Differential Dynamical Systems

Revised Edition

James D. Meiss

University of Colorado
Boulder, Colorado

Society for Industrial and Applied Mathematics
Philadelphia

Library of Congress Cataloging-in-Publication Data
Names: Meiss, J. D.
Title: Differential dynamical systems / James D. Meiss, University of Colorado, Boulder, Colorado.
Description: Revised. | Philadelphia : Society for Industrial and Applied Mathematics, [2017] | Series: Mathematical modeling and computation ; 22 | Includes bibliographical references and index.
Identifiers: LCCN 2016052904 (print) | LCCN 2016053307 (ebook) | ISBN 9781611974638 (print) | ISBN 9781611974645 (ebook)
Subjects: LCSH: Differentiable dynamical systems--Mathematical models.
Classification: LCC QA614.8 .M45 2017 (print) | LCC QA614.8 (ebook) | DDC 515/.39--dc23
LC record available at https://lccn.loc.gov/2016052904

Contents

Preface to the Revised Edition

For this revised edition, I have corrected many typographical errors and improved a number of figures (3.2, 3.8, 4.1-3, 4.15, 5.7, 8.3, 8.5, 8.6, 8.7,8.14, 8.19, 9.7). The exposition in Section 3.2 has been reordered and made more complete, and I have added a couple of new of theorems (e.g., Theorem 7.3, Birkhoff Transitivity and Theorem 3.5, Completeness of C^0). Thanks to the sharp eyes and careful thinking of a number of readers, the statements and/or proofs of a number of other theorems have been improved (e.g., 3.24, 4.6, 4.8, 4.42, 4.23, 4.46, 5.9, 5.10, 5.11, and 7.12). Finally there are are several new exercises (3.3, 3.5, 4.16 and 7.9).

James Meiss
Boulder, Colorado
September 2016

Preface

On one level, this text can be viewed as suitable for a traditional course on ordinary differential equations (ODEs). Since differential equations are the basis for models of any physical systems that exhibit smooth change, students in all areas of the mathematical sciences and engineering require the tools to understand the methods for solving these equations. It is traditional for this exposure to start during the second year of training in calculus, where the basic methods of solving one- and two-dimensional (primarily linear) ODEs are studied. The typical reader of this text will have had such a course, as well as an introduction to analysis where the theoretical foundations (the ε's and δ's) of calculus are elucidated. The material for this text has been developed over a decade in a course given to upper-division undergraduates and beginning graduate students in applied mathematics, engineering, and physics at the University of Colorado. In a one-semester course, I typically cover most of the material in Chapters 1–6 and add a selection of sections from later chapters.

There are a number of classic texts for a traditional differential equations course, for example (Coddington and Levinson 1955; Hirsch and Smale 1974; Hartman 2002). Such courses usually begin with a study of linear systems; we begin there as well in Chapter 2. Matrix algebra is fundamental to this treatment, so we give a brief discussion of eigenvector methods and an extensive treatment of the matrix exponential. The next stage in the traditional course is to provide a foundation for the study of nonlinear differential equations by showing that, under certain conditions, these equations have solutions (existence) and that there is only one solution that satisfies a given initial condition (uniqueness). The theoretical underpinning of this result, as well as many other results in applied mathematics, is the majestic contraction mapping theorem. Chapter 3 provides a self-contained introduction to the analytic foundations needed to understand this theorem. Once this tool is concretely understood, students see that many proofs quickly yield to its power. It is possible to omit §§3.3–3.5, as most of the material is not heavily used in later chapters, although at least passing acquaintance with Theorem 3.19 and Lemma 3.28 (Grönwall) is to be encouraged.

However, this text does not aim to cover only the material in such a traditional ODE course; rather, it aspires to serve as an introduction to the more modern theory of dynamical systems. The emphasis is on obtaining a *qualitative* understanding of the properties of *differential* dynamical systems, namely, those evolution rules that describe smooth evolution in time.[1] The primary concept of this study, the *flow*, is introduced in Chapter 4. The qualitative theory is often concerned with questions of shape and asymptotic behavior that lead us to use topological notions such as conjugacy in the classification of dynamics.

[1] This is not to say that the dynamical systems that we study are always *differentiable*—vector fields need not be smooth.

The classification of dynamical behavior begins with the simplest orbits, equilibria and periodic orbits. As Henri Poincaré noted in his classic *New Methods in Celestial Mechanics*, (1892, Vol. 1, §36),

> *what renders these periodic solutions so precious to us is that they are, so to speak, the only breach through which we may attempt to penetrate an area hitherto deemed inaccessible.*

Only in the demonstration that dynamics in the neighborhood of some of these orbits is conjugate to their linearization is it seen that the predisposition of applied scientists to concentrate on linear systems has any value whatsoever.

The local classification of equilibria leads to the theory of invariant manifolds in Chapter 5. The stable and unstable manifolds, proved to exist for a hyperbolic saddle, give rise to one prominent mechanism for chaos—heteroclinic intersection. The center manifold theorem is also important preparation for the treatment of bifurcations in Chapter 8.

As mathematicians, allow yourselves to become entranced by the exceptions to the validity of linearization, namely, with those orbits that are nonhyperbolic. It is in the study of these exceptions that we find the most beautiful dynamics—even in the case of the phase plane, to which we return in Chapter 6. The first three sections of this chapter are fundamental; §§6.4–6.8 can be omitted in favor of later chapters. As we see in Chapter 8, the exceptional cases form the organizing centers for the behavior of systems undergoing changing parameters. A qualitative change in behavior under a small change of parameters is called a bifurcation. A complete exegesis of theory of bifurcations requires a full text on its own, and there are many excellent texts appropriate for a more advanced class (Guckenheimer and Holmes 1983; Golubitsky and Schaeffer 1985; Kuznetsov 1995). We introduce the reader to the basic ideas of normal forms and treat codimension-one and -two bifurcations.

Perhaps the most exciting recent developments in dynamical systems are those that show that even simple systems can behave in complicated ways, namely, the phenomena of *chaos*. In Chapter 7, we introduce the reader to the concepts necessary for understanding chaos: Lyapunov exponents, transitivity, fractals, etc. We also give an extensive discussion of Melnikov's method for the onset of chaos in Chapter 8. A more advanced treatment of chaotic dynamics requires a discussion of discrete dynamics (mappings) and can be found in texts such as (Katok and Hasselblatt 1999; Robinson 1999; Wiggins 2003).

The final chapter treats the subject closest to this author's heart: Hamiltonian dynamics. Since the basic models of physics all have a Hamiltonian (or Lagrangian) formulation, it is worthwhile to become familiar with them. While a traditional physics text treats these on a concrete level, this book provides an introduction to some of the geometrical aspects of Hamiltonian dynamics, including a discussion of their variational foundation, spectral properties, the KAM theorem, and transition to chaos. Again, there are several advanced texts that go much further, for example (Arnold 1978; Lichtenberg and Lieberman 1992; Meyer and Hall 1992).

While the proofs of many of the classical theorems are included, this text is not just an abstract treatment of ODEs but an attempt to place the theory in the context of its many applications to physics, biology, chemistry, and engineering. Examples in such areas as population modeling, fluid convection, electronics, and mechanics are discussed throughout the text, and especially in Chapter 1. The exercises introduce the reader to many more. Furthermore, to develop a geometrical understanding of dynamics, each student must experiment; we provide some examples of simple codes written in Maple, Mathematica, and MATLAB in the appendix, and we use the exercises to encourage the student to explore further. There are several texts that focus completely on using one or more of tools like these to explore dynamics (Lynch 2001; Baumann 2004).

I hope that this book conveys a bit of my amazement with the beauty and utility of this field. Dynamical systems is the perfect combination of analysis, geometry, and physical intuition. Central questions in dynamics have been formulated for centuries, and although some have been solved in the past few years, many await solution by the next generation.

> *It is far better to foresee even without certainty than not to foresee at all.*
> (Henri Poincaré, *The Foundations of Science*)

James Meiss
Boulder, Colorado
March 2007

Chapter 1

Introduction

> *It is not nature that imposes [time and space] upon us, it is we who impose them upon nature because we find them convenient.* (Henri Poincaré 1914)

This book is about dynamical systems governed by ordinary differential equations (ODEs). Although a typical reader will have seen differential equations in previous courses, we use this chapter to discuss their origins, give some examples of where they occur, and introduce a few of the classical techniques for finding their solutions.

1.1 ▪ Modeling

To construct a mathematical model of a physical system, one must decide on the realm in which the model lives. Since it would be impossible to describe everything in the universe, a model must include only a limited number of variables. The set of values that these variables can take makes up the *phase space* of the model. In this book we will study systems for which the phase space is finite dimensional—that is, the state of the model can be described by the values of finitely many variables. Typically, the *state* of the system will be denoted by x and the phase space by M; sometimes M will be the Euclidean space $\mathbb{R}^n$, and x a vector in that space; however, it is also common for the phase space to be a manifold. The main point is that for a given model with a phase space M, the modeler asserts that the system can be *completely* described by the variables $x \in M$ together with a set of constants that define the *parameters* of the model. For example a simple, planar pendulum has a fixed length and mass and is acted on by a constant gravitational field. The values of these constants describe the parameters of the system. The phase space M consists of possible values of the pendulum's position, represented by an angle, and of its angular velocity. Thus M is the two-dimensional cylinder, and the dynamics corresponds to smooth motion on M.

Models of systems that undergo evolution are called *dynamical systems*: in a dynamical system the state depends upon a special scalar quantity that is called *time*, denoted t. As we will discuss further in Chapter 4, there are many possible formulations of dynamical systems. When t can take all values on the real line, $\mathbb{R}$, and the state x changes continuously with t, the appropriate dynamical model is often a *differential equation*.

1.2 ▪ What Are Differential Equations?

> *Data aequatione quotcunque fluentes quantitates involvente, fluxiones invenire; et vice versa.* (Isaac Newton, as an anagram in a letter to Leibniz, 1677)[2]

As Newton realized, many aspects of the natural world can be accurately described by differential equations (fluxions). Indeed, the theory of gravitation consists essentially of the statement that gravitationally interacting bodies move according to a system of differential equations. In his letter to Leibniz, quoted above, Newton stated the fundamental problem: how does one "solve" a differential equation, or in Newton's terminology, "find the fluent quantities"? Although Newton and his contemporaries found some solutions to some of his equations, this is a problem that has occupied mathematicians and scientists ever since its conception.

Differential equations are relations between a function and its derivatives. When the function depends upon a single variable, the resulting differential equation is *ordinary* as opposed to *partial* (i.e., ordinary versus partial derivatives). Only the former case will be treated in this book. In our applications the independent variable usually represents *time*, so we call it t. For the moment, let us call the set of dependent variables $\mathbf{y}$; this "vector" is assumed to be a point in some space N. Often $N = \mathbb{R}^d$, Euclidean space with d dimensions, but N could also be a manifold such as a torus or cylinder (in which case the vector notation is not really appropriate). Mathematically, the fact that the function $\mathbf{y}$ maps its domain $\mathbb{R}$ to its range N is denoted

$$\mathbf{y} : \mathbb{R} \to N.$$

The set of values $C = \{\mathbf{y}(t) : t \in \mathbb{R}\}$ is a curve in N. The derivative of $\mathbf{y}$ with respect to t will be denoted $d\mathbf{y}/dt$ or $\dot{\mathbf{y}}$. An ODE is a relation among t, $\mathbf{y}$, and a finite number of derivatives of $\mathbf{y}$:

$$\mathbf{F}\left(t, \mathbf{y}, \frac{d\mathbf{y}}{dt}, \ldots, \frac{d^k\mathbf{y}}{dt^k}\right) = 0. \tag{1.1}$$

If the space N has d dimensions, then the relation (1.1) defines a *system* of d ODEs. The ODE is of kth *order* if F depends on the kth derivative of $\mathbf{y}$ but no higher derivative. Newton's problem can be restated in modern terms as follows: How can one find a function, $\mathbf{y} = \mathbf{u}(t)$, or, if possible the set of all possible functions, that makes $\mathbf{F} = 0$ an identity?

When (1.1) can be solved explicitly for the highest derivative term, the ODE becomes

$$\frac{d^k\mathbf{y}}{dt^k} = \mathbf{G}\left(t, \mathbf{y}, \dot{\mathbf{y}}, \ddot{\mathbf{y}}, \ldots, \frac{d^{k-1}\mathbf{y}}{dt^{k-1}}\right).$$

When this is not possible, the differential equation is *implicit*. In this case the coefficient of the highest derivative typically vanishes on some subset of the phase space and the ODE is said to have "singularities." In his classic book (Ince 1956), Edward Ince discusses some of the interesting things that can happen for the implicit case.

Any explicit ODE can be easily rewritten as a *system* of first-order equations by defining new variables,

$$\mathbf{x}_1 \equiv \mathbf{y}, \quad \mathbf{x}_2 \equiv \frac{d\mathbf{y}}{dt}, \quad \mathbf{x}_i \equiv \frac{d^{i-1}\mathbf{y}}{dt^{i-1}}, \quad \mathbf{x}_k \equiv \frac{d^{k-1}\mathbf{y}}{dt^{k-1}}.$$

[2] Given an equation involving any number of fluent quantities to find the fluxions, and conversely.

The resulting system consists of k first-order equations in the $\mathbf{x}_i$, written as

$$\begin{aligned}\frac{d\mathbf{x}_i}{dt} &= \mathbf{x}_{i+1}, \ i = 1, 2, \dots, k-1, \\ \frac{d\mathbf{x}_k}{dt} &= \mathbf{G}(t, \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k).\end{aligned} \tag{1.2}$$

Note that there are other ways of converting a system to first order, and these may be more convenient in applications (see, e.g., Exercise 5).

Since each $\mathbf{x}_i$ in (1.2) represents d variables, there are really $n = kd$ variables. Thus, each kth order system of ODEs on the d-dimensional space N is really a system of $n = kd$ first-order ODEs on the n-dimensional phase space $M = N^k$. Equation (1.2) is a special case of the general system of first order ODEs,

$$\frac{dx_i}{dt} = f_i(t, x_1, x_2, \dots, x_{n-1}, x_n), \ i = 1, 2, \dots, n,$$

or, more compactly,

$$\dot{x} = f(t, x). \tag{1.3}$$

Here we adopt the notation that x represents a set of variables, that is, a point in the phase space M of dimension n: the bold vector notation, $\mathbf{x}$, will no longer be used; it is replaced by carefully indicating the domain and range of our functions, e.g., $x : \mathbb{R} \to M$. The quantity $f(t, x)$ represents the velocity at time t at point x; consequently, $f : \mathbb{R} \times M \to \mathbb{R}^n$. Since any (explicit) differential equation can be written as a first-order system, (1.3) is the object that we will study.

The special case that f does not depend explicitly on time is called

> ▷ *autonomous*: A differential equation that does not depend explicitly on time.

In this case (1.3) becomes

$$\dot{x} = f(x). \tag{1.4}$$

For the system (1.4) the function $f : M \to \mathbb{R}^n$ specifies the velocity at each point in the phase space M; it is called a *vector field*. A vector field assigns to each point in space a velocity—the direction and speed of motion through that point. It is often visualized by plotting the values of f on a grid of points in the phase space as small vectors. The fact that $f(x)$ is a vector is reflected by the fact that if we change the units of time, replacing t by $\tau = t/c$, then the differential equation for the new function $x(\tau)$ becomes $dx/d\tau = cf(x)$. Consequently scalar multiplication is sensible for f; however, it is not appropriate for x if the components of x represent, say, angles.

Example 1.1. It is quite easy to use a computer algebra system such as Mathematica, Maple, or MATLAB to create a plot that represents a vector field. For example, consider the vector field

$$f(x) = \begin{pmatrix} \sin(xy) - y \\ y + x \end{pmatrix}. \tag{1.5}$$

In Figure 1.1 we show a plot generated using Mathematica on a 20×20 grid of arrows whose maximum length is scaled to one—see the appendix for the simple commands in Mathematica, Maple, and MATLAB for making such plots.

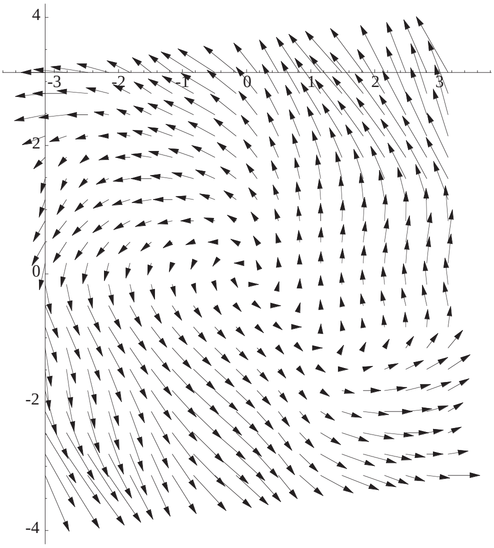

Figure 1.1. *The vector field* (1.5) *plotted by Mathematica.*

A solution to the differential equation corresponds to a curve that moves in the direction of the arrow at each point in the phase space. Much more will be said about vector fields and their properties in Chapter 4. ■

To reiterate, we say that a differential equation consists of a phase space M together with a vector field, $f : M \to \mathbb{R}^n$.

In principle, there is no reason to study nonautonomous systems since they can be rewritten as autonomous systems at the expense of introducing an additional variable, say, $x_{n+1} = t$. In this case x_{n+1} obeys the trivial equation $\dot{x}_{n+1} = 1$, so that upon replacing x by (x, x_{n+1}) and f by $(f, 1)$, (1.3) reduces to (1.4) with the dimension increased by one. However, there are some situations where it can be worthwhile to treat nonautonomous systems separately; for example, we will see in Chapter 3 that fewer assumptions on the smoothness of the dependence of f on t than on x are needed to show that the solutions of (1.3) are well behaved.

Physical problems for ODEs often require that we find a solution of an ODE that starts at a specific initial state. This is called the

> ▷ *initial value problem*: Find a solution $x(t)$ of (1.4) that satisfies a specific initial value $x(t_o) = x_o$ at a given time t_o.

In some cases one might be interested in finding a solution of an ODE that satisfies conditions at both an initial time and a final time; this is a “boundary value problem.” These more commonly arise in the context of partial differential equations (PDEs), or of minimization problems, and will not be studied in this book.

A nascent modeler confronted with a set of ODEs might hope to find the “complete” set of solutions, or find the

> ▷ *general solution*: A solution $x(t; c)$ of (1.4) that depends on a set of parameters c is the *general solution* if for *any* initial value x_o there is some choice of c such that $x(0; c) = x_o$.

Hence, one goal of the theory of ODEs would be to find analytically the general solution of an ODE system. It is perhaps surprising that this goal is essentially unattainable—the solutions of ODE systems can have incredible complexity. There is one case, however, where the general solution can always be found: the autonomous, linear case; this will be the subject of Chapter 2.

1.3 ▪ One-Dimensional Dynamics

> *Nothing puzzles me more than time and space; and yet nothing troubles me less, as I never think about them.* (Charles Lamb in a letter to Thomas Manning, 1810)

Dynamics in one dimension is much easier than that in higher dimensions primarily because motion on the line must be ordered (as we will discuss further in Chapter 4). Solving autonomous differential equations on $\mathbb{R}$ is no more difficult than antidifferentiation.

Any one-dimensional, autonomous initial value problem $\dot{x} = f(x)$, $x(0) = x_o$, can be integrated by the method of "separation of variables." This method is implemented by dividing both sides of the ODE by the function f and integrating the result:

$$\int_0^t \frac{\dot{x}(s)ds}{f(x(s))} = \int_0^t ds \quad \Rightarrow \quad \int_{x_o}^x \frac{du}{f(u)} = t. \tag{1.6}$$

Here we use a dummy integration variable s to avoid confusing it with the limit t. The second form is obtained by using the substitution $u = x(s)$, noting that $du = \dot{x}(s)ds$. In a formal sense, (1.6) constitutes the general solution of the ODE.

Example 1.2. One of the simplest nonlinear ODEs is the "logistic equation"

$$\dot{x} = rx(1-x), \tag{1.7}$$

which is a simple model for the growth of a population. Here $x = N/K$, where $N(t)$ is the number of individuals in a population at time t and K is the "carrying capacity" of the environment; see §1.4. The coefficient $r = b—d$ is the difference between the birth and death rates of the population when it is small compared to the carrying capacity. As N approaches K the population growth rate decreases, approaching zero at $N = K$, or equivalently at $x = 1$. This represents the fact that all the individuals are competing for a finite set of resources, and so the net growth rate must decrease as the population grows. For this application $x \geq 0$, so the phase space is the set $M = \mathbb{R}^+ \cup \{0\}$, the set of positive real values together with zero.

If the interval $[x, x_o]$ does not contain 0 or 1, then $1/f(u)$ exists and the integration represented by (1.6) can be easily done for (1.7) by the method of partial fractions. This results in

$$\ln\left|\frac{x}{1-x}\right| - \ln\left|\frac{x_o}{1-x_o}\right| = rt.$$

In this case, combining and then inverting the logarithms gives the explicit solution

$$x(t) = \frac{x_o}{x_o + (1-x_o)e^{-rt}}. \tag{1.8}$$

Since $1/f(x)$ does not exist for the cases $x = 0$ or 1, these cases must be studied separately. In both cases $\dot{x} = 0$, and so x does not change; therefore, there are two additional

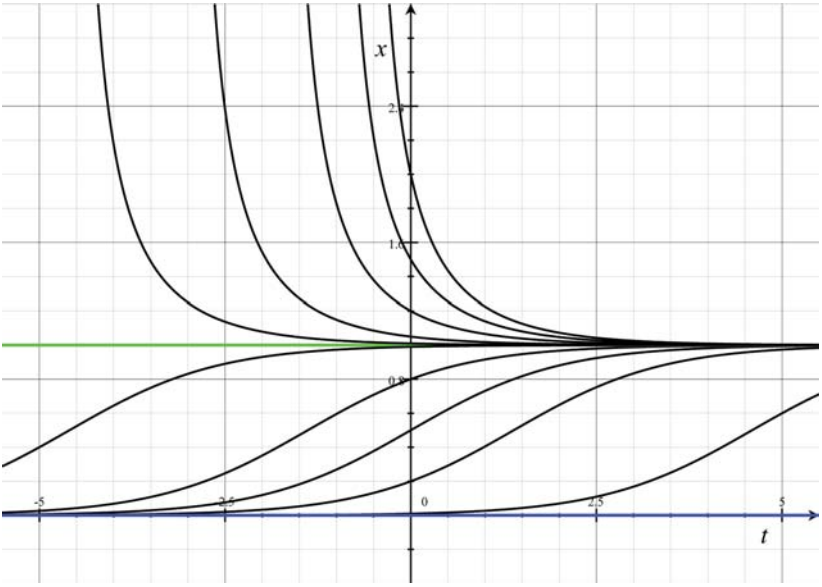

Figure 1.2. *Solutions of the logistic differential equation* (1.7) *as a function of time for* $r = 1$. *The green and blue lines at* $x = 1$ *and* $x = 0$ *are equilibria. All other solutions with* $x(0) > 0$ *asymptotically approach* $x = 1$ *as* $t \to \infty$.

equilibrium solutions, $x(t) \equiv 0$ and $x(t) \equiv 1$. Their validity can be seen from the ODE by direct substitution; for example, $\frac{d}{dt}(1) = f(1) = 0$. Note that the solution (1.8) actually works for $x_o = 0$ or 1. We have therefore proved that (1.8) is the general solution of (1.7). The solutions are sketched in Figure 1.2. ■

While (1.6) is the formal solution to a one-dimensional ODE, this integral cannot always be computed analytically; in this case one says that the ODE has been solved *up to a quadrature.*[3] Even if the integral can be done, the result is a formula for $t(x)$, not for $x(t)$, so that the solution is implicit. This implicit solution often cannot be analytically inverted.

Example 1.3. Consider the initial value problem

$$\dot{x} = f(x) = -\frac{x}{1+x^2}, \quad x(0) = x_o.$$

As before, one solution can immediately be found: since $f(0) = 0$, if x vanishes it does not change; in consequence, one solution is $x(t) \equiv 0$. Solutions that do not move are called *equilibria*. Under the assumption that $x \neq 0$, (1.6) can be easily integrated and the constant of integration eliminated in favor of x_o, giving

$$\ln|x| + \tfrac{1}{2}x^2 = -t + \ln|x_o| + \tfrac{1}{2}x_o^2. \tag{1.9}$$

This solution is valid for any $x_o \neq 0$, but it cannot be explicitly inverted to obtain $x(t; x_o)$ since the functions are transcendental. Nevertheless, the implicit solution (1.9)

[3] The Greeks used the word *quadrature* for the process of constructing a square with the same area as another figure, for example, a circle. More generally, it has the meaning of finding the area under any curve. The idea is that the value of an integral is known in principle as the limit of the Riemann sums, even though an explicit formula in terms of elementary functions may not exist.

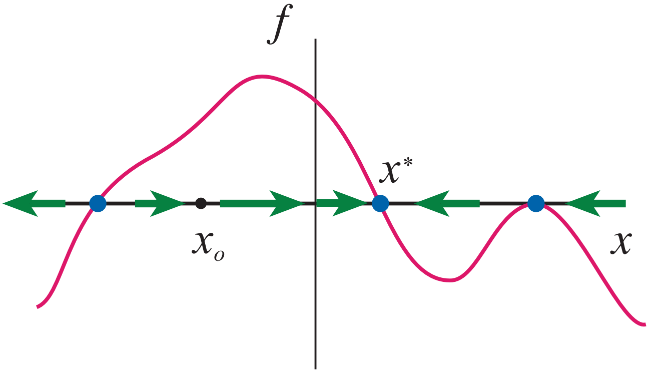

Figure 1.3. *Qualitative motion for a one-dimensional vector field with three equilibria. The size and direction of the arrows indicate the velocity.*

together with the solution $x(t;0)=0$ make up the general solution. The usefulness of this general solution is debatable. ■

Even when the integration (1.6) cannot be done or $t(x)$ cannot be inverted explicitly, graphical analysis can be used to extract most of the information that is important about a system.

In general, $f(x)$ represents the velocity at the point x in phase space. For the one-dimensional case there are only three qualitatively distinct cases: positive velocity, $f(x) > 0$, negative velocity, $f(x) < 0$, or equilibrium, $f(x) = 0$. The graph $\{(x,y): y = f(x)\}$ directly displays the intervals of initial conditions for which these conditions apply; see Figure 1.3. If $f(x_o) = \dot{x} > 0$, the motion is to the right and $x(t)$ grows. Indeed, $x(t)$ continues to increase monotonically so long as $f(x(t)) > 0$. If x^* is the first zero of f above x_o and $f(x) > 0$ on $[x_o, x^*)$, then $x(t) \underset{t\to\infty}{\to} x^*$. To see this, recall that every monotone, bounded function has a limit;[4] suppose that the limit is not x^* but rather $x(t) \to \xi < x^*$. Then, by continuity $f(x(t)) \to 0$ as $x \to \xi$, but this contradicts the fact that x^* was assumed to be the first zero above x_o. Similarly, if there are no zeros of f above x_o, then $x(t) \to \infty$ as t increases. A similar analysis applies on intervals where $f(x) < 0$.

Our conclusion is that the dynamics of a one-dimensional, autonomous ODE are extremely simple: trajectories move monotonically toward equilibrium or to infinity.

Example 1.4. Consider again the logistic equation (1.7), but now allow both positive and negative values of x. From the graph of f, it is easy to immediately extract certain qualitative properties. Since $f(x) < 0$ for all $x < 0$, the solution $x(t)$ for $x_o < 0$ must decrease monotonically with time and is unbounded: $x(t) \to -\infty$. When $0 < x < 1$, $f(x) > 0$, so the solution grows monotonically. As x approaches one, $f(x) \to 0$, so the motion slows, and the solution must limit to the value $x = 1$ as $t \to \infty$. Finally if $x > 1$, the solution decreases monotonically to 1. In conclusion, the equilibrium $x = 1$ is an *attractor*: all points $x \in \mathbb{R}^+$ asymptotically approach one as $t \to \infty$. We will formally define attractors in Chapter 4. ■

It is noteworthy that these conclusions can be obtained in a few lines of reasoning—a process that is much shorter than that leading to the analytical solution. Moreover,

[4] More generally, see Theorem 3.1, the Bolzano–Weierstrass Theorem.

even when we are given the solution (1.8), additional work must be done to extract these results since its form is complicated.

The use of geometrical methods to obtain *qualitative* information about a dynamical system without finding explicit solutions is a theme that will recur throughout this text.

1.4 ▪ Examples

Population Dynamics

When a biological population consists of many individuals, it is convenient to represent its number by a continuous function $N(t)$, although predictions that depend upon there being a nonintegral number of say, "rabbits," are suspect (unless you are cooking a portion for a meal). In a similar vein, when $N \gg 1$, discretely occurring birth and death events can be approximated by a population growth rate, so that $\dot{N} = b(N) - d(N)$, where b is the birth rate and d is the death rate. When the population is isolated (immigration does not occur) and mutation and speciation are neglected (the population maintains its identity and can reproduce only by births arising from existing members), then $b(0) = d(0) = 0$. Consequently, the ODE can be written in the form

$$\dot{N} = N r(N)$$

so that $r(N)$ is the net growth rate per individual. Such equations have the felicitous feature that if N is initially positive, it can never become negative—which would in any case be a gross violation of biology.

The logistic model, (1.7), corresponds to a simple version of the function r in which the net growth rate decreases linearly as the population grows, reflecting increased competition among the individuals for resources. For the logistic model the zero point K, $r(K) = 0$, is called the *carrying capacity* of the environment; if $N > K$, the population is too large to be sustained, and the death rate exceeds the birth rate.

Competition between species can be easily included in our model upon supposing that there are a number of species with populations N_i, $i = 1, \ldots, n$. The net growth rates of each species may depend upon the populations in the other species if they compete for the same resource or if one species serves as a food source (is prey) for another (a predator). The general model will have the form

$$\dot{N}_i = N_i r_i(N_1, N_2, \ldots, N_n), \; i = 1, \ldots, n.$$

In the spirit of the logistic model, it is interesting to consider the case that the per-individual growth rates r_i depend linearly on the populations; such models are called *Lotka–Volterra systems*. For example, when there are two species competing for resources, the model becomes

$$\begin{aligned} \dot{N}_1 &= N_1(a - bN_1 - cN_2), \\ \dot{N}_2 &= N_2(d - eN_1 - fN_2). \end{aligned} \tag{1.10}$$

The coefficients (a, b, c, d, e, f) are typically positive; (a, d) represent net growth rates when the populations are small, (b, f) represent intraspecies competition, and (c, e) represent interspecies competition. This model will be studied in §1.5 when we examine dynamics in two dimensions more generally. In certain parameter regimes the two species will be seen to stably coexist, while in others one species always drives the other to extinction.

In contrast, when one species is a food source for the other, it is reasonable to suppose that if the prey are scarce, then the predators will die off; that is the net birth rate $d < 0$; however, the prey, who may be feeding on plentiful vegetation, will have a net positive birth rate $a > 0$. Neglecting intraspecies competition, the model becomes

$$\begin{aligned}\dot{N}_1 &= N_1(\alpha - \beta N_2),\\ \dot{N}_2 &= N_2(-\gamma + \delta N_1),\end{aligned}$$

where again the parameters are positive. Solutions of this model have been compared to data collected by fur trappers for snowshoe hares (N_1) and Canadian lynx (N_2) over about a century, beginning in 1845. These populations are observed to oscillate in time (as the model predicts) with a period of approximately a decade. However, this model does not take into account the important effect of the trappers themselves.

Each additional species adds a dimension to the phase space. For example, the three-species food-chain model proposed by Rosenweig (1973),

$$\begin{aligned}\dot{R} &= R\left(1 - \frac{R}{K}\right) - x_c y_c \frac{CR}{R + R_o},\\ \dot{C} &= -x_c C\left(1 - y_c \frac{R}{R + R_o}\right) - x_p y_p \frac{PC}{C + C_o},\\ \dot{P} &= -x_p P\left(1 - y_p \frac{C}{C + C_o}\right),\end{aligned} \tag{1.11}$$

has been much studied. Here R, C, and P represent populations of the resource, consumer (of the resource), and predator (of the consumer), respectively. The resource has a simple logistic intraspecies competition; the remaining terms all correspond to interspecies competition. These nonlinear terms do not reverse sign like the logistic term but instead saturate when the populations become large compared to the "saturation densities" R_o and C_o. The saturation models the fact, for example, that an animal has only a finite need for food. The coefficients x_i and y_i represent "mass specific metabolic rates" for the consumer or predator.

Much more about biological modeling is contained in the excellent text (Murray 1993).

Mechanical Systems

A mechanical system consisting of a set of rigid pieces interacting through forces can be modeled by a system of Newtonian equations using $\mathbf{F} = m\mathbf{a}$. For example, suppose there are d components that can be idealized as points at locations $q_i \in \mathbb{R}^3$ with masses m_i, $i = 1, 2, \ldots, d$. If the force is assumed to be due to some potential energy $V(q_1, \ldots, q_d) = V(q)$, then in Cartesian coordinates the equations have the form

$$m_i \ddot{q}_i = -\frac{\partial}{\partial q_i} V(q).$$

These equations can be converted into a first-order system by defining the momenta $p_i = m_i \dot{q}_i$ so that

$$\begin{aligned}\dot{q}_i &= \frac{p_i}{m_i},\\ \dot{p}_i &= -\frac{\partial}{\partial q_i} V(q).\end{aligned} \tag{1.12}$$

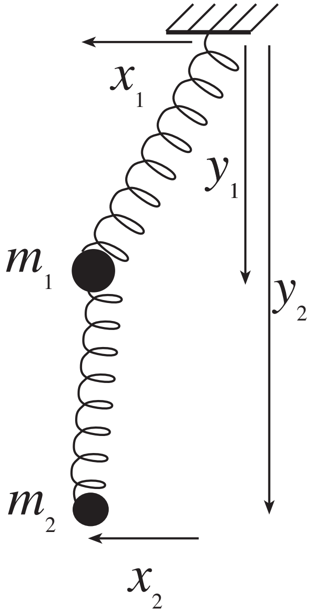

Figure 1.4. *Coupled harmonic springs.*

Example 1.5. Consider a pair of coupled springs in the plane as shown in Figure 1.4: a mass, m_1 at position $q_1 = (x_1, y_1)$ hangs below a fixed support at the origin, $(0,0)$, on a linear spring with spring constant k_1. It is connected to a second mass, m_2, at position $q_2 = (x_2, y_2)$ by a second spring with spring constant k_2. We let positive y be downward. If a spring is assumed (somewhat artificially!) to have zero natural length, then its potential energy is proportional to the square of its length, and the total spring potential energy is $V_s(q_1, q_2) = \frac{k_1}{2}|q_1|^2 + \frac{k_2}{2}|q_1 - q_2|^2$. If the force due to gravity is assumed constant (the distances moved are small compared to the earth's radius), the gravitational potential energy is $V_g = -m_1 g y_1 - m_2 g y_2$. Newton's equations of motion for this system then have the form

$$\begin{aligned}
m_1\ddot{x}_1 &= -k_1 x_1 - k_2(x_1 - x_2),\\
m_2\ddot{x}_2 &= -k_2(x_2 - x_1),\\
m_1\ddot{y}_1 &= -k_1 y_1 - k_2(y_1 - y_2) + m_1 g,\\
m_2\ddot{y}_2 &= -k_2(y_2 - y_1) + m