrightarrow{z} \mathbf{R}^2 \xrightarrow{W} \mathbf{R}$$

to obtain

$$\frac{d}{dt} W(z(t)) = DW(z(t))(z'(t)) = x(t)x'(t) + y(t)y'(t);$$

suppressing t, this is equal to

$$x(y - f(x)) - yx = -xf(x)$$

by (1). Here J could be any interval of real numbers in the domain of z.

The statement of the proposition has an interpretation for the electric circuit that gave rise to (1) and which we will pursue later: energy decreases along the solution curves according to the power dissipated in the resistor.

In circuit theory, a resistor whose characteristic is the graph of $f: \mathbf{R} \to \mathbf{R}$, is called *passive* if its characteristic is contained in the set consisting of (0, 0) and the interior of the first and third quadrant (Fig. A for example). Thus in the case of a passive resistor $-xf(x)$ is negative except when $x = 0$.

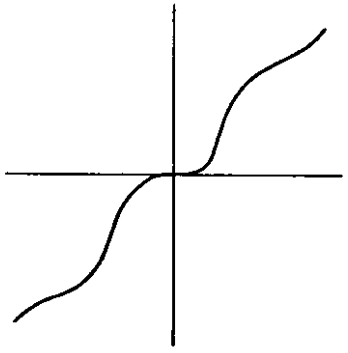


FIG. A

From Theorem 2 of Chapter 9, Section 3, it follows that the origin is asymptotically stable and its basin of attraction is the whole plane. Thus the word *passive* correctly describes the dynamics of such a circuit.

§3. Van der Pol's Equation

The goal here is to continue the study of Lienard's equation for a certain function f.

(1)
$$\frac{dx}{dt} = y - f(x), \qquad f(x) = x^3 - x,$$

$$\frac{dy}{dt} = -x.$$

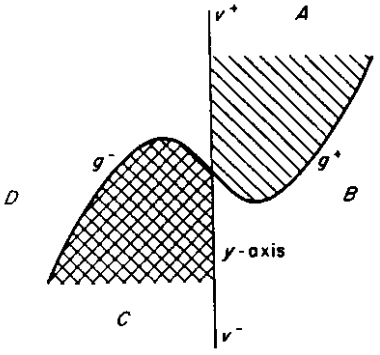


FIG. A

This is called *Van der Pol's equation*; equivalently

$$\frac{dx}{dt} = y - x^3 + x, \tag{2}$$

$$\frac{dy}{dt} = -x.$$

In this case we can give a fairly complete phase portrait analysis.

Theorem *There is one nontrivial periodic solution of* (1) *and every nonequilibrium solution tends to this periodic solution. "The system oscillates."*

We know from the previous section that (2) has a unique equilibrium at (0, 0), and it is a source. The next step is to show that every nonequilibrium solution "rotates" in a certain sense around the equilibrium in a clockwise direction. To this end we divide the (x, y) plane into four disjoint regions (open sets) A, B, C, D in Fig. A. These regions make up the complement of the curves

$$y - f(x) = 0, \tag{3}$$

$$-x = 0.$$

These curves (3) thus form the boundaries of the four regions. Let us make this more precise. Define four curves

$$v^+ = \{(x, y) \mid y > 0, x = 0\},$$

$$g^+ = \{(x, y) \mid x > 0, y = x^3 - x\},$$

$$v^- = \{(x, y) \mid y < 0, x = 0\},$$

$$g^- = \{(x, y) \mid x < 0, y = x^3 - x\}.$$

These curves are disjoint; together with the origin they form the boundaries of the four regions.

Next we see how the vector field (x', y') of (1) behaves on the boundary curves. It is clear that $y' = 0$ at (0, 0) and on $v^+ \cup v^-$, and nowhere else; and $x' = 0$ exactly on $g^+ \cup g^- \cup (0, 0)$. Furthermore the vector (x', y') is horizontal on $v^+ \cup v^-$ and points right on v^+, and left on v^- (Fig. B). And (x', y') is vertical on $g^+ \cup g^-$, pointing downward on g^+ and upward on g^-. In each region A, B, C, D the signs of x' and y' are constant. Thus in A, for example, we have $x' > 0$, $y' < 0$, and so the vector field always points into the fourth quadrant.

The next part of our analysis concerns the nature of the flow in the interior of the regions. Figure B suggests that trajectories spiral around the origin clockwise. The next two propositions make this precise.

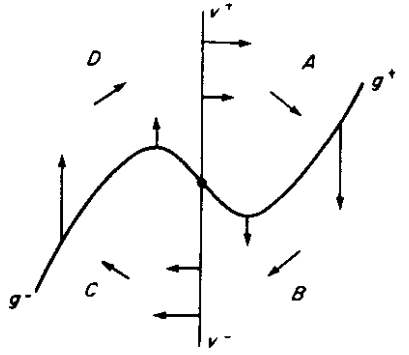


FIG. B

Proposition 1 *Any trajectory starting on v^+ enters A. Any trajectory starting in A meets g^+; furthermore it meets g^+ before it meets v^-, g^- or v^+.*

Proof. See Fig. B. Let $(x(t), y(t))$ be a solution curve to (1). If $(x(0), y(0)) \in v^+$, then $x(0) = 0$ and $y(0) > 0$. Since $x'(0) > 0$, $x(t)$ increases for small t and so $x(t) > 0$ which implies that $y(t)$ decreases for small t. Hence the curve enters A. Before the curve leaves A (if it does), x' must become 0 again, so the curve must cross g^+ before it meets v^-, g^- or v^+. Thus the first and last statements of the proposition are proved.

It remains to show that if $(x(0), y(0)) \in A$ then $(x(t), y(t)) \in g^+$ for some $t > 0$. Suppose not.

Let $P \subset \mathbf{R}^2$ be the compact set bounded by (0, 0) and v^+, g^+ and the line $y = y(0)$ as in Fig. C. The solution curve $(x(t), y(t))$, $0 \le t < \beta$ is in P. From Chapter 8, it follows since $(x(t), y(t))$ does not meet g^+, it is defined for all $t > 0$.

Since $x' > 0$ in A, $x(t) \ge a$ for $t > 0$. Hence from (1), $y'(t) \le -a$ for $t > 0$.

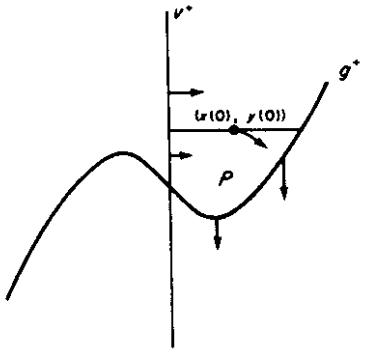


FIG. C

For these values of t, then

$$y(t) = \int_0^t y'(s)\, ds \leq y(0) - at.$$

This is impossible, unless our trajectory meets g^+, proving Proposition 1.

Similar arguments prove (see Fig. D):

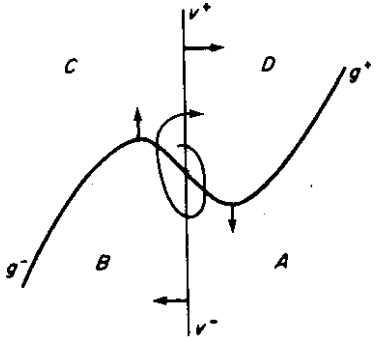


FIG. D. Trajectories spiral clockwise.

Proposition 2 *Every trajectory is defined for (at least) all $t \geq 0$. Except for $(0, 0)$, each trajectory repeatedly crosses the curves v^+, g^+, v^-, g^-, in clockwise order, passing among the regions A, B, C, D in clockwise order.*

To analyze further the flow of the Van der Pol oscillator we define a map

$$\sigma: v^+ \to v^+$$

as follows. Let $p \in v^+$; the solution curve $t \to \phi_t(p)$ through p is defined for all $t \geq 0$. There will be a smallest $t_1(p) = t_1 > 0$ such that $\phi_{t_1}(p) \in v^+$. We put $\sigma(p) = \phi_{t_1}(p)$. Thus $\sigma(p)$ is the first point after p on the trajectory of p (for $t > 0$) which is again on v^+ (Fig. E). The map $p \to t_1(p)$ is continuous; while this should be intuitively clear, it follows rigorously from Chapter 11. Hence σ is also continuous. Note that σ is one to one by uniqueness of solutions.

The importance of this *section map* $\sigma: v^+ \to v^+$ comes from its intimate relationship to the phase portrait of the flow. For example:

Proposition 3 *Let $p \in v^+$. Then p is a fixed point of σ (that is, $\sigma(p) = p$) if and only if p is on a periodic solution of (1) (that is, $\phi_t(p) = p$ for some $t \neq 0$). Moreover every periodic solution curve meets v^+.*

Proof. If $\sigma(p) = p$, then $\phi_{t_1}(p) = p$, where $t_1 = t_1(p)$ is as in the definition of σ. Suppose on the other hand that $\sigma(p) \neq p$. Let $v^* = v^+ \cup (0, 0)$. We observe first that σ extends to a map $v^* \to v^*$ which is again continuous and one to one, sending $(0, 0)$ to itself. Next we identify v^* with $\{y \in \mathbf{R} \mid y \geq 0\}$ by assigning to each point its y-coordinate. Hence there is a natural order on v^*: $(0, y) < (0, z)$ if $y < z$. It follows from the intermediate value theorem that $\sigma: v^* \to v^*$ is *order preserving*. If $\sigma(p) > p$, then $\sigma^2(p) > \sigma(p) > p$ and by induction $\sigma^n(p) > p$, $n = 1, 2, \ldots$. This means that the trajectory of p never crosses v^+ again at p. Hence $\phi_t(p) \neq p$ for all $t \neq 0$. A similar argument applies if $\sigma(p) < p$. Therefore if $\sigma(p) \neq p$, p is not on a periodic trajectory. The last statement of Proposition 3 follows from Proposition 2 which implies that every trajectory (except $(0, 0)$) meets v^+.

For every point $p \in v^+$ let $t_2(p) = t_2$ be the smallest $t > 0$ such that $\phi_t(p) \in v^-$. Define a continuous map

$$\alpha: v^+ \to v^-,$$

$$\alpha(p) = \phi_{t_2}(p).$$

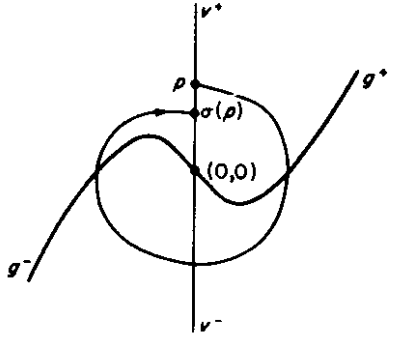


FIG. E. The map $\sigma: v^+ \to v^+$.

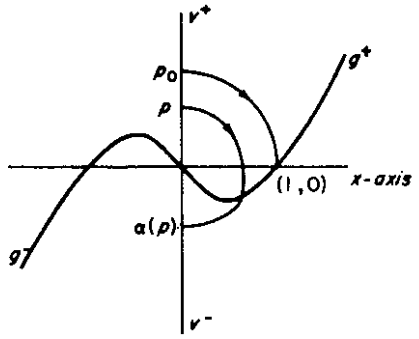


FIG. F. The map $\alpha: v^+ \to v^-$.

See Fig. F. The map α is also one to one by uniqueness of solutions and thus monotone.

Using the methods in the proof of Proposition 1 it can be shown that there is a unique point $p_0 \in v^+$ such that the solution curve

$$\{\phi_t(p_0) \mid 0 \leq t \leq t_2(p_0)\}$$

intersects the curve g^+ at the point $(1, 0)$ where g^+ meets the x-axis. Let $r = |p_0|$.

Define a continuous map

$$\delta: v^+ \to \mathbf{R},$$

$$\delta(p) = \tfrac{1}{2}(|\alpha(p)|^2 - |p|^2)$$

where $|p|$ means the usual Euclidean norm of the vector p. Further analysis of the flow of (1) is based on the following rather delicate result:

Proposition 4 (a) $\delta(p) > 0$ *if* $0 < |p| < r$;
(b) $\delta(p)$ *decreasers monotonely to* $-\infty$ *as* $|p| \to \infty$, $|p| \geq r$.

Part of the graph of $\delta(p)$ as a function of $|p|$ is shown schematically in Fig. G. The intermediate value theorem and Proposition 4 imply that *there is a unique* $q_0 \in v^+$ *with* $\delta(q_0) = 0$.

We will prove Proposition 4 shortly; first we use it to complete the proof of the main theorem of this section. We exploit the *skew symmetry* of the vector field

$$g(x, y) = (y - x^3 + x, -x)$$

given by the right-hand side of (2), namely,

$$g(-x, -y) = -g(x, y).$$

This means that if $t \to (x(t), y(t))$ is a solution curve, so is $t \to (-x(t), -y(t))$. Consider the trajectory of the unique point $q_0 \in v^+$ such that $\delta(q_0) = 0$. This

point has the property that $|\alpha(q_0)| = |q_0|$, hence that

$$\phi_{t_2}(q_0) = -q_0.$$

From skew symmetry we have also

$$\phi_{t_2}(-q_0) = -(-q_0) = q_0;$$

hence putting $\lambda = 2t_2 > 0$ we have

$$\phi_\lambda(q_0) = q_0.$$

Thus q_0 lies on a nontrivial periodic trajectory γ.

Since δ is monotone, similar reasoning shows that the trajectory through q_0 is the unique nontrivial periodic solution.

To investigate other trajectories we define a map $\beta: v^- \to v^+$, sending each point of v^- to the first intersection of its trajectory (for $t > 0$) with v^+. By symmetry

$$\beta(p) = -\alpha(-p).$$

Note that $\sigma = \beta\alpha$.

We identify the y-axis with the real numbers in the y-coordinate. Thus if p, $q \in v^+ \cup v^-$ we write $p > q$ if p is above q. Note that α and β reverse this ordering while σ preserves it.

Now let $p \in v_+$, $p > q_0$. Since $\alpha(q_0) = -q_0$ we have $\alpha(p) < -q_0$ and $\sigma(p) > q_0$. On the other hand, $\delta(p) < 0$ which means the same thing as $\alpha(p) > -p$. Therefore $\sigma(p) = \beta\alpha(p) < p$. We have shown that $p > q_0$ implies $p > \sigma(p) > q_0$. Similarly $\sigma(p) > \sigma^2(p) > q_0$ and by induction $\sigma^n(p) > \sigma^{n+1}(p) > q_0$, $n = 1, 2, \ldots$.

The sequence $\sigma^n(p)$ has a limit $q_1 \geq q_0$ in v^+. Note that q_1 is a fixed point of σ, for by continuity of σ we have

$$\sigma(q_1) - q_1 = \lim_{n\to\infty} \sigma(\sigma^n(p)) - q_1$$

$$= q_1 - q_1 = 0.$$

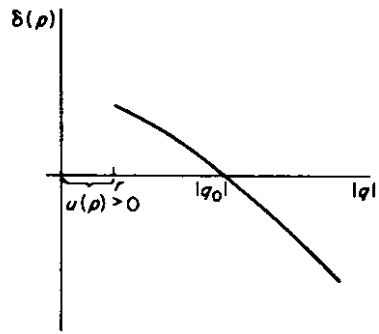


FIG. G

Since σ has only one fixed point $q_1 = q_0$. This shows that *the trajectory of p spirals toward γ as $t \to \infty$*. The same thing is true if $p < q_0$; the details are left to the reader. Since every trajectory except $(0, 0)$ meets v^+, the proof of the main theorem is complete.

It remains to prove Proposition 4.

We adopt the following notation. Let $\gamma\colon [a, b] \to \mathbf{R}^2$ be a C^1 curve in the plane, written $\gamma(t) = (x(t), y(t))$. If $F\colon \mathbf{R}^2 \to \mathbf{R}$ is C^1, define

$$\int_\gamma F(x, y) = \int_a^b F(x(t), y(t))\, dt.$$

It may happen that $x'(t) \neq 0$ for $a \leq t \leq b$, so that along γ, y is a function of x, $y = y(x)$. In this case we can change variables:

$$\int_a^b F(x(t), y(t))\, dt = \int_{x(a)}^{x(b)} F(x, y(x)) \frac{dt}{dx}\, dx;$$

hence

$$\int_\gamma F(x, y) = \int_{x(a)}^{x(b)} \frac{F(x, y(x))}{dx/dt}\, dx.$$

Similarly if $y'(t) \neq 0$.

Recall the function

$$W(x, y) = \tfrac{1}{2}(x^2 + y^2).$$

Let $\gamma(t) = (x(t), y(t))$, $0 \leq t \leq t_2 = t_2(p)$ be the solution curve joining $p \in v^+$ to $\alpha(p) \in v^-$. By definition $\delta(p) = W(x(t_2), y(t_2)) - W(x(0), y(0))$. Thus

$$\delta(p) = \int_0^{t_2} \frac{d}{dt} W(x(t), y(t))\, dt$$

By the proposition of Section 2 we have

$$\delta(p) = \int_0^{t_2} -x(t)(x(t)^2 - x(t))\, dt;$$

$$\delta(p) = \int_0^{t_2} x(t)^2(1 - x(t)^2)\, dt.$$

This immediately proves (a) of Proposition 4 because the integrand is positive for $0 < x(t) < 1$.

We may rewrite the last equality as

$$\delta(p) = \int_\gamma x^2(1 - x^2).$$

We restrict attention to points $p \in v^+$ with $|\, p \,| > r$. We divide the corresponding

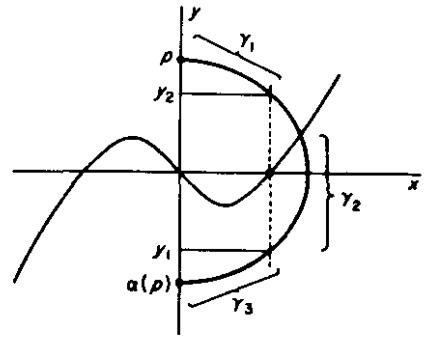


FIG. H

solution curve γ into three curves γ_1, γ_2, γ_3 as in Fig. H. Then

$$\delta(p) = \delta_1(p) + \delta_2(p) + \delta_3(p),$$

where

$$\delta_i(p) = \int_{\gamma_i} x^2(1 - x^2), \qquad i = 1, 2, 3.$$

Notice that along γ_1, $y(t)$ is a function of $x(t)$. Hence

$$\delta_1(p) = \int_0^1 \frac{x^2(1 - x^2)}{dx/dt}\, dx$$

$$= \int_0^1 \frac{x^2(1 - x^2)}{y - f(x)}\, dx,$$

where $f(x) = x^3 - x$. As p moves up the y-axis, $y - f(x)$ increases (for (x, y) on γ_1). Hence $\delta_1(p)$ decreases as $|\, p \,| \to \infty$. Similarly $\delta_3(p)$ decreases as $|\, p \,| \to \infty$.

On γ_2, x is a function of y, and $x \geq 1$. Therefore, since $dy/dt = -x$,

$$\delta_2(p) = \int_{y_2}^{y_1} -x(y)(1 - x(y)^2)\, dy$$

$$= \int_{y_1}^{y_2} x(y)(1 - x(y)^2)\, dy < 0.$$

As $|\, p \,|$ increases, the domain $[y_1, y_2]$ of integration becomes steadily larger. The function $y \to x(y)$ depends on p; we write it $x_p(y)$. As $|\, p \,|$ increases, the curves γ_2 move to the right; hence $x_p(y)$ increases and so $x_p(y)(1 - x_p(y)^2)$ decreases. It follows that $\delta_2(p)$ decreases as $|\, p \,|$ increases; and evidently $\lim_{|p| \to \infty} \delta_2(p) = -\infty$. This completes the proof of Proposition 4

PROBLEMS

1. Find the phase portrait for the differential equation

$$x' = y - f(x), \qquad f(x) = x^3,$$
$$y' = -x.$$

2. Give a proof of Proposition 2.

3. (Hartman [9, Chapter 7, Theorem 10.2]) Find the phase portrait of the following differential equation and in particular show there is a unique nontrivial periodic solution:

$$x' = y - f(x),$$
$$y' = -g(x),$$

where all of the following are assumed:
(i) f, g are C^1;
(ii) $g(-x) = -g(x)$ and $xg(x) > 0$ for all $x \neq 0$;
(iii) $f(-x) = -f(x)$ and $f(x) < 0$ for $0 < x < a$;
(iv) for $x > a$, $f(x)$ is positive and increasing;
(v) $f(x) \to \infty$ as $x \to \infty$.
(*Hint*: Imitate the proof of the theorem in Section 3.)

4. (Hard!) Consider the equation

$$x' = y - f(x), \qquad f: \mathbf{R} \to \mathbf{R},\ C^1,$$
$$y' = -x.$$

Given f, how many periodic solutions does this system have? This would be interesting to know for many broad classes of functions f. Good results on this would probably make an interesting research article.

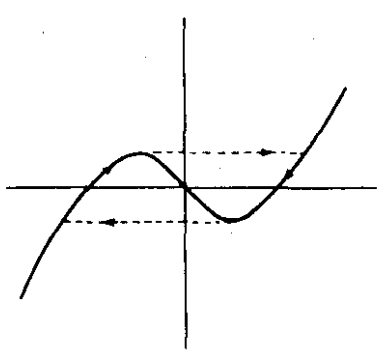


FIG. I

5. Consider the equation

$$x' = \mu(y - (x^3 - x)), \qquad \mu > 0,$$
$$y' = -x.$$

It has a unique nontrivial periodic solution γ_μ by Problem 3. Show that as $\mu \to \infty$, γ_μ tends to the closed curve consisting of two horizontal line segments and two arcs on $y = x^3 - x$ as in Fig. I.

§4. Hopf Bifurcation

Often one encounters a differential equation *with parameter*. Precisely, one is given a C^1 map $g_\mu: W \to E$ where W is an open set of the vector space E and μ is allowed to vary over some parameter space, say $\mu \in J = [-1, 1]$. Furthermore it is convenient to suppose that g_μ is differentiable in μ, or that the map

$$J \times W \to E, \qquad (\mu, x) \to g_\mu(x)$$

is C^1.

Then one considers the differential equation

$$x' = g_\mu(x) \qquad \text{on} \quad W. \tag{1}$$

One is especially concerned how the phase portrait of (1) changes as μ varies. A value μ_0 where there is a basic structural change in this phase portrait is called a *bifurcation* point. Rather than try to develop any sort of systematic bifurcation theory here, we will give one fundamental example, or a realization of what is called Hopf bifurcation.

Return to the circuit example of Section 1, where we now suppose that the resistor characteristic depends on a parameter μ and is denoted by $f_\mu: \mathbf{R} \to \mathbf{R}$, $-1 \leq \mu \leq 1$. (Maybe μ is the temperature of the resistor.) The physical behavior of the circuit is then described by the differential equation on $\mathbf{R}^2$:

$$\frac{dx}{dt} = y - f_\mu(x), \tag{2}$$
$$\frac{dy}{dt} = -x.$$

Consider as an example the special case where f_μ is described by

$$f_\mu(x) = x^3 - \mu x. \tag{2a}$$

Then we apply the results of Sections 2 and 3 to see what happens as μ is varied from -1 to 1.

For each μ, $-1 \leq \mu \leq 0$, the resistor is passive and the proposition of Section 2 implies that all solutions tend asymptotically to zero as $t \to \infty$. Physically the circuit is dead, in that after a period of transition all the currents and voltages

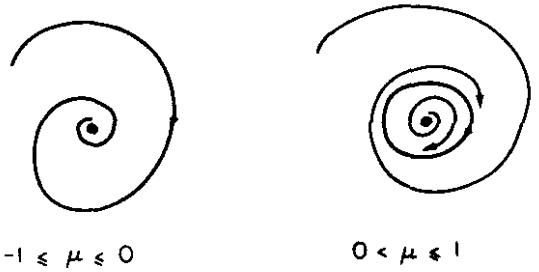


FIG. A. Bifurcation.

stay at 0 (or as close to 0 as we want). But note that as μ crosses 0, the circuit becomes alive. It will begin to oscillate. This follows from the fact that the analysis of Section 3 applies to (2) when $0 < \mu \leq 1$; in this case (2) will have a unique periodic solution γ_μ and the origin becomes a source. In fact every nontrivial solution tends to γ_μ as $t \to \infty$. Further elaboration of the ideas in Section 3 can be used to show that $\gamma_\mu \to 0$ as $\mu \to 0, \mu > 0$.

For (2), $\mu = 0$ is the bifurcation value of the parameter. The basic structure of the phase portrait changes as μ passes through the value 0. See Fig. A.

The mathematician E. Hopf proved that for fairly general one-parameter families of equations $x' = f_\mu(x)$, there must be a closed orbit for $\mu > \mu_0$ if the eigenvalue character of an equilibrium changes suddenly at μ_0 from a sink to a source.

PROBLEMS

1. Find all values of μ which are the bifurcation points for the linear differential equation:

$$\frac{dx}{dt} = \mu x + y,$$

$$\frac{dy}{dt} = x - 2y.$$

2. Prove the statement in the text that $\gamma_\mu \to 0$ as $\mu \to 0, \mu > 0$.

§5. More General Circuit Equations

We give here a way of finding the ordinary differential equations for a class of electrical networks or circuits. We consider networks made up of resistors, capacitors, and inductors. Later we discuss briefly the nature of these objects, called the *branches* of the circuit; at present it suffices to consider them as devices with two

terminals. The circuit is formed by connecting together various terminals. The connection points are called *nodes*.

Toward giving a mathematical description of the network, we define in $\mathbf{R}^3$ a *linear graph* which corresponds to the network. This linear graph consists of the following data:

(a) A finite set A of points (called nodes) in $\mathbf{R}^3$. The number of nodes is denoted by a, a typical node by α.

(b) A finite set B of line segments in $\mathbf{R}^3$ (called branches). The end points of a branch must be nodes. Distinct branches can meet only at a node. The number of branches is b; a typical branch is denoted by β.

We assume that each branch β is *oriented* in the sense that one is given a direction from one terminal to the other, say from a $(-)$ terminal β^- to a $(+)$ terminal β^+. The *boundary* of $\beta \in B$ is the set $\partial\beta = \beta^+ \cup \beta^-$

For the moment we ignore the exact nature of a branch, whether it is a resistor, capacitor, or inductor.

We suppose also that the set of nodes and the set of branches are ordered, so that it makes sense to speak of the kth branch, and so on.

A *current state* of the network will be some point $i = (i_1, \ldots, i_b) \in \mathbf{R}^b$ where i_k represents the current flowing through the kth branch at a certain moment. In this case we will often write $\mathcal{S}$ for $\mathbf{R}^b$.

The *Kirchhoff current law* or KCL states that the amount of current flowing into a node at a given moment is equal to the amount flowing out. The water analogy of Section 1 makes this plausible. We want to express this condition in a mathematical way which will be especially convenient for our development. Toward this end we construct a linear map $d\colon \mathcal{S} \to \mathfrak{D}$ where $\mathfrak{D}$ is the Cartesian space $\mathbf{R}^a$ (recall a is the number of nodes).

If $i \in \mathcal{S}$ is a current state and α is a node we define the αth coordinate of $di \in \mathfrak{D}$ to be

$$(di)_\alpha = \sum_{\beta \in B} \epsilon_{\alpha\beta} i_\beta,$$

where

$$\epsilon_{\alpha\beta} = \begin{cases} 1 & \text{if } \beta^+ = \alpha, \\ -1 & \text{if } \beta^- = \alpha, \\ 0 & \text{otherwise.} \end{cases}$$

One may interpret $(di)_\alpha$ as the net current flow into node α when the circuit is in the current state i.

Theorem 1 *A current state $i \in \mathcal{S}$ satisfies KCL if and only if $di = 0$.*

Proof. It is sufficient to check the condition for each node $\alpha \in A$. Thus $(di)_\alpha = 0$ if and only if

$$\sum_{\beta \in B} \epsilon_{\alpha\beta} i_\beta = 0,$$

or from the definition of $\epsilon_{\alpha\beta}$,

$$\sum_{\substack{\beta \in B \\ \beta^+ = \alpha}} i_\beta = \sum_{\substack{\beta \in B \\ \beta^- = \alpha}} i_\beta.$$

This last is just the expression of KCL at the node α. This proves the theorem.

Next, a *voltage state* of our network is defined to be a point $v = (v_1, \ldots, v_b) \in \mathbf{R}^b$, where in this context we denote $\mathbf{R}^b$ by $\mathcal{V}$. The kth coordinate v_k represents the voltage drop across the kth branch. The *Kirchhoff voltage law* (KVL) may be stated as asserting that there is a real function on the set of nodes, a *voltage potential* (given, for example, by voltmeter readings), $V: A \to R$, so that $v_\beta = V(\beta^+) - V(\beta^-)$ for each $\beta \in \mathcal{B}$.

To relate KCL to KVL and to prove what is called Tellegen's theorem in network theory, we make a short excursion into linear algebra. Let E, F be vector spaces whose dual vector spaces (Chapter 9) are denoted by E^*, F^*, respectively. If $u: E \to F$ is a linear transformation, then its *adjoint* or dual is a linear map $u^*: F^* \to E^*$ defined by $u^*(x)(y) = x(u(y))$, where $x \in F^*$, $y \in E$. (Here $u^*(x)$ is an element of E^* and maps $E \to \mathbf{R}$.)

Now let ϕ be the natural bilinear map defined on the Cartesian product vector space $E \times E^*$ with values in $\mathbf{R}$: if $(e, e^*) \in E \times E^*$, then $\phi(e, e^*) = e^*(e)$.

Proposition *Let $u: E \to F$ be a linear map and let $K = (\operatorname{Ker} u) \times ?\operatorname{Im} u^*) \subset E \times E^*$. Then ϕ is zero on K.*

Proof. Let $(e, e^*) \in K$ so that $u(e) = 0$ and $e^* = u^*y$ for some $y \in F^*$. Then

$$\phi(e, e^*) = \phi(e, u^*y) = (u^*y)(e) = y(u(e)) = 0.$$

This proves the proposition.

Remark. A further argument shows that $\dim K = \dim E$.

We return to the analysis of the voltage and current states of a network. It turns out to be useful, as we shall see presently, to identify the space $\mathcal{V}$ with the dual space $\mathcal{I}^*$ of $\mathcal{I}$. Mathematically this is no problem since both $\mathcal{V}$ and $\mathcal{I}^*$ are naturally isomorphic to $\mathbf{R}^b$. With this identification, the voltage which a voltage state $v \in \mathcal{I}^*$ assigns to the kth branch β is just $v(i_\beta)$, where $i_\beta \in \mathcal{I}$ is the vector where the kth coordinate is 1 and where other coordinates are 0.

We can now express KVL more elegantly:

Theorem 2 *A voltage state $v \in \mathcal{I}^*$ satisfies KVL if and only if it is in the image of the adjoint $d^*: \mathcal{D}^* \to \mathcal{I}^*$ of $d: \mathcal{I} \to \mathcal{D}$.*

Proof. Suppose v satisfies Kirchhoff's voltage law. Then there is a voltage potential V mapping the set of nodes to the real numbers, with $v(\beta) = V(\beta^+) - V(\beta^-)$ for each branch β. Recalling that $\mathcal{D} = \mathbf{R}^a$, a = number of nodes, we define

a linear map $\hat{V}: \mathcal{D} \to \mathbf{R}$ by

$$\hat{V}(x_1, \ldots, x_a) = \sum_{i=1}^{a} x_i V(\alpha_i).$$

Thus $\hat{V} \in \mathcal{D}^*$.

To see that $d^*\hat{V} = v$, consider first the current state $i_\beta \in \mathcal{I}$ defined above just before Theorem 2. Then

$$\begin{aligned}(d^*\hat{V})i_\beta &= V(di_\beta) \\ &= V(\beta^+) - V(\beta^-) \\ &= v(\beta).\end{aligned}$$

Since the states i_β, $\beta \in B$ form a basis for $\mathcal{I}$, this shows that $v = d^*\hat{V}$. Hence v is in the image of $\mathcal{I}^*$.

Conversely, assume that $v = d^*W$, $W \in \mathcal{D}^*$. For the kth node α define $V(\alpha) = W(f_\alpha)$, where $f_\alpha \in \mathcal{D}$ has kth coordinate 1 and all other coordinates 0. Then V is a voltage potential for v since the voltage which v assigns to the branch β is

$$\begin{aligned}v(i_\beta) &= d^*W(i_\beta) \\ &= W(f_{\beta^+}) - W(f_{\beta^-}) \\ &= V(\beta^+) - V(\beta^-).\end{aligned}$$

This completes the proof of Theorem 2.

The *space of unrestricted states* of the circuit is the Cartesian space $\mathcal{I} \times \mathcal{I}^*$. Those states which satisfy KCL and KVL constitute a linear subspace $K \subset \mathcal{I} \times \mathcal{I}^*$. By Theorems 1 and 2,

$$K = \operatorname{Ker} d \times \operatorname{Im} d^* \subset \mathcal{I} \times \mathcal{I}^*.$$

An actual or physical state of the network must lie in K.

The *power* ϕ in a network is a real function defined on the big state space $\mathcal{I} \times \mathcal{I}^*$ and in fact is just the natural pairing discussed earlier. Thus if $(i, v) \in \mathcal{I} \times \mathcal{I}^*$, the power $\phi(i, v) = v(i)$ or in terms of Cartesian coordinates

$$\phi(i, v) = \sum_{\beta} i_\beta v_\beta,$$

$$i = (i_1, \ldots, i_b), \qquad v = (v_1, \ldots, v_b).$$

The previous proposition gives us

Theorem 3 (Tellegen's theorem) *The power is zero on states satisfying Kirchhoff's laws.*

Mathematically this is the same thing as saying that $\phi: \mathcal{I} \times \mathcal{I}^* \to \mathbf{R}$ restricted to K is zero.

Now we describe in mathematical terms the three different types of devices in the network: the resistor, inductor, and capacitor. These devices impose conditions on the state, or on how the state changes in time, in the corresponding branch.

Each resistor ρ imposes a relation on the current and voltage in its branch. This relation might be an equation of the form $F_\rho(i_\rho, v_\rho) = 0$; but for simplicity we will assume that (i_ρ, v_ρ) satisfy $f_\rho(i_\rho) = v_\rho$ for some real-valued C^1 function f_ρ of a real variable. Thus f is a "generalized Ohm's law." The graph of f_ρ in the (i_ρ, v_ρ) plane is called the *characteristic* of the ρth resistor and is determined by the physical properties of the resistor. (Compare Section 1.) For example, a battery is a resistor in this context, and its characteristic is of the form $\{(i_\rho, v_\rho) \in \mathbf{R}^2 \mid v_\rho = \text{constant}\}$.

An inductor or capacitor does not impose conditions directly on the state, but only on how the state in that branch changes in time. In particular let λ be an inductor branch with current, voltage in that branch denoted by i_λ, v_λ. Then the λth inductor imposes the condition:

$$L_\lambda(i_\lambda)\frac{di_\lambda}{dt} = v_\lambda. \tag{1a}$$

Here L_λ is determined by the inductor and is called the inductance. It is assumed to be a C^1 positive function of i_λ.

Similarly a capacitor in the γth branch defines a C^1 positive function $v_\gamma \to C_\gamma(v_\gamma)$ called the capacitance; and the current, voltage in the γth branch satisfy

$$C_\gamma(v_\gamma)\frac{dv_\gamma}{dt} = i_\gamma. \tag{1b}$$

We now examine the resistor conditions more carefully. These are conditions on the states themselves and have an effect similar to Kirchhoff's laws in that they place physical restrictions on the space of all states, $\mathcal{I} \times \mathcal{I}^*$. We define Σ to be the subset of $\mathcal{I} \times \mathcal{I}^*$ consisting of states that satisfy the two Kirchhoff laws and the resistor conditions. This space Σ is called the space of *physical states* and is described by

$$\Sigma = \{(i, v) \in \mathcal{I} \times \mathcal{I}^* \mid (i, v) \in K, f_\rho(i_\rho) = v_\rho, \rho = 1, \ldots, r\}.$$

Here (i_ρ, v_ρ) denotes the components of i, v in the ρth branch and ρ varies over the resistor branches, r in number.

Under rather generic conditions, Σ will be a *manifold*, that is, the higher dimensional analog of a surface. Differential equations can be defined on manifolds; the capacitors and inductors in our circuit will determine differential equations on Σ whose corresponding flow Φ_t: $\Sigma \to \Sigma$ describes how a state changes with time.

Because we do not have at our disposal the notions of differentiable manifolds, we will make a simplifying assumption before proceeding to the differential equations of the circuit. This is the assumption that the space of currents in the inductors and voltages in the capacitors may be used to give coordinates to Σ. We make this more precise.

Let $\mathcal{L}$ be the space of all currents in the inductor branches, so that $\mathcal{L}$ is naturally isomorphic to $\mathbf{R}^l$, where l is the number of inductors. A point i of $\mathcal{L}$ will be denoted by $i = (i_1, \ldots, i_l)$ where i_λ is the current in the λth branch. There is a natural map (a projection) i_L: $\mathcal{I} \to \mathcal{L}$ which just sends a current state into its components in the inductors.

Similarly we let $\mathcal{C}^*$ be the space of all voltages in the capacitor branches so that $\mathcal{C}^*$ is isomorphic to $\mathbf{R}^c$, where c is the number of capacitors. Also v_C: $\mathcal{I}^* \to \mathcal{C}^*$ will denote the corresponding projection.

Consider the map $i_L \times v_C$: $\mathcal{I} \times \mathcal{I}^* \to \mathcal{L} \times \mathcal{C}^*$ restricted to $\Sigma \subset \mathcal{I} \times \mathcal{I}^*$. Call this map π: $\Sigma \to \mathcal{L} \times \mathcal{C}^*$. (It will help in following this rather abstract presentation to follow it along with the example in Section 1.)

Hypothesis *The map* π: $\Sigma \to \mathcal{L} \times \mathcal{C}^*$ *has an inverse which is a* C^1 *map*

$$\mathcal{L} \times \mathcal{C}^* \to \Sigma \subset \mathcal{I} \times \mathcal{I}^*.$$

Under this hypothesis, we may identify the space of physical states of the network with the space $\mathcal{L} \times \mathcal{C}^*$. This is convenient because, as we shall see, the differential equations of the circuit have a simple formulation on $\mathcal{L} \times \mathcal{C}^*$. In words the hypothesis may be stated: the current in the inductors and the voltages in the capacitors, via Kirchhoff's laws and the laws of the resistor characteristics, determine the currents and voltages in all the branches.

Although this hypothesis is strong, it makes some sense when one realizes that the "dimension" of Σ should be expected to be the same as the dimension of $\mathcal{L} \times \mathcal{C}^*$. This follows from the remark after the proposition on dim K, and the fact that Σ is defined by r additional equations.

To state the equations in this case we define a function P: $\mathcal{I} \times \mathcal{I}^* \to \mathbf{R}$ called the *mixed potential*. We will follow the convention that indices ρ refer to resistor branches and sums over such ρ means summation over the resistor branches. Similarly λ is used for inductor branches and γ for capacitor branches. Then P: $\mathcal{I} \times \mathcal{I}^* \to \mathbf{R}$ is defined by

$$P(i, v) = \sum_\gamma i_\gamma v_\gamma + \sum_\rho \int f_\rho(i_\rho)\, di_\rho.$$

Here the integral refers to the indefinite integral so that P is defined only up to an arbitrary constant. Now P by restriction may be considered as a map P: $\Sigma \to \mathbf{R}$ and finally by our hypothesis may even be considered as a map

$$P\colon \mathcal{L} \times \mathcal{C}^* \to \mathbf{R}.$$

(By an "abuse of language" we use the same letter P for all three maps.)

Now assume we have a particular circuit of the type we have been considering. At a given instant t_0 the circuit is in a particular current-voltage state. The states will change as time goes on. In this way a curve in $\mathcal{I} \times \mathcal{I}^*$ is obtained, depending on the initial state of the circuit.

The components $i_\beta(t)$, $v_\beta(t)$, $\beta \in B$ of this curve must satisfy the conditions imposed by Kirchhoff's laws and the resistor characteristics; that is, they must be in Σ. In addition at each instant of time the components di_λ/dt and dv_γ/dt of the tangent vectors of the curve must satisfy the relations imposed by (1a) and (1b). A curve satisfying these conditions we call a *physical trajectory*.

If the circuit satisfies our special hypothesis, each physical trajectory is identified with a curve in $\mathcal{L} \times \mathcal{C}^*$. The following theorem says that the curves so obtained are exactly the solution curves of a certain system of differential equations in $\mathcal{L} \times \mathcal{C}^*$:

Theorem 4 (Brayton–Moser) *Each physical trajectory of an electrical circuit satisfying the special hypothesis is a solution curve of the system*

$$L_\lambda(i_\lambda)\frac{di_\lambda}{dt} = -\frac{\partial P}{\partial i_\lambda},$$

$$C_\gamma(v_\gamma)\frac{dv_\gamma}{dt} = \frac{\partial P}{\partial v_\gamma},$$

where λ and γ run through all inductors and capacitors of the circuit respectively. Conversely, every solution curve to these equations is a physical trajectory.

Here P is the map $\mathcal{L} \times \mathcal{C}^* \to \mathbf{R}$ defined above. The right-hand sides of the differential equations are thus functions of all the i_λ, v_γ.

Proof. Consider an *arbitrary* C^1 curve in $\mathcal{L} \times \mathcal{C}^*$. Because of our hypothesis we identify $\mathcal{L} \times \mathcal{C}^*$ with $\Sigma \subset \mathcal{g} \times \mathcal{g}^*$; hence we write the curve

$$t \to (i(t), v(t)) \in \mathcal{g} \times \mathcal{g}^*.$$

By Kirchhoff's law (Theorem 1) $i(t) \in \operatorname{Ker} d$. Hence $i'(t) \in \operatorname{Ker} d$. By Theorem 2 $v(t) \in \operatorname{Im} d^*$. By Tellegen's theorem, for all t

$$\sum_{\beta \in \mathcal{B}} v_\beta(t) i_\beta(t) = 0.$$

We rewrite this as

$$\sum v_\rho i_\rho + \sum v_\lambda i_\lambda + \sum v_\gamma i_\gamma = 0.$$

From Leibniz' rule we get

$$\sum v_\gamma i_\gamma' = (\sum v_\gamma i_\gamma)' - \sum i_\gamma v_\gamma'.$$

Substituting this into the preceding equation gives

$$-\sum v_\lambda i_\lambda' + \sum i_\gamma v_\gamma' = (\sum i_\gamma v_\gamma)' + \sum v_\rho i_\rho' = \frac{dP}{dt},$$

from the definition of P and the generalized Ohm's laws. By the chain rule

$$\frac{dP}{dt} = \sum \frac{\partial P}{\partial i_\lambda} i_\lambda' + \sum \frac{\partial P}{\partial v_\gamma} v_\gamma'.$$

From the last two equations we find

$$\sum \left(\frac{\partial P}{\partial i_\lambda} + v_\lambda\right) i_\lambda' + \sum \left(\frac{\partial P}{\partial v_\gamma} - i_\gamma\right) v_\gamma' = 0.$$

Since i_λ' and v_λ' can take any values,

$$\frac{\partial P}{\partial i_\lambda} = -v_\lambda, \qquad \frac{\partial P}{\partial v_\gamma} = i_\gamma.$$

The theorem now follows from (1a) and (1b).

Some remarks on this theorem are in order. First, one can follow this development for the example of Section 1 to bring the generality of the above down to earth. Secondly, note that if there are either no inductors or no capacitors, the Brayton–Moser equations have many features of gradient equations and much of the material of Chapter 9 can be applied; see Problem 9. In the more general case the equations have the character of a gradient with respect to an indefinite metric.

We add some final remarks on an energy theorem. Suppose for simplicity that all the L_λ and C_γ are constant and let

$$W: \mathcal{L} \times \mathcal{C}^* \to \mathbf{R}$$

be the function $W(i, v) = \frac{1}{2}\sum_\lambda L_\lambda i_\lambda^2 + \frac{1}{2}\sum_\gamma C_\gamma v_\gamma^2$. Thus W has the form of a norm square and its level surfaces are generalized ellipsoids; W may be interpreted as the energy in the inductor and capacitor branches. Define P_r: $\mathcal{L} \times \mathcal{C}^* \to \mathbf{R}$ (power in the resistors) to be the composition

$$\mathcal{L} \times \mathcal{C}^* \to \Sigma \subset \mathcal{g} \times \mathcal{g}^* \xrightarrow{\tilde{P}_r} \mathbf{R},$$

where $\tilde{P}_r(i, v) = \sum i_\rho v_\rho$ (summed over resistor branches). We state without proof:

Theorem 5 *Let ϕ: $I \to \mathcal{L} \times \mathcal{C}^*$ be any solution of the equations of the previous theorem. Then*

$$\frac{d}{dt}(W\phi(t)) = -P_r(\phi(t)).$$

Theorem 5 may be interpreted as asserting that in a circuit the energy in the inductors and capacitors varies according to power dissipated in the resistors.

See the early sections where W appeared and was used in the analysis of Lienard's equation. Theorem 5 provides criteria for asymptotic stability in circuits.

PROBLEMS

1. Let N be a finite set and $P \subset N \times N$ a symmetric binary relation on N (that is, $(x, y) \in P$ if $(y, x) \in P$). Suppose $x \neq y$ for all $(x, y) \in P$. Show that there is a linear graph in $\mathbf{R}^3$ whose nodes are in one-to-one correspondence with N, such that the two nodes corresponding to x, y are joined by a branch if and only if $(x, y) \in N$.

2. Show that Kirchhoff's voltage law as stated in the text is equivalent to the following condition ("the voltage drop around a loop is zero"): Let $\alpha_0, \alpha_1, \ldots, \alpha_k = \alpha_0$ be nodes such that α_m and α_{m-1} are end points of a branch β_m, $m = 1, \ldots, k$. Then

$$\sum_{m=1}^{k} \epsilon_m v(\beta_m) = 0,$$

where $\epsilon_m = \pm 1$ according as $(\beta_m)^+ = \alpha_m$ or α_{m-1}.

3. Prove that dim K = dim E (see the proposition in the text and the remark after it).

4. Prove Theorem 5.

5. Consider resistors whose characteristic is of the form $F(i_\rho, v_\rho) = 0$, where F is a real-valued C^1 function. Show that an *RLC* circuit (Fig. A) with this kind of resistor satisfies the special hypothesis if and only if the resistor is current controlled, that is, F has the form

$$F(i_\rho, v_\rho) = v_\rho - f(i_\rho).$$

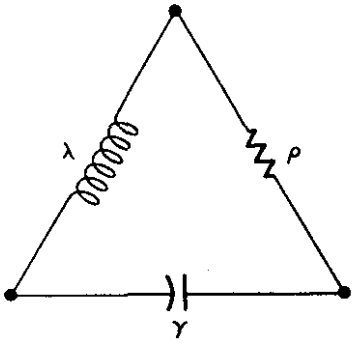


FIG. A

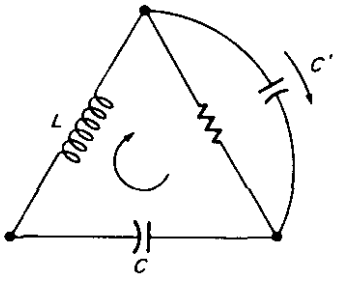


FIG. B

6. Show that the differential equations for the circuit in Fig. B are given by:

$$L \frac{di_\lambda}{dt} = -(v_\gamma + v_{\gamma'}),$$

$$C \frac{dv_\gamma}{dt} = i_\lambda,$$

$$C' \frac{dv_{\gamma'}}{dt} = i_\lambda - f(v_{\gamma'}).$$

Here $i = f(v)$ gives the resistor characteristic.

7. Suppose given a circuit satisfying the basic hypothesis of this section and all the other assumptions except that the characteristic of one resistor is given by a voltage-controlled characteristic $i = f(v)$, not necessarily current controlled. Show that if the corresponding term of the mixed potential P is replaced by $\int vf'(v)\, dv$, then Theorem 4 is still true.

8. Find the differential equations for this circuit (Brayton) (Fig. C). Here ||| denotes a battery (resistor with characteristic: v = const.), ⁄\⁄\⁄ denotes a

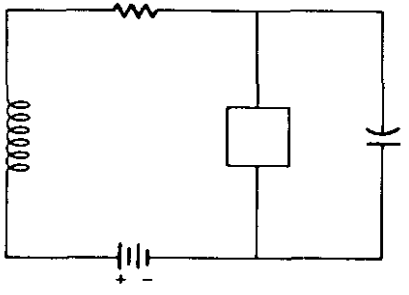


FIG. C

linear resistor, and the box is a resistor with characteristic given by $i = f(v)$. Find the mixed potential and the phase portrait for some choice of f. See Problem 7.

9. We refer to the Brayton–Moser equations. Suppose there are no capacitors.
   (a) Show that the function P: $\mathcal{L} \to \mathbf{R}$ decreases along nonequilibrium trajectories of the Brayton–Moser equations.
   (b) Let n be the number of inductors. If each function L_γ is a constant, find an inner product on $\mathbf{R}^n = \mathcal{L}$ which makes the vector

$$\left(\frac{1}{L_1}\frac{\partial P}{\partial i_1}, \ldots, \frac{1}{L_n}\frac{\partial P}{\partial i_n}\right)$$

   the gradient of P in the sense of Chapter 9, Section 5.

Notes

This chapter follows to a large extent "Mathematical foundations of electrical circuits" by Smale in the *Journal of Differential Geometry* [22]. The undergraduate text on electrical circuit theory by Desoer and Kuh [5] is excellent for a treatment of many related subjects. Hartman's book [9], mentioned also in Chapter 11, goes extensively into the material of our Sections 2 and 3 with many historical references. Lefschetz's book *Differential Equations, Geometrical Theory* [14] also discusses these nonlinear planar equations. Van der Pol himself related his equation to heartbeat and recently E. C. Zeeman has done very interesting work on this subject. For some physical background of circuit theory, one can see *The Feynman Lectures on Physics* [6].

Chapter 11

The Poincaré–Bendixson Theorem

We have already seen how periodic solutions in planar dynamical systems play an important role in electrical circuit theory. In fact the periodic solution in Van der Pol's equation, coming from the simple circuit equation in the previous chapter, has features that go well beyond circuit theory. This periodic solution is a "limit cycle," a concept we make precise in this chapter.

The Poincaré–Bendixson theorem gives a criterion for the detection of limit cycles in the plane; this criterion could have been used to find the Van der Pol oscillator. On the other hand, this approach would have missed the uniqueness.

Poincaré–Bendixson is a basic tool for understanding planar dynamical systems but for differential equations in higher dimensions it has no generalization or counterpart. Thus after the first two rather basic sections, we restrict ourselves to planar dynamical systems. The first section gives some properties of the limiting behavior of orbits on the level of abstract topological dynamics while in the next section we analyze the flow near nonequilibrium points of a dynamical system.

Throughout this chapter we consider a dynamical system on an open set W in a vector space E, that is, the flow ϕ_t defined by a C^1 vector field f: $W \to E$.

§1. Limit Sets

We recall from Chapter 9, Section 3 that $y \in W$ is an *ω-limit point* of $x \in W$ if there is a sequence $t_n \to \infty$ such that $\lim_{n\to\infty} \phi_{t_n}(x) = y$. The set of all ω-limit points of y is the *ω-limit set* $L_\omega(y)$. We define *α-limit points* and the *α-limit set* $L_\alpha(y)$ by replacing $t_n \to \infty$ with $t_n \to -\infty$ in the above definition. By a *limit set* we mean a set of the form $L_\omega(y)$ or $L_\alpha(y)$.

Here are some examples of limit sets. If $\bar{x}$ is an asymptotically stable equilibrium, it is the ω-limit set of every point in its basin (see Chapter 9, Section 2). Any

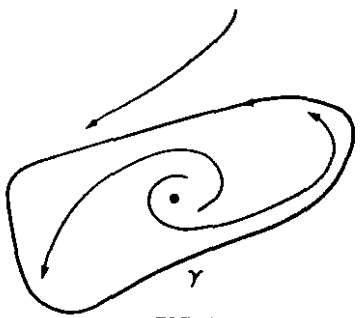


FIG. A

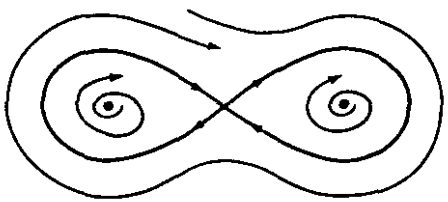


FIG. B

equilibrium is its own α-limit set and ω-limit set. A closed orbit is the α-limit and ω-limit set of every point on it. In the Van der Pol oscillator there is a unique closed orbit γ; it is the ω-limit of every point except the origin (Fig. A). The origin is the α-limit set of every point inside γ. *If y is outside γ, then $L_\alpha(y)$ is empty.*

There are examples of limit sets that are neither closed orbits nor equilibria, for example the figure 8 in the flow suggested by Fig. B. There are three equilibria, two sources, and one saddle. The figure 8 is the ω-limit set of all points outside it. The right half of the 8 is the ω-limit set of all points inside it except the equilibrium, and similarly for the left half.

In three dimensions there are extremely complicated examples of limit sets, although they are not easy to describe. In the plane, however, limit sets are fairly simple. In fact Fig. B is typical, in that one can show that a limit set other than a closed orbit or equilibrium is made up of equilibria and trajectories joining them. The Poincaré–Bendixson *theorem* says that *if* a compact limit set in the plane contains no equilibria it is a closed orbit.

We recall from Chapter 9 that a limit set is closed in W, and is invariant under the flow. We shall also need the following result:

Proposition (a) *If x and z are on the same trajectory, then $L_\omega(x) = L_\omega(z)$; similarly for α-limits.*

(b) *If D is a closed positively invariant set and $z \in D$, then $L_\omega(z) \subset D$; similarly for negatively invariant sets and α-limits.*

(c) *A closed invariant set, in particular a limit set, contains the α-limit and ω-limit sets of every point in it.*

Proof. (a) Suppose $y \in L_\omega(x)$, and $\phi_s(x) = z$. If $\phi_{t_n}(x) \to y$, then $\phi_{t_n - s}(z) \to y$. Hence $y \in L_\omega(z)$.

(b) If $t_n \to \infty$ and $\phi_{t_n}(z) \to y \in L_\omega(z)$, then $t_n \geq 0$ for sufficiently large n so that $\phi_{t_n}(z) \in D$. Hence $y \in \bar{D} = D$.

(c) Follows from (b).

PROBLEMS

1. Show that a compact limit set is connected (that is, not the union of two disjoint non-empty closed sets.

2. Identify $\mathbf{R}^4$ with $\mathbf{C}^2$ having two complex coordinates (w, z), and consider the linear system

$$(*) \qquad w' = 2\pi i w,$$

$$z' = 2\pi\theta i z,$$

where θ is an *irrational* real number.

(a) Put $\alpha = e^{2\pi\theta i}$ and show that the set $\{\alpha^n \mid n = 1, 2, \ldots\}$ is dense in the unit circle $C = \{z \in \mathbf{C} \,|\, |z| = 1\}$.

(b) Let ϕ_t be the flow of $(*)$. Show that for n an integer,

$$\phi_n(w, z) = (w, \alpha^n z).$$

(c) Let (w_0, z_0) belong to the torus $C \times C \subset \mathbf{C}^2$. Use (a), (b) to show that

$$L_\omega(w_0, z_0) = L_\alpha(w_0, z_0) = C \times C.$$

(d) Find L_ω and L_α of an arbitrary point of $\mathbf{C}^2$.

3. Find a linear system on $\mathbf{R}^{2k} = \mathbf{C}^k$ such that if a belongs to the k-torus $C \times \cdots \times C \subset \mathbf{C}^k$, then

$$L_\omega(a) = L_\alpha(a) = C^k.$$

4. In Problem 2, suppose instead that θ is *rational*. Identify L_ω and L_α of every point.

5. Let X be a nonempty compact invariant set for a C^1 dynamical system. Suppose that X is *minimal*, that is, X contains no compact invariant nonempty proper subset. Prove the following:

(a) Every trajectory in X is dense in X;

(b) $L_\alpha(x) = L_\omega(x) = X$ for each $x \in X$;

(c) For any (relatively) open set $U \subset X$, there is a number $P > 0$ such that for any $x \in X$, $t_0 \in \mathbf{R}$, there exists t such that $\phi_t(x) \in U$ and $|t - t_0| < P$;

(d) For any x, y in X there are sequences $t_n \to \infty$, $s_n \to -\infty$ such that

$$|t_n - t_{n+1}| < 2P, \qquad |s_n - s_{n+1}| < 2P,$$

and

$$\phi_{t_n}(x) \to y, \qquad \phi_{s_n}(x) \to y.$$

6. Let X be a closed invariant set for a C^1 dynamical system on $\mathbf{R}^n$, such that $\phi_t(x)$ is defined for all $t \in \mathbf{R}$, $x \in X$. Suppose that $L_\alpha(x) = L_\omega(x) = X$ for all $x \in X$. Prove that X is compact.

§2. Local Sections and Flow Boxes

We consider again the flow ϕ_t of the C^1 vector field $f: W \to E$. Suppose the origin $0 \in E$ belongs to W.

A *local section* at 0 of f is an open set S containing 0 in a hyperplane $H \subset E$ which is transverse to f. By a *hyperplane* we mean a linear subspace whose dimension is one less than dim E. To say that $S \subset H$ is *transverse* to f means that $f(x) \notin H$ for all $x \in S$. In particular $f(x) \neq 0$ for $x \in S$.

Our first use of a local section at 0 will be to construct a "flow box" in a neighborhood of 0. A flow box gives a complete description of a flow in a neighborhood of any nonequilibrium point of any flow, by means of special (nonlinear) coordinates. The description is simple: points move in parallel straight lines at constant speed.

We make this precise as follows. A *diffeomorphism* $\Psi: U \to V$ is a differentiable map from one open set of a vector space to another with a differentiable inverse. A *flow box* is a diffeomorphism

$$\mathbf{R} \times H \supset N \xrightarrow{\Psi} W$$

of a neighborhood N of $(0, 0)$ onto a neighborhood of 0 in W, which transforms the vector field $f: W \to E$ into the constant vector field $(1, 0)$ on $\mathbf{R} \times H$. The flow of f is thereby converted to a simple flow on $\mathbf{R} \times H$:

$$\psi_s(t, y) = (t + s, y).$$

The map Ψ is defined by

$$\Psi(t, y) = \phi_t(y),$$

for (t, y) in a sufficiently small neighborhood of $(0, 0)$ in $\mathbf{R} \times H$. One appeals to Chapter 15 to see that Ψ is a C^1 map. The derivative of Ψ at $(0, 0)$ is easily computed to be the linear map which is the identity on $0 \times H$, and on $\mathbf{R} = \mathbf{R} \times 0$ it sends 1 to $f(0)$. Since $f(0)$ is transverse to H, it follows that $D\Psi(0, 0)$ is an isomorphism. Hence by *the inverse function theorem* Ψ *maps an open neighborhood* N of $(0, 0)$ diffeomorphically onto a neighborhood V of 0 in E. We take N of the form

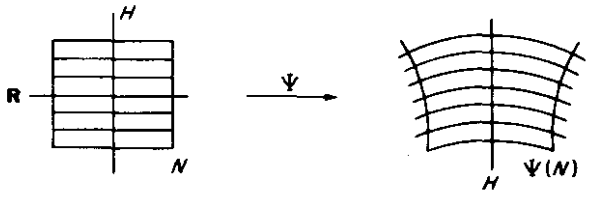


FIG. A. The flow box.

$S \times (-\sigma, \sigma)$, where $S \subset H$ is a section at 0 and $\sigma > 0$. In this case we sometimes write $V_\sigma = \Psi(N)$ and call V_σ a *flow box* at (or about) 0 in E. See Fig. A. An important property of a flow box is that if $x \in V_\sigma$, then $\phi_t(x) \in S$ for a unique $t \in (-\sigma, \sigma)$.

From the definition of Ψ it follows that if $\Psi^{-1}(p) = (s, y)$, then $\Psi^{-1}(\phi_t(p)) = (s + t, y)$ for sufficiently small $|s|, |t|$.

We remark that a flow box can be defined about any nonequilibrium point x_0. The assumption that $x_0 = 0$ is no real restriction since if x_0 is any point, one can replace $f(x)$ by $f(x - x_0)$ to convert the point to 0.

If S is a local section, the trajectory through a point z_0 (perhaps far from S) may reach $0 \in S$ in a certain time t_0; see Fig. B. We show that in a certain local sense, t_0 is a continuous function of z_0. More precisely:

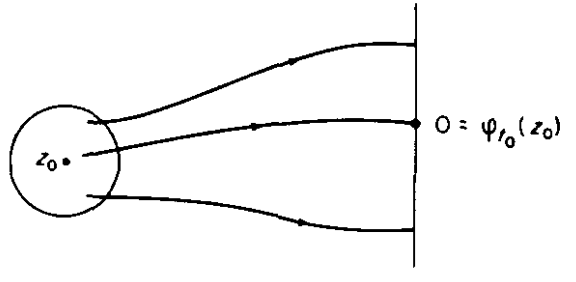


FIG. B

Proposition *Let S be a local section at 0 as above, and suppose $\phi_{t_0}(z_0) = 0$. There is an open set $U \subset W$ containing z_0 and a unique C^1 map $\tau: U \to \mathbf{R}$ such that $\tau(z_0) = t_0$ and*

$$\phi_{\tau(x)}(x) \in S$$

for all $x \in U$.

Proof. Let $h: E \to \mathbf{R}$ be a linear map whose kernel H is the hyperplane containing S. Then $h(f(0)) \neq 0$. The function

$$G(x, t) = h\phi_t(x)$$

is C^1, and

$$\frac{\partial G}{\partial t}(z_0, t_0) = h(f(0)) \neq 0.$$

By the implicit function theorem there is a unique C^1 map $x \to \tau(x) \in \mathbf{R}$ defined on a neighborhood U_1 of z_0 in W such that $\tau(z_0) = t_0$ and $G(x, \tau(x)) \equiv 0$. Hence $\phi_{\tau(x)}(x) \in H$; if $U \subset U_1$ is a sufficiently small neighborhood of z_0 then $\phi_{\tau(x)}(x) \in S$. This proves the proposition.

For later reference note that

$$D\tau(z_0) = -\left[\frac{\partial G}{\partial t}(z_0, t_0)\right]^{-1} \frac{\partial G}{\partial x}(z_0, t_0) = -\left[\frac{\partial G}{\partial t}(z_0, t_0)\right]^{-1} \cdot h \cdot D\phi_{t_0}(z_0).$$

§3. Monotone Sequences in Planar Dynamical Systems

We now restrict our discussion to planar dynamical systems.

Let $x_0, x_1, \ldots$ be a finite or infinite sequence of distinct points on the solution curve $C = \{\phi_t(x_0) \mid 0 \leq t \leq \alpha\}$. We say the sequence is *monotone along the trajectory* if $\phi_{t_n}(x_0) = x_n$ with $0 \leq t_1 < \cdots \leq \alpha$.

Let $y_0, y_1, \ldots$ be a finite or infinite sequence of points on a line segment I in $\mathbf{R}^2$. We say the sequence is *monotone along* I if the vector $y_n - y_0$ is a scalar multiple $\lambda_n(y_1 - y_0)$ with $1 < \lambda_2 < \lambda_3 < \ldots$ $n = 2, 3, \ldots$. Another way to say this is that y_n is between y_{n-1} and y_{n+1} in the natural order along I, $n = 1, 2, \ldots$.

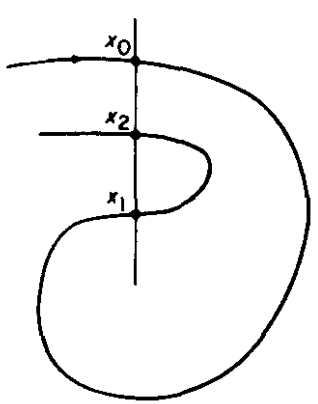


FIG. A

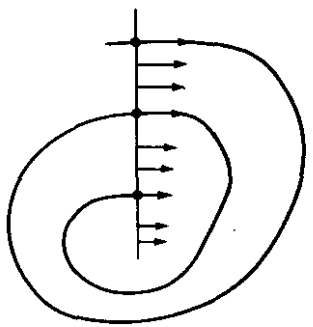


FIG. B

A sequence of points may be on the intersection of a solution curve and a segment I; they may be monotone along the solution curve but not along the segment, or vice versa; see Fig. A. However, *this is impossible if the segment is a local section.* Figure B shows an example; we suggest the reader experiment with paper and pencil!

Proposition 1 *Let S be a local section of a C^1 planar dynamical system and $y_0, y_1, y_2, \ldots$ a sequence of distinct points of S that are on the same solution curve C. If the sequence is monotone along C, it is also monotone along S.*

Proof. It suffices to consider three points y_0, y_1, y_2. Let Σ be the simple closed curve made up of the part B of C between y_0 and y_1 and the segment $T \subset S$ between y_0 and y_1. Let D be the closed bounded region bounded by Σ. We suppose that the trajectory of y_1 *leaves* D at y_1 (Fig. C); if it enters, the argument is similar.

We assert that at any point of T the trajectory leaves D. For it either leaves or enters because, T being transverse to the flow, it crosses the boundary of D. The set of points in T whose trajectory leaves D is a nonempty open subset $T_- \subset T$, by

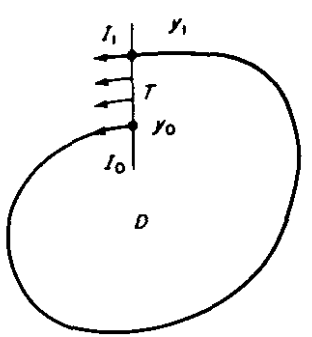


FIG. C

continuity of the flow; the set $T_+ \subset T$ where trajectories enter D is also open in T. Since T and T_+ are disjoint and $T = T_- \cup T_+$, it follows from connectedness of the interval that T_+ must be empty.

It follows that the complement of D is positively invariant. For no trajectory can enter D at a point of T; nor can it cross B, by uniqueness of solutions.

Therefore $\phi_t(y_1) \in \mathbf{R}^2 - D$ for all $t > 0$. In particular, $y_2 \in S - T$.

The set $S - T$ is the union of two half open intervals I_0 and I_1 with y_0 an endpoint of I_0 and y_1 an endpoint of I_1. One can draw an arc from a point $\phi_\epsilon(y_1)$ (with $\epsilon > 0$ very small) to a point of I_1, without crossing Σ. Therefore I_1 is outside D. Similarly I_0 is inside D. It follows that $y_2 \in I_1$ since it must be outside D. This shows that y_1 is between y_0 and y_2 in I, proving Proposition 1.

We come to an important property of limit points.

Proposition 2 *Let $y \in L_\omega(x) \cup L_\alpha(x)$. Then the trajectory of y crosses any local section at not more than one point.*

Proof. Suppose y_1 and y_2 are distinct points on the trajectory of y and S is a local section containing y_1 and y_2. Suppose $y \in L_\omega(x)$ (the argument for $L_\alpha(x)$ is similar). Then $y_k \in L_\omega(x)$, $k = 1, 2$. Let $V_{(k)}$ be flow boxes at y_k defined by some intervals $J_k \subset S$; we assume J_1 and J_2 disjoint (Fig. D). The trajectory of x enters $V_{(k)}$ infinitely often; hence it crosses J_k infinitely often. Hence there is a sequence

$$a_1, b_1, a_2, b_2, a_3, b_3, \ldots,$$

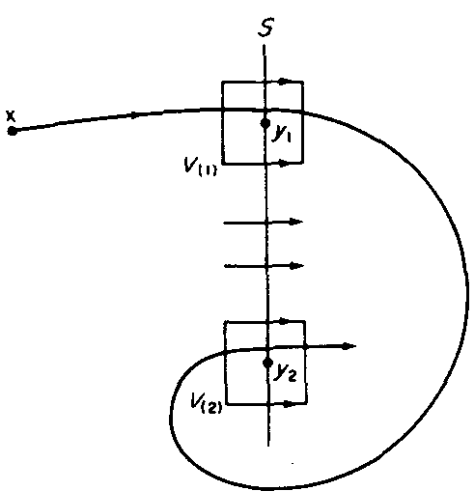


FIG. D

which is monotone along the trajectory of x, with $a_n \in J_1$, $b_n \in J_2$, $n = 1, 2, \ldots$. But such a sequence cannot be monotone along S since J_1 and J_2 are disjoint, contradicting Proposition 1.

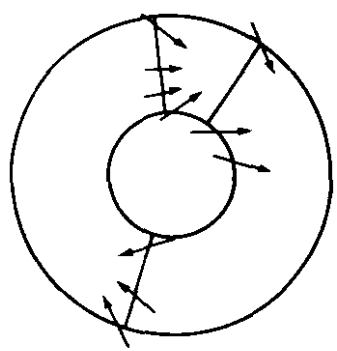
PROBLEMS

1. Let $A \subset \mathbf{R}^2$ be the annulus

$$A = \{z \in \mathbf{R}^2 \mid 1 \leq |z| \leq 2\}.$$

Let f be a C^1 vector field on a neighborhood of A which points inward along the two boundary circles of A. Suppose also that every radial segment of A is local section (Fig. E). Prove there is a periodic trajectory in A.

FIG. E



(*Hint*: Let S be a radial segment. Show that if $z \in S$ then $\phi_t(z) \in S$ for a smallest $t = t(z) > 0$. Consider the map $S \to S$ given by $z \mapsto \phi_{t(z)}(z)$.)

2. Show that a closed orbit of a planar C^1 dynamical system meets a local section in at most one point.

3. Let $W \subset \mathbf{R}^2$ be open and let $f: W \to \mathbf{R}^2$ be a C^1 vector field with no equilibria. Let $J \subset W$ be an open line segment whose end points are in the boundary of W. Suppose J is a *global section* in the sense that f is transverse to J, and for any $x \in W$ there exists $s < 0$ and $t > 0$ such that $\phi_s(x) \in J$ and $\phi_t(x) \in J$. Prove the following statements.
   (a) For any $x \in J$ let $\tau(x) \in \mathbf{R}$ be the smallest positive number such that $F(x) = \phi_{\tau(x)} \in J$; this map $F: J \to J$ is C^1 and has a C^1 inverse.
   (b) A point $x \in J$ lies on a closed orbit if and only if $F(x) = x$.
   (c) Every limit set is a closed orbit.

4. Let x be a *recurrent* point of a C^1 planar dynamical system, that is, there is a sequence $t_n \to \pm\infty$ such that

$$\phi_{t_n}(x) \to x.$$

(a) Prove that either x is an equilibrium or x lies on a closed orbit.
(b) Show by example that there can be a recurrent point for higher dimensional systems that is not an equilibrium and does not lie on a closed orbit.

§4. The Poincaré–Bendixson Theorem

By a *closed orbit* of a dynamical system we mean the image of a nontrivial periodic solution. Thus a trajectory γ is a closed orbit if γ is not an equilibrium and $\phi_p(x) = x$ for some $x \in \gamma$, $p \neq 0$. It follows that $\phi_{np}(y) = y$ for all $y \in \gamma$, $n = 0, \pm 1, \pm 2, \ldots$.

In this section we complete the proof of a celebrated result:

Theorem (Poincaré–Bendixson) *A nonempty compact limit set of a C^1 planar dynamical system, which contains no equilibrium point, is a closed orbit.*

Proof. Assume $L_\omega(x)$ is compact and $y \in L_\omega(x)$. (The case of α-limit sets is similar.) We show first that the trajectory of y is a closed orbit.

Since y belongs to the compact invariant set $L_\omega(x)$ we know that $L_\omega(y)$ is a nonempty subset of $L_\omega(x)$. Let $z \in L_\omega(y)$; let S be a local section at z, and N a flow box neighborhood of z about some open interval J, $z \in J \subset S$. By Proposition 2 of the previous section, the trajectory of y meets S at exactly one point. On the other hand, there is a sequence $t_n \to \infty$ such that $\phi_{t_n}(y) \to z$; hence infinitely many $\phi_{t_n}(y)$ belong to V. Therefore we can find $r, s \in \mathbf{R}$ such that $r > s$ and

$$\phi_r(y) \in S \cap V, \qquad \phi_s(y) \in S \cap V.$$

It follows that $\phi_r(y) = \phi_s(y)$; hence $\phi_{r-s}(y) = y$, $r - s > 0$. Since $L_\omega(x)$ contains no equilibrium, y belongs to closed orbit.

It remains to prove that if γ is a closed orbit in $L_\omega(x)$ then $\gamma = L_\omega(x)$. It is enough to show that

$$\lim_{t\to\infty} d(\phi_t(x), \gamma) = 0,$$

where $d(\phi_t(x), \gamma)$ is the distance from x to the compact set γ (that is, the distance from $\phi_t(x)$ to the nearest point of γ).

Let S be a local section at $z \in \gamma$, so small that $S \cap \gamma = z$. By looking at a flow box V_ϵ near z we see that there is a sequence $t_0 < t_1 < \cdots$ such that

$$\phi_{t_n}(x) \in S,$$

$$\phi_{t_n}(x) \to z,$$

$$\phi_t(x) \notin S \quad \text{for} \quad t_{n-1} < t < t_n, \quad n = 1, 2, \ldots.$$

Put $x_n = \phi_{t_n}(x)$. By Proposition 1, Section 3, $x_n \to z$ monotonically in S.

There exists an upper bound for the set of positive numbers $t_{n+1} - t_n$. For suppose $\phi_\lambda(z) = z$, $\lambda > 0$. Then for x_n sufficiently near z, $\phi_\lambda(x_n) \in V_\epsilon$ and hence

$$\phi_{\lambda+t}(x_n) \in S$$

for some $t \in [-\epsilon, \epsilon]$. Thus

$$t_{n+1} - t_n \leq \lambda + \epsilon.$$

Let $\beta > 0$. From Chapter 8, there exists $\delta > 0$ such that if $|x_n - u| < \delta$ and $|t| \leq \lambda + \epsilon$ then $|\phi_t(x_n) - \phi_t(u)| < \beta$.

Let n_0 be so large that $|x_n - z| < \delta$ for all $n \geq n_0$. Then

$$|\phi_t(x_n) - \phi_t(z)| < \beta$$

if $|t| \leq \lambda + \epsilon$ and $n \geq n_0$. Now let $t \geq t_{n_0}$. Let $n \geq n_0$ be such that

$$t_n \leq t \leq t_{n+1}.$$

Then

$$\begin{aligned} d(\phi_t(x), \gamma) &\leq |\phi_t(x) - \phi_{t-t_n}(z)| \\ &= |\phi_{t-t_n}(x_n) - \phi_{t-t_n}(z)| \\ &< \beta \end{aligned}$$

since $|t - t_n| \leq \lambda + \epsilon$. The proof of the Poincaré–Bendixson theorem is complete.

PROBLEMS

1. Consider a C^1 dynamical system in $\mathbf{R}^2$ having only a finite number of equilibria.
   (a) Show that every limit set is either a closed orbit or the union of equilibria and trajectories $\phi_t(x)$ such that $\lim_{t\to\infty} \phi_t(x)$ and $\lim_{t\to-\infty} \phi_t(x)$ are equilibria.
   (b) Show by example (draw a picture) that the number of distinct trajectories in $L_\omega(x)$ may be infinite.
2. Let γ be a closed orbit of a C^1 dynamical system on an open set in $\mathbf{R}^2$. Let λ be the period of γ. Let $\{\gamma_n\}$ be a sequence of closed orbits; suppose the period

of γ_n is λ_n. If there are points $x_n \in \gamma_n$ such that $x_n \to x \in \gamma$, prove that $\lambda_n \to \lambda$. (This result can be false for higher dimensional systems. It is true, however, that if $\lambda_n \to \mu$, then μ is an integer multiple of λ.)

§5. Applications of the Poincaré–Bendixson Theorem

We continue to suppose given a planar dynamical system.

Definition A *limit cycle* is a closed orbit γ such that $\gamma \subset L_\omega(x)$ or $\gamma \subset L_\alpha(x)$ for some $x \notin \gamma$. In the first case γ is called an ω-limit cycle; in the second case, an α-limit cycle.

In the proof of the Poincaré–Bendixson theorem it was shown that limit cycles enjoy a certain property not shared by other closed orbits: if γ is an ω-limit cycle, there exists $x \notin \gamma$ such that

$$\lim_{t\to\infty} d(\phi_t(x), \gamma) = 0.$$

For an α-limit cycle replace ∞ by $-\infty$. Geometrically this means that some trajectory spirals toward γ as $t \to \infty$ (for ω-limit cycles) or as $t \to -\infty$ (for α-limit cycles). See Fig. A.

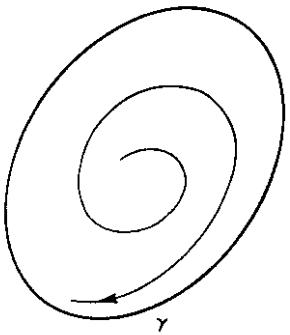
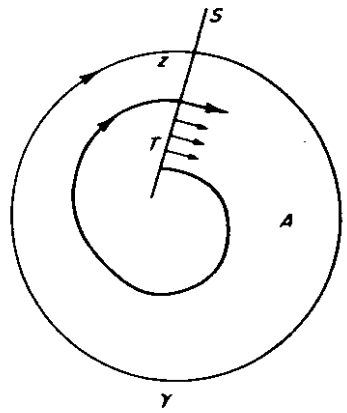


FIG. A. γ is an ω-limit cycle.

Limit cycles possess a kind of one-sided stability. Suppose γ is an ω-limit cycle and let $\phi_t(x)$ spiral toward γ as $t \to \infty$. Let S be a local section at $z \in \gamma$. Then there will be an interval $T \subset S$ disjoint from γ bounded by $\phi_{t_0}(x)$, $\phi_{t_1}(x)$, with $t_0 < t_1$ and not meeting the trajectory of x for $t_0 < t < t_1$ (Fig. B). The region A bounded by γ, T and the curve

$$\{\phi_t(x) \mid t_0 \le t \le t_1\}$$

FIG. B



is positively invariant, as is the set $B = A - \gamma$. It is easy to see that $\phi_t(y)$ spirals toward γ for all $y \in B$. A useful consequence of this is

Proposition 1 *Let γ be an ω-limit cycle. If $\gamma = L_\omega(x)$, $x \notin \gamma$ then x has a neighborhood V such that $\gamma = L_\omega(y)$ for all $y \in V$. In other words, the set*

$$A = \{y \mid \gamma = L_\omega(y)\} - \gamma$$

is open.

Proof. For sufficiently large $t > 0$, $\phi_t(x)$ is in the interior of the set A described above. Hence $\phi_t(y) \in A$ for y sufficiently close to x. This implies the proposition.

A similar result holds for α-limit cycles.

Theorem 1 *A nonempty compact set K that is positively or negatively invariant contains either a limit cycle or an equilibrium.*

Proof. Suppose for example that K is positively invariant. If $x \in K$, then $L_\omega(x)$ is a nonempty subset of K; apply Poincaré–Bendixson.

The next result exploits the spiraling property of limit cycles.

Proposition 2 *Let γ be a closed orbit and suppose that the domain W of the dynamical system includes the whole open region U enclosed by γ. Then U contains either an equilibrium or a limit cycle.*

Proof. Let D be the compact set $U \cup \gamma$. Then D is invariant since no trajectory from U can cross γ. If U contains no limit cycle and no equilibrium, then, for any $x \in U$,

$$L_\omega(x) = L_\alpha(x) = \gamma$$

by Poincaré–Bendixson. If S is a local section at a point $z \in \gamma$, there are sequences $t_n \to \infty$, $s_n \to -\infty$ such that

$$\phi_{t_n}(x) \in S, \qquad \phi_{t_n}(x) \to z,$$

and

$$\phi_{s_n}(x) \in S, \qquad \phi_{s_n}(x) \to z.$$

But this leads to a contradiction of the proposition in Section 3 on monotone sequences.

Actually this last result can be considerably sharpened:

Theorem 2 *Let γ be a closed orbit enclosing an open set U contained in the domain W of the dynamical system. Then U contains an equilibrium.*

Proof. Suppose U contains no equilibrium. If $x_n \to x$ in U and each x_n lies on a closed orbit, then x must lie on a closed orbit. For otherwise the trajectory of x would spiral toward a limit cycle, and by Proposition 1 so would the trajectory of some x_n.

Let $A \geq 0$ be the greatest lower bound of the areas of regions enclosed by closed orbits in U. Let $\{\gamma_n\}$ be a sequence of closed orbits enclosing regions of areas A_n such that $\lim_{n\to\infty} A_n = A$. Let $x_n \in \gamma_n$. Since $\gamma \cup U$ is compact we may assume $x_n \to x \in U$. Then if U contains no equilibrium, x lies on a closed orbit β of area $A(\beta)$. The usual section argument shows that as $n \to \infty$, γ_n gets arbitrarily close to β and hence the area $A_n - A(\beta)$, of the region between γ_n and β, goes to 0. Thus $A(\beta) = A$.

We have shown that if U contains no equilibrium, it contains a closed orbit β enclosing a region of minimal area. Then the region enclosed by β contains neither an equilibrium nor a closed orbit, contradicting Proposition 2.

The following result uses the spiraling properties of limit cycles in a subtle way.

Theorem 3 *Let H be a first integral of a planar C^1 dynamical system (that is, H is a real-valued function that is constant on trajectories). If H is not constant on any open set, then there are no limit cycles.*

Proof. Suppose there is a limit cycle γ; let $c \in \mathbf{R}$ be the constant value of H on γ. If $x(t)$ is a trajectory that spirals toward γ, then $H(x(t)) \equiv c$ by continuity of H. In Proposition 1 we found an open set whose trajectories spiral toward γ; thus H is constant on an open set.

PROBLEMS

1. The celebrated *Brouwer fixed point theorem* states that any continuous map f of the closed unit ball

$$D^n = \{x \in \mathbf{R}^n \mid |x| = 1\}$$

into itself has a fixed point (that is, $f(x) = x$ for some x).
   (a) Prove this for $n = 2$, assuming that f is C^1, by finding an equilibrium for the vector field $g(x) = f(x) - x$.
   (b) Prove Brouwer's theorem for $n = 2$ using the fact that any continuous map is the uniform limit of C^1 maps.

2. Let f be a C^1 vector field on a neighborhood of the annulus

$$A = \{x \in \mathbf{R}^2 \mid 1 \leq |x| \leq 2\}.$$

Suppose that f has no zeros and that f is transverse to the boundary, pointing inward.
   (a) Prove there is a closed orbit. (Notice that the hypothesis is weaker than in Problem 1, Section 3.)
   (b) If there are exactly seven closed orbits, show that one of them has orbits spiraling toward it from both sides.

3. Let $f: \mathbf{R}^2 \to \mathbf{R}^2$ be a C^1 vector field with no zeros. Suppose the flow ϕ_t generated by f preserves area (that is, if S is any open set, the area of $\phi_t(S)$ is independent of t). Show that every trajectory is a closed set.

4. Let f be a C^1 vector field on a neighborhood of the annulus A of Problem 2. Suppose that for every boundary point x, $f(x)$ is a nonzero vector tangent to the boundary.
   (a) Sketch the possible phase portraits in A under the further assumption that there are no equilibria and no closed orbits besides the boundary circles. Include the case where the boundary trajectories have opposite orientations.
   (b) Suppose the boundary trajectories are oppositely oriented and that the flow preserves area. Show that A contains an equilibrium.

5. Let f and g be C^1 vector fields on $\mathbf{R}^2$ such that $\langle f(x), g(x) \rangle = 0$ for all x. If f has a closed orbit, prove that g has a zero.

6. Let f be a C^1 vector field on an open set $W \subset \mathbf{R}^2$ and $H: W \to \mathbf{R}$ a C^1 function such that

$$DH(x)f(x) = 0$$

for all x. Prove that:
   (a) H is constant on solution curves of $x' = f(x)$;

(b) $DH(x) = 0$ if x belongs to a limit cycle;
(c) If x belongs to a compact invariant set on which DH is never 0, then x lies on a closed orbit.

Notes

P. Hartman's *Ordinary Differential Equations* [9], a good but advanced book, covers extensively the material in this chapter.

It should be noted that our discussion implicitly used the fact that a closed curve in $\mathbf{R}^2$ which does not intersect itself must separate $\mathbf{R}^2$ into two connected regions, a bounded one and an unbounded one. This theorem, the Jordan curve theorem, while naïvely obvious, needs mathematical proof. One can be found in Newman's *Topology of Plane Sets* [17].

Chapter 12

Ecology

In this chapter we examine some nonlinear two dimensional systems that have been used as mathematical models of the growth of two species sharing a common environment. In the first section, which treats only a single species, various mathematical assumptions on the growth rate are discussed. These are intended to capture mathematically, in the simplest way, the dependence of the growth rate on food supply and the negative effects of overcrowding.

In Section 2, the simplest types of equations that model a predator–prey ecology are investigated: the object is to find out the long-run qualitative behavior of trajectories. A more sophisticated approach is used in Section 3 to study two competing species. Instead of explicit formulas for the equations, certain qualitative assumptions are made about the form of the equations. (A similar approach to predator and prey is outlined in one of the problems.) Such assumptions are more plausible than any set of particular equations can be; one has correspondingly more confidence in the conclusions reached.

An interesting phenomenon observed in Section 3 is bifurcation of behavior. Mathematically this means that a slight quantitative change in initial conditions leads to a large qualitative difference in long-term behavior (because of a change of ω-limit sets). Such bifurcations, also called "catastrophes," occur in many applications of nonlinear systems; several recent theories in mathematical biology have been based on bifurcation theory.

§1. One Species

The birth rate of a human population is usually given in terms of the number of births per thousand in one year. The number one thousand is used merely to avoid decimal places; instead of a birth rate of 17 per thousand one could just as

well speak of 0.017 per individual (although this is harder to visualize). Similarly, the period of one year is also only a convention; the birth rate could just as well be given in terms of a week, a second, or any other unit of time. Similar remarks apply to the death rate and to the *growth rate*, or birth rate minus death rate. The growth rate is thus the net change in population per unit of time divided by the total population at the beginning of the time period.

Suppose the population $y(t)$ at time t changes to $y + \Delta y$ in the time interval $[t, t + \Delta t]$. Then the (average) growth rate is

$$\frac{\Delta y}{y(t)\,\Delta t}.$$

In practice $y(t)$ is found only at such times $t_0, t_1, \ldots$ when population is counted; and its value is a nonnegative integer. We assume that y is extended (by interpolation or some other method) to a nonnegative real-valued function of a real variable. We assume that y has a continuous derivative.

Giving in to an irresistible mathematical urge, we form the limit

$$\lim_{\Delta t \to 0} \frac{\Delta y}{y\,\Delta t} = \frac{y'(t)}{y(t)}.$$

This function of t is the *growth rate* of the population at time t.

The simplest assumption is that of a *constant* growth rate α. This is the case if the number of births and deaths in a small time period Δt have a fixed ratio to the total population. These ratios will be linear functions of Δt, but independent of the size of the population. Thus the net change will be $\alpha y\,\Delta t$ where α is a constant; hence

$$\alpha = \frac{y'}{y} = \frac{d}{dt}\log y;$$

integrating we obtain the familiar formula for *unlimited growth*:

$$y(t) = e^{\alpha t}y(0).$$

The growth rate can depend on many things. Let us assume for the moment that it depends only on the per capita food supply σ, and that $\sigma \geq 0$ is constant. There will be a minimum σ_0 necessary to sustain the population. For $\sigma > \sigma_0$, the growth rate is positive; for $\sigma < \sigma_0$, it is negative; while for $\sigma = \sigma_0$, the growth rate is 0. The simplest way to ensure this is to make the growth rate a linear function of $\sigma - \sigma_0$:

$$\alpha = a(\sigma - \sigma_0), \qquad a > 0.$$

Then

$$\frac{dy}{dt} = a(\sigma - \sigma_0)y(t). \tag{1}$$

Here a and σ_0 are constants, dependent only on the species, and σ is a parameter,

dependent on the particular environment but constant for a given ecology. (In the next section σ will be another species satisfying a second differential equation.)

The preceding equation is readily solved:

$$y(t) = \exp[ta(\sigma - \sigma_0)]y(0).$$

Thus the population must increase without limit, remain constant, or approach 0 as a limit, depending on whether $\sigma > \sigma_0$, $\sigma = \sigma_0$, or $\sigma < \sigma_0$. If we recall that actually fractional values of $y(t)$ are meaningless, we see that for all practical purposes "$y(t) \to 0$" really means that the population dies out in a finite time.

In reality, a population cannot increase without limit; at least, this has never been observed! It is more realistic to assume that when the population level exceeds a certain value η, the growth rate is negative. We call this value η, the *limiting population*. Note that η is not necessarily an upper bound for the population. Reasons for the negative growth rate might be insanity, decreased food supply, overcrowding, smog, and so on. We refer to these various unspecified causes as *social phenomena*. (There may be positive social phenomena; for example, a medium size population may be better organized to resist predators and obtain food than a small one. But we ignore this for the moment.)

Again making the simplest mathematical assumptions, we suppose the growth rate is proportional to $\eta - y$:

$$\alpha = c(\eta - y), \qquad c > 0 \text{ a constant.}$$

Thus we obtain the *equation of limited growth*:

$$\frac{dy}{dt} = c(\eta - y)y; \qquad c > 0, \quad \eta > 0. \tag{2}$$

Note that this suggests

$$\frac{\Delta y}{\Delta t} = c\eta y - cy^2.$$

This means that during the period Δt the population change is $cy^2\,\Delta t$ less than it would be without social phenomena. We can interpret cy^2 as a number proportional to the average number of encounters between y individuals. Hence cy^2 is a kind of social friction.

The equilibria of (2) occur at $y = 0$ and $y = \eta$. The equilibrium at η is asymptotically stable (if $c > 0$) since the derivative of $c(\eta - y)y$ at η is $-c\eta$, which is negative. The basin of η is $\{y \mid y > 0\}$ since $y(t)$ will increase to η as a limit if $0 < y(0) < \eta$, and decrease to η as a limit if $\eta < y(0)$. (This can be seen by considering the sign of dy/dt.)

A more realistic model of a single species is

$$y' = M(y)y.$$

Here the variable growth rate M is assumed to depend only on the total population y.

It is plausible to assume as before that there is a limiting population η such that $M(\eta) = 0$ and $M(y) < 0$ for $y > \eta$. If very small populations behave like the unlimited growth model, we assume $M(0) > 0$.

PROBLEMS

1. A population $y(t)$ is governed by an equation

$$y' = M(y)y.$$

Prove that:
(a) equilibria occur at $y = 0$ and whenever $M(y) = 0$;
(b) the equilibrium at $y = 0$ is unstable;
(c) an equilibrium $\xi > 0$ is asymptotically stable if and only if there exists $\epsilon > 0$ such that $M > 0$ on the interval $[\xi - \epsilon, \xi)$ and $M < 0$ on $(\xi, \xi + \epsilon]$.

2. Suppose the population of the United States obeys limited growth. Compute the limiting population and the population in the year 2000, using the following data:

Year	Population
1950	150,697,361
1960	179,323,175
1970	203,184,772

§2. Predator and Prey

We consider a predator species y and its prey x. The prey population is the total food supply for the predators at any given moment. The total food consumed by the predators (in a unit of time) is proportional to the number of predator–prey encounters, which we assume proportional to xy. Hence the per capita food supply for the predators at time t is proportional to $x(t)$. Ignoring social phenomena for the moment, we obtain from equation (1) of the preceding section:

$$y' = a(x - \sigma_0)y,$$

where $a > 0$ and $\sigma_0 > 0$ are constants. We rewrite this as

$$y' = (Cx - D)y; \qquad C > 0, \quad D > 0.$$

Consider next the growth rate of the prey. In each small time period Δt, a certain number of prey are eaten. This number is assumed to depend only on the two populations, and is proportional to Δt; we write it as $f(x, y)\,\Delta t$. What should we postulate about $f(x, y)$?

It is reasonable that $f(x, y)$ be proportional to y: twice as many cats will eat twice as many mice in a small time period. We also assume $f(x, y)$ is proportional to x: if the mouse population is doubled, a cat will come across a mouse twice as often. Thus we put $f(x, y) = \beta xy$, β a positive constant. (This assumption is less plausible if the ratio of prey to predators is very large. If a cat is placed among a sufficiently large mouse population, after a while it will ignore the mice.)

The prey species is assumed to have a constant per capita food supply available, sufficient to increase its population in the absence of predators. Therefore the prey is subject to a differential equation of the form

$$x' = Ax - Bxy.$$

In this way we arrive at the *predator–prey equations* of Volterra and Lotka:

$$(1) \qquad \begin{aligned} x' &= (A - By)x, \\ y' &= (Cx - D)y. \end{aligned} \qquad A, B, C, D > 0.$$

This system has equilibria at $(0, 0)$ and $z = (D/C, A/B)$. It is easy to see that $(0, 0)$ is a saddle, hence unstable. The eigenvalues at $(D/C, A/B)$ are pure imaginary, however, which gives no information about stability.

We investigate the phase portrait of (1) by drawing the two lines

$$x' = 0: \qquad y = \frac{A}{B},$$

$$y' = 0: \qquad x = \frac{D}{C}.$$

These divide the region $x > 0, y > 0$ into four quadrants (Fig. A). In each quadrant the signs of x' and y' are constant as indicated.

The positive x-axis and the positive y-axis are each trajectories as indicated in Fig. A. The reader can make the appropriate conclusion about the behavior of the population.

Otherwise each solution curve $(x(t), y(t))$ moves counterclockwise around z from one quadrant to the next. Consider for example a trajectory $(x(t), y(t))$ starting at a point

$$x(0) = u > \frac{D}{C} > 0,$$

$$y(0) = v > \frac{A}{B} > 0$$

in quadrant I. There is a maximal interval $[0, \tau) = J$ such that $(x(t), y(t)) \in$

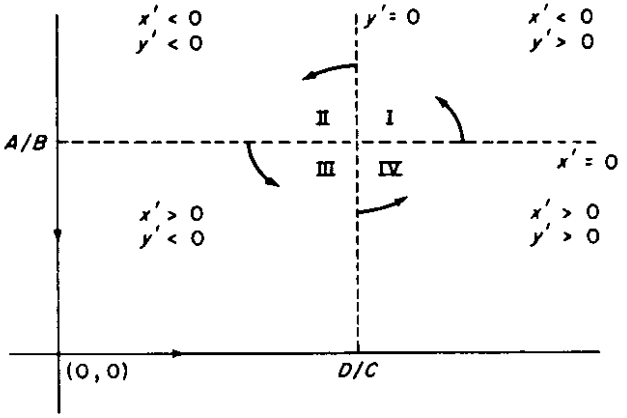


FIG. A

quadrant I for $0 \leq t < \tau$ (perhaps $\tau = \infty$). Put

$$A - Bv = -r < 0,$$

$$Cu - D = s > 0.$$

As long as $t \in J$, $x(t)$ is decreasing and $y(t)$ is increasing. Hence

$$\frac{d}{dt} \log x(t) = \frac{x'}{x} = A - By \leq -r,$$

$$\frac{d}{dt} \log y(t) = \frac{y'}{y} = Cx - D \geq s.$$

Therefore

$$\frac{D}{C} \leq x(t) \leq ue^{-rt}, \tag{2}$$

$$\frac{A}{B} \geq y(t) \geq ve^{st}, \tag{3}$$

for $0 \leq t < \tau$. From the second inequality of (2) we see that τ is finite. From (2) and (3) we see that for $t \in J$, $(x(t), y(t))$ is confined to the compact region

$$\frac{D}{C} \leq x(t) \leq u,$$

$$\frac{A}{B} \leq y(t) \leq ve^{s\tau}.$$

Therefore (Chapter 8) $(x(\tau), y(\tau))$ is defined and in the boundary of that region; since $x(t)$ is decreasing, $x(\tau) = D/C$. Thus the trajectory enters quadrant II. Similarly for other quadrants.

We cannot yet tell whether trajectories spiral in toward z, spiral toward a limit cycle, or spiral out toward "infinity" and the coordinate axes. Let us try to find a Liapunov function H.

Borrowing the trick of *separation of variables* from partial differential equations, we look for a function of the form

$$H(x, y) = F(x) + G(y).$$

We want $\dot{H} \leq 0$, where

$$\dot{H}(x, y) = \frac{d}{dt} H(x(t), y(t))$$

$$= \frac{dF}{dx} x' + \frac{dG}{dy} y'.$$

Hence

$$\dot{H}(x, y) = x \frac{dF}{dx} (A - By) + y \frac{dG}{dy} (Cx - D).$$

We obtain $\dot{H} \equiv 0$ provided

$$\frac{x\, dF/dx}{Cx - D} \equiv \frac{y\, dG/dy}{By - A}.$$

Since x and y are independent variables, this is possible if and only if

$$\frac{x\, dF/dx}{Cx - D} = \frac{y\, dG/dy}{By - A} = \text{constant}.$$

Putting the constant equal to 1 we get

$$\frac{dF}{dx} = C - \frac{D}{x}, \tag{4}$$

$$\frac{dG}{dy} = B - \frac{A}{y};$$

integrating we find

$$F(x) = Cx - D \log x,$$

$$G(y) = By - A \log y.$$

Thus the function

$$H(x, y) = Cx - D \log x + By - A \log y,$$

defined for $x > 0$, $y > 0$, is constant on solution curves of (1).

By considering the signs of $\partial H/\partial x$ and $\partial H/\partial y$ it is easy to see that the equilibrium $z = (D/C, A/B)$ is an absolute minimum for H. It follows that H (more precisely, $H - H(z)$) is a Liapunov function (Chapter 9). Therefore *z is a stable equilibrium.*

We note next that *there are no limit cycles*; this follows from Chapter 11 because H is not constant on any open set.

We now prove

Theorem 1 *Every trajectory of the Volterra–Lotka equations* (1) *is a closed orbit (except the equilibrium z and the coordinate axes).*

Proof. Consider a point $w = (u, v)$, $u > 0$, $v > 0$; $w \neq z$. Then there is a doubly infinite sequence $\cdots < t_{-1} < t_0 < t_1 < \cdots$ such that $\phi_{t_n}(w)$ is on the line $x = D/C$, and

$$t_n \to \infty \qquad \text{as} \quad n \to \infty,$$

$$t_n \to -\infty \qquad \text{as} \quad n \to -\infty.$$

If w is not in a closed orbit, the points $\phi_{t_n}(w)$ are monotone along the line $x = D/C$ (Chapter 11). Since there are no limit cycles, either

$$\phi_{t_n}(w) \to z \qquad \text{as} \quad n \to \infty,$$

or

$$\phi_{t_n}(w) \to z \qquad \text{as} \quad n \to -\infty.$$

Since H is constant on the trajectory of w, this implies that $H(w) = H(z)$. But this contradicts minimality of $H(z)$.

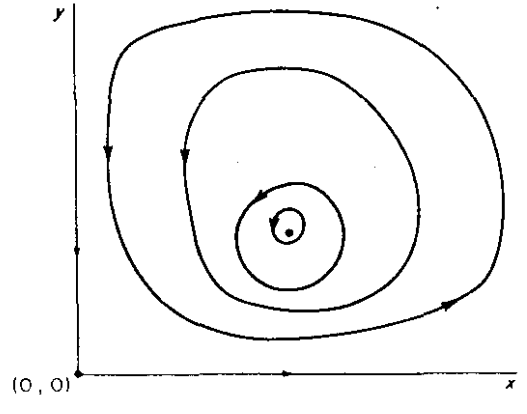


FIG. B. Phase portrait of (1).

We now have the following (schematic) phase portrait (Fig. B). Therefore, for any given initial populations $(x(0), y(0))$ with $x(0) \neq 0$, and $y(0) \neq 0$, other than z, the populations of predator and prey will oscillate cyclically.

No matter what the numbers of prey and predator are, neither species will die out, nor will it grow indefinitely. On the other hand, except for the state z, which is improbable, the populations will not remain constant.

Let us introduce social phenomena of Section 1 into the equations (1). We obtain the following *predator–prey equations of species with limited growth*:

$$x' = (A - By - \lambda x)x,$$

$$y' = (Cx - D - \mu y)y. \tag{5}$$

The constants A, B, C, D, λ, μ are all positive.

We divide the upper-right quadrant Q ($x > 0$, $y > 0$) into sectors by the two lines

$$L: \quad A - By - \lambda x = 0;$$

$$M: \quad Cx - D - \mu y = 0.$$

Along these lines $x' = 0$ and $y' = 0$, respectively. There are two possibilities, according to whether these lines intersect in Q or not. If not (Fig. C), the predators die out and the prey population approaches its limiting value A/λ (where L meets the x-axis).

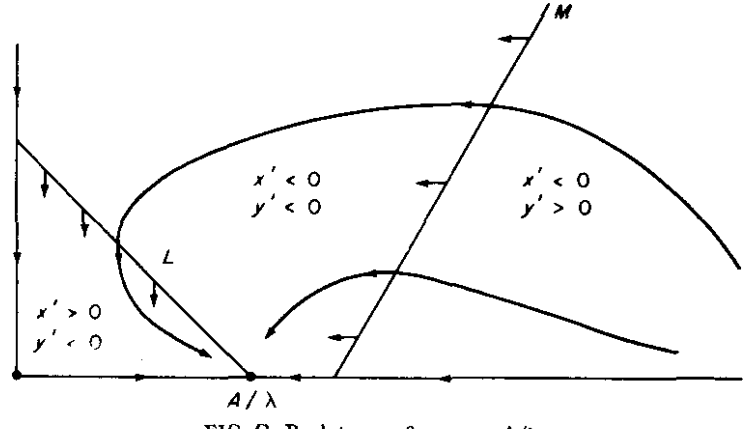


FIG. C. Predators $\to 0$; prey $\to A/\lambda$.

This is because it is impossible for both prey and predators to increase at the same time. If the prey is above its limiting population it must decrease and after a while the predator population also starts to decrease (when the trajectory crosses M). After that point the prey can never increase past A/λ, and so the predators continue to decrease. If the trajectory crosses L, the prey increases again (but not past A/λ), while the predators continue to die off. In the limit the predators disappear and the prey population stabilizes at A/λ.

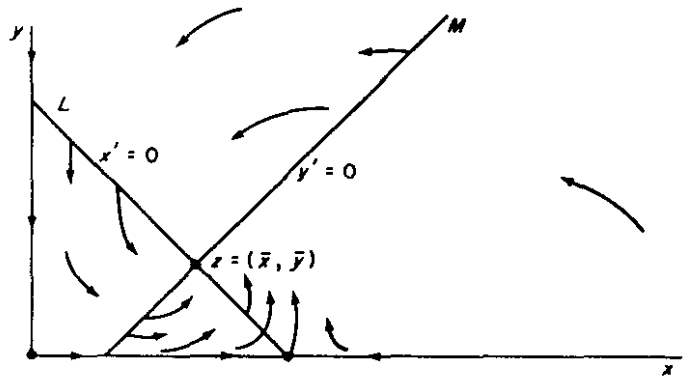


FIG. D

Suppose now that L and M cross at a point $z = (\bar{x}, \bar{y})$ in the quadrant Q (Fig. D); of course z is an equilibrium. The linear part of the vector field (3) at z is

$$\begin{bmatrix} -\lambda\bar{x} & -B\bar{x} \\ C\bar{y} & -\mu\bar{y} \end{bmatrix}.$$

The characteristic polynomial has positive coefficients. Both roots of such a polynomial have negative real parts. Therefore z *is asymptotically stable.*

Note that in addition to the equilibria at z and $(0, 0)$, there is also an equilibrium, a saddle, at the intersection of the line L with the x-axis.

It is not easy to determine the basin of z; nor do we know whether there are any limit cycles. Nevertheless we can obtain some information.

Let L meet the x-axis at $(p, 0)$ and the y-axis at $(0, q)$. Let Γ be a rectangle whose corners are

$$(0, 0), \qquad (\bar{p}, 0), \qquad (0, \bar{q}), \qquad (\bar{p}, \bar{q})$$

with $\bar{p} > p$, $\bar{q} > q$, and $(\bar{p}, \bar{q}) \in M$ (Fig. E). Every trajectory at a boundary point of Γ either enters Γ or is part of the boundary. Therefore Γ *is positively invariant.* Every point in Q is contained in such a rectangle.

By the Poincaré–Bendixson theorem the ω-limit set of any point (x, y) in Γ, with $x > 0$, $y > 0$, must be a limit cycle or one of the three equilibria $(0, 0)$, z or $(p, 0)$. We rule out $(0, 0)$ and $(p, 0)$ by noting that x' is increasing near $(0, 0)$; and y' is increasing near $(p, 0)$. Therefore $L_\omega(\mu)$ is either z or a limit cycle in Γ. By a consequence of the Poincaré–Bendixson theorem any limit cycle must surround z.

We observe further that any such rectangle Γ contains *all* limit cycles. For a limit cycle (like any trajectory) must enter Γ, and Γ is positively invariant.

Fixing $(\bar{p}, \bar{q})$ as above, it follows that *for any initial values* $(x(0), y(0))$, *there exists* $t_0 > 0$ *such that*

$$x(t) < \bar{p}, \quad y(t) < \bar{q} \qquad \text{if} \quad t \geq t_0.$$

One can also find eventual lower bounds for $x(t)$ and $y(t)$.

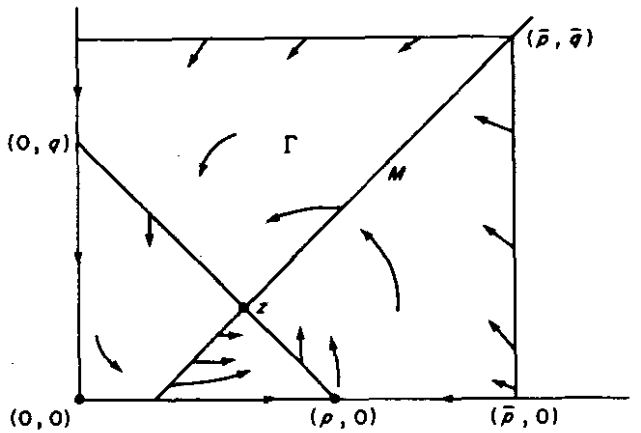


FIG. E

We also see that in the long run, a trajectory either approaches z or else spirals down to a limit cycle.

From a practical standpoint a trajectory that tends toward z is indistinguishable from z after a certain time. Likewise a trajectory that approaches a limit cycle γ can be identified with γ after it is sufficiently close.

The conclusion is that *any ecology of predators and prey which obeys equations* (2) *eventually settles down to either a constant or periodic population. There are absolute upper bounds that no population can exceed in the long run, no matter what the initial populations are.*

PROBLEM

Show by examples that the equilibrium in Fig. D can be either a spiral sink or a node. Draw diagrams.

§3. Competing Species

We consider now two species x, y which compete for a common food supply. Instead of analyzing specific equations we follow a different procedure: we consider a large class of equations about which we assume only a few qualitative features. In this way considerable generality is gained, and little is lost because specific equations can be very difficult to analyze.

The equations of growth of the two species are written in the form

$$x' = M(x, y)x,$$
$$y' = N(x, y)y, \tag{1}$$

where the growth rates M and N are C^1 functions of nonnegative variables x, y. The following assumptions are made:

(a) If either species increases, the growth rate of the other goes down. Hence

$$\frac{\partial M}{\partial y} < 0 \quad \text{and} \quad \frac{\partial N}{\partial x} < 0.$$

(b) If either population is very large, neither species can multiply. Hence there exists $K > 0$ such that

$$M(x, y) \leq 0 \quad \text{and} \quad N(x, y) \leq 0 \quad \text{if} \quad x \geq K \quad \text{or} \quad y \geq K.$$

(c) In the absence of either species, the other has a positive growth rate up to a certain population and a negative growth rate beyond it. Therefore there are constants $a > 0$, $b > 0$ such that

$$M(x, 0) > 0 \quad \text{for} \quad x < a \quad \text{and} \quad M(x, 0) < 0 \quad \text{for} \quad x > a,$$
$$N(0, y) > 0 \quad \text{for} \quad y < b \quad \text{and} \quad N(0, y) < 0 \quad \text{for} \quad y > b.$$

By (a) and (c) each vertical line $x \times \mathbf{R}$ meets the set $\mu = M^{-1}(0)$ exactly once if $0 \leq x \leq a$ and not at all if $x > a$. By (a) and the implicit function theorem μ is the graph of a nonnegative C^1 map f: $[0, a] \to \mathbf{R}$ such that $f^{-1}(0) = a$. Below the curve μ, $M > 0$ and above it $M < 0$ (Fig. A).

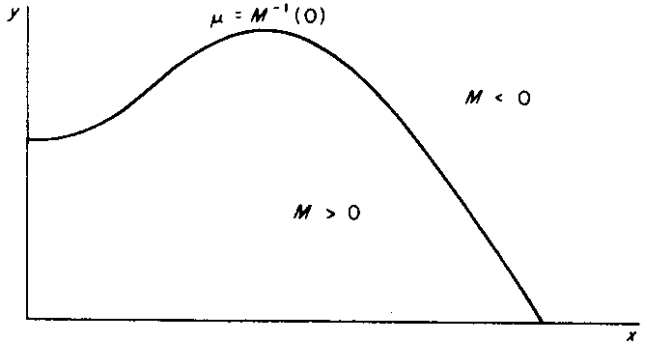


FIG. A

In the same way the set $\nu = N^{-1}(0)$ is a smooth curve of the form

$$\{(x, y) \mid x = g(y)\},$$

where g: $[0, b] \to \mathbf{R}$ is a nonnegative C^1 map with $g^{-1}(0) = b$. The function N is positive to the left of ν and negative to the right.

Suppose μ and ν do not intersect and that μ is below ν. Then a phase portrait can be found in a straightforward way following methods of the previous section. The equilibria are $(0, 0)$, $(a, 0)$ and $(0, b)$. All orbits tend to one of the three equilibria but most to the asymptotically stable equilibrium $(0, b)$. See Fig. B.

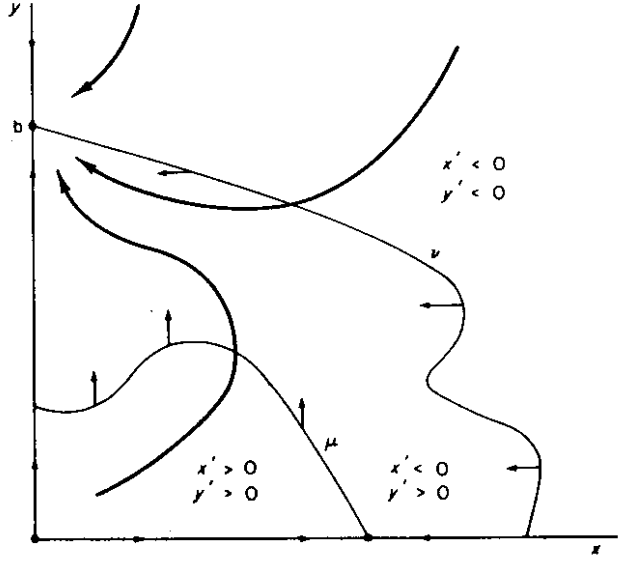


FIG. B

Suppose now that μ and ν intersect. We make the assumption that $\mu \cap \nu$ is a finite set, and at each intersection point, μ and ν *cross transversely*, that is, they have distinct tangent lines. This assumption could be dispensed with but it simplifies the topology of the curves. Moreover M and N can be approximated arbitrarily closely by functions whose zero sets have this property. In a sense which can be made precise, this is a "generic" property.

The curves μ and ν and the coordinate axes bound a finite number of connected open sets in the upper right quadrant: these are sets where $x' \neq 0$ and $y' \neq 0$. We call these open sets *basic regions* (Fig. C). They are of four types:

$$\text{I:} \quad x' > 0, \quad y' > 0;$$
$$\text{II:} \quad x' < 0, \quad y' > 0;$$
$$\text{III:} \quad x' < 0, \quad y' < 0;$$
$$\text{IV:} \quad x' > 0, \quad y' < 0.$$

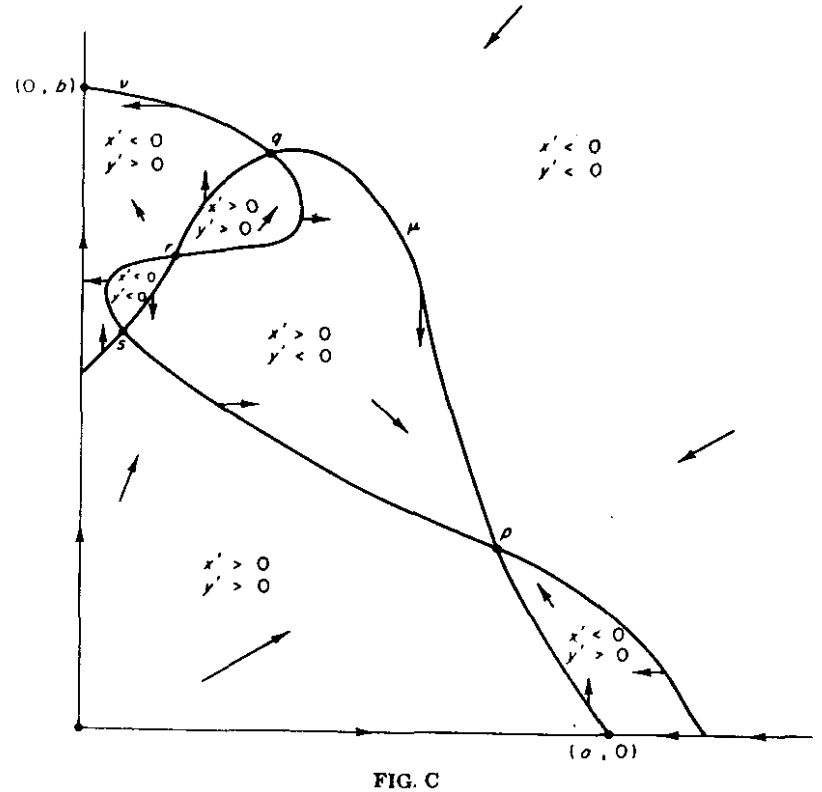


FIG. C

The boundary ∂B of a basic region B is made up of points of the following types: points of $\mu \cap \nu$, called *vertices*; points on μ or ν but not on both nor on the coordinate axes, called *ordinary* boundary points; and points on the axes.

A vertex is an equilibrium; the other equilibria are at $(0, 0)$, $(a, 0)$, and $(0, b)$. At an ordinary boundary point $w \in \partial B$, the vector (x', y') is either vertical (if $w \in \mu$) or horizontal (if $w \in \nu$). It points either into or out of B since μ has no vertical tangents and ν has no horizontal tangents. We call w an *inward* or *outward* point of ∂B, accordingly.

The following technical result is the key to analyzing equation (1):

Lemma *Let B be a basic region. Then the ordinary boundary points of B are either all inward or all outward.*

Proof. If the lemma holds for B, we call B *good*.

Let p be a vertex of B where μ and ν cross. Then p is on the boundary of four basic regions, one of each type. Types II and IV, and types I and III, are diagonally opposite pairs.

Let $\mu_0 \subset \mu$ and $\nu_0 \subset \nu$ be the open arcs of ordinary boundary points having p as a common end point. If $\mu_0 \cup \nu_0$ consists entirely of inward or entirely of outward points of ∂B, we call p *good for* B; otherwise p is *bad for* B. It is easy to see that if p is good for B, it is good for the other three basic regions adjacent to p, and similarly for bad (Fig. D). This is because (x', y') reverses direction as one proceeds along μ or ν past a crossing point. Hence it makes sense to call a vertex simply *good* or *bad*.

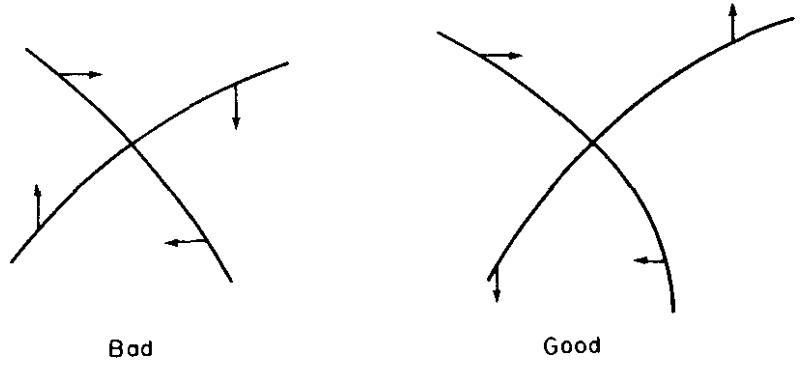


FIG. D

Consider first of all the region B_0 whose boundary contains $(0, 0)$. This is of type I $(x' > 0, y' > 0)$. If q is an ordinary point of $\mu \cap \partial B_0$, we can connect q to a point inside B_0 by a path which avoids ν. Along such a path $y' > 0$. Hence (x', y') points upward *out* of B_0 at q since μ is the graph of a function. Similarly at an ordinary point r of $\nu \cap \partial B_0$, (x', y') points to the right, *out* of B_0 at r. Hence B_0 is good, and so every vertex of B_0 is good.

Next we show that if B is a basic region and ∂B contains one good vertex p of $\mu \cap \nu$, then B is good. We assume that near p, the vector field along ∂B points into B; we also assume that in B, $x' < 0$ and $y' > 0$. (The other cases are similar.) Let $\mu_0 \subset \mu$, $\nu_0 \subset \nu$ be arcs of ordinary boundary points of B adjacent to p (Fig. E). For example let r be any ordinary point of $\partial B \cap \mu$ and q any ordinary point of μ_0. Then $y' > 0$ at q. As we move along μ from q to r the sign of y' changes each time we cross ν. The number of such crossings is *even* because r and q are on the same side of ν. Hence $y' > 0$ at r. This means that (x', y') points up at r. Similarly, $x' < 0$ at every ordinary point of $\nu \cap \partial B$. Therefore along μ the vector (x', y') points up; along ν it points left. Then B lies *above* μ and *left* of ν. Thus B is good.

This proves the lemma, for we can pass from any vertex to any other along μ, starting from a good vertex. Since successive vertices belong to the boundary of a common basic region, each vertex in turn is proved good. Hence all are good.

As a consequence of the lemma, *each basic region, and its closure, is either positively or negatively invariant.*

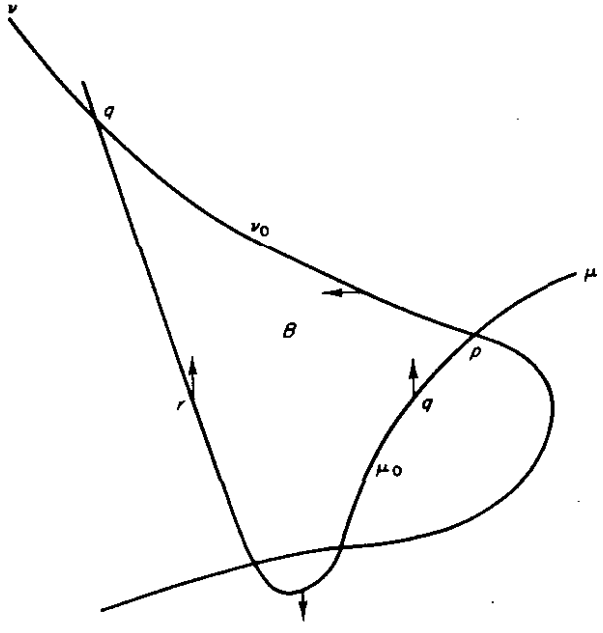


FIG. E

What are the possible ω-limit points of the flow (1)? There are no closed orbits. For a closed orbit must be contained in a basic region, but this is impossible since $x(t)$ and $y(t)$ are monotone along any solution curve in a basic region. Therefore all ω-limit points are equilibria.

We note also that each trajectory is defined for all $t \geq 0$, because any point lies in a large rectangle Γ spanned by $(0, 0)$, $(x_0, 0)$, $(0, y_0)$, (x_0, y_0) with $x_0 > a$, $y_0 > b$; such a rectangle is compact and positively invariant (Fig. F). Thus we have shown:

Theorem *The flow ϕ_t of (1) has the following property: for all $p = (x, y)$, $x \geq 0$, $y \geq 0$, the limit*

$$\lim_{t \to \infty} \phi_t(p)$$

exists and is one of a finite number of equilibria.

We conclude that *the populations of two competing species always tend to one of a finite number of limiting populations.*

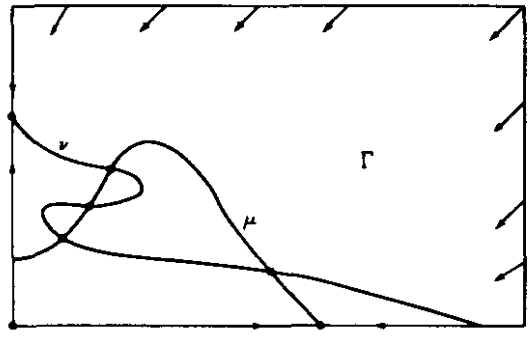


FIG. F

Examining the equilibria for stability, one finds the following results. A vertex where μ and ν each have negative slope, but μ is steeper, is asymptotically stable (Fig. G). One sees this by drawing a small rectangle with sides parallel to the axes around the equilibrium, putting one corner in each of the four adjacent regions. Such a rectangle is positively invariant; since it can be arbitrarily small, the equilibrium is asymptotically stable. Analytically this is expressed by

$$\text{slope of } \mu = -\frac{M_x}{M_y} < \text{slope of } \nu = -\frac{N_x}{N_y} < 0,$$

where $M_x = \partial M/\partial x$, $M_y = \partial M/\partial y$, and so on, at the equilibrium, from which a computation yields eigenvalues with negative real parts. Hence we have a sink.

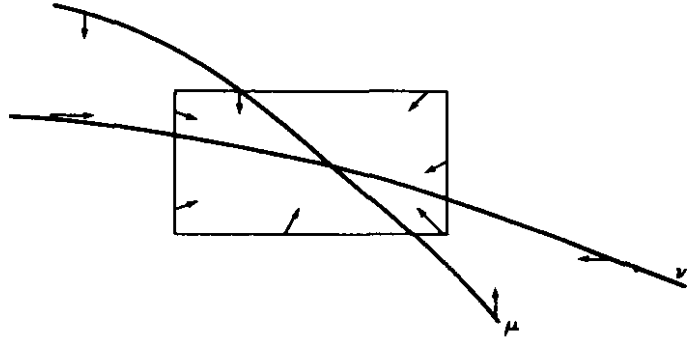


FIG. G

A case by case study of the different ways μ and ν can cross shows that the only other asymptotically stable equilibrium is $(b, 0)$ when $(b, 0)$ is above μ, or $(a, 0)$ when $(a, 0)$ is to the right of ν. All other equilibria are unstable. For example, q in Fig. H is unstable because arbitrarily near it, to the left, is a trajectory with x

decreasing; such a trajectory tends toward $(0, b)$. Thus in Fig. H, $(0, b)$ and p are asymptotically stable, while q, r, s and $(a, 0)$ are unstable. Note that r is a source.

There must be at least one asymptotically stable equilibrium. If $(0, b)$ is not one, then it lies under μ; and if $(a, 0)$ is not one, it lies over μ. In that case μ and ν cross, and the first crossing to the left of $(a, 0)$ is asymptotically stable.

Every trajectory tends to an equilibrium; it is instructive to see how these ω-limits change as the initial state changes. Let us suppose that q is a saddle. Then it can be shown that exactly two trajectories α, α' approach q, the so-called *stable manifolds* of q, or sometimes *separatrices* of q. We concentrate on the one in the unbounded basic region B_∞, labeled α in Fig. H.

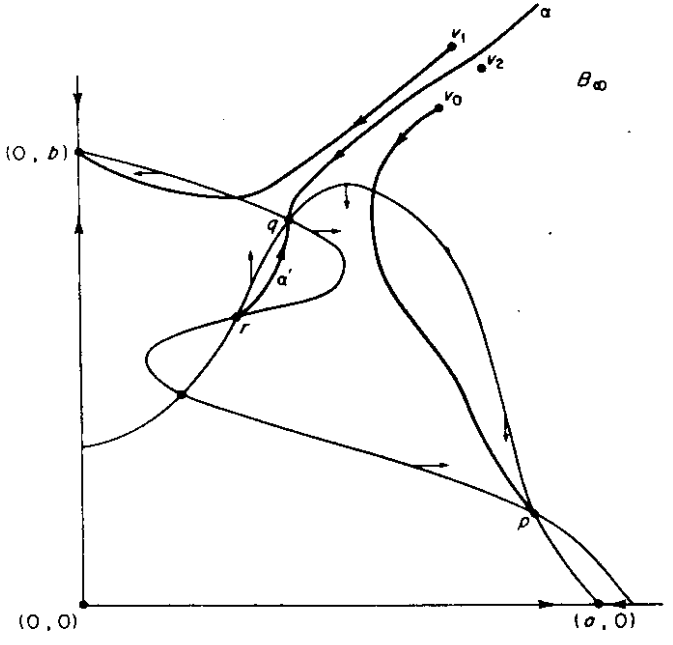


FIG. H. Bifurcation of behavior.

All points of B_∞ to the left of α end up at $(0, b)$, while points to the right go to p. As we move across α this limiting behavior changes radically. Let us consider this *bifurcation of behavior* in biological terms.

Let v_0, v_1 be states in B_0, very near each other but separated by α; suppose the trajectory of v_0 goes to p while that of v_1 goes to $(0, b)$. The point $v_0 = (x_0, y_0)$ represents an ecology of competing species which will eventually stabilize at p.

Note that both populations are positive at p. Suppose that some unusual event occurs, not accounted for by our model, and the state of the ecology changes suddenly from v_0 to v_1. Such an event might be introduction of a new pesticide, importation of additional members of one of the species, a forest fire, or the like. Mathematically the event is a jump from the basin of p to that of $(0, b)$.

Such a change, even though quite small, is an ecological catastrophe. For the trajectory of v_1 has quite a different fate: it goes to $(0, b)$ and the x species is wiped out!

Of course in practical ecology one rarely has Fig. H to work with. Without it, the change from v_0 to v_1 does not seem very different from the insignificant change from v_0 to a near state v_2, which also goes to p. The moral is clear: in the absence of comprehensive knowledge, a deliberate change in the ecology, even an apparently minor one, is a very risky proposition.

PROBLEMS

1. The equations

$$x' = x(2 - x - y),$$
$$y' = y(3 - 2x - y)$$

satisfy conditions (a) through (d) for competing species. Explain why these equations make it mathematically possible, but extremely unlikely, for both species to survive.

2. Two species x, y are in *symbiosis* if an increase of either population leads to an increase in the growth rate of the other. Thus we assume

$$\begin{aligned} x' &= M(x, y)x \\ y' &= N(x, y)y \end{aligned} \qquad (x \geq 0,\ y \geq 0)$$

with

$$\frac{\partial M}{\partial y} > 0 \quad \text{and} \quad \frac{\partial N}{\partial x} > 0.$$

We also suppose that the total food supply is limited; hence for some $A > 0$, $B > 0$ we have

$$M(x, y) < 0 \quad \text{if} \quad x > A,$$
$$N(x, y) < 0 \quad \text{if} \quad y > B.$$

If both populations are very small, they both increase; hence

$$M(0, 0) > 0 \quad \text{and} \quad N(0, 0) > 0.$$

Assuming that the intersections of the curves $M^{-1}(0)$, $N^{-1}(0)$ are finite, and all are *transverse*, show that:

(a) every trajectory tends to an equilibrium in the region $0 < x < A$, $0 < y < B$;
(b) there are no sources;
(c) there is at least one sink;
(d) if $\partial M/\partial x < 0$ and $\partial N/\partial y < 0$, there is a unique sink z, and $z = L_\omega(x, y)$ for all $x > 0$, $y > 0$.

3. Prove that under plausible hypotheses, two mutually destructive species cannot coexist in the long run.

4. Let y and x denote predator and prey populations. Let

$$x' = M(x, y)x,$$
$$y' = N(x, y)y,$$

where M and N satisfy the following conditions.

(i) If there are not enough prey, the predators decrease. Hence for some $b > 0$

$$N(x, y) < 0 \quad \text{if} \quad x < b.$$

(ii) An increase in the prey improves the predator growth rate; hence $\partial N/\partial x > 0$.
(iii) In the absence of predators a small prey population will increase; hence $M(0, 0) > 0$.
(iv) Beyond a certain size, the prey population must decrease; hence there exists $A > 0$ with $M(x, y) < 0$ if $x > A$.
(v) Any increase in predators decreases the rate of growth of prey; hence $\partial M/\partial y < 0$.
(vi) The two curves $M^{-1}(0)$, $N^{-1}(0)$ intersect transversely, and at only a finite number of points.

Show that if there is some (u, v) with $M(u, v) > 0$ and $N(u, v) > 0$ then there is either an asymptotically stable equilibrium or an ω-limit cycle. Moreover, if the number of limit cycles is finite and positive, one of them must have orbits spiraling toward it from both sides.

5. Show that the analysis of equation (1) is essentially the same if (c) is replaced by the more natural assumptions: $M(0, 0) > 0$, $N(0, 0) > 0$, and $M(x, 0) < 0$ for $x > A$, $N(0, y) < 0$ for $y > B$.

Notes

There is a good deal of experimental and observational evidence in support of the general conclusions of this chapter—that predator–prey ecologies oscillate while competitor ecologies reach an equilibrium. In fact Volterra's original study

was inspired by observation of fish populations in the Upper Adriatic. A discussion of some of this material is found in a paper by E. W. Montroll *et al.*, "On the Volterra and other nonlinear models" [16]. See also the book *The Struggle for Existence* by U. D'Ancona [4].

A very readable summary of some recent work is in "The struggle for life, I" by A. Rescigno and I. Richardson [21]. Much of the material of this chapter was adapted from their paper.

A recent book by René Thom [24] on morphogenesis uses very advanced theories of stability and bifurcation in constructing mathematical models of biological processes.

Chapter 13

Periodic Attractors

Here we define asymptotic stability for closed orbits of a dynamical system, and an especially important kind called a periodic attractor. Just as sinks are of major importance among equilibria in models of "physical" systems, so periodic attractors are the most important kind of oscillations in such models. As we shall show in Chapter 16, such oscillations persist even if the vector field is perturbed.

The main result is that a certain eigenvalue condition on the derivative of the flow implies asymptotic stability. This is proved by the same method of local sections used earlier in the Poincaré–Bendixson theorem. This leads to the study of "discrete dynamical systems" in Section 2, a topic which is interesting by itself.

§1. Asymptotic Stability of Closed Orbits

Let $f\colon W \to \mathbf{R}^n$ be a C^1 vector field on an open set $W \subset \mathbf{R}^n$; the flow of the differential equation

$$x' = f(x) \tag{1}$$

is denoted by ϕ_t.

Let $\gamma \subset W$ be a closed orbit of the flow, that is, a nontrivial periodic solution curve. We call γ *asymptotically stable* if for every open set $U_1 \subset W$, with $\gamma \subset U_1$ there is an open set U_2, $\gamma \subset U_2 \subset U_1$ such that $\phi_t(U_2) \subset U_1$ for all $t > 0$ and

$$\lim_{t\to\infty} d(\phi_t(x), \gamma) = 0.$$

Here $d(x, \gamma)$ means the minimum distance from x to a point of γ.

The closed orbit in the Van der Pol oscillator was shown to be asymptotically stable. On the other hand, the closed orbits of the harmonic oscillator are not since an asymptotically stable closed orbit is evidently isolated from other closed orbits.

We say a point $x \in W$ has *asymptotic period* $\lambda \in \mathbf{R}$ if

$$\lim_{t\to\infty} |\phi_{\lambda+t}(x) - \phi_t(x)| = 0.$$

Theorem 1 *Let γ be an asymptotically stable closed orbit of period λ. Then γ has a neighborhood $U \subset W$ such that every point of U has asymptotic period λ.*

Proof. Let U be the open set U_2 in the definition of asymptotically stable with $W^0 = U_1$. Let $x \in U$ and fix $\epsilon > 0$. There exists δ, $0 < \delta \leq \epsilon$, such that if $z \in \gamma$ and $|y - z| < \delta$, then $|\phi_\lambda(y) - \phi_\lambda(z)| < \epsilon$ (by continuity of the flow). Of course $\phi_\lambda(z) = z$. Since $d(\phi_t(x), \gamma) \to 0$ as $t \to \infty$, there exists $t_0 \geq 0$ such that if $t \geq t_0$, there is a point $z_t \in \gamma$ such that $|\phi_t(x) - z_t| < \delta$. Keeping in mind $\phi_\lambda(z_t) = z_t$ we have for $t \geq t_0$:

$$\begin{aligned} |\phi_{\lambda+t}(x) - \phi_t(x)| &\leq |\phi_\lambda\phi_t(x) - \phi_\lambda(z_t)| + |\phi_\lambda(z_t) - \phi_t(x)| \\ &\leq \epsilon + \delta \leq 2\epsilon. \end{aligned}$$

This proves the theorem.

The significance of Theorem 1 is that after a certain time, trajectories near an asymptotically stable closed orbit behave as if they themselves had the same period as the closed orbit.

The only example we have seen of an asymptotic closed orbit occurs in a two dimensional system. This is no accident; planar systems are comparatively easy to analyze, essentially because solution curves locally separate the plane.

The theorem below is analogous to the fact that an equilibrium $\bar{x}$ is asymptotically stable if the eigenvalues of $Df(\bar{x})$ have negative real part. It is not as convenient to use since it requires information about the solutions of the equation, not merely about the vector field. Nevertheless it is of great importance.

Theorem 2 *Let γ be a closed orbit of period λ of the dynamical system (1). Let $p \in \gamma$. Suppose that $n - 1$ of the eigenvalues of the linear map $D\phi_\lambda(p)\colon E \to E$ are less than 1 in absolute value. Then γ is asymptotically stable.*

Some remarks on this theorem are in order. First, it assumes that ϕ_λ is differentiable. In fact, $\phi_t(x)$ is a C^1 function of (t, x); this is proved in Chapter 16. Second, the condition on $D\phi_\lambda(p)$ is independent of $p \in \gamma$. For if $q \in \gamma$ is a different point, let $r \in \mathbf{R}$ be such that $\phi_r(p) = q$. Then

$$\begin{aligned} D\phi_\lambda(p) &= D(\phi_{-r}\phi_\lambda\phi_r)(p) \\ &= D\phi_r(p)^{-1}D\phi_\lambda(q)D\phi_r(p), \end{aligned}$$

which shows that $D\phi_\lambda(p)$ is similar to $D\phi_\lambda(q)$. Third, note that 1 is always an eigenvalue of $D\phi_\lambda(p)$ since

$$D\phi_\lambda(p)f(p) = f(p).$$

The eigenvalue condition in Theorem 2 is stronger than asymptotic stability. If it holds, we call γ a *periodic attractor*. Not only do trajectories near a periodic attractor γ have the same asymptotic period as γ, but they are asymptotically "in phase" with γ. This is stated precisely in the following theorem.

Theorem 3 *Let γ be a periodic attractor. If $\lim_{t\to\infty} d(\phi_t(x), \gamma) = 0$, then there is a unique point $z \in \gamma$ such that $\lim_{t\to\infty} |\phi_t(x) - \phi_t(z)| = 0$.*

This means that any point sufficiently near to γ has the same fate as a definite point of γ.

It can be proved (not easily) that the closed orbit in the Van der Pol oscillator is a periodic attractor (see the Problems).

The proofs of Theorems 2 and 3 occupy the rest of this chapter. The proof of Theorem 2 depends on a *local section* S to the flow at p, analogous to those in Chapter 10 for planar flows: S is an open subset of an $(n-1)$-dimensional subspace transverse to the vector field at p. Following trajectories from one point of S to another, defines a C^1 map $h: S_0 \to S$, where S_0 is open in S and contains p. We call h the *Poincaré map*. The following section studies the "discrete dynamical system" $h: S_0 \to S$. In particular $p \in S_0$ is shown to be an asymptotically stable fixed point of h, and this easily implies Theorem 2.

PROBLEM

Let γ be a closed orbit of period $\lambda > 0$ in a planar dynamical system $x' = f(x)$. Let $p \in \gamma$.

(a) If

$$|\operatorname{Det} D\phi_\lambda(p)| < 1,$$

then γ is a periodic attractor, and conversely.

(b) Using the methods of Chapter 10, Section 3, and *Liouville's formula* (a proof of Liouville's formula may be found in Hartman's book [9])

$$\operatorname{Det} D\phi_\lambda(p) = \exp\left\{\int_0^\lambda \operatorname{Tr}\ Df(\phi_t p)\, dt\right\},$$

show that the closed orbit in the Van der Pol oscillator is a periodic attractor.

§2. Discrete Dynamical Systems

An important example of a discrete dynamical system (precise definition later) is a C^1 map $g: W \to W$ on an open set W of vector space which has a C^1 inverse $g^{-1}: W \to W$. Such a map is called a *diffeomorphism* of W. If W represents a "state space" of some sort, then $g(x)$ is the state of the system 1 unit of time after it is in state x. After 2 units of time it will be in state $g^2(x) = g(g(x))$; after n units, in state $g^n(x)$. Thus instead of a continuous family of states $\{\phi_t(x) \mid t \in \mathbf{R}\}$ we have the discrete family $\{g^n(x) \mid n \in \mathbf{Z}\}$, where $\mathbf{Z}$ is the set of integers.

The diffeomorphism might be a linear operator $T: E \to E$. Such systems are studied in linear algebra. We get rather complete information about their structure from the canonical form theorems of Chapter 6.

Suppose $T = e^A$, $A \in L(E)$. Then T is the "time one map" of the linear flow e^{tA}. If this continuous flow e^{tA} represents some natural dynamical process, the discrete flow $T^n = e^{nA}$ is like a series of photographs of the process taken at regular time intervals. If these intervals are very small, the discrete flow is a good approximation to the continuous one. A motion picture, for example, is a discrete flow that is hard to distinguish from a continuous one.

The analogue of an equilibrium for a discrete system $g: E \to E$ is a *fixed point* $\bar{x} = g(\bar{x})$. For a linear operator T, the origin is a fixed point. If there are other fixed points, they are eigenvectors belonging to the eigenvalue 1.

We shall be interested in stability properties of fixed points. The key example is a *linear contraction*: an operator $T \in L(E)$ such that

$$\lim_{n\to\infty} T^n x = 0 \tag{1}$$

for all $x \in E$. The time one map of a contracting flow is a linear contraction.

Proposition *The following statements are equivalent:*

(a) *T is a linear contraction;*
(b) *the eigenvalues of T have absolute values less than 1;*
(c) *there is a norm on E, and $\mu < 1$, such that*

$$|Tx| \le \mu |x|$$

for all $x \in E$.

Proof. If some real eigenvalue λ has absolute value $|\lambda| \ge 1$, (1) is not true if x is an eigenvector for λ. If $|\lambda| \ge 1$ and λ is complex, a similar argument about the complexification of T shows that T is not a contraction. Hence (a) implies (b). That (c) implies (a) is obvious; it remains to prove (b) implies (c).

We embed E in its complexification E_C, extending T to a complex linear operator T_C on E_C (Chapter 4). It suffices to find a norm on E_C as in (c) (regarding E_C as a real vector space), for then (c) follows by restricting this norm to E.

Recall that E_C is the direct sum of the generalized eigenspaces V_λ of T_C, which are invariant under T_C. It suffices to norm each of these subspaces; if $x = \sum x_\lambda$, $x_\lambda \in V_\lambda$, then we define $|x| = \max\{|x_\lambda|\}$. Thus we may replace E_C by V_λ, or what is the same thing, assume that T has only one eigenvalue λ.

A similar argument reduces us to the case where the Jordan form of T_C has only

one elementary Jordan block

$$\begin{bmatrix} \lambda & & & \\ 1 & \cdot & & \\ & \cdot & \cdot & \\ & & 1 & \lambda \end{bmatrix}.$$

For any $\epsilon > 0$ there is another basis $\{e_1, \ldots, e_m\}$ giving T_C the "ϵ-Jordan form"

$$B = \begin{bmatrix} \lambda & & & \\ \epsilon & \cdot & & \\ & \cdot & \cdot & \\ & & \epsilon & \lambda \end{bmatrix}.$$

This was proved in Chapter 7. Give E_C the max norm for this basis:

$$|\sum \alpha_j e_j| = \max\{|\alpha_j|\},$$

where $\alpha_1, \ldots, \alpha_m$ are arbitrary complex numbers. Then if $|\lambda| < 1$ and ϵ is sufficiently small, (c) is satisfied. This completes the proof of Proposition 1.

We now define a *discrete dynamical system* to be a C^1 map $g: W \to E$ where W is an open set in a vector space E. If $W \neq E$, it is possible that g^2 is not defined at all points of W, or even at any points of W. (This last case is of course uninteresting as a dynamical system.)

A *fixed point* $\bar{x} = g(\bar{x})$ of such a system is *asymptotically stable* if every neighborhood $U \subset W$ of $\bar{x}$ contains a neighborhood U_1 of $\bar{x}$ such that $g^n(U_1) \subset U$ for $n \geq 0$ and

$$\lim_{n\to\infty} g^n(x) = \bar{x}$$

for all $x \in U_1$. It follows the Proposition that 0 is asymptotically stable for a linear contraction.

In analogy with continuous flows we define a *sink* of a discrete dynamical system g to mean an *equilibrium* (that is, fixed point) at which the eigenvalues of Dg have absolute value less than 1.

The main result of this section is:

Theorem *Let $\bar{x}$ be a fixed point of a discrete dynamical system $g: W \to E$. If the eigenvalues of $Dg(\bar{x})$ are less than 1 in absolute value, $\bar{x}$ is asymptotically stable.*

Proof. We may assume $\bar{x} = 0 \in E$. Give E a norm such that for some $\mu < 1$,

$$|Dg(0)x| \leq \mu |x|$$

for all $x \in E$. Let $0 < \epsilon < 1 - \mu$. By Taylor's theorem there is a neighborhood $V \subset W$ of 0 so small that if $x \in V$, then

$$|g(x) - Dg(0)x| \leq \epsilon |x|.$$

Hence

$$\begin{aligned} |g(x)| &\leq |Dg(0)x| + \epsilon |x| \\ &\leq \mu |x| + \epsilon |x|. \end{aligned}$$

Putting $\nu = \mu + \epsilon < 1$ we have $|g(x)| \leq \nu |x|$ for $x \in V$. Given a neighborhood U of 0, choose $r > 0$ so small that the ball U_1 of radius r about 0 lies in U. Then $|g^n x| \leq \nu^n |x|$ for $x \in U_1$; hence $g^n x \in U$, and $g^n x \to 0$ as $x \to \infty$. This completes the proof.

The preceding argument can be slightly modified to show that in the specified norm,

$$|g(x) - g(y)| \leq \mu |x - y|, \qquad \mu < 1,$$

for all x, y in some neighborhood of 0 in W.

§3. Stability and Closed Orbits

We consider again the flow ϕ_t of a C^1 vector field $f: W \to E$. Let $\gamma \subset W$ be a closed orbit and suppose $0 \in \gamma$.

Suppose S is a section at 0. If $\lambda > 0$ is the period of γ, then as t increases past λ, the solution curve $\phi_t(0)$ crosses S at 0. If x is sufficiently near 0, there will be a time $\tau(x)$ near λ when $\phi_{\tau(x)}(x)$ crosses S. In this way a map

$$g: U \to S,$$

$$g(x) = \phi_{\tau(x)}(x)$$

is obtained, U being a neighborhood of 0. In fact, by Section 2 of Chapter 11, there is such a U and a unique C^1 map $\tau: U \to R$ such that $\phi_{\tau(x)}(x) \in S$ for all x in U and $\tau(0) = \lambda$.

Now let U, τ be as above and put $S_0 = S \cap U$. Define a C^1 map

$$g: S_0 \to S,$$

$$g(x) = \phi_{\tau(x)}(x).$$

Then g is a discrete dynamical system with a fixed point at 0. See Fig. A. We call g a *Poincaré map*. Note that the Poincaré map may not be definable at all points of S (Fig. B).

There is an intimate connection between the dynamical properties of the flow near γ and those of the Poincaré map near 0. For example:

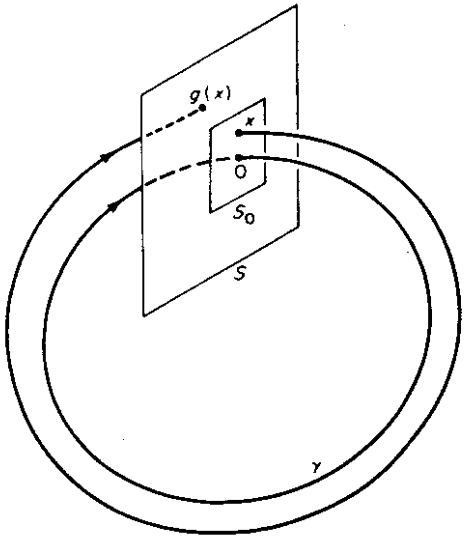


FIG. A. A Poincaré map $g: S_0 \to S$.

Proposition 1 *Let $g: S_0 \to S$ be a Poincaré map for γ as above. Let $x \in S_0$ be such that $\lim_{n\to\infty} g^n(x) = 0$. Then*

$$\lim_{t\to\infty} d(\phi_t(x), \gamma) = 0.$$

Proof. Let $g^n(x) = x_n \in S$. Since $g^{n+1}(x)$ is defined, $x_n \in S_0$. Put $r(x_n) = \lambda_n$. Since $x_n \to 0$, $\lambda_n \to \lambda$ (the period of γ). Thus there is an upper bound r for $\{|\lambda_n| \mid n \geq 0\}$. By continuity of the flow, as $n \to \infty$,

$$|\phi_s(x_n) - \phi_s(0)| \to 0$$

uniformly in $s \in [0, r]$. For any $t > 0$, there exist $s(t) \in [0, r]$, and an integer

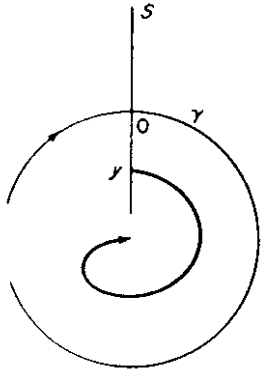


FIG. B. The Poincaré map is not defined at y.

$n(t) \geq 0$ such that

$$\phi_t(x) = \phi_{s(t)}(x_{n(t)})$$

and $n(t) \to \infty$ as $t \to \infty$. Therefore for $t > 0$

$$d(\phi_t(x), \gamma) \leq |\phi_t(x) - \phi_{s(t)}(0)|$$
$$= |\phi_{s(t)}x_{n(t)} - \phi_{s(t)}(0)|,$$

which goes to 0 as $t \to 0$.

Keeping the same notation, we also have:

Proposition 2 *If 0 is a sink for g, then γ is asymptotically stable.*

Proof. Let U be any neighborhood of γ in W; we must find U_1, a neighborhood of γ in U, such that $\phi_t(U_1) \subset U$ for all $t \geq 0$ and

$$\lim_{t\to\infty} d(\phi_t(x), \gamma) = 0$$

for all $x \in U_1$.

Let $N \subset U$ be a neighborhood of γ so small that if $x \in N$ and $|t| < 2\lambda$, then $\phi_t(x) \in U$ (where λ is the period of γ).

Let $H \subset E$ be the hyperplane containing the local section S. Since 0 is a sink, the main result of Section 2 says that H has a norm such that for some $\mu < 1$, and some neighborhood V of 0 in S_0, it is true that

$$|g(x)| \leq \mu |x|$$

for all $x \in V$. Let $\rho > 0$ be so small that the ball B_ρ in H around 0 of radius ρ is contained in $V \cap N$; and such that $r(x) < 2\lambda$ if $x \in B_\rho$.

Define

$$U_1 = \{\phi_t(x) \mid x \in B_\rho, t \geq 0\}.$$

See Fig. C. Then U_1 is a neighborhood of γ which is positively invariant. Moreover $U_1 \subset U$. For let $y \in U_1$. Then $y = \phi_t(x)$ for some $x \in B_\rho$, $t \geq 0$. We assert that (t, x) can be chosen so that $0 < t \leq r(x)$. For put $g^n(x) = x_n$. Then $x_n \in V$ for all $n \geq 0$. There exists n such that y is between x_n and x_{n+1} on the trajectory of x; since $x_n \in V$, $r(x_n) < 2\lambda$; and $y = \phi_t(x) = \phi_t(x_n)$ for $0 \leq t < 2\lambda$. Then $y \in U$ because $x_n \in N$.

Finally, $d(\phi_t(y), \gamma) \to 0$ as $t \to \infty$ for all $y \in U$. For we can write, for given y,

$$y = \phi_s(x), \qquad x \in V.$$

Since $g^n(x) \to 0$, the result follows from Proposition 1.

The following result links the derivative of the Poincaré map to that of the flow. We keep the same notation.

Proposition 3 *Let the hyperplane $H \subset E$ be invariant under $D\phi_\lambda(0)$. Then*

$$Dg(0) = D\phi_\lambda(0) | H.$$

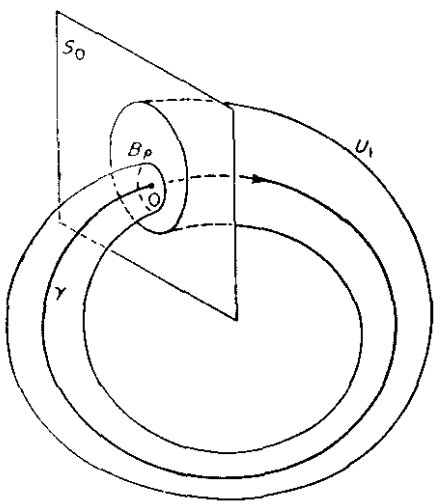


FIG. C. U_1 is positively invariant.

Proof. Let $\tau: S_0 \to \mathbf{R}$ be the C^1 map such that $\tau(0) = \lambda$ and $g(x) = \phi(\tau(x), x)$ By the remark at the end of Section 2, Chapter 11, we have

$$D\tau(0) = -\left[\frac{\partial G}{\partial t}(0, \lambda)\right]^{-1} \cdot h \cdot D\phi_\lambda(0) | H.$$

Since $D\phi_\lambda(0)(H) = H = \operatorname{Ker} h$, $D\tau(0) = 0$. Hence by the chain rule

$$Dg(0) = D\phi_\lambda(0) | H + \frac{\partial \phi}{\partial t}(\lambda, 0) D\tau(0)$$

$$= D\phi_\lambda(0) | H.$$

It is easy to see that the derivatives of any two Poincaré maps, for different sections at 0, are similar.

We now have all the ingredients for the proof of Theorem 2 of the first section. Suppose γ is a closed orbit of period λ as in that theorem. We may assume $0 \in \gamma$.

We choose an $(n-1)$-dimensional subspace H of E as follows. H is like an eigenspace corresponding to the eigenvalues of $D\phi_\lambda(0)$ with absolute value less than 1. Precisely, let $B \subset E_{\mathbf{C}}$ be the direct sum of the generalized eigenspaces belonging to these eigenvalues for the complexification $(D\phi_\lambda(0))_{\mathbf{C}}: E_{\mathbf{C}} \to E_{\mathbf{C}}$, and let $H = B \cap E$. Then H is an $(n-1)$-dimensional subspace of E invariant under $D\phi_\lambda(0)$ and the restriction $D\phi_\lambda(0) | H$ is a linear contraction.

Let $S \subset H$ be a section at 0 and $g: S_0 \to S$ a Poincaré map. The previous proposition implies that the fixed point $0 \in S_0$ is a sink for g. By Proposition 2, γ is asymptotically stable.

To prove Theorem 3, it suffices to consider a point $x \in S_0$ where $g: S_0 \to S$ is the Poincaré map of a local section at $0 \in \gamma$ (since every trajectory starting near γ intersects S_0).

If $\phi_{n\lambda}(x)$ is defined and sufficiently near 0 for $n = 1, \ldots, k$, then

$$\phi_{n\lambda}(x) = \phi_{t_n}(g^n x),$$

where

$$t_n = t_{n-1} + \tau(g^{n-1}x) - \lambda.$$

For some $\nu < 1$ and some norm on E we have

$$| g^n x | \leq \nu | g^{n-1} x |;$$

and using $D\tau(0) = 0$, we know that for any $\epsilon > 0$,

$$| t_n - t_{n-1} | \leq \epsilon | g^{n-1} x | \leq \epsilon \nu^{n-1} | x |$$

if $| x |$ is sufficiently small. Thus

$$| t_n | \leq | t_0 | + \epsilon \sum_{k=0}^{\infty} \nu^k = \frac{\epsilon}{1-\nu}.$$

Hence if ϵ is sufficiently small, the sequence $\phi_{n\lambda}(x)$ stays near 0 and can be continued for all positive integers n, and the above inequalities are valid for all n. It follows that the sequence $\{t_n\}$ is Cauchy and converges to some $s \in \mathbf{R}$. Thus $\phi_{n\lambda}(x)$ converges to $\phi_s(0) = z \in \gamma$. This implies Theorem 3 of Section 1.

PROBLEMS

1. Show that the planar system

$$x' = (1 - x^2 - y^2)x - y,$$

$$y' = x + (1 - x^2 - y^2)y$$

has a unique closed orbit γ and compute its Poincaré map. Show that γ is a periodic attractor. (*Hint*: Use polar coordinates.)

2. Let X denote either a closed orbit or an equilibrium. If X is asymptotically stable, show that for every $\lambda > 0$ there is a neighborhood U of X such that if $p \in U - X$, then $\phi_t(p) \neq p$ for all $t \in [0, \lambda]$.

3. Show that a linear flow cannot have an asymptotically stable closed orbit.

4. Define the concepts of *stable closed orbit* of a flow, and *stable fixed point* of a discrete dynamical system. Prove the following:
   (a) A closed orbit is stable if and only if its Poincaré map has a stable fixed point at 0.
   (b) If a closed orbit γ of period λ is stable then no eigenvalue of $D\phi_\lambda(p)$, $p \in \gamma$, has absolute value more than one, but the converse can be false.

5. (a) Let p be an asymptotically stable fixed point of a discrete dynamical system $g: W \to E$. Show that p has arbitrarily small compact neighborhoods $V \subset W$ such that $g(V) \subset \operatorname{int} V$ and $\bigcap_{n\geq 0} g^n(V) = p$.
   (b) State and prove the analogue of (a) for closed orbits.

6. Let $g: \mathbf{R} \to \mathbf{R}$ be the map

$$g(x) = ax + bx^2 + cx^3, \qquad a \neq 0.$$

Investigate the fixed point 0 for stability and asymptotic stability (see Problem 4). Consider separately the cases $|a| < 1, |a| = 1, |a| > 1$.

7. (The Contracting Map Theorem) Let $X \subset \mathbf{R}^n$ be a nonempty closed set and $f: X \to X$ a continuous map. Suppose f has a Lipschitz constant $\alpha < 1$. Prove that f has unique fixed point p, and $\lim_{n\to\infty} f^n(x) = p$ for all $x \in X$. (*Hint*: Consider the sequence $f^n(x)$.)

Chapter 14

Classical Mechanics

The goal of this very short chapter is to do two things: (1) to give a statement of the famous n-body problem of celestial mechanics and (2) to give a brief introduction to Hamiltonian mechanics. We give a more abstract treatment of Hamiltonian theory than is given in physics texts; but our method exhibits invariant notions more clearly and has the virtue of passing easily to the case where the configuration space is a manifold.

§1. The n-Body Problem

We give a description of the n-body "problem" of celestial mechanics; this extends the Kepler problem of Chapter 2. The basic example of this mechanical system is the solar system with the sun and planets representing the n bodies. Another example is the system consisting of the earth, moon, and sun. We are concerned here with Newtonian gravitational forces; on the other hand, the Newtonian n-body problem is the prototype of other n-body problems, with forces other than gravitational.

The data, or parameters of this system, are n positive numbers representing the masses of the n bodies. We denote these numbers by $m_1, \ldots, m_n$.

The first goal in understanding a mechanical system is to define the *configuration* space, or space of generalized positions. In this case a configuration will consist precisely of the positions of each of the n bodies. We will write x_i for the position of the ith body so that x_i is a point in Euclidean three space (the space in which we live) denoted by E. Now E is isomorphic to Cartesian space $\mathbf{R}^3$ but by no natural isomorphism. However E does have a natural notion of inner product and associated norm; the notions of length and perpendicular make sense in the space in which we live, while any system of coordinate axes is arbitrary.

Thus Euclidean three space, the configuration space of one body, is a three-dimensional vector space together with an inner product.

The configuration space M for the n-body problem is the Cartesian product of E with itself n times; thus $M = (E)^n$ and $x = (x_1, \ldots, x_n)$, where $x_i \in E$ is the position of the ith body. Note that x_i denotes a point in E, not a number.

One may deduce the space of states from the configuration space as the space T_M of all tangent vectors to all possible curves in M. One may think of T_M as the product $M \times M$ and represent a state as $(x, v) \in M \times M$, where x is a configuration as before and $v = (v_1, \ldots, v_n)$, $v_i \in E$ being the velocity of the ith body. A state of the system gives complete information about the system at a given moment and (at least in classical mechanics) determines the complete life history of the state.

The determination of this life history goes via the ordinary differential equations of motion, Newton's equations in this instance. Good insights into these equations can be obtained by introducing kinetic and potential energy.

The *kinetic energy* is a function $K: M \times M \to \mathbf{R}$ on the space of states which is given by

$$K(x, v) = \frac{1}{2} \sum_{i=1}^{n} m_i \mid v_i \mid^2.$$

Here the norm of v_i is the Euclidean norm on E. One may also consider K to be given directly by an inner product B on M by

$$B(v, w) = \frac{1}{2} \sum_{1}^{n} m_i \langle v_i, w_i \rangle,$$

$$K(x, v) = B(v, v).$$

It is clear that B defines an inner product on M where $\langle v_i, w_i \rangle$ means the original inner product on E.

The *potential energy* V is a function on M defined by

$$V(x) = \sum_{i<j} \frac{m_i m_j}{\mid x_i - x_j \mid}.$$

We suppose that the gravitational constant is 1 for simplicity. Note that this function is not defined at any "collision" (where $x_i = x_j$). Let Δ_{ij} be the subspace of collisions of the ith and jth bodies so that

$$\Delta_{ij} = \{x \in M \mid x_i = x_j\}, \; i < j.$$

Thus Δ_{ij} is a linear subspace of the vector space M. Denote the space of all collisions by $\Delta \subset M$ so that $\Delta = \bigcup \Delta_{ij}$. Then properly speaking, the domain of the potential energy is $M - \Delta$:

$$V: M - \Delta \to \mathbf{R}.$$

We deal then with the space of noncollision states which is $(M - \Delta) \times M$.

Newton's equations are second order equations on $M - \Delta$ which may be written

$$m_i \ddot{x}_i = -\text{grad}_i V(x) \qquad \text{for} \quad i = 1, \ldots, n.$$

Here the partial derivative $D_i V$ of V with respect to x_i is a map from $M - \Delta$ to $L(E, \mathbf{R})$; then the inner product on E converts $D_i V(x)$ to a vector which we call grad$_i$ $V(x)$. The process is similar to the definition of gradient in Chapter 9. Thus the equations make sense as written.

One may rewrite Newton's equations in such a way that they become a first order system on the space of states $(M - \Delta) \times M$:

$$\dot{x}_i = v_i,$$

$$m_i \dot{v}_i = -\text{grad}_i V(x), \quad \text{for} \quad i = \ldots, n.$$

The flow obtained from this differential equation then determines how a state moves in time, or the life history of the n bodies once their positions and velocities are given. Although there is a vast literature of several centuries on these equations, no clear picture has emerged. In fact it is still not even clear what the basic questions are for this "problem."

Some of the questions that have been studied include: Is it true that almost all states do not lead to collisions? To what extent are periodic solutions stable? How to show the existence of periodic solutions? How to relate the theory of the n-body problem to the orbits in the solar system?

Our present goal is simply to put Newton's equations into the framework of this book and to see how they fit into the more abstract framework of Hamiltonian mechanics.

We put the n-body problem into a little more general setting. The key ingredients are:

(1) Configuration space Q, an open set in a vector space E (in the above case $Q = M - \Delta$ and $E = M$).

(2) A C^2 function $K: Q \times E \to \mathbf{R}$, kinetic energy, such that $K(x, v)$ has the form $K(x, v) = K_x(v, v)$, where K_x is an inner product on E (in the above case K_x was independent of x, but in problems with constraints, K_x depends on x).

(3) A C^2 function $V: Q \to \mathbf{R}$, potential energy.

The triple (Q, K, V) is called a *simple mechanical system*, and $Q \times E$ the *state space* of the system. Given a simple mechanical system (Q, K, V) the *energy* or total energy is the function $e: Q \times E \to \mathbf{R}$ defined by $e(x, v) = K(x, v) + V(x)$.

For a simple mechanical system, one can canonically define a vector field on $Q \times E$ which gives the equations of motion for the states (points of $Q \times E$). We will see how this can be done in the next section.

Examples of simple mechanical systems beside the n-body problem include a particle moving in a conservative central force field, a harmonic oscillator, and a frictionless pendulum. If one extends the definition of simple mechanical systems to permit Q to be a manifold, then a large part of classical mechanics may be analyzed in this framework.

§2. Hamiltonian Mechanics

We shall introduce Hamiltonian mechanics from a rather abstract point of view, and then relate it to the Newtonian point of view. This abstract development proceeds quite analogously to the modern treatment of gradients using inner products; now however the inner product is replaced by a "symplectic form." So we begin our discussion by defining this kind of form.

If F is a vector space, a *symplectic form* Ω on F is a real-valued bilinear form that is antisymmetric and nondegenerate. Thus

$$\Omega: F \times F \to \mathbf{R}$$

is a bilinear map that is *antisymmetric*: $\Omega(u, v) = -\Omega(v, u)$, and *nondegenerate*, which means that the map

$$\Phi_\Omega = \Phi: F \to F^*$$

is an isomorphism. Here Φ is the linear map from F to F^* defined by

$$\Phi(u)(v) = \Omega(u, v), \qquad u, v \in F.$$

It turns out that the existence of a symplectic form on F implies that the dimension of F is even (see the Problems).

We give an example of such a form Ω_0 on every even dimensional vector space. If F is an even dimensional vector space, we may write F in the form $F = E \times E^*$, the Cartesian product of a vector space E and its dual E^*. Then an element f of F is of the form (v, w) where v, w are vectors of E, E^*, respectively. Now if $f = (v, w)$, $f^0 = (v^0, w^0)$ are two vectors of F, we define

$$\Omega_0(f, f^0) = w^0(v) - w(v^0).$$

Then it is easy to check that Ω_0 is a symplectic form on F. The nondegeneracy is obtained by showing that if $\alpha \neq 0$, then one may find β such that $\Omega_0(\alpha, \beta) \neq 0$. Note that Ω_0 does not depend on a choice of coordinate structure on E, so that it is natural on $E \times E^*$.

If one chooses a basis for E, and uses the induced basis on E^*, Ω_0 is expressed in coordinates by

$$\Omega_0((v, w), (v^0, w^0)) = \sum w_i^0 v_i - \sum w_i v_i^0.$$

It can be shown that every symplectic form is of this type for some representation of F as $E \times E^*$.

Now let U be an open subset of a vector space F provided with a symplectic form Ω. There is a prescription for assigning to any C^2 function $H: U \to \mathbf{R}$, a C_1 vector field X_H on U called the *Hamiltonian vector field* of H. In this context H is called a *Hamiltonian* or a Hamiltonian function. To obtain X_H let $DH: U \to F^*$ be the derivative of H and simply write

$$X_H(x) = \Phi^{-1}DH(x), \qquad x \in U, \tag{1}$$

where Φ^{-1} is the inverse of the isomorphism $\Phi: F \to F^*$ defined by Ω above. (1) is equivalent to saying $\Omega(X_H(x), y) = DH(x)(y)$, all $y \in F$. Thus $X_H: U \to F$ is a C^1 vector field on U; the differential equations defined by this vector field are called *Hamilton's equations*. By using coordinates we can compare these with what are called Hamilton's equations in physics books.

Let Ω_0 be the symplectic form on $F = E \times E^*$ defined above and let $x = (x_1, \ldots, x_n)$ represent points of E and $y = (y_1, \ldots, y_n)$ points of E^* for the dual coordinate structures on E and E^*. Let $\Phi_0: F \to F^*$ be the associated isomorphism.

For any C^2 function $H: U \to \mathbf{R}$,

$$DH(x, y) = \sum_1^n \frac{\partial H}{\partial x_i}\, dx_i + \sum_1^n \frac{\partial H}{\partial y_i}\, dy_i.$$

From this, one has that $\Phi_0^{-1}DH(x, y)$ is the vector with components

$$X_H(x, y) = \left(\frac{\partial H}{\partial y_1}, \ldots, \frac{\partial H}{\partial y_n}, -\frac{\partial H}{\partial x_1}, \ldots, -\frac{\partial H}{\partial x_n}\right).$$

This is seen as follows. Observe that (suppressing (x, y))

$$\Phi_0(X_H) = DH$$

or

$$\Omega_0(X_H, \omega) = DH(\omega) \qquad \text{for all} \quad \omega \in F.$$

By letting ω range over the standard basis elements of $\mathbf{R}^{2n}$, one confirms the expression for X_H. The differential equation defined by the vector field X_H is then:

$$x_i' = \frac{\partial H}{\partial y_i}, \qquad i = 1, \ldots, n,$$

$$y_i' = -\frac{\partial H}{\partial x_i}, \qquad i = 1, \ldots, n.$$

These are the usual expressions for Hamilton's equations.

Continuing on the abstract level we obtain the "conservation of energy" theorem. The reason for calling it by this name is that in the mechanical models described

in this setting, H plays the role of energy, and the solution curves represent the motions of states of the system.

Theorem (Conservation of Energy) *Let U be an open set of a vector space F, $H: U \to \mathbf{R}$ any C^2 function and Ω a symplectic form on F. Then H is constant on the solution curves defined by the vector field X_H.*

Proof. If $\phi_t(x)$ is a solution curve of the vector field X_H, then it has to be shown that

$$\frac{d}{dt} H(\phi_t(x)) = 0, \qquad \text{all} \quad x, t.$$

This expression by the chain rule is

$$DH(\phi_t(x)) \left(\frac{d}{dt} \phi_t(x)\right) = DH(X_H).$$

But $DH(X_H)$ is simply, by the definition of X_H, $\Omega(X_H, X_H)$ which is 0 since Ω is antisymmetric. This ends the proof.

It is instructive to compare this development with that of a gradient dynamical system. These are the same except for the character of the basic bilinear form involved; for one system it is an inner product and for the other it is a symplectic form. The defining function is constant on solution curves for the Hamiltonian case, but except at equilibria, it is increasing for the gradient case.

From the point of view of mechanics, the Hamiltonian formulation has the advantage that the equations of motion are expressed simply and without need of coordinates, starting just from the energy H. Furthermore, conservation laws follow easily and naturally, the one we proved being the simplest example. Rather than pursue this direction however, we turn to the question of relating abstract Hamiltonian mechanics to the more classical approach to mechanics. We shall see how the energy of a simple mechanical system can be viewed as a Hamiltonian H; the differential equations of motion of the system are then given by the vector field X_H.

Thus to a given simple mechanical system (Q, K, V), we will associate a Hamiltonian system $H: U \to \mathbf{R}$, $U \subset F$, Ω a symplectic form on F in a natural way.

Recall that configuration space Q is an open set in a vector space E and that the state space of the simple mechanical system is $Q \times E$. The space of generalized momenta or *phase space* of the system is $Q \times E^*$, where E^* is the dual vector space of Q.

The relation between the state space and the phase space of the system is given by the *Legendre transformation* $\lambda: Q \times E \to Q \times E^*$. To define λ, first define a linear isomorphism $\lambda_q: E \to E^*$, for each $q \in Q$, by

$$\lambda_q(v)w = 2K_q(v, w); \qquad v \in E, \quad w \in E.$$

Then set

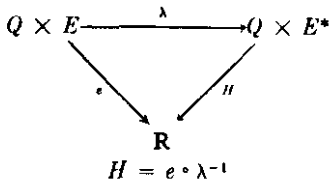
$$\lambda(q, v) = (q, \lambda_q(v)).$$

Consider the example of a simple mechanical system of a particle with mass m moving in Euclidean three space E under a conservative force field given by potential energy V. In this case state space is $E \times E$ and $K: E \times E \to \mathbf{R}$ is given by $K(q, v) = \frac{1}{2}m \mid v \mid^2$. Then $\lambda: E \times E \to E \times E^*$ is given by $\lambda_q(v) = p \in E^*$, where $p(w) = 2K_q(v, w)$; or

$$p(w) = m\langle v, w \rangle$$

and $\langle\ ,\ \rangle$ is the inner product on E. In a Cartesian coordinate system on E, $p = mv$, so that the image p of v under λ is indeed the classical momentum, "conjugate" to v.

Returning to our simple mechanical system in general, note that the Legendre transformation has an inverse, so that λ is a diffeomorphism from the state space to the phase space. This permits one to transfer the energy function e on state space to a function H on phase space called the *Hamiltonian* of a *simple mechanical system*. Thus we have



$$H = e \circ \lambda^{-1}$$

The final step in converting a simple mechanical system to a Hamiltonian system is to put a symplectic form on $F = E \times E^* \supset Q \times E^* = U$. But we have already constructed such a form Ω_0 in the early part of this section. Using (q, p) for variables on $Q \times E^*$, then Hamilton's equations take the form in coordinates

$$q_i' = \frac{\partial H}{\partial p_i}, \qquad i = 1, \ldots, n,$$

$$p_i' = -\frac{\partial H}{\partial q_i}, \qquad i = 1, \ldots, n.$$

Since for a given mechanical system H (interpreted as total energy) is a known function of p_i, q_i, these are ordinary differential equations. The basic assertion of Hamiltonian mechanics is that they describe the motion of the system.

The justification for this assertion is twofold. On one hand, there are many cases where Hamilton's equations are equivalent to Newton's; we discuss one below. On the other hand, there are common physical systems to which Newton's laws do not directly apply (such as a spinning top), but which fit into the framework of "simple mechanical systems," especially if the configuration space is allowed to be a surface or higher dimensional manifold. For many such systems, Hamilton's equations have been verified experimentally.

It is meaningless, however, to try to deduce Hamilton's equations from Newton's on the abstract level of simple mechanical system (Q, K, V). For there is no identification of the elements of the "configuration space" Q with any particular physical or geometrical parameters.

Consider as an example the special case above where $K(q, v) = \frac{1}{2} \sum m_i v_i^2$ in Cartesian coordinates. Then $m_i v_i = p_i$ and $H(p, q) = \sum (p_i^2/2m_i) + V(q)$; Hamilton's equations become

$$q_i' = \frac{p_i}{m_i},$$

$$p_i' = -\frac{\partial V}{\partial q_i}.$$

Differentiating the first and combining these equations yield

$$mq_i'' = -\frac{\partial V}{\partial q_i}.$$

These are the familiar Newton's equations, again. Conversely, Newton's equations imply Hamilton's in this case.

PROBLEMS

1. Show that if the vector space F has a symplectic form Ω on it, then F has even dimension. *Hint*: Give F an inner product $\langle\ ,\ \rangle$ and let $A: F \to F$ be the operator defined by $\langle Ax, y\rangle = \Omega(x, y)$. Consider the eigenvectors of A.

2. (Lagrange) Let (Q, K, V) be a simple mechnnical system and X_H the associated Hamiltonian vector field on phase space. Show that $(q, 0)$ is an equilibrium for X_H if and only if $DV(q) = 0$; and $(q, 0)$ is a stable equilibrium if q is an isolated minimum of V. (*Hint*: Use conservation of energy.)

3. Consider the second order differential equation in one variable

$$\ddot{x} + f(x) = 0,$$

where $f: \mathbf{R} \to \mathbf{R}$ is C^2 and if $f(x) = 0$, then $f'(x) \neq 0$. Describe the orbit structure of the associated system in the plane

$$\dot{x} = v$$

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

when $f(x) = x - x^3$. Discuss this phase-portrait in general. (*Hint*: Consider

$$H(x, v) = \tfrac{1}{2}v^2 + \int_0^x f(t)\, dt$$

and show that H is constant on orbits. The critical points of H are at $v = 0$, $f(x) = 0$; use $H_{xx} = f'(x)$, $H_{vv} = 1$.)

4. Consider the equation

$$\ddot{x} + g(x)\dot{x} + f(x) = 0,$$

where $g(x) > 0$, and f is as in Problem 3. Describe the phase portrait (the function g may be interpreted as coming from friction in a mechanical problem).

Notes

One modern approach to mechanics is Abraham's book, *Foundations of Mechanics* [1]. Wintner's *Analytical Foundations of Celestial Mechanics* [25] has a very extensive treatment of the n-body problem.

Chapter 15

Nonautonomous Equations and Differentiability of Flows

This is a short technical chapter which takes care of some unfinished business left over from Chapter 8 on fundamental theory. We develop existence, uniqueness, and continuity of solutions of nonautonomous equations $x' = f(t, x)$. Even though our main emphasis is an autonomous equations, the theory of nonautonomous linear equations $x' = A(t)x$ is needed as a technical device in establishing differentiability of flows. The variational equation along a solution of an autonomous equation is an equation of this type.

§1. Existence, Uniqueness, and Continuity for Nonautonomous Differential Equations

Let E be a normed vector space, $W \subset \mathbf{R} \times E$ an open set, and $f: W \to E$ a continuous map. Let $(t_0, u_0) \in W$. A solution to the initial value problem

$$x'(t) = f(t, x), \qquad x(t_0) = u_0 \tag{1}$$

is a differentiable curve $x(t)$ in E defined for t in some interval J having the following properties:

$$t_0 \in J \quad \text{and} \quad x(t_0) = u_0,$$

$$(t, x(t)) \in W, \quad x'(t) = f(t, x(t))$$

for all $t \in J$.

We call the function $f(t, x)$ *Lipschitz in* x if there is a constant $K \geq 0$ such that

$$|f(t, x_1) - f(t, x_2)| \leq K |x_1 - x_2|$$

for all (t, x_1) and (t, x_2) in W.

The fundamental local theorem for nonautonomous equations is:

Theorem 1 *Let $W \subset \mathbf{R} \times E$ be open and $f: W \to E$ a continuous map that is Lipschitz in x. If $(t_0, u_0) \in W$, there is an open interval J containing t and a unique solution to (1) defined on J.*

The proof is the same as that of the fundamental theorem for autonomous equations (Chapter 8), the extra variable t being inserted where appropriate.

The theorem applies in particular to functions $f(t, x)$ that are C^1, or even continuously differentiable only in x; for such an f is locally Lipschitz in x (in the obvious sense). In particular we can prove:

Theorem 2 *Let $A: J \to L(E)$ be a continuous map from an open interval J to the space of linear operators on E. Let $(t_0, u_0) \in J \times E$. Then the inital value problem*

$$x' = A(t)x, \qquad x(t_0) = u_0$$

has a unique solution on all of J.

Proof. It suffices to find a solution on every compact interval; by uniqueness such solutions can be continued over J. If $J_0 \subset J$ is compact, there is an upper bound K to the norms of the operators $A(t)$, $t \in J_0$. Such an upper bound is a Lipschitz constant in x for $f \mid J_0 \times E$, and Theorem 1 can be used to prove Theorem 2.

As in the autonomous case, solutions of (1) are continuous with respect to initial conditions if $f(t, x)$ is locally Lipschitz in x. We leave the precise formulation and proof of this fact to the reader.

A different kind of continuity is continuity of solutions as functions of the *data* $f(t, x)$. That is, if $f: W \to E$ and $g: W \to E$ are both Lipschitz in x, and $|f - g|$ is uniformly small, we expect solutions to $x' = f(t, x)$ and $y' = g(t, y)$, having the same initial values, to be close. This is true; in fact we have the following more precise result.

Theorem 3 *Let $W \subset \mathbf{R} \times E$ be open and $f, g: W \to E$ continuous. Suppose that for all $(t, x) \in W$,*

$$|f(t, x) - g(t, x)| < \epsilon.$$

Let K be a Lipschitz constant in x for $f(t, x)$. If $x(t)$, $y(t)$ are solutions to

$$x' = f(t, x),$$

$$y' = g(t, y),$$

respectively, on some interval J, and $x(t_0) = y(t_0)$, then

$$|x(t) - y(t)| \leq \frac{\epsilon}{K}(\exp(K|t - t_0|) - 1)$$

for all $t \in J$.

Proof. For $t \in J$ we have

$$x(t) - y(t) = \int_{t_0}^{t} [x'(s) - y'(s)]\,ds$$

$$= \int_{t_0}^{t} [f(s, x(s)) - g(s, y(s))]\,ds.$$

Hence

$$|x(t) - y(t)| \leq \int_{t_0}^{t} |f(s, x(s)) - f(s, y(s))|\,ds$$

$$+ \int_{t_0}^{t} |f(s, y(s)) - g(s, y(s))|\,ds$$

$$\leq \int_{t_0}^{t} K|x(s) - y(s)|\,ds + \int_{t_0}^{t} \epsilon\,ds.$$

Let $u(t) = |x(t) - y(t)|$. Then

$$u(t) \leq K\int_{t_0}^{t}\left[u(s) + \frac{\epsilon}{K}\right]ds,$$

$$u(t) + \frac{\epsilon}{K} \leq \frac{\epsilon}{K} + K\int_{t_0}^{t}\left[u(s) + \frac{\epsilon}{K}\right]ds.$$

It follows from Gronwall's inequality (Chapter 8) that

$$u(t) + \frac{\epsilon}{K} \leq \frac{\epsilon}{K}\exp(K|t - t_0|),$$

which yields the theorem.

§2. Differentiability of the Flow of Autonomous Equations

Consider an autonomous differential equation

$$x' = f(x), \qquad f: W \to E, \qquad W \text{ open in } E, \tag{1}$$

where f is assumed C^1. Our aim is to show that the flow

$$(t, x) \to \phi(t, x) = \phi_t(x)$$

defined by (1) is a C^1 function of two variables, and to identify $\partial\phi/\partial x$.

To this end let $y(t)$ be a particular solution of (1) for t in some open interval J. Fix $t_0 \in J$ and put $y(t_0) = y_0$. For each $t \in J$ put

$$A(t) = Df(y(t));$$

thus $A: J \to L(E)$ is continuous. We define a nonautonomous linear equation

$$u' = A(t)u. \tag{2}$$

This is the *variational equation of* (1) *along the solution* $y(t)$.

From Section 1 we know that (2) has a solution on all of J for every initial condition $u(t_0) = u_0$.

The significance of (2) is that *if u_0 is small, then the map*

$$t \to y(t) + u(t)$$

is a good approximation to the solution $x(t)$ of (1) *with initial value $x(t_0) = y_0 + u_0$.*

To make this precise we introduce the following notation. If $\xi \in E$, let the map

$$t \to u(t, \xi)$$

be the solution to (2) which sends t_0 to ξ. If ξ and $y_0 + \xi \in W$, let the map

$$t \to y(t, \xi)$$

be the solution to (1) which sends t_0 to $y_0 + \xi$. (Thus $y(t, \xi) = \phi_{t-t_0}(y_0 + \xi)$.)

Proposition *Let $J_0 \subset J$ be a compact interval containing t_0. Then*

$$\lim_{\xi \to 0} \frac{|y(t, \xi) - y(t) - u(t, \xi)|}{|\xi|} = 0$$

uniformly in $t \in J_0$.

This means that for every $\epsilon > 0$, there exists $\delta > 0$ such that if $|\xi| \leq \delta$, then

$$|y(t, \xi) - (y(t) + u(t, \xi))| \leq \epsilon|\xi| \tag{3}$$

for all $t \in J_0$. Thus as $\xi \to 0$, the curve $t \to y(t) + u(t, \xi)$ is a better and better approximation to $y(t, \xi)$. In many applications $y(t) + u(t, \xi)$ is used in place of $y(t, \xi)$; this is convenient because $u(t, \xi)$ is linear in ξ.

We will prove the proposition presently. First we use (3) to prove:

Theorem 1 *The flow $\phi(t, x)$ of* (1) *is C^1; that is, $\partial\phi/\partial t$ and $\partial\phi/\partial x$ exist and are continuous in (t, x).*

Proof. Of course $\partial\phi(t, x)/\partial t$ is just $f(\phi_t(x))$, which is continuous. To compute $\partial\phi/\partial x$ we have, for small ξ,

$$\phi(t, y_0 + \xi) - \phi(t, y_0) = y(t, \xi) - y(t).$$

The proposition now implies that $\partial\phi(t, y_0)/\partial x \in L(E)$ is the linear map $\xi \to u(t, \xi)$. The continuity of $\partial\phi/\partial x$ is a consequence of the continuity in initial conditions and data of solutions to the variational equation (2).

Denoting the flow again by $\phi_t(x)$, we note that for each t the derivative $D\phi_t(x)$ of the map ϕ_t at $x \in W$ is the same as $\partial\phi(t, x)/\partial x$. We call this the *space derivative* of the flow, as opposed to the *time derivative* $\partial\phi(t, x)/\partial t$.

The proof of the preceding theorem actually computes $D\phi_t(x)$ as the solution to an initial value problem in the vector space $L(E)$: *for each $x_0 \in W$ the space derivative of the flow satisfies*

$$\frac{d}{dt}(D\phi_t(x_0)) = Df(\phi_t(x_0))D\phi_t(x_0),$$

$$D\phi_0(x_0) = I.$$

Here we regard x_0 as a parameter. An important special case is that of an equilibrium $\bar{x}$ so that $\phi_t(\bar{x}) \equiv \bar{x}$. Putting $Df(\bar{x}) = A \in L(E)$, we get

$$\frac{d}{dt}(D\phi_t(\bar{x})) = AD\phi_t(\bar{x}),$$

$$D\phi_0(\bar{x}) = I.$$

The solution to this is

$$D\phi_t(\bar{x}) = e^{tA}.$$

This means that *in a neighborhood of an equilibrium the flow is approximately linear.*

We now prove the proposition. For simplicity we take $t_0 = 0$. The integral equations satisfied by $y(t, \xi)$, $y(t)$, and $u(t, \xi)$ are

$$y(t) = y_0 + \int_0^t f(y(s))\,ds,$$

$$y(t, \xi) = y_0 + \xi + \int_0^t f(y(s, \xi))\,ds,$$

$$u(t, \xi) = \xi + \int_0^t Df(y(s))u(s, \xi)\,ds.$$

From these we get

$$|y(t, \xi) - y(t) - u(t, \xi)| \le \int_0^t |f(y(s, \xi) - f(y(s)) - Df(y(s))u(s, \xi)|\,ds. \tag{4}$$

The Taylor estimate for f says

$$f(y) - f(z) = Df(x)(y - z) + R(z, y - z),$$

$$\lim_{z \to y} R(z, y - z)/|y - z| = 0$$

uniformly in y for y in a given compact set. We apply this to $y = y(s, \xi)$, $z = y(s)$; from linearity of $Df(y(s))$ and (4) we get

$$|y(t, \xi) - y(t) - u(t, \xi)| \le \int_0^t |Df(y(s))[y(s, \xi) - y(s) - u(s, \xi)]|\,ds + \int_0^t |R(y(s), y(s, \xi) - y(s))|\,ds. \tag{5}$$

Denote the left side of (5) by $g(t)$ and put

$$N = \max\{\|Df(y, s)\| \mid s \in J_0\}.$$

Then from (5) we get

$$g(t) \le N\int_0^t g(s)\,ds + \int_0^t |R(y(s), y(s, \xi) - y(s))|\,ds. \tag{6}$$

Fix $\epsilon > 0$ and pick $\delta_0 > 0$ so small that

$$|R(y(s), y(s, \xi) - y(s))| \le \epsilon\,|y(s, \xi) - y(s)| \tag{7}$$

if $|y(s, \xi) - y(s)| \le \delta_0$ and $s \in J_0$.

From Chapter 8, Section 4 there are constants $K \ge 0$ and $\delta_1 > 0$ such that

$$|y(s, \xi) - y(s)| \le |\xi|\,e^{Ks} \le \delta_0 \tag{8}$$

if $|\xi| \le \delta_1$ and $s \in J_0$.

Assume now that $|\xi| \le \delta_1$. From (6), (7), and (8) we find, for $t \in J_0$,

$$g(t) \le N\int_0^t g(s)\,ds + \int_0^t \epsilon\,|\xi|\,e^{Ks}\,ds,$$

whence

$$g(t) \le N\int_0^t g(s)\,ds + C\epsilon\,|\xi|$$

for some constant C depending only on K and the length of J_0. Applying Gronwall's

inequality we obtain

$$g(t) \leq Ce^{Nt} |\xi|$$

if $t \in J_0$ and $|\xi| \leq \delta_1$. (Recall that δ_1 depends on ϵ.) Since ϵ is any positive number, this shows that $g(t)/|\xi| \to 0$ uniformly in $t \in J_0$, which proves the proposition.

We show next that the flow enjoys the same degree of differentiability as does the data.

A function $f: W \to E$ is called C^r, $1 \leq r < \infty$ if it has r continuous derivatives. For $r \geq 2$ this is equivalent to: f is C^1 and $Df: W \to L(E)$ is C^{r-1}. If f is C^r for all $r \geq 1$, we say f is C^∞. We let C^0 mean "continuous."

Theorem 2 *Let $W \subset E$ be open and let $f: W \to E$ be C^r, $1 \leq r \leq \infty$. Then the flow $\phi: \Omega \to E$ of the differential equation*

$$x' = f(x)$$

is also C^r.

Proof. We induct on r, the case $r = 1$ having been proved in Theorem 1.

We may suppose $r < \infty$ for the proof.

Suppose, as the inductive hypothesis, that $r \geq 2$ and that the flow of every differential equation

$$\xi' = F(\xi),$$

with C^{r-1} data F, is C^{r-1}.

Consider the differential equation on $E \times E$ defined by the vector field

$$F: W \times E \to E \times E, \qquad F(x, u) = (f(x), Df(x)w),$$

$$\frac{d}{dt}(x, u) = F(x, u),$$

or equivalently,

$$x' = f(x), \qquad u' = Df(x)u. \tag{9}$$

Since F is C^{r-1}, the flow Φ of (9) is C^{r-1}. But this flow is just

$$\Phi(t, (x, u)) = (\phi(t, x), D\phi_t(x)u),$$

since the second equation in (9) is the variational equation of the first equation. Therefore $\partial\phi/\partial x$ is a C^{r-1} function of (t, x), since $\partial\phi/\partial x = D\phi_t(x)$. Moreover $\partial\phi/\partial t$ is C^{r-1} (in fact, C^r in t) since

$$\frac{\partial\phi}{\partial t} = f(\phi(t, x)).$$

It follows that ϕ is C^r since its first partial derivatives are C^{r-1}.

PROBLEMS

1. Let $A: \mathbf{R} \to L(E)$ be continuous and let $P: \mathbf{R} \to L(E)$ be the solution to the initial value problem

$$P' = A(t)P, \qquad P(0) = P_0 \in L(E).$$

Show that

$$\operatorname{Det} P(t) = (\operatorname{Det} P_0) \exp\left[\int_0^t \operatorname{Tr} A(s)\, ds\right].$$

2. Show that if f is C^r, some r with $0 \leq r \leq \infty$, and $x(t)$ is a solution to $x' = f(x)$, then x is a C^{r+1} function.

Chapter 16

Perturbation Theory and Structural Stability

This chapter is an introduction to the problem: What effect does changing the differential equation itself have on the solution? In particular, we find general conditions for equilibria to persist under small perturbations of the vector field. Similar results are found for periodic orbits. Finally, we discuss briefly more global problems of the same type. That is to say, we consider the question: When does the phase portrait itself persist under perturbations of the vector field? This is the problem of structural stability.

§1. Persistence of Equilibria

Let W be an open set in a vector space E and $f: W \to E$ a C^1 vector field. By a *perturbation* of f we simply mean another C^1 vector field on W which we think of as being "C^1 close to f," that is,

$$|f(x) - g(x)| \qquad \text{and} \qquad \|Df(x) - Dg(x)\|$$

are small for all $x \in W$.

To make this more precise, let $\mathcal{U}(W)$ be the set of all C^1 vector fields on W. If E has a norm, we define the *C^1-norm* $\| h \|_1$ of a vector field $h \in \mathcal{U}(W)$ to be the least upper bound of all the numbers

$$|h(x)|, \quad \|Dh(x)\|; \qquad x \in W.$$

We allow the possibility $\| h \|_1 = \infty$ if these numbers are unbounded.

A *neighborhood* of $f \in \mathcal{U}(W)$ is any subset $\mathfrak{N} \subset \mathcal{U}(W)$ that contains a set of the form

$$\{g \in \mathcal{U}(W) \mid \| g - f \|_1 < \epsilon\}$$

for some $\epsilon > 0$ and some norm on E.

The set $\mathcal{U}(W)$ has the formal properties of a vector space under the usual operations of addition and scalar multiplication of vector-valued functions. The C^1 norm has many of the same formal properties as the norms for vector spaces defined earlier, namely,

$$\| h \|_1 \geq 0,$$

$$\| h \|_1 = 0 \qquad \text{if and only if} \qquad h = 0,$$

$$\| h + g \|_1 \leq \| h \|_1 + \| g \|_1,$$

where if $\| h \|_1$ or $\| g \|_1$ is infinite, the obvious interpretation is made.

We can now state our first perturbation theorem.

Theorem 1 *Let $f: W \to E$ be a C^1 vector field and $\bar{x} \in W$ an equilibrium of $x' = f(x)$ such that $Df(\bar{x}) \in L(E)$ is invertible. Then there exists a neighborhood $U \subset W$ of $\bar{x}$ and a neighborhood $\mathfrak{N} \subset \mathcal{U}(W)$ of f such that for any $g \in \mathfrak{N}$ there is a unique equilibrium $\bar{y} \in U$ of $y' = g(y)$. Moreover, if E is normed, for any $\epsilon > 0$ we can choose $\mathfrak{N}$ so that $| \bar{y} - \bar{x} | < \epsilon$.*

Theorem 1 applies to the special case where $\bar{x}$ is a hyperbolic equilibrium, that is, the eigenvalues of $Df(\bar{x})$ have nonzero real parts. In this case, the *index* $\text{ind}(\bar{x})$ of $\bar{x}$ is the number of eigenvalues (counting multiplicities) of $Df(\bar{x})$ having negative real parts. If $\dim E = n$, then $\text{ind}(\bar{x}) = n$ means $\bar{x}$ is a sink, while $\text{ind}(\bar{x}) = 0$ means it is a source. We can sharpen Theorem 1 as follows:

Theorem 2 *Suppose that $\bar{x}$ is a hyperbolic equilibrium. In Theorem 1, then, $\mathfrak{N}$, U can be chosen so that if $g \in \mathfrak{N}$, the unique equilibrium $\bar{y} \in U$ of $y' = g(y)$ is hyperbolic and has the same index as $\bar{x}$.*

Proof. This follows from a theorem in Chapter 7 and Theorem 1.

The proof of Theorem 1 has nothing to do with differential equations; rather, it depends on the following result about C^1 maps:

Proposition *Let $f: W \to E$ be C^1 and suppose $x_0 \in W$ is such that the linear operator $Df(x_0): E \to E$ is invertible. Then there is a neighborhood $\mathfrak{N} \subset \mathcal{U}(W)$ of f and an open set $U \subset W$ containing x_0 such that if $g \in \mathfrak{N}$, then*

(a) *$g \mid U$ is one-to-one, and*
(b) *$f(x_0) \in g(U)$.*

Theorem 1 follows by taking $x_0 = \bar{x}$ and $f(\bar{x}) = 0$, for then $g(\bar{y}) = 0$ for a unique $\bar{y} \in U$. To make $|\bar{y} - \bar{x}| < \epsilon$ (assuming E is normed now) we simply replace W by $W_0 = \{x \in W \mid |x - \bar{x}| < \epsilon\}$. The proposition guarantees that $\mathfrak{N}$ can be chosen so that U, and hence $\bar{y}$, is in W_0 for any $g \in \mathfrak{N}$.

It remains to prove the proposition. In the following lemmas we keep the same notation.

Lemma 1 *Assume E is normed. Let*

$$\nu > \|Df(x_0)^{-1}\|.$$

Let $V \subset W$ be an open ball around x_0 such that

$$\|Df(y)^{-1}\| < \nu, \tag{1}$$

and

$$\|Df(y) - Df(z)\| < 1/\nu \tag{2}$$

for all $y, z \in V$. Then $f \mid V$ is one-to-one.

Proof. If $y \in V$ and $u \in E$ is nonzero, then

$$u = Df(y)^{-1}(Df(y)u);$$

hence

$$|u| \leq \|Df(y)^{-1}\| \, |Df(y)u|,$$

so, from (1),

$$|Df(y)(u)| > \frac{|u|}{\nu}. \tag{3}$$

Now let y, z be distinct points of V with $z = y + u$. Note that since V is a ball, $y + tu \in V$ for all $t \in [0, 1]$. Define a C^1 map $\varphi: [0, 1] \to E$ by

$$\varphi(t) = f(y + tu).$$

Then

$$\varphi(0) = f(y), \qquad \varphi(1) = f(z).$$

By the chain rule,

$$\varphi'(t) = Df(y + tu)u.$$

Hence

$$f(z) - f(y) = \int_0^1 Df(y + tu)u \, dt$$

$$= \int_0^1 Df(y)u \, dt + \int_0^1 [Df(y + tu) - Df(y)]u \, dt.$$

Therefore

$$|f(z) - f(y)| \geq |Df(y)u| - \int_0^1 \|Df(y + tu) - Df(y)\| \, |u| \, dt.$$

From (3) and (2) we then get

$$|f(y) - f(z)| > \frac{|u|}{\nu} - \frac{|u|}{\nu} = 0.$$

Thus $f(y) \neq f(z)$. This proves Lemma 1.

Lemma 2 *Suppose E is a normed vector space with norm defined by an inner product. Let $B \subset W$ be a closed ball around x_0 with boundary ∂B, and $f: W \to E$ a C^1 map. Suppose $Df(y)$ is invertible for all $y \in B$. Let*

$$\min\{|f(y) - f(x_0)| \mid y \in \partial B\} > 2\delta > 0.$$

Then $w \in f(B)$ if $|w - f(x_0)| < \delta$.

Proof. Since B is compact, there exists $y_0 \in B$ at which the function

$$H: B \to \mathbf{R},$$

$$H(y) = \tfrac{1}{2}|f(y) - w|^2$$

takes a minimal value. Note that y_0 cannot be in ∂B, for if $y \in \partial B$, then

$$|f(y) - w| \geq |f(y) - f(x_0)| - |f(x_0) - w|$$

$$> 2\delta - \delta.$$

Hence

$$|f(y) - w| > \delta > |f(x_0) - w|,$$

showing that $|f(y) - w|$ is not minimal if $y \in \partial B$.

Since the norm on E comes from an inner product, $\frac{1}{2}|x|^2$ is differentiable; its derivative at x is the linear map $z \to \langle x, z \rangle$. By the chain rule, H is differentiable and its derivative at y_0 is the linear map

$$z \to DH(y_0)z = \langle f(y_0) - w, Df(y_0)z \rangle.$$

Since y_0 is a critical point of H and y_0 in an interior point of B, $DH(y_0) = 0$. Since $Df(y_0)$ is invertible, there exists $v \in E$ with

$$Df(y_0)v = f(y_0) - w.$$

Then

$$0 = DH(y_0)v$$

$$= \langle f(y_0) - w, f(y_0) - w \rangle$$

$$= |f(y_0) - w|^2.$$

Therefore $f(y_0) = w$, proving Lemma 2.

Note that the proof actually shows that

$$w \in f(B - \partial B).$$

To prove the proposition we give E a norm coming from an inner product. The subset of invertible operators in the vector space $L(E)$ is open. Therefore there exists $\alpha > 0$ such that $A \in L(E)$ is invertible if

$$\| A - Df(x_0) \| < \alpha.$$

Since the map $x \to Df(x)$ is continuous, there is a neighborhood $N_1 \subset W$ of x_0 such that if $x \in N_1$, then

$$\| Df(x) - Df(x_0) \| < \tfrac{1}{2}\alpha.$$

It follows that if $g \in \mathcal{U}(W)$ is such that

$$\| Dg(x) - Df(x) \| < \tfrac{1}{2}\alpha$$

for all $x \in N_1$, then $Dg(x)$ is invertible for all $x \in N_1$. The set of such g is a neighborhood $\mathfrak{N}_1$ of f.

Let $\nu > \| Df(x_0)^{-1} \|$. The map $A \to A^{-1}$ from invertible operators to $L(E)$, is continuous (use the formula in Appendix I for the inverse of a matrix). It follows that f has a neighborhood $\mathfrak{N}_2 \subset \mathfrak{N}_1$ and x_0 has a neighborhood $N_2 \subset N_1$ such that if $g \in \mathfrak{N}_2$ and $y \in N_2$, then

$$\| Dg(y)^{-1} \| < \nu.$$

We can find still smaller neighborhoods, $\mathfrak{N}_3 \subset \mathfrak{N}_2$ of f and $N_3 \subset N_2$ of x_0, such that if $g \in \mathfrak{N}_3$ and $y, z \in N_3$, then

$$\| Dg(y) - Dg(z) \| < \frac{1}{\nu}.$$

It now follows from Lemma 1 that for any ball $V \subset N$ and any $g \in \mathfrak{N}_3$, $g \mid V$ is one-to-one.

Fix a ball $V \subset N_3$ around x_0. Let $B \subset V$ be a closed ball around x_0 and choose $\delta > 0$ as in Lemma 2. There is a neighborhood $\mathfrak{N} \subset \mathfrak{N}_3$ of f such that if $g \in \mathfrak{N}$, then

$$\min\{ | g(y) - g(x_0) | \mid y \in \partial B \} > 2\delta > 0.$$

It follows that if $| w - g(x_0) | < \delta$ and $g \in \mathfrak{N}$, then $w \in g(B)$. The proposition is now proved using this $\mathfrak{N}$ and taking $U = V$.

We have not discussed the important topic of nonautonomous perturbations. Problem 2 shows that in a certain sense the basin of attraction of a sink persists under small nonautonomous perturbations.

PROBLEMS

1. Show that the stable and unstable manifolds of a hyperbolic equilibrium of a linear differential equation $x' = Ax$ vary continuously with linear perturbations of $A \in L(E)$. That is, suppose $E^u \oplus E^s$ is the invariant splitting of E such that $e^{tA}: E^u \to E^u$ is an expanding linear flow and $e^{tA}: E^s \to E^s$ is contracting. Given $\epsilon > 0$, there exists $\delta > 0$ such that if $\| B - A \| < \delta$, then B leaves invariant a splitting $F^u \oplus F^s$ of E such that $e^{tB} \mid F^u$ is expanding, $e^{tB} \mid F^s$ is contracting, and there is a linear isomorphism $T: E \to E$ such that $T(E^u) = F^u$, $T(E^s) = F^s$, and $\| T - I \| < \epsilon$.
2. Let $W \subset \mathbf{R}^n$ be an open set and $0 \in W$ an asymptotically stable equilibrium of a C^1 vector field $f: W \to \mathbf{R}^n$. Assume that 0 has a strict Liapunov function. Then 0 has a neighborhood $W_0 \subset W$ with the following property. For any $\epsilon > 0$ there exists $\delta > 0$ such that if $g: \mathbf{R} \times W \to \mathbf{R}$ is C^1 and $| g(t, x) - f(x) | < \delta$ for all (t, x), then every solution $x(t)$ to $x' = g(t, x)$ with $x(t_0) \in W$ satisfies $x(t) \in W$ for all $t \geq t_0$ and $| x(t) | < \epsilon$ for all t greater than some t_1. (*Hint*: If V is a strict Liapunov function for 0, then $(d/dt)(V(x(t))$ is close to $(d/dt)(V(y(t))$, where $y' = f(y)$. Hence $(d/dt)(V(x(t)) < 0$ if $| x(t) |$ is not too small. Imitate the proof of Liapunov's theorem.)

§2. Persistence of Closed Orbits

In this section we consider a dynamical system ϕ_t defined by a C^1 vector field $f: W \to E$ where $W \subset E$ is an open set. We suppose that there is a closed orbit $\gamma \subset W$ of period $\lambda > 0$. For convenience we assume the origin $0 \in E$ is in γ. The main result is:

Theorem 1 *Let $u: S_0 \to S$ be a Poincaré map for a local section S at 0. Let $U \subset W$ be a neighborhood of γ. Suppose that 1 is not an eigenvalue of $Du(0)$. Then there exists a neighborhood $\mathfrak{N} \subset \mathcal{U}(W)$ of f such that every vector field $g \in \mathfrak{N}$ has a closed orbit $\beta \subset U$.*

The condition on the Poincaré map in Theorem 1 is equivalent to the condition that the eigenvalue 1 of $D\phi_\lambda(0)$ has multiplicity 1. Unfortunately, no equivalent condition on the vector field f is known.

Proof of the theorem. Let $\tau: S_0 \to \mathbf{R}$ be the C^1 map such that $\tau(0) = \lambda$ and

$$u(x) = \phi_{\tau(x)}(x).$$

We may assume that the closure of S_0 is a compact subset of S. Let $\alpha > 0$. There exists $\delta_0 > 0$ such that if $g \in \mathfrak{N}(W)$ and $|g(x) - f(x)| < \delta_0$ for all $x \in S_0$, then, first, S will be a local section at 0 for g, and second, there is a C^1 map $\sigma\colon S_0 \to \mathbf{R}$ such that

$$|\sigma(x) - \tau(x)| < \alpha,$$

$$\psi_{\sigma(x)}(x) \in S,$$

and

$$|\psi_{\sigma(x)}(x) - u(x)| < \alpha,$$

where ψ_t is the flow of g.

Put

$$\psi_{\sigma(x)}(x) = v(x).$$

Then

$$v\colon S_0 \to S$$

is a C^1 map which is a kind of Poincaré map for the flow ψ_t.

Given any $t_0 > 0$ and any compact set $K \subset W$, and any $\nu > 0$ we can ensure that

$$\|D\phi_t(x) - D\psi_t(x)\| < \nu$$

for all $t \in [-t_0\ t_0]$, $x \in K$, provided we make $\|g - f\|_1$ small enough. This follows from continuity of solutions of differential equations as functions of the original data and initial conditions and the expression of $\partial\psi_t(x)/\partial x$ as solutions of the nonautonomous equation in $L(E)$,

$$\frac{dA}{dt} = Dg(y(t))A(t),$$

where $y' = g(y)$. (See Chapter 15.)

From this one can show that provided $\|g - f\|_1$ is small enough, one can make $|u(x) - v(x)|$ and $\|Du(x) - Dv(x)\|$ as small as desired for all $x \in S_0$.

A fixed point $x = v(x)$ of v lies on a closed orbit of the flow ψ_t. We view such a fixed point as a zero of the C^1 map

$$\eta\colon S_0 \to H, \qquad \eta(x) = v(x) - x,$$

where H is the hyperplane containing S.

Let $\xi\colon S_0 \to H$ be the C^1 map

$$\xi(x) = u(x) - x$$

so that $\xi(0) = 0$. Now

$$D\xi(0) = Du(0) - I,$$

where $I\colon H \to H$ is the identity. Since 1 is not an eigenvalue of $Du(0)$ we know that 0 is not an eigenvalue of $D\xi(0)$, that is, $D\xi(0)$ *is invertible*. From the proposition in the preceding section we can find a neighborhood $\mathfrak{M} \subset \mathcal{U}(S_0)$ of ξ such that any map in $\mathfrak{M}$ has a unique zero $y \in S_0$. If $\|g - f\|_1$ is sufficiently small, $\eta \in \mathfrak{M}$.

Hence η has a unique zero $y \in S_0$; and y lies on a closed orbit β of g. Moreover, we can make y so close to 0 that $\beta \subset U$. This proves Theorem 1.

The question of the uniqueness of the closed orbit of the perturbation is interesting. It is *not* necessarily unique; in fact, it is possible that *all* points of U lie on closed orbits of f! But it is true that closed orbits other than γ will have periods much bigger than γ. In fact, given $\epsilon > 0$, there exists $\delta > 0$ so small that if $0 < d(x, \gamma) < \delta$ and $\phi_t(x) = x$, $t > 0$, then $t > 2\lambda - \epsilon$. The same will hold true for sufficiently small perturbations of γ: the fixed point y of v that we found above lies on a closed orbit β of g whose period is within ϵ of λ; while any other closed orbit of g that meets S_0 will have to circle around β *several times* before it closes up. This follows from the relation of closed orbits to the sections; see Fig. A.

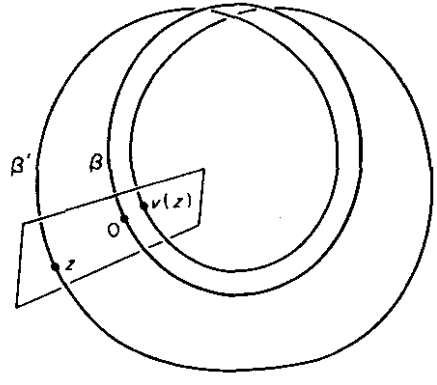


FIG. A. A closed orbit β' near a hyperbolic closed orbit β.

There is one special case where the uniqueness of the closed orbit of the perturbation can be guaranteed: if γ is a periodic attractor and g is sufficiently close to f, then β will also be a periodic attractor; hence every trajectory that comes near β winds closer and closer to β as $t \to \infty$ and so cannot be a closed orbit.

Similarly, if γ is a periodic repeller, so is β, and again uniqueness holds.

Consider next the case where γ is a *hyperbolic closed orbit*. This means that the derivative at $0 \in \gamma$ of the Poincaré map has no eigenvalues of absolute value 1. In this case a weaker kind of uniqueness obtains: there is a neighborhood $V \subset U$ of γ such that if $\mathfrak{N}$ is small enough, every $g \in \mathfrak{N}$ will have a unique closed orbit *that is entirely contained in* V. It is possible, however, for every neighborhood of a hyperbolic closed orbit to intersect other closed orbits, although this is hard to picture.

We now state without proof an important approximation result. Let $B \subset \mathbf{R}^n$ be a closed ball and ∂B its boundary sphere.

Theorem 2 *Let $W \subset \mathbf{R}^n$ be an open set containing B and $f\colon W \to \mathbf{R}^n$ a C^1 vector field which is transverse to ∂B at every point of ∂B. Let $\mathfrak{N} \subset \mathcal{U}(W)$ be any neighborhood*

of f. Then there exists $g \in \mathfrak{N}$ such that:

(a) *if $\bar{x} \in B$ is an equilibrium of g, then $\bar{x}$ is hyperbolic;*
(b) *if $\gamma \subset B$ is a closed orbit of g, then γ is hyperbolic.*

The condition that f be transverse to ∂B is not actually necessary, and in fact, B can be replaced by any compact subset of W.

PROBLEMS

1. Show that the eigenvalue condition in the main theorem of this section is necessary.
2. Let γ be a periodic attractor of $x' = f(x)$. Show there is a C^1 real-valued function $V(x)$ on a neighborhood of γ such that $V \geq 0$, $V^{-1}(0) = \gamma$, and $(d/dt)(V(x(t))) < 0$ if $x(t)$ is a solution curve not in γ. (*Hint*: Let $z(t)$ be the solution curve in γ such that $x(t) - z(t) \to 0$ as $t \to \infty$; see Chapter 13, Section 1, Theorem 3. Consider $\int_0^T |x(t) - z(t)|^2\, dt$ for some large constant T.)
3. Let $W \subset \mathbf{R}^n$ be open and let γ be a periodic attractor for a C^1 vector field $f: W \to \mathbf{R}^n$. Show that γ has a neighborhood U with the following property. For any $\epsilon > 0$ there exists $\delta > 0$ such that if $g: \mathbf{R} \times W \to \mathbf{R}^n$ is C^1 and $|g(t, x) - f(x)| < \delta$, then every solution $x(t)$ to $x' = g(t, x)$ with $x(t_0) \in U$ satisfies $x(t) \in U$ for all $t \geq t_0$ and $d(x(t), \gamma) < \epsilon$ for all t greater than some t_1. (*Hint*: Problem 2, and Problem 2 of Section 1.)

§3. Structural Stability

In the previous sections we saw that certain features of a flow may be preserved under small perturbations. Thus if a flow has a sink or attractor, any nearby flow will have a nearby sink; similarly, for periodic attractors.

It sometimes happens that any nearby flow is topologically the same as a given flow, that is, for any sufficiently small perturbation of the flow, a homeomorphism exists that carries each trajectory of the original flow onto a trajectory of the perturbation. (A *homeomorphism* is simply a continuous map, having a continuous inverse.) Such a homeomorphism sets up a one-to-one correspondence between equilibria of the two flows, closed orbits, and so on. In this case the original flow (or its vector field) is called *structurally stable*.

Here is the precise definition of structural stability, at least in the restricted setting of vector fields which point in on the unit disk (or ball) in $\mathbf{R}^n$. Let

$$D^n = \{x \in \mathbf{R}^n \mid |x| \leq 1\}$$

and

$$\partial D^n = \{x \in \mathbf{R}^n \mid |x| = 1\}.$$

Consider C^1 vector fields $f: W \to \mathbf{R}^n$ defined on some open set W containing D^n such that $\langle f(x), x\rangle < 0$ for each x in ∂D^n. Such an f is called *structurally stable* on D^n if there exists a neighborhood $\mathfrak{N} \subset \mathfrak{V}(W)$ such that if $g: W \to \mathbf{R}^n$ is in $\mathfrak{N}$, then flows of f and g are topologically equivalent on D^n. This means there exists a homeomorphism $h: D^n \to D^n$ such that for each $x \in D^n$,

$$h(\{\phi_t(x) \mid t \geq 0\}) = \{\psi_t(h(x)) \mid t \geq 0\},$$

where ψ_t is the flow of g; and if x is not an equilibrium, h preserves the orientation of the trajectory. (The orientation of the trajectory is simply the direction that points move along the curve as t increases.)

This is a very strong condition on a vector field. It means that the flow ϕ_t cannot have any "exceptional" dynamical features in D^n. For example, it can be shown that if $\bar{x} \in \operatorname{int} D^n$ is an equilibrium, then it must be hyperbolic; the basic reason is that linear flows with such equilibria are generic.

The harmonic oscillator illustrates the necessity of this condition as follows. Suppose that $f: W \to \mathbf{R}^2$, with $W \supset D^2$, is a vector field which in some neighborhood of 0 is given by

$$x' = Ax, \qquad A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

By arbitrary slight perturbation, the matrix A can be changed to make the origin either a sink, saddle, or source. Since these have different dynamic behavior, the flows are not topologically the same. Hence f is not structurally stable. In contrast, it is known that the Van der Pol oscillator is structurally stable.

The following is the main result of this section. It gives an example of a class of structurally stable systems. (See Fig. A.)

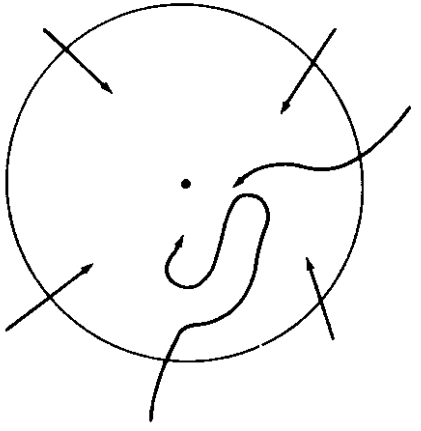


FIG. A. A structurally stable vector field.

Theorem 1 *Let f: $W \to \mathbf{R}^n$ be a C^1 vector field on an open set $W \supset D^n$ with the following properties*:

(a) *f has exactly one equilibrium $0 \in D^n$, and 0 is a sink;*
(b) *f points inward along the boundary ∂D^n of D^n, that is,*

$$\langle f(x), x\rangle < 0 \quad \text{if} \quad x \in \partial D^n.$$

(c) *$\lim_{t\to\infty} \phi_t(x) = 0$ for all $x \in D^n$, where ϕ_t is the flow of f.*

Then f is structurally stable on D^n.

Before proving this we mention three other results on structural stability. These concern a C^1 vector field f: $W \to \mathbf{R}^2$ where $W \subset \mathbf{R}^2$ is a neighborhood of D^2. The first is from the original paper on structural stability by Pontryagin and Andronov.

Theorem 2 *Suppose f points inward on D^2. Then the following conditions taken together are equivalent to structural stability on D^2:*

(a) *the equilibria in D^2 are hyperbolic;*
(b) *each closed orbit in D^2 is either a periodic attractor or a periodic repeller (that is, a periodic attractor for the vector field $-f(x)$);*
(c) *no trajectory in D^2 goes from saddle to saddle.*

The necessity of the third condition is shown by breaking a saddle connection as in Fig. B(a) by an approximation as in Fig. B(b).

A good deal of force is given to Theorem 2 by the following result of Peixoto; it implies that structural stability on D^2 is a generic condition. Let $\mathcal{U}_0(W)$ be the set of C^1 vector fields on W that point inward on ∂D^2.

Theorem 3 *The set*

$$\mathcal{S} = \{f \in \mathcal{U}_0(W) \mid f \text{ is structurally stable on } D^2\}$$

is dense and open. That is, every element of $\mathcal{S}$ has a neighborhood in $\mathcal{U}_0(W)$ contained in $\mathcal{S}$, and every open set in $\mathcal{U}_0(W)$ contains a vector field which is structurally stable on D^2.

Unfortunately, it has been shown that there can be no analogue of Theorem 3 for dimensions greater than 2. Nevertheless, there are many interesting vector fields that are structurally stable, and the subject continues to inspire a lot of research.

In the important case of gradient dynamical systems, there is an analogue of Theorem 3 for higher dimensions as follows. Consider in $\mathcal{U}(D^n)$ the set $\operatorname{grad}(D^n)$ of gradient vector fields that point inward on D^n.

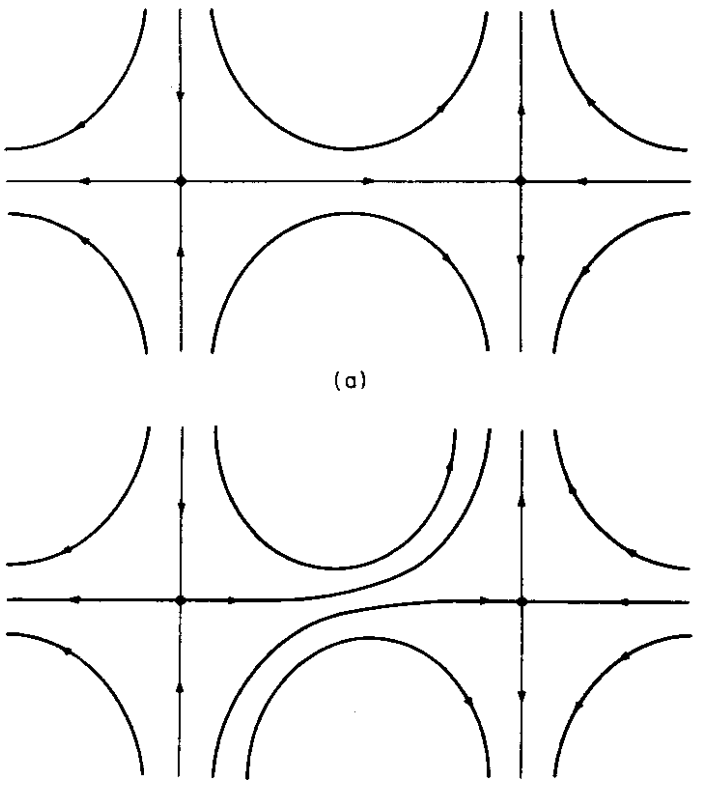


FIG. B. (a) Flow near a saddle connection; (b) breaking a saddle connection.

Theorem 4 *The set of structurally stable systems contained in* $\operatorname{grad}(D^n)$ *is open and dense in* $\operatorname{grad}(D^n)$.

We turn to the proof of Theorem 1. In outline it proceeds as follows. A vector field g sufficiently close to f is shown to have a unique equilibrium $a \in D^n$ near 0; moreover, all trajectories of g in D^n tend toward a. Once this is known, the homeomorphism h: $D^n \to D^n$ is defined to be the identity on ∂D^n; for each $x \in \partial D^n$ it maps the f-trajectory of x onto the g-trajectory of x preserving the parametrization; and $h(0) = a$.

The proof is based on the following result which is interesting in itself. In Section 1 we showed the persistence of a hyperbolic equilibrium under small perturbations. In the special case of a sink we have a sharper result showing that the basin of attraction retains a certain size under perturbation.

Proposition *Let $0 \in E$ be a sink for a C^1 vector field $f: W \to E$ where W is an open set containing 0. There exists an inner product on E, a number $r > 0$, and a neighborhood $\mathfrak{N} \subset \mathcal{U}(W)$ of f such that the following holds: for each $g \in \mathfrak{N}$ there is a sink $a = a(g)$ for g such that the set*

$$B_r = \{x \in E \mid |x| \leq r\}$$

contains a, is in the basin of a, and is positively invariant under the flow of g.

Proof. From Chapter 9 we give E an inner product with the following property. For some $\nu < 0$ and $2r > 0$ it is true that

$$\langle f(x), x\rangle < \nu |x|^2$$

if $0 < |x| < 2r$. It follows that B_r is in the basin of 0, and that $f(x)$ points inward along ∂B_r. It is clear that f has a neighborhood $\mathfrak{N}_0 \subset \mathcal{U}(W)$ such that if $g \in \mathfrak{N}_0$, then also $g(x)$ points inward along ∂B_r.

Let $0 < \epsilon < r$ and put $s = r + \epsilon$. If $|y| < \epsilon$, then the closed ball $B_\epsilon(y)$ about y with radius ϵ satisfies:

$$B_r \subset B_\epsilon(y) \subset B_{2r}.$$

Let $\nu < \mu < 1$. We assert that if $\|g - f\|_1$ is sufficiently small, then the sink a of g will be in B_ϵ, and moreover,

$$\langle g(x), x - a\rangle \leq \mu |x - a|^2 \tag{1}$$

if $x \in B_\epsilon(a)$. To see this, write

$$\begin{aligned}\langle g(x), x - a\rangle &= \langle f(x - a), x - a\rangle + \langle g(x) - f(x - a), x - a\rangle \\ &\leq \nu |x - a|^2 + \langle g(x) - f(x - a), x - a\rangle.\end{aligned}$$

The map $\alpha(x) = g(x) - f(x - a)$ vanishes at a. The norm of its derivative at x is estimated thus:

$$\|D\alpha(x)\| \leq \|Dg(x) - Df(x)\| + \|Df(x) - Df(x - a)\|;$$

as $\|g - f\|_1 \to 0$, $\|Dg(x) - Df(x)\| \to 0$ uniformly for $|x| \leq 2r$; and also $x - a \to 0$, so $\|Df(x) - Df(x - a)\| \to 0$ uniformly for $|x| \leq 2r$. Thus if $\|g - f\|_1$ is small enough, $\|D\alpha(x)\| \leq \mu - \nu$, and $\mu - \nu$ is a Lipschitz constant for α; hence

$$|\alpha(x)| = |\alpha(x) - \alpha(a)| \leq (\mu - \nu)|x - a|.$$

Consequently, if $\|g - f\|_1$ is sufficiently small, say, less than $\delta > 0$,

$$\begin{aligned}\langle g(x), x - a\rangle &\leq \nu |x - a|^2 + \langle \alpha(x), x - a\rangle \\ &\leq \nu |x - a|^2 + (\mu - \nu)|x - a|^2 \\ &= \mu |x - a|^2\end{aligned}$$

as required.

Put $\mathfrak{N}_1 = \{g \in \mathcal{U}(W) \mid \|g - f\|_1 < \delta\}$, and set $\mathfrak{N} = \mathfrak{N}_0 \cap \mathfrak{N}_1$. Suppose $g \in \mathfrak{N}$, with sink $a \in B_\epsilon$. By (1) the set $B_\epsilon(a)$ is in the basin of a. Since $B_r \subset B_\epsilon(a)$, and $g(x)$ points inward along ∂B_r, the proof is complete.

We now prove Theorem 1. Since D^n is compact and $f(x)$ points inward along the boundary, no solution curve can leave D^n. Hence D^n is positively invariant. Choose $r > 0$ and $\mathfrak{N} \subset \mathcal{U}(W)$ as in the proposition. Let $\mathfrak{N}_0 \subset \mathfrak{N}$ be a neighborhood of f so small that if $g \in \mathfrak{N}_0$, then $g(x)$ points inward along ∂D^n. Let ψ_t be the flow of $g \in \mathfrak{N}_0$. Note that D^n is also positively invariant for ψ_t.

For every $x \in D^n - \operatorname{int} B_r$, there is a neighborhood $U_x \subset W$ of x and $t_x > 0$ such that if $y \in U_x$ and $t \geq t_x$, then

$$|\phi_t(y)| < r.$$

By compactness of ∂D^n a finite number $U_{x_1}, \ldots, U_{x_k}$ of the sets U_x cover ∂D^n. Put

$$t_0 = \max\{t_{x_1}, \ldots, t_{x_k}\}.$$

Then $\phi_t(D^n - \operatorname{int} B_r) \subset B_r$, if $t \geq t_0$. It follows from continuity of the flow in f (Chapter 15) that f has a neighborhood $\mathfrak{N}_1 \subset \mathfrak{N}$ such that if $g \in \mathfrak{N}_1$, then

$$\psi_t(D^n - \operatorname{int} B_r) \subset B_r \qquad \text{if} \quad t \geq t_0.$$

This implies that

$$\lim_{t\to\infty} \psi_t(x) = a \qquad \text{for all} \quad x \in D^n.$$

For let $x \in D^n$; then $y = \psi_{t_0}(x) \in B_r$, and $B_r \subset$ basin of a under ψ_t.

It also implies that every $y \in D^n - a$ is of the form $\psi_t(x)$ for some $x \in \partial D^n$ and $t \geq 0$. For otherwise $L_a(y)$ is not empty; but if $z \in L_a(y)$, then $\psi_t(z) \to a$ as $t \to \infty$, hence $y = a$.

Fix $g \in \mathfrak{N}_1$. We have proved so far that the map

$$\Psi: [0, \infty) \times \partial D^n \to D^n,$$

$$\Psi(t, x) = \psi_t(x)$$

has $D^n - a$ for its image. And the map

$$\Phi: [0, \infty) \times \partial D^n \to D^n,$$

$$\Phi(x, t) = \phi_t(x)$$

has $D^n - 0$ as its image. We define

$$h: D^n \to D^n,$$

$$h(y) = \begin{cases} \Psi\Phi^{-1}(y) & \text{if} \quad y \neq 0, \\ a & \text{if} \quad y = 0. \end{cases}$$

Another way of saying this is that h maps $\phi_t(x)$ to $\psi_t(x)$ for $x \in \partial D^n$, $t \geq 0$, and $h(0) = a$; therefore h maps trajectories of ϕ to trajectories of ψ, preserving orientation. Clearly, $h(D^n) = D^n$. The continuity of h is verified from continuity of the flows, and by reversing the role of the flow and its perturbation one obtains a con-

tinuous inverse to h. Thus h is a homeomorphism; the proof of Theorem 1 is complete.

PROBLEMS

1. Show that if $f: \mathbf{R}^2 \to \mathbf{R}^2$ is structurally stable on D^2 and $f(0) = 0$, then 0 is a hyperbolic equilibrium.

2. Let $\gamma \subset \mathbf{R}^n$, $n \geq 2$ be the circle

$$\gamma = \{x \in \mathbf{R}^n \mid x_1^2 + x_2^2 = 4,\ x_k = 0 \quad \text{for} \quad k > 2\}.$$

Let

$$N = \{x \in \mathbf{R}^n \mid d(x, \gamma) \leq 1\}.$$

Let $W \subset \mathbf{R}^n$ be a neighborhood of N and $f: W \to \mathbf{R}^n$ a C^1 vector field. Suppose $f(x)$ points into N for all x in $\partial N = \{x \in \mathbf{R}^n \mid d(x, \gamma) = 1\}$. If γ is a periodic attractor and $\gamma = L_\omega(x)$ for all $x \in N$, prove that f is structurally stable on N. (See Fig. C for $n = 3$.)

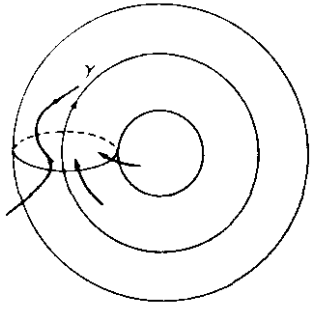


FIG. C

3. If $f \in \mathcal{U}(W)$ is structurally stable on $D^n \subset \mathbf{R}^n$, show that f has a neighborhood $\mathfrak{N}$ such that every $g \in \mathfrak{N}$ is structurally stable.

4. Show that Theorem 1 can be sharpened as follows. For every $\epsilon > 0$ there is a neighborhood $\mathfrak{N}$ of f such that if $g \in \mathfrak{N}$ the homeomorphism h (in the definition of structural stability) can be chosen so that $| h(x) - x | < \epsilon$ for all $x \in D^n$.

5. Find necessary and sufficient conditions that a vector field $f: \mathbf{R} \to \mathbf{R}$ be structurally stable on a compact interval.

6. Let A be an operator on $\mathbf{R}^n$ such that the linear flow e^{tA} is hyperbolic. Find $\epsilon > 0$ such that if B is an operator on $\mathbf{R}^n$ satisfying $\| B - A \| < \epsilon$, then there is a homeomorphism of $\mathbf{R}^n$ onto itself that takes each trajectory of the differential equation $x' = Ax$ onto a trajectory of $y' = By$.

Afterword

This book is only an introduction to the subject of dynamical systems. To proceed further requires the treatment of differential equations on manifolds; and the formidable complications arising from infinitely many closed orbits must be faced.

This is not the place to develop the theory of manifolds, but we can try to indicate their use in dynamical systems. The surface S of the unit ball in $\mathbf{R}^3$ is an example of the two-dimensional manifold. A vector field on $\mathbf{R}^3$ might be tangent to S at all points of S; if it is, then S is invariant under the flow. In this way we get an example of a dynamical system on the manifold S (see Fig. A).

The compactness of S implies that solution curves of such a system are defined for all $t \in \mathbf{R}$. This is in fact true for all flows on compact manifolds, and is one reason for the introduction of manifolds.

Manifolds arise quite naturally in mechanics. Consider for example a simple mechanical system as in Chapter 14. There is the Hamiltonian function $H: U \to \mathbf{R}$, where U is an open subset of a vector space. The "conservation of energy" theorem states that H is constant on trajectories. Another way of saying the same thing is that if $H(x) = c$, then the whole trajectory of x lies in the subset $H^{-1}(c)$. For "most" values of c this subset is a submanifold of U, just as the sphere S in $\mathbf{R}^3$ can be viewed as $H^{-1}(1)$ where $H(x, y, z) = x^2 + y^2 + z^2$. The dimension of $H^{-1}(c)$ is one less than that of U. Other first integrals cut down the dimension even further. In the planar Kepler problem, for example, the state space is originally an open subset U of $\mathbf{R}^4$. The flow conserves both total energy H and angular momentum h. For all values of c, d the subset $\{x \in U \mid H(x) = c,\ h(x) = d\}$ is a manifold that is invariant under the flow.

Manifolds also arise in mechanical problems with constraints. A pendulum in three dimensions has a configuration space consisting of the 2-sphere S, and its state space is the manifold of tangent vectors to S. The configuration space of a

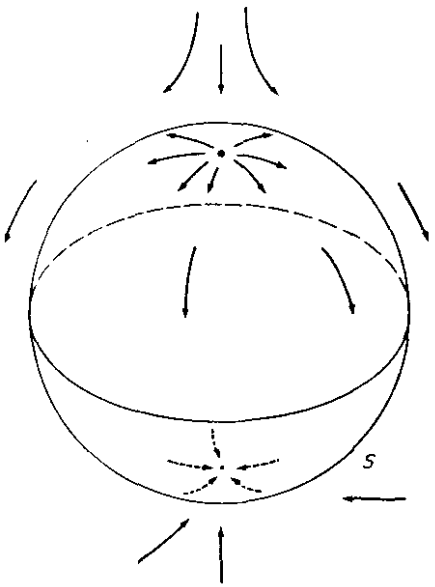


FIG. A. A vector field tangent to S.

rigid body with one point fixed is a compact three-dimensional manifold, the set of rotations of Euclidean three space.

The topology (global structure) of a manifold plays an important role in the analysis of dynamical systems on the manifold. For example, a dynamical system on the two sphere S must have an equilibrium; this can be proved using the Poincaré–Bendixson theorem.

The mathematical treatment of electrical circuit theory can be extended if manifolds are used. The very restrictive special hypothesis in Chapter 10 was made in order to avoid manifolds. That hypothesis is that the physical states of a circuit (obeying Kirchhoff's and generalized Ohm's laws) can be parametrized by the inductor currents and capacitor voltages. This converts the flow on the space of physical states into a flow on a vector space. Unfortunately this assumption excludes many circuits. The more general theory simply deals with the flow directly on the space of physical states, which is a manifold under "generic" hypotheses on the circuit.

Manifolds enter into differential equations in another way. The set of points whose trajectories tend to a given hyperbolic equilibrium form a submanifold called the stable manifold of the equilibrium. These submanifolds are a key to any deep global understanding of dynamical systems.

Our analysis of the long-term behavior of trajectories has been limited to the simplest kinds of limit sets, equilibria and closed orbits. For some types of systems these are essentially all that can occur, for example gradient flows and planar systems. But to achieve any kind of general picture in dimensions higher than two, one

must confront limit sets which can be extremely complicated, even for structurally stable systems. It can happen that a compact region contains infinitely many periodic solutions with periods approaching infinity. Poincaré was dismayed by his discovery that this could happen even in the Newtonian three-body problem, and expressed despair of comprehending such a phenomenon.

In spite of the prevalence of such systems it is not easy to prove their existence, and we cannot go into details here. But to give some idea of how they arise in apparently simple situations, we indicate in Fig. B a discrete dynamical system in the plane. Here the rectangle $ABCD$ is sent to its image $A'B'C'D'$ in the most obvious way by a diffeomorphism f of $\mathbf{R}^2$; thus $f(A) = A'$, and so on. It can be shown that f will have infinitely many periodic points, and that this property is preserved by perturbations. (A point p is periodic if $f^n(p) = p$ for some $n > 0$.) Considering $\mathbf{R}^2$ as embedded in $\mathbf{R}^3$, one can construct a flow in $\mathbf{R}^3$ transverse to $\mathbf{R}^2$ whose time one map leaves $\mathbf{R}^2$ invariant and is just the diffeomorphism f in $\mathbf{R}^2$. Such a flow has closed orbits through the periodic points of f.

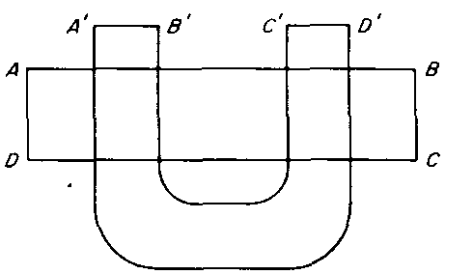


FIG. B

In spite of Poincaré's discouragement there has been much progress in recent years in understanding the global behavior of fairly general types of dynamical systems, including those exhibiting arbitrarily long closed orbits. On the other hand, we are far from a clear picture of the subject and many interesting problems are unsolved.

The following books are recommended to the reader who wishes to see how the subject of dynamical systems has developed in recent years. They represent a good cross section of current research: *Proceedings of Symposia in Pure Mathematics Volume XIV, Global Analysis* [3] and *Dynamical Systems* [19]. See also Nitecki's *Differentiable Dynamics* [18].

Appendix I

Elementary Facts

This appendix collects various elementary facts that most readers will have seen before.

1. Set Theoretic Conventions

We use extensively maps, or functions, from one set X to another Y, which we write

$$f: X \to Y \qquad \text{or} \qquad X \xrightarrow{f} Y.$$

Thus the map f assigns to each element $x \in X$ (that is, x belongs to X) an element $f(x) = y$ of Y. In this case we often write $x \to y$ or $x \to f(x)$. The *identity* map $i: X \to X$ is defined by $i(x) = x$ and if Q is a subset of X, $Q \subset X$, the inclusion map $\alpha: Q \to X$ is defined by $\alpha(q) = q$. If $f: X \to Y$, and $g: Y \to Z$ are two maps, the composition $g \circ f$ (or sometimes written gf) is defined by $g \circ f(x) = g(f(x))$. The map $f: X \to Y$ is said to be one-to-one if whenever $x, x' \in X$, $x \neq x'$, then $f(x) \neq f(x')$. The *image* of f is the set described as

$$\operatorname{Im} f = \{y \in Y \mid y = f(x), \text{ some } x \in X\}.$$

Then f is *onto* if $\operatorname{Im} f = Y$. An inverse g (or f^{-1}) of f is a map $g: Y \to X$ such that $g \circ f$ is the identity map on X and $f \circ g$ is the identity on Y. If the image of f is Y and f is one to one, then f has an inverse and conversely.

If $f: X \to Y$ is a map and $Q \subset X$, then $f \mid Q: Q \to Y$ denotes the *restriction* of f to Q so $f \mid Q(q) = f(q)$.

We frequently use the *summation* sign:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n,$$

where the x_i are elements of a vector space. If there is not much ambiguity, the limits are omitted:

$$\sum x_i = x_1 + \cdots + x_n.$$

2. Complex Numbers

We recall the elements of complex numbers $\mathbf{C}$. We are not interested in complex analysis in itself; but sometimes the use of complex numbers simplifies the study of real differential equations.

The set of complex numbers $\mathbf{C}$ is the Cartesian plane $\mathbf{R}^2$ considered as a vector space, together with a product operation.

Let i be the complex number $i = (0, 1)$ in coordinates on $\mathbf{R}^2$. Then every complex number z can be written uniquely in the form $z = x + iy$ where x, y are real numbers. Complex numbers are added as elements of $\mathbf{R}^2$, so if $z = x + iy$, $z' = x' + iy'$, then $z + z' = (x + x') + i(y + y')$: the rules of addition carry over from $\mathbf{R}^2$ to $\mathbf{C}$.

Multiplication of complex numbers is defined as follows: if $z = x + iy$ and $z' = x' + iy'$, then $zz' = (xx' - yy') + i(xy' + x'y)$. Note that $i^2 = -1$ (or "$i = \sqrt{-1}$") with this definition of product and this fact is an aid to remembering the product definition. The reader may check the following properties of multiplication:

(a) $zz' = z'z$.
(b) $(zz')z'' = z(z'z'')$.
(c) $1z = z$ (here $1 = 1 + i \cdot 0$).
(d) If $z = x + iy$ is not 0, then

$$z^{-1}z = zz^{-1} = 1, \qquad \text{where} \quad z^{-1} = \frac{x - iy}{x^2 + y^2}.$$

(e) If z is *real* (that is, $z = x + i \cdot 0$), then multiplication by z coincides with scalar multiplication in $\mathbf{R}^2$. If z and z' are both real, complex multiplication specializes to ordinary multiplication.
(f) $(z + z')w = zw + z'w$, $z, z', w \in \mathbf{C}$.

The *complex conjugate* of a complex number $z = x + iy$ is the complex number $\bar{z} = x - iy$. Thus conjugation is a map $\sigma: \mathbf{C} \to \mathbf{C}$, $\sigma(z) = \bar{z}$, which has as its set of fixed points the real numbers; that is to say $\bar{z} = z$ if and only if z is real. Simple

properties of conjugation are:

$$\bar{\bar{z}} = z,$$

$$\overline{(z + z')} = \bar{z} + \bar{z}',$$

$$\overline{zz'} = \bar{z}\bar{z}'.$$

The absolute value of a complex number $z = x + iy$ is

$$| z | = (z\bar{z})^{1/2} = (x^2 + y^2)^{1/2}.$$

Then

$$| z | = 0 \quad \text{if and only if} \quad z = 0,$$

$$| z + z' | \leq | z | + | z' |,$$

$$| zz' | = | z | \, | z' |,$$

and $| z |$ is the ordinary absolute value if z is real.

Suppose a complex number z has absolute value 1. Then on $\mathbf{R}^2$ it is on the unit circle (described by $x^2 + y^2 = 1$) and there is a $\theta \in \mathbf{R}$ such that $z = \cos\theta + i \sin\theta$. We define the symbol $e^{i\theta}$ by

$$e^{i\theta} = \cos\theta + i \sin\theta,$$

$$e^{a+ib} = e^a e^{ib}.$$

This use of the exponential symbol can be justified by showing that it is consistent with a convergent power series representation of e^t. Here one takes the power series of e^{a+ib} as one does for ordinary real exponentials; thus

$$e^{a+ib} = \sum_{n=0}^{\infty} \frac{(a + ib)^n}{n!}.$$

One can operate with complex exponentials by the same rules as for real exponentials.

3. Determinants

One may find a good account of determinants in Lang's *Second Course in Calculus*. [12]. Here we just write down a couple of facts that are useful.

First we give a general expression for a determinant. Let $A = [a_{ij}]$ be the $(n \times n)$ matrix whose entry in the ith row and jth column is a_{ij}. Denote by A_{ij} the $(n - 1) \times (n - 1)$ matrix obtained by deleting the ith row and jth column. Then if i is a fixed integer, $1 \leq i \leq n$, the determinant satisfies

$$\text{Det } A = (-1)^{i+1} a_{i1} \text{ Det } A_{i1} + \cdots + (-1)^{i+n} a_{in} \text{ Det } A_{in}.$$

Thus the expression on the right does not depend on i and furthermore gives a way of finding (or defining) Det A inductively. The determinant of a 2×2 matrix $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ is $ad - bc$. For a 3×3 matrix

$$A = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$

one obtains

$$\text{Det } (A) = a_{11} \text{ Det} \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix} - a_{12} \text{ Det} \begin{bmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{bmatrix} + a_{13} \text{ Det} \begin{bmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}.$$

Recall that if Det $A \neq 0$, then A has an inverse. One way of finding this inverse is to solve explicitly the system of equations $Ax = y$ for x obtaining $x = By$; then B is an inverse A^{-1} for A.

If Det $A \neq 0$, one has the formula

$$A^{-1} = \text{transpose of} \left[\frac{(-1)^{i+j} \text{ Det } A_{ij}}{\text{Det } A}\right].$$

It follows easily from the recursive definition that the determinant of a triangular matrix is the product of the diagonal entries.

4. Two Propositions on Linear Algebra

The purpose of this section is to prove Propositions 1 and 3 of Section 1B, Chapter 3.

Proposition 1 *Every vector space F has a basis, and every basis of F has the same number of elements. If $\{e_1, \ldots, e_k\} \subset F$ is an independent subset that is not a basis, by adjoining to it suitable vectors $e_{k+1}, \ldots, e_m$, one can form a basis $e_1, \ldots, e_m$.*

The proof goes by some easy lemmas.

Lemma 1 *A system of n linear homogeneous equations in $n + 1$ unknowns always has a nontrivial solution.*

The proof of Lemma 1 is done by the process of elimination of one unknown to obtain a system of $n - 1$ equations in n unknowns. Then one is finished by induction (the first case, $n = 2$, being obvious). The elimination is done by using the first equation to solve for one variable as a linear combination of the rest. The

expression obtained is substituted in the remaining equations to make the reduction.

Lemma 2 *Let $\{e_1, \ldots, e_n\}$ be a basis for a vector space F. If $v_1, \ldots, v_m$ are linearly independent elements of F, then $m \leq n$.*

Proof. It is sufficient to show that $m \neq n + 1$. Suppose otherwise. Then each v_i is a linear combination of the e_i,

$$v_i = \sum_{k=1}^{n} a_{ik}e_k, \qquad i = 1, \ldots, n + 1.$$

By Lemma 1, the system of equations

$$\sum_{i=1}^{n+1} x_i a_{ik} = 0, \qquad k = 1, \ldots, n,$$

has a nontrivial solution $x = (x_1, \ldots, x_{n+1})$. Then

$$\sum x_i v_i = \sum_i x_i \sum_k a_{ik}e_k = \sum_k \sum_i x_i a_{ik} e_k = 0,$$

so that the v_i are linearly dependent. This contradiction proves Lemma 2.

From Lemma 2 we obtain the part of Proposition 1 which says that two bases have the same number of elements. If $\{e_1, \ldots, e_n\}$ and $\{v_1, \ldots, v_m\}$ are the two bases, then the lemma says $m \leq n$. An interchange yields $n \leq m$.

Say that a set $S = \{v_1, \ldots, v_m\}$ of linearly independent elements of F is *maximal* if for every v in F, $v \notin S$, the set $\{v, v_1, \ldots, v_m\}$ is dependent.

Lemma 3 *A maximal set of linearly independent elements $B = \{v_1, \ldots, v_m\}$ in a vector space F is a basis.*

Proof. We have to show that any $v \in F$, $v \notin B$, is a linear combination of the v_i. But by hypothesis $v, v_1, \ldots, v_m$ are dependent so that one can find numbers x, x_i not all zero such that $\sum x_i v_i + xv = 0$. Then $x \neq 0$ since the v_i are independent. Thus $v = \sum (-x_i/x)v_i$. This proves Lemma 3.

Proposition 1 now goes easily. Recall F is a linear subspace of $\mathbf{R}^n$ (our definition of vector space!). If $F \neq 0$, let v_1 be any nonzero element. If $\{v_1\}$ is not a basis, one can find $v_2 \in F$, v_2 not in the space spanned by $\{v_1\}$. Then v_1, v_2 are independent and if $\{v_1, v_2\}$ is maximal, we are finished by Lemma 3. Otherwise we continue the process. The process must stop with a maximal set of linearly independent elements $\{v_1, \ldots, v_m\}$, $m \leq n$ by Lemma 2. This gives us a basis for F. The rest of the proof of the proposition proceeds in exactly the same manner.

Proposition 3 *Let $T: E \to F$ be a linear map. Then*

$$\dim(\text{Im } T) + \dim(\text{Ker } T) = \dim E.$$

In particular, suppose $\dim E = \dim F$. *Then the following are equivalent statements:*

(a) $\text{Ker } T = 0$;
(b) $\text{Im } T = F$;
(c) *T is an isomorphism.*

Proof. The second part follows from the first part (and things said in Section 1 of Chapter 3).

To prove the first part of the proposition let $f_1, \ldots, f_k$ be a basis for Im T. Choose $e_1, \ldots, e_k$ such that $Te_i = f_i$. Let $g_1, \ldots, g_l$ be a basis for Ker T. It is sufficient to show that

$$\{e_1, \ldots, e_k, g_1, \ldots, g_l\}$$

is a basis for E since $k = \dim \text{Im } T$ and $l = \dim \text{Ker } T$.

First, these elements are independent: for if $\sum \lambda_i e_i + \sum M_i g_i = 0$, application of T yields $\sum \lambda_i Te_i = \sum \lambda_i f_i = 0$. Then the $\lambda_i = 0$ since the f_i are independent. Thus $\sum M_j g_j = 0$ and the $M_j = 0$ since the g_j are independent.

Second, E is spanned by the e_i and g_j, that is, every element of E can be written as a linear combination of the e_i and the g_j. Let e be any element of E. Define $v = \sum \lambda_i e_i$, where $Te = \sum \lambda_i f_i$ defines the λ_i. Then $e = (e - v) + v$. Now $T(e - v) = 0$ so $e - v \in \text{Ker } T$ and thus $(e - v)$ can be written as a linear combination of the g_j.

Appendix II

Polynomials

1. The Fundamental Theorem of Algebra

Let

$$p(z) = a_n z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0, \qquad a_n \neq 0,$$

be a polynomial of degree $n \geq 1$ with complex coefficients $a_0, \ldots, a_n$. Then $p(z) = 0$ for at least one $z \in \mathbf{C}$.

The proof is based on the following basic property of polynomials.

Proposition 1 $\lim_{|z|\to\infty} |\, p(z)\,| = \infty$.

Proof. For $z \neq 0$ we can write

$$\frac{p(z)}{z^n} = a_n + \sum_{k=1}^{n} \frac{a_{n-k}}{z^k}.$$

Hence

$$\frac{|\, p(z)\,|}{|\, z\,|^n} \geq |\, a_n\,| - \sum_{k=1}^{n} \frac{|\, a_{n-k}\,|}{|\, z\,|^k}. \tag{1}$$

Therefore there exists $L > 0$ such that if $|\, z\,| \geq L$, then the right-hand side of (1) is $\geq \frac{1}{2}|\, a_n\,| > 0$, and hence

$$\frac{|\, p(z)\,|}{|\, z\,|^n} \geq \tfrac{1}{2}|\, a_n\,|,$$

from which the proposition follows.

Proposition 2 $|p(z)|$ *attains a minimum value.*

Proof. For each $k > 0$ define the compact set

$$D_k = \{z \in \mathbf{C} \mid |\, z\,| \leq k\}.$$

The continuous function $|\, p(z)\,|$ attains a minimum value

$$v_k = |\, p(z_k)\,|, \qquad z_k \in D_k,$$

on D_k. (z_k may not be unique.) By Proposition 1 there exists $k > 0$ such that

$$|\, p(z)\,| \geq v_1 \quad \text{if} \quad |\, z\,| \geq k. \tag{2}$$

We may take $k \geq 1$. Then v_k is the minimum value of $|\, p(z)\,|$, for if $z \in D_k$, then $|\, p(z)\,| \leq v_k$, while if $z \notin D_k$, $|\, p(z)\,| \geq v_1$ by (2); and $v_1 \geq v_k$ since $D_1 \subset D_k$.

Proof of theorem. Let $|\, p(z_0)\,|$ be minimal. The function

$$q(z) \equiv p(z + z_0)$$

is a polynomial taking the same values as p, hence it suffices to prove that q has a root. Clearly, $|\, q(0)\,|$ is minimal. Hence we may assume that

(3) *the minimum value of* $|\, p(z)\,|$ *is* $|\, p(0)\,| = |\, a_0\,|$.

We write

$$p(z) = a_0 + a_k z^k + z^{k+1} r(z), \qquad a_k \neq 0, \quad k \geq 1,$$

where r is a polynomial of degree $n - k - 1$ if $k < n$ and $r = 0$ otherwise.

We choose w so that

$$a_0 + a_k w^k = 0. \tag{4}$$

In other words, w is a kth root of $-a_0/a_k$. Such a root exists, for if

$$\frac{-a_0}{a_k} = \rho(\cos\theta + i \sin\theta),$$

then we can take

$$w = \rho^{1/k}\left(\cos\left(\frac{\theta}{k}\right) + i \sin\left(\frac{\theta}{k}\right)\right).$$

We now write, for $0 < t < 1$,

$$\begin{aligned} p(tw) &= (1 - t^k)a_0 + t^k(a_0 + a_k w^k) + (tw)^{k+1} r(tw) \\ &= (1 - t^k)a_0 + (tw)^{k+1} r(tw). \end{aligned}$$

Hence

$$\begin{aligned} |\, p(tw)\,| &\leq |\, a_0\,| - t^k |\, a_0\,| + t^{k+1} |\, w^{k+1} r(tw)\,| \\ &= |\, a_0\,| - t^k(|\, a_0\,| - t\,|\, w^{k+1} r(tw)\,|). \end{aligned}$$

But if $|a_0| > 0$, for t sufficiently small, we have

$$|a_0| - t|w^{k+1}r(tw)| > 0,$$

such a value of t makes

$$|p(tw)| < |a_0|.$$

This contradicts minimality of $|p(0)| = |a_0|$. Hence $|p(0)| = 0$.

Corollary *A polynomial p of degree n can be factored:*

$$p(z) = (z - \lambda_1)\cdots(z - \lambda_n),$$

where $p(\lambda_k) = 0$, $k = 1, \ldots, n$, and $p(z) \neq 0$ for $z \neq \lambda_k$.

Proof. For any $\lambda \in \mathbf{C}$ we have

$$p(z) = p((z - \lambda) + \lambda)$$

$$= \sum_{k=0}^{n} a_k((z - \lambda) + \lambda)^k.$$

Expanding by the binomial theorem, we have

$$p(z) = \sum_{k=0}^{n}\sum_{j=0}^{k}\binom{k}{j}a_k(z - \lambda)^j\lambda^{k-j}.$$

Every term on the right with $j > 0$ has a factor of $z - \lambda$; hence

$$p(z) = (z - \lambda)q(z) + \sum_{k=0}^{n} a_k\lambda^k$$

or

$$p(z) = (z - \lambda)q(z) + p(\lambda)$$

for some polynomial $q(z)$ of degree $n - 1$ (which depends on λ). In particular, if $p(\lambda_1) = 0$, which must be true for some λ_1, we have

$$p(z) = (z - \lambda_1)q_1(z).$$

Since q_1 has a root λ_2, we write

$$p(z) = (z - \lambda_1)(z - \lambda_2)q_2(z)$$

and so on.

The complex numbers $\lambda_1, \ldots, \lambda_n$ are the *roots* of p. If they are distinct, p has *simple roots*. If λ appears k times among $\{\lambda_1, \ldots, \lambda_n\}$, λ is a root of *multiplicity* k, or a *k-fold root*. This is equivalent to $(z - \lambda)^k$ being a factor of $p(z)$.

Appendix III

On Canonical Forms

The goal of this appendix is to prove three results of Chapter 6: Theorem 1 and the uniqueness of the $S + N$ decomposition, of Section 1; and Theorem 1 of Section 3.

1. A Decomposition Theorem

Theorem 1 (Section 1, Chapter 6) *Let T be an operator on V where V is a complex vector space, or V is real and T has real eigenvalues. Then V is the direct sum of the generalized eigenspaces of T. The dimension of each generalized eigenspace equals the multiplicity of the corresponding eigenvalue.*

For the proof we consider thus an operator $T: V \to V$, where we suppose that V is a complex vector space.

Define subspaces for each nonnegative integer j as follows:

$$K_j(T) = K_j = \operatorname{Ker} T^j; \qquad N = \bigcup_j K_j;$$

$$L_j(T) = Lj = \operatorname{Im} T^j; \qquad M = \bigcap_j L_j.$$

Then

$$0 = K_0 \subset K_1 \subset \cdots \subset K_j \subset K_{j+1} \subset \cdots \subset N;$$

$$V = L_0 \supset L_1 \supset \cdots \supset L_j \subset L_{j+1} \supset \cdots \supset M.$$

Choose n and m so that

$$K_j = K_n \quad \text{if } j \geq n,$$

$$L_j = L_m \quad \text{if } j \geq m,$$

which is possible since V is finite dimensional. Put

$$N(T) = N = K_n, \qquad M(T) = M = L_m.$$

Clearly, N and M are invariant.

Lemma $V = N \oplus M$.

Proof. Since $TM = L_{m+1} = M$, $T \mid M$ is invertible; also, $T^n(M) = M$ and $T^n x \neq 0$ for nonzero x in M. Since $T^n(N) = 0$, we have $N \cap M = 0$. If $x \in V$ is any vector, let $T^m x = y \in M$. Since $T^m \mid M$ is invertible, $T^m x = T^m z$, $z \in M$. Put $x = (x - z) + z$. Since $x - z \in N$, $z \in M$, this proves the lemma.

Let $\alpha_1, \ldots, \alpha_q$ be the distinct eigenvalues of T. For each eigenvalue α_k define subspaces

$$N_k = N(T - \alpha_k I) = \bigcup_{i \geq 0} \operatorname{Ker}(T - \alpha_k I)^i,$$

$$M_k = M(T - \alpha_k I) = \bigcap_{i \geq 0} \operatorname{Im}(T - \alpha_k I)^i.$$

Clearly, these subspaces are invariant under T.

By the lemma,

$$V = N_1 \oplus M_1.$$

Proposition $V = N_1 \oplus \cdots \oplus N_q$.

Proof. We use induction on the dimension d of V, the cases $d = 0$ or 1 being trivial. Suppose $d > 1$ and assume the theorem for any space of smaller dimension. In particular, *the theorem is assumed to hold for* $T \mid M_1 \colon M_1 \to M_1$.

It therefore suffices to prove that the eigenvalues of $T \mid M_1$ are $\alpha_2, \ldots, \alpha_q$, and that

$$N(T - \alpha_k I \mid M_1) = N(T - \alpha_k I), \qquad \text{all} \quad k > 1. \tag{1}$$

We first prove that

$$\operatorname{Ker}((T - \alpha_1 I) \mid N_k) = 0, \qquad \text{all} \quad k > 1. \tag{2}$$

Suppose $(T - \alpha_1 I)x = 0$ and $x \neq 0$. Then $Tx = \alpha_1 x$; hence

$$(T - \alpha_k I)x = (\alpha_1 - \alpha_k)x.$$

But then

$$(T - \alpha_k I)^j x = (\alpha_1 - \alpha_k)^j x \neq 0$$

for all $j \geq 0$, so $x \notin N_k$.

Since N_k is invariant under $T - \alpha_1 I$ we have

$$(T - \alpha_1 I)N_k = N_k,$$

by (2). Therefore $N_k \subset \operatorname{Im}(T - \alpha_1 I)^j$, all $j \geq 0$, $k > 1$. This shows that

$$N_k \subset M_1, \qquad \text{all} \quad k > 1.$$

This implies that $\alpha_2, \ldots, \alpha_q$ are eigenvalues of $T \mid M_1$. It is now clear that the eigenvalues of $T \mid M_1$ are precisely $\alpha_2, \ldots, \alpha_q$ since α_1 is not, and any eigenvalue of $T \mid M_1$ is also an eigenvalue of T. The proposition is proved.

We can now prove Theorem 1. Let n_k be the multiplicity of α_k as a root of the characteristic polynomial of T. Then $T \mid N_k \colon N_k \to N_k$ has the unique eigenvalue α_k (the proof is like that of (2) above), and in fact the lemma implies that α_k has multiplicity n_k as an eigenvalue of $T \mid N_k$. Thus the degree of the characteristic polynomial of $T \mid N_k$ is $n_k = \dim N_k$.

The generalized eigenspace of $T \colon V \to V$ belonging to α_k is defined by $E_k = E(T, \alpha_k) = \operatorname{Ker}(T - \alpha_k)^{n_k}$. Then, clearly, $E_k \subset N_k$.

In fact, it follows that $E_k = N_k$ from the definition of N_k and Lemma 2 of the next section (applied to $T - \alpha_k$). This finishes the proof of the theorem if V is complex. But everything said above is valid for an operator on a real vector space provided its eigenvalues are real. The theorem is proved.

2. Uniqueness of S and N

Theorm *Let T be a linear operator on a vector space E which is complex if T has any nonreal eigenvalues. Then there is only one way of expressing T as $S + N$, where S is diagonalizable, N is nilpotent, and $SN = NS$.*

Proof. Let $E_k = E(\lambda_k, T)$, $k = 1, \ldots, r$, be the generalized eigenspaces of T. Then $E = E_1 \oplus \cdots \oplus E_r$ and $T = T_1 \oplus \cdots \oplus T_r$, where $T_k = T \mid E_k$. Note that E_k is invariant under every operator that commutes with T.

Since S and N both commute with S and N, they both commute with T. Hence E_k is invariant under S and N.

Put $S_k = \lambda_k I \in L(E_k)$, and $N_k = T_k - S_k$. It suffices to show that $S \mid E_k = S_k$, for then $N \mid E_k = E_k$, proving the uniqueness of S and N.

Since S is diagonalizable, so is $S \mid E_k$ (Problem 17 of Chapter 6, Section 2). Therefore $S \mid E_k - \lambda_k I$ is diagonalizable; in other words $S \mid E_k - S_k$ is diagonalizable. This operator is the same as $N_k - N \mid E_k$. Since $N \mid E_k$ commutes with $\lambda_k I$ and with T_k, it also commutes with N_k. It follows that $N_k - N \mid E_k$ is nilpotent (use the binomial theorem). Thus $S \mid E_k - S_k$ is represented by a nilpotent diagonal matrix. The only such matrix is O; thus $S \mid E_k = S_k$ and the theorem is proved.

3. Canonical Forms for Nilpotent Operators

The goal is to prove the following theorem.

Theorem 1 (Section 3, Chapter 6) *Let N be a nilpotent operator on a real or complex vector space V. Then V has a basis giving N a matrix of the form*

$$A = \operatorname{diag}\{A_1, \ldots, A_r\},$$

where A_j is an elementary nilpotent block, and the size of A_k is a nonincreasing function of k. The matrices $A_1, \ldots, A_r$ are uniquely determined by the operator N.

In this section, V is a real or complex vector space.

A subspace $W \subset V$ is a *cyclic* subspace of an operator T on V if $T(W) \subset W$ and there is a vector $x \in W$ such that W is spanned by the vector $T^n x$, $n = 0, 1, \ldots$. We call such an x a *cyclic vector* for W.

Any vector x generates a cyclic subspace, for the iterates of x under T, that is, x, Tx, T^2x, . . . generate a subspace which is evidently cyclic. We denote this subspace by $Z(x)$ or $Z(x, T)$.

Suppose $N: V \to V$ is a nilpotent operator. For each $x \in V$ there is a smallest positive integer n, denoted by $\operatorname{nil}(x)$ or $\operatorname{nil}(x, N)$, such that $N^n x = 0$. If $x \neq 0$, then $N^k x \neq 0$ for $0 \leq k < \operatorname{nil}(x)$.

Lemma 1 *Let $\operatorname{nil}(x, N) = n$. Then the vectors $N^k x$, $0 \leq k \leq n-1$, form a basis for $Z(x, N)$.*

Proof. They clearly span $Z(x)$. If they are dependent, there is a relation $\sum^{n-1} a_k N^k(x) = 0$ with not all $a_k = 0$. Let j be the smallest index, $0 \leq j \leq n-1$ such that $a_j \neq 0$. Then

$$\begin{aligned} 0 &= N^{n-j-1}\left(\sum_{k=j}^{n-1} a_k N^k x\right) \\ &= \sum_{k=j}^{n-1} a_k N^{n+k-j-1} x \\ &= a_j N^{n-1} x + \sum_{k=j+1}^{n-1} a_k N^{n+k-j-1} x \\ &= a_j N^{n-1} x \end{aligned}$$

since $n + k - j - 1 \geq n$ if $k \geq j + 1$. Thus $a_j N^{n-1} x = 0$, so $N^{n-1} x = 0$ because $a_j \neq 0$. But this contradicts $n = \operatorname{nil}(x, N)$.

This result proves that *in the basis* $\{x, Nx, \ldots, N^{n-1}x\}$, $n = \operatorname{nil}(x)$, the nilpotent operator $N \mid Z(x)$ has the matrix

$$\begin{bmatrix} 0 & & & \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}$$

with ones below the diagonal, zeros elsewhere. This is where the ones below the diagonal in the canonical form come from.

An argument similar to the proof of Lemma 1 shows:

if $\sum_{k=0}^{q} a_k N^k x = 0$, then $a_k = 0$ for $k < \operatorname{nil}(x, N)$.

It is convenient to introduce the notation $p(T)$ to denote the operator $\sum_{k=0}^{q} a_k T^k$ if p is the polynomial

$$p(t) = a_q t^q + \cdots + a_1 t + a_0,$$

where t is an indeterminate (that is, an "unknown"). Then the statement proved above can be rephrased:

Lemma 2 *Let $n = \operatorname{nil}(x, N)$. If $p(t)$ is a polynomial such that $p(N)x = 0$, then t^n divides $p(t)$, that is, there is a polynomial $p_1(t)$ such that $p(t) = t^n p_1(t)$.*

We now prove the existence of a canonical form for a nilpotent operator N. In view of the matrix discussed above for $N \mid Z(x)$, this amounts to proving:

Proposition *Let $N: V \to V$ be a nilpotent operator. Then V is a direct sum of cyclic subspaces.*

The proof goes by induction on $\dim V$, the case $\dim V = 0$ being trivial. If $\dim V > 0$, then $\dim N(V) < \dim V$, since N has a nontrivial kernel. Therefore there are nonzero vectors $y_1, \ldots, y_r$ in $N(V)$ such that

$$N(V) = Z(y_1) \oplus \cdots \oplus Z(y_r).$$

Let $x_j \in V$ be a nonzero vector with

$$Nx_j = y_j, \qquad j = 1, \ldots, r.$$

We prove *the subspaces $Z(x_1), \ldots, Z(x_r)$ are independent.*

Observe that $\operatorname{nil}(x_j) \geq 2$ since

$$Nx_j = y_j \neq 0.$$

If the subspaces $Z(x_j)$ are not independent, there are vectors $u_j \in Z(x_j)$, not all zero, such that $\sum_{j=1}^{r} u_j = 0$. Therefore $\sum_j Nu_j = 0$. Since $Nu_j \in N(Z(x_j)) =$

$Z(y_j)$ and the $Z(y_j)$ are independent by assumption, it follows that $u_j \in \operatorname{Ker} N$, $j = 1, \ldots, r$. Now each u_j has the form

$$\sum_{k=0}^{n_j-1} a_{jk} N^k x_j, \qquad n_j = \operatorname{nil}(x_j).$$

Hence $u_j = p_j(N)x_j$ for the polynomial $p_j(t) = \sum_{k=0}^{n_j-1} a_{jk}t^k$. Therefore $Nu_j = p_j(N)y_j = 0$. By Lemma 2, $p_j(t)$ is divisible by t^m if $m \leq \operatorname{nil}(y_j)$. Since $1 \leq \operatorname{nil}(y_j)$, we can write

$$p_j(t) = s_j(t)t$$

for some polynomial $s_j(t)$.

But now, substituting N for t, we have

$$\begin{aligned} u_j &= s_j(N)Nx_j \\ &= s_j(N)y_j \in Z(y_j). \end{aligned}$$

Therefore $u_j = 0$ since the $Z(y_j)$ are independent.

We now show that

$$V = Z(x_1) \oplus \cdots \oplus Z(x_r) \oplus L \tag{1}$$

with $L \subset \operatorname{Ker} N$. Let $\operatorname{Ker} N = K$ and let L be a subspace of K such that

$$K = (K \cap N(V)) \oplus L.$$

Then L is independent from the $Z(x_j)$. To see this, let $v \in (\oplus Z(x_j)) \cap L$. Then $v \in (\oplus Z(x_j)) \cap K$, and by an argument similar to the one above, this implies $v \in N(V)$. But $N(V) \cap L = 0$, hence $v = 0$.

It is clear that every cyclic subspace in K, and hence in L, is one dimensional. Therefore $L = Z(w_1) \oplus \cdots \oplus Z(w_s)$, where $\{w_1, \ldots, w_s\}$ is a basis for L. Finally,

$$V = Z(x_1) \oplus \cdots \oplus Z(x_r) \oplus Z(w_1) \oplus \cdots \oplus Z(w_s).$$

This proposition implies the theorem, except for the question of uniqueness of the matrices $A_1, \ldots, A_r$. This uniqueness is equivalent to the assertion that the operator N determines the sizes of the blocks A_i (or the dimensions of the cyclic subspaces). This is done by induction on $\dim V$.

Consider the restriction of N to its image $N(V) = F$:

$$N \mid F: F \to F.$$

It is easy to see that if V is the direct sum of cyclic subspaces $Z_1 \oplus \cdots \oplus Z_r \oplus W_1$, where $W_1 \subset \operatorname{Ker} N$, and Z_k is generated by x_k, $\dim Z_k > 1$, then $N(V)$ is the direct sum

$$N(Z_1) \oplus \cdots \oplus N(Z_r),$$

where $N(Z_k)$ is cyclic, generated by $N(x_k)$, and $\dim N(Z_k) = \dim Z_k - 1$. Since $\dim N(F) < \dim V$, the numbers $\{\dim Z_k - 1\}$ are determined by $N \mid F$, hence by N. It follows that $\{\dim Z_k\}$ are also determined by N.

This finishes the proof of the theorem.

Appendix IV

The Inverse Function Theorem

In this appendix we prove the inverse function theorem and the implicit function theorem.

Inverse function theorem *Let W be an open set in a vector space E and let $f: W \to E$ be a C^1 map. Suppose $x_0 \in W$ is such that $Df(x_0)$ is an invertible linear operator on E. Then x_0 has an open neighborhood $V \subset W$ such that $f \mid V$ is a diffeomorphism onto an open set.*

Proof. By continuity of $Df: W \to L(E)$ there is an open ball $V \subset W$ about x_0 and a number $\nu > 0$ such that if $y, z \in V$, then $Df(y)$ is invertible,

$$\| Df(y)^{-1} \| < \nu,$$

and

$$\| Df(y) - Df(z) \| < \nu^{-1}.$$

It follows from Lemma 1 of Chapter 16, Section 1, that $f \mid V$ is one-to-one. Moreover, Lemma 2 of that section implies that $f(V)$ is an open set.

The map $f^{-1}: f(V) \to V$ is continuous. This follows from local compactness of $f(V)$. Alternatively, in the proof of Lemma 1 it is shown that if y and z are in V, then

$$| y - z | \leq \nu | f(y) - f(z) |;$$

hence, putting $f(y) = a$ and $f(z) = b$, we have

$$| f^{-1}(a) - f^{-1}(b) | \leq \nu | a - b |,$$

which proves f^{-1} continuous.

It remains to prove that f^{-1} is C^1. The derivative of f^{-1} at $a = f(x) \in f(V)$ is $Df(x)^{-1}$. To see this, we write, for $b = f(y) \in f(V)$:

$$f^{-1}(b) - f^{-1}(a) - Df(x)^{-1}(b - a) = y - x - Df(x)^{-1}(f(y) - f(x)).$$

Now

$$f(y) - f(x) = Df(x)(y - x) + R(y, x),$$

where

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

Hence

$$|\, y - x - Df(x)^{-1}(f(y) - f(x)) \,| = |\, y - x - Df(x)^{-1}(Df(y - x) + R(y, x)) \,|$$

$$= |\, Df(x)^{-1}(R(y, x) \,|.$$

Hence

$$\frac{|\, y - x - Df(x)^{-1}(f(y) - f(x)) \,|}{|\, f(y) - f(x) \,|} = \frac{|\, Df(x)^{-1}(R(y, x) \,|}{|\, f(y) - f(x) \,|}$$

$$\leq \frac{\nu\, |\, R(y, x) \,|}{|\, y - x \,|} \Big/ \frac{|\, f(y) - f(x) \,|}{|\, y - x \,|}.$$

This clearly goes to 0 as $|\, f(y) - f(x) \,|$ goes to 0. Therefore $D(f^{-1})(a) = [Df(f^{-1}a)]^{-1}$. Thus the map $D(f^{-1}) : f(V) \to L(E)$ is the composition: f^{-1}, followed by Df, followed by the inversion of invertible operators. Since each of these maps is continuous, so is $D(f^{-1})$.

Remark. *Induction on $r = 1, 2, \ldots$ shows also that if f is C^r, then f^{-1} is C^r.*

Implicit function theorem *Let $W \subset E_1 \times E_2$ be an open set in the Cartesian product of two vector spaces. Let $F: W \to E_2$ be a C^1 map. Suppose $(x_0, y_0) \in W$ is such that the linear operator*

$$\frac{\partial F}{\partial y}(x_0, y_0) : E_2 \to E_2$$

is invertible. Put $F(x_0, y_0) = c$. Then there are open sets $U \subset E_1$, $V \subset E_2$ with

$$(x_0, y_0) \in U \times V \subset W$$

and a unique C^1 map

$$g: U \to V$$

such that

$$F(x, g(x)) = c$$

for all $x \in U$, and moreover, $F(x, y) \neq c$ if $(x, y) \in U \times V$ and $y \neq g(x)$.

Before beginning the proof we remark that the conclusion can be rephrased thus: *the graph of g is the set*

$$F^{-1}(c) \cap (U \times V).$$

Thus $F^{-1}(c)$ is a "hypersurface" in a neighborhood of (x_0, y_0).

To prove the implicit function theorem we apply the inverse function theorem to the map

$$f: W \to E_1 \times E_2,$$

$$f(x, y) = (x, F(x, y)).$$

The derivative of f at $(x, y) \in W$ is the linear map

$$Df(x, y) : E_1 \times E_2 \to E_1 \times E_2,$$

$$(\xi, \eta) \to \left(\xi, \frac{\partial F}{\partial x}(x, y)\xi + \frac{\partial F}{\partial y}(x, y)\eta\right).$$

It is easy to find an inverse to this if $\partial F(x, y)/\partial y$ is invertible. Thus $Df(x_0, y_0)$ is invertible. Hence there is an open set $U_0 \times V \subset W$ containing (x_0, y_0) such that f restricts to a diffeomorphism of $U_0 \times V$ onto an open set $Z \subset E_1 \times E_2$.

Choose open sets $U \subset U_0$, $Y \subset E_2$ such that $x_0 \in U$, $c \in Y$, and

$$U \times Y \subset Z.$$

The inverse of $f: U_0 \times V \to Z$ preserves the first coordinate because f preserves it. The restriction of $(f \mid U_0 \times V)^{-1}$ to $U \times Y$ is thus a C^1 map of the form

$$h: U \times Y \to U_0 \times V,$$

$$h(x, w) = (x, \varphi(x, w)),$$

where

$$\varphi: U \times Y \to V$$

is C^1.

Define a C^1 map

$$g: U \to V,$$

$$g(x) = \varphi(x, c).$$

From the relation $f \circ h =$ identity of $U \times Y$ we obtain, for $x \in U$:

$$(x, c) = fh(x, c)$$

$$= (x, Fh(x, c))$$

$$= (x, F(x, \varphi(x, c)))$$

$$= (x, F(x, g(x))).$$

Thus

$$F(x, g(x)) = c$$

for all $x \in U$. Since f is one-to-one on $U \times V$, if $y \neq g(x)$, then

$$f(x, y) \neq f(x, g(x));$$

hence

$$(x, F(x, y)) \neq (x, F(x, g(x))) = (x, c),$$

so $F(x, y) \neq c$. This completes the proof of the implicit function theorem.

We note that if F is C^r, g is C^r.
From the identity

$$F(x, g(x)) = c,$$

we find from the chain rule that for all x in U:

$$\frac{\partial F}{\partial x}(x, g(x)) + \frac{\partial F}{\partial y}(x, g(x)) Dg(x) = 0.$$

This yields the formula

$$Dg(x) = -\left[\frac{\partial F}{\partial y}(x, g(x))\right]^{-1} \frac{\partial F}{\partial x}(x, g(x)).$$

References

1. R. Abraham, *Foundations of Mechanics* (New York: Benjamin, 1967).
2. R. Bartle, *The Elements of Real Analysis* (New York: Wiley, 1964).
3. S. S. Chern and S. Smale (eds.), *Proceedings of the Symposium in Pure Mathematics XIV, Global Analysis* (Providence, Rhode Island: Amer. Math. Soc., 1970).
4. U. D'Ancona, *The Struggle for Existence* (Leiden, The Netherlands: Brill, 1954).
5. C. Desoer and E. Kuh, *Basic Circuit Theory* (New York: McGraw-Hill, 1969).
6. R. Feynman, R. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vol. 1 (Reading, Massachusetts: Addison-Wesley, 1963).
7. N. S. Goel, S. C. Maitra, and E. W. Montroll, *Nonlinear Models of Interacting Populations* (New York: Academic Press, 1972).
8. P. Halmos, *Finite Dimensional Vector Spaces* (Princeton, New Jersey: Van Nostrand, 1958).
9. P. Hartman, *Ordinary Differential Equations* (New York: Wiley, 1964).
10. S. Lang, *Calculus of Several Variables* (Reading, Massachusetts: Addison-Wesley, 1973).
11. S. Lang, *Analysis I* (Reading, Massachusetts: Addison-Wesley, 1968).
12. S. Lang, *Second Course in Calculus*, 2nd ed. (Reading, Massachusetts: Addison-Wesley, 1964).
13. J. La Salle and S. Lefschetz, *Stability by Liapunov's Direct Method with Applications* (New York: Academic Press, 1961).
14. S. Lefschetz, *Differential Equations, Geometric Theory* (New York: Wiley (Interscience), 1957).
15. L. Loomis and S. Sternberg, *Advanced Calculus* (Reading, Massachusetts: Addison-Wesley, 1968).
16. E. W. Montroll, On the Volterra and other nonlinear models, *Rev. Mod. Phys.* **43** (1971).
17. M. H. A. Newman, *Topology of Plane Sets* (London and New York: Cambridge Univ. Press, 1954).
18. Z. Nitecki, *Differentiable Dynamics* (Cambridge, Massachusetts: MIT Press, 1971).
19. M. M. Peixoto (ed.), *Dynamical Systems* (New York: Academic Press, 1973).
20. L. Pontryagin, *Ordinary Differential Equations* (Reading, Massachusetts: Addison-Wesley, 1962).
21. A. Rescigno and I. Richardson, The struggle for life; I, Two species, *Bull. Math. Biophysics* **29** (1967), 377–388.
22. S. Smale, On the mathematical foundations of electrical circuit theory, *J. Differential Geom.* **7** (1972), 193–210.

23. J. Synge and B. Griffiths, *Principles of Mechanics* (New York: McGraw-Hill, 1949).
24. R. Thom, *Stabilité Structurelle et Morphogénèse: Essai d'une théorie générale des modèles* (Reading, Massachusetts: Addison-Wesley, 1973).
25. A. Wintner, *The Analytical Foundations of Celestial Mechanics* (Princeton, New Jersey: Princeton Univ. Press, 1941).
26. E. Zeeman, Differential equations for heartbeat and nerve impulses, in *Dynamical Systems* (M. M. Peixoto, ed.), p. 683 (New York: Academic Press, 1973).

Answers to Selected Problems

Chapter 1

Section 2, page 12

2. (a) (k_1e^t, k_2e^t, k_3e^t)
   (b) $(k_1e^t, k_2e^{-2t}, k_3)$
   (c) $(k_1e^t, k_2e^{-2t}, k_3e^{2t})$
6. $A = \text{diag}\,\{a_1, \ldots, a_n\}$ and $a_i < 0$, $i = 1, \ldots, n$.
8. (b) Any solutions u, v such that $u(0)$ and $v(0)$ are independent vectors.

Chapter 2

Page 27

1. $F(x) = -Kx$: $V(x) = \dfrac{K \| x \|^2}{2}$, $x \in \mathbf{R}^2$

   $$m\frac{dx^2}{dt^2} = -\text{grad } V = -Kx$$

   "Most" initial conditions means the set of $(x, v) \in \mathbf{R}^2 \times \mathbf{R}^2$ such that v is not collinear with x.

2. (a) with $V(x, y) = -\dfrac{x^3}{3} - \dfrac{2y^3}{3}$ and (c) with $V(x, y) = \dfrac{x^2}{2}$

7. *Hint*: Use (4) Section 6.

Chapter 3

Section 3, page 54

1. (a) $x(t) = 0, y(t) = 3e^{2t}$

2. $A = \begin{bmatrix} \frac{5}{3} & \frac{1}{3} \\ 2 & 0 \end{bmatrix}$

4. All eigenvalues are positive.

6. (b) $b > 0$

Section 4, page 60

1. (d) $x = 3e^t \cos 2t + 9e^t \sin 2t$
$y = 3e^t \sin 2t - 9e^t \cos 2t.$

Chapter 4

Section 1, page 65

2. $\dim E = \dim E_{\mathbf{C}}$ and $\dim F \geq \dim F_{\mathbf{R}}$

3. $F \supset R_{\mathbf{CR}}$

Section 2, page 69

1. (a) Basis for E is given by $(0, -\sqrt{2}, \sqrt{2})$ and $(1, -2, -1)$.

Matrix is $\begin{bmatrix} 0 & -\sqrt{2} \\ \sqrt{2} & 0 \end{bmatrix}$

Section 3, page 73

Introduce the new basis $(1, 0, 0)$, $(0, -\sqrt{2}, \sqrt{2})$, $(1, -2, -1)$, and new coordinates (y_1, y_2, y_3) related to the old by

$$x_1 = y_1 + y_2,$$
$$x_2 = -\sqrt{2}\, y_2 - 2y_3,$$
$$x_3 = \sqrt{2}\, y_2 - y_3.$$

In the new coordinates the differential equation becomes

$$y_1' = y,$$
$$y_2' = -\sqrt{2}\, y_3,$$
$$y_3' = \sqrt{2}\, y_2.$$

The general solution is

$$y_1 = Ce^t,$$
$$y_2 = A\cos(\sqrt{2}\, t) + B\sin(\sqrt{2}\, t),$$
$$y_3 = -B\cos(\sqrt{2}\, t) + A\sin(\sqrt{2}\, t).$$

Therefore

$$x_1 = Ce^t - B\cos(\sqrt{2}\, t) + A\sin(\sqrt{2}\, t),$$
$$x_2 = (2B - A\sqrt{2})\cos(\sqrt{2}\, t) - (B\sqrt{2} + 2A)\sin(\sqrt{2}\, t),$$
$$x_3 = (B + A\sqrt{2})\cos(\sqrt{2}\, t) + (B\sqrt{2} - A)\sin(\sqrt{2}\, t).$$

(The authors solved this problem in only two days.)

Chapter 5

Section 2, page 81

3. $A = 1, B = \sqrt{n}$

4. (a) $\sqrt{2}$ (b) $\frac{1}{2}$ (c) 1 (d) $\frac{1}{2}$

6. (a) and (d)

Section 3, page 87

1. Suppose $C \| S \| \leq N(S) \leq D \| S \|$. Then

$$N(ST) \leq D \| ST \| \leq D \| S \| \cdot \| T \| \leq \frac{D}{C^2} N(S)N(T).$$

3. *Hint*: Note $\dfrac{|Tx|}{|x|} = \dfrac{|Ty|}{|y|} = |Ty|$ if $y = \dfrac{x}{|x|}$.

4. (a) The norm is 1.

7. *Hint*: Use geometric series.

$$\sum_{i=0}^{\infty} x^i = \frac{1}{1-x} \quad \text{for} \quad 0 < x < 1, \quad \text{with } x = \| I - T \|.$$

13. *Hint*: Show that all the terms in the power series for e^A leave E invariant.

Section 4, page 97

1. (a) $x(t) = (K_2 - tK_1)e^{2t}$,
   $y(t) = K_1e^{2t}$.
   (b) $x(t) = e^{2t}(K_1 \cos t - K_2 \sin t)$,
   $y(t) = e^{2t}(K_2 \cos t + K_1 \sin t)$.
2. (a) $x(t) = (2t + 1)e^{2t}$,
   $y(t) = -2e^{2t}$.
   (b) $x(t) = 2e^{2t} \sin t$,
   $y(t) = -2e^{2t} \cos t$.
4. *Hint*: Consider A restricted to eigenspaces of λ and use result of Problem 3.
9. (a) sink (b) source (c) source
   (d) none of these (f) none of these
10. (a) Only if $a < -2$ are there any values of such k and in this case for $k > \sqrt{-2a}$.
    (b) No values of k.
14. *Hint*: There is a real eigenvalue. Study T on its eigenspace.

Section 5, page 102

1. (a) $x(t) = \frac{1}{17}[-4 \cos t + \sin t] - \frac{4}{17}e^{4t} + e^{4t}k$.

   (b) $x(t) = -\frac{1}{16}[4t + 1] + \frac{e^{4t}}{16} + e^{4t}k$.

   (c) $x(t) = A \cos t + B \sin t$.
   $y(t) = -A \sin t + B \cos t + 2t$.

Section 6, page 107

2. (a) $s(t) = \cos 2t$. (b) $s(t) = -e^{t-1} + e^{2t-2}$.
3. (a) $\cos\sqrt{3}\,t$, $\sin\sqrt{3}\,t$ (b) $\exp\sqrt{3}\,t$, $\exp -\sqrt{3}\,t$
4. *Hint*: Check cases (a), (b), (c) of the theorem.
8. $a = 0$, $b > 0$; period is $\sqrt{b}/2\pi$.

Chapter 6

Section 2, page 120

1. (a) Generalized 1-eigenspace spanned by $(1, 0)$, $(0, 1)$;

$$S = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad N = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

   (b) Generalized 1-eigenspace spanned by $(1, 0)$; generalized (-1)-eigenspace spanned by $(1, 2)$;

$$S = \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix}, \qquad N = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

2. If the rth power of the matrix is $[b_{ij}]$, then $b_{ij} = 0$ for $i < j + r$ $(r = 1, 2, \ldots)$.
3. The only eigenvalue is 0.
4. (c) $\begin{bmatrix} e^t & 0 & 0 \\ e^t - e^{2t} & e^{2t} & 0 \\ -e^t + e^{2t} & 0 & e^{2t} \end{bmatrix}$
5. Consider the $S + N$ decomposition.
6. A preserves each generalized eigenspace E_λ; hence it suffices to consider the restrictions of A and T to E_λ. If $T = S + N$, then $S \mid E_\lambda = \lambda I$ which commutes with A. Thus S and T both commute with A; so therefore does $N = T - S$.
8. Use the Cayley–Hamilton theorem.
15. Consider bases of the kernel and the image.

Section 3, page 126

1. Canonical forms:
   (a) $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} 0 & 1 & & \\ 0 & 0 & & \\ & & 0 & 1 \\ & & 0 & 0 \end{bmatrix}$
3. Assume that N is in nilpotent canonical form. Let b denote the number of blocks and s the maximal number of rows in a block. Then $bs \leq n$; also $b = n - r$ and $s \leq k$.
4. Similar pairs are (a), (d) and (b), (c).

Section 4, page 132

1. (a) $\begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}$ (c) $\begin{bmatrix} 1 + i & 1 \\ 0 & 1 + i \end{bmatrix}$
4. For $n = 3$: $\begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix}$, $\begin{bmatrix} a & 1 & 0 \\ 0 & a & 0 \\ 0 & 0 & b \end{bmatrix}$, $\begin{bmatrix} \alpha & -\beta & 0 \\ \beta & \alpha & 0 \\ 0 & 0 & c \end{bmatrix}$.

6. If $Ax = \mu x$, $x \neq 0$, then $0 = q(A)x = q(\mu)x$.
8. Show that A and A^t have the same Jordan form if A is a complex matrix, and the same real canonical form if A is real.

Section 5, page 136

1. (a) Let every eigenvalue have real part $< -b$ with $b > a > 0$. Let $A = S + N$ with S semisimple and N nilpotent. In suitable coordinates $\| e^{tS} \| \leq e^{-tb}$, $\| e^{tN} \| \leq Ct^n$. Then $\| e^{tA} \| \leq ce^{-tb}t^n$, and so $e^{ta} \| e^{tA} \| \to 0$ as $t \to \infty$. Let $s > 0$ be so large that $e^{ta} \| e^{tA} \| < 1$ for $t \geq s$. Put $k = \min(\| e^{tA} \|^{-1})$ for $0 \leq t \leq s$.
2. If x is an eigenvector belonging to an eigenvalue with nonzero real part, then the solution $e^{tA}x$ is not periodic. If ib, ic are pure imaginary eigenvalues, $b \neq \pm c$, and z, $w \in \mathbf{C}^n$ are corresponding eigenvectors, then the real part of $e^{tA}(z + w)$ is a nonperiodic solution.

Section 6, page 141

1. $s(t) = e^{-t}$.
2. (a) In (7), $A = B = 0$. Hence $s(0) = C$, $s'(0) = D$, $s^{(2)}(0) = -c$, $s^{(3)}(0) = -D$.

Chapter 7

Section 1, page 150

3. (a) Use $e^{tB}e^{tA} = e^{t(B+A)}$.

Section 2, page 153

2. Use the theorem of this section and Theorems 1 and 2 of Section 1.
3. Use Problem 2.

Section 3, page 157

1. (a) dense, open (b) dense (c) dense, open
   (e) open (f) open (g) dense, open

Chapter 8

Page 177

1. (a) $f(x) = x + 2$.

$$u_0(t) = 2,$$

$$u_1(t) = 2 + \int_0^t f(u_0(s))\, ds = 2 + \int_0^t 4\, ds$$

$$= 2 + 4t,$$

$$u_2(t) = 2 + \int_0^t (4 + 4s)\, ds = 2 + 4t + 2t^2,$$

$$u_3(t) = 2 + \int_0^t (4 + 4s + 2s^2)\, ds = 2 + 4t + 2t^2 + \frac{2t^3}{3}.$$

By induction

$$u_n(t) = 4\left(1 + t + \frac{t^2}{2!} + \cdots + \frac{t^n}{n!}\right) - 2.$$

Hence

$$x(t) = \lim_{n \to \infty} u_n(t) = 4e^t - 2.$$

(b)

$$u_0(t) = 0,$$

$$u_1(t) = \int_0^t 0\, ds = 0,$$

$$u_n(t) = 0$$

for all n: Hence $x(t) = 0$.

(c) $x(t) = t^{-1}$.

4. (a) 1
   (b) $\dfrac{|f(x) - f(0)|}{|x - 0|} \to \infty$ as $x \to 0$; no Lipschitz constant.
   (c) 1
5. (a) For $0 < c < \beta$ let

$$x(t) = \begin{cases} 0, & 0 \leq t \leq c \\ \frac{1}{27}(t - c)^3, & c \leq t \leq \beta. \end{cases}$$

Chapter 9

Section 1, page 185

2. For example, $f(x) = -x^3$ $(x \in \mathbf{R})$.
3. *Hint*: Use a special inner product on $\mathbf{R}^n$. Compute the rate of change of $|x(t)|^2$ where $x(t)$ is a solution such that $x(0)$ is (the real part of) an eigenvector for $Df(0)$ having positive real part; take $x(0)$ very small.
4. Use (b) of the theorem of Section 1.

Section 2, page 191

2. (a), (b), (c)
3. *Hint*: Look at the Jordan form of A. It suffices to consider an elementary Jordan block.

Section 3, page 199

1. $x^2 + y^2$ is a strict Liapunov function.
8. $V^{-1}[0, c]$ is positively invariant. The ω-limit set of any point of $V^{-1}[0, c]$ consists entirely of equilibria in $V^{-1}[0, c]$; hence it is just $\bar{x}$.

Section 4, page 204

2. Let $x' = -\text{grad } V(x)$. Then V decreases along trajectories, so that V is constant on a recurrent trajectory. Hence, a recurrent trajectory consists entirely of equilibrium points, and so is a constant.
3. (a) Each set $V^{-1}(-\infty, c]$ is positively invariant.
   (b) Use Theorem 3.

Section 5, page 209

1. *Hint*: Find eigenvectors.
2. Let $Ax = \lambda x$, $Ay = \mu y$, $\lambda \neq \mu$, $\mu \neq 0$. Then $\langle x, y \rangle = \mu^{-1}\langle x, Ay \rangle = \mu^{-1}\langle Ax, y \rangle = \lambda\mu^{-1}\langle x, y \rangle$, and $\lambda\mu^{-1} \neq 1$.
5. $Ax = \text{grad } \frac{1}{2}\langle x, Ax \rangle$.

Chapter 10

Section 1, page 215

1. $L\dfrac{dx}{dy} = y, \qquad C\dfrac{dy}{dt} = -x + f(y);$

   $x = i_L, \qquad y = v_C.$

Section 3, page 226

1. Every solution is periodic! *Hint*: If $(x(t), y(t))$ is a solution, so is $(-x(-t), y(-t))$.

Section 4, page 228

1. $\mu = -2, \qquad \mu = -1 \pm 2\sqrt{7}\,.$

Section 5, page 237

8. $L\dfrac{dx}{dt} = -y - R_x + E, \qquad C\dfrac{dy}{dt} = x - f(y)$

   $x = i_L, \qquad y = v_C$

Chapter 11

Section 1, page 241

1. *Hint*: If the limit set L is not connected, find disjoint open sets U_1, U_2 containing L. Then find a bounded sequence of points x_i on the trajectory with $x_i \notin U_1$, $x_i \notin U_2$.
4. *Hint*: Every solution is periodic.

Section 3, page 247

2. *Hint*: Apply Proposition 2.
4. *Hints*: (a) If x is not an equilibrium, take a local section at x. (b) See Problem 2 of Section 1.

Section 4, page 249

2. *Hint*: Let $y \in \gamma$. Take a local section at y and apply Proposition 1 of the previous section.

Section 5, page 253

2. *Hints*: (a) Use Poincaré–Bendixson. (b) Do the problem for $2n + 1$ closed orbits; use induction on n.
5. *Hint*: Let U be the region bounded by a closed orbit γ of f. Then g is transverse to the boundary γ of U. Apply Poincaré–Bendixson.

Chapter 13

Section 1, page 278

(a) *Hint*: Show that the given condition is equivalent to the existence of an eigenvalue α of $D\phi_\lambda(x)$ with $|\alpha| < 1$. Apply Theorem 2.

Section 3, page 285

3. *Hint*: If v is periodic of period λ, then so is rv for all $r > 0$.
5. *Hints*: (a) Do the problem first in case p is zero and g is linear. Then use Taylor's formula for the general case. (b) Apply the result in (a) after taking a local section.

Chapter 15

Section 2, page 303

2. This is pretty trivial. Since x' is the C^r function f, then x is C^{r+1}.

Chapter 16

Section 1, page 309

1. *Hint*: If B is close to A, each eigenvalue of B having negative real part will be close to a similar eigenvalue λ of A. Arguing as in the proof that S_1 is open in Theorem 1 of Chapter 7, Section 3, show that the sum of the multiplicities of these eigenvalues μ_i of B near λ equals the multiplicity of λ. Then show that bases for the generalized eigenspaces of the μ_i can be chosen near corresponding bases for λ.

Section 3, page 318

1. Suppose $Df(0)$ has 0 as an eigenvalue, let $g_\epsilon(x) = f(x) + \epsilon x$, $\epsilon \neq 0$. For $|\epsilon|$ sufficiently small, one of $g_{-\epsilon}$, g_ϵ will be a saddle and the other a source or sink; hence f cannot have the same phase portrait as both $g_{-\epsilon}$ and g_ϵ. If $Df(0)$ has $\pm\lambda i$, $\lambda > 0$, as an eigenvalue, then $g_{-\epsilon}$ is a sink and $g_{+\epsilon}$ is a source.
6. *Hint*: First consider the case where e^{tA} is a contraction or expansion. Then use Problem 1 of Section 1.

Subject Index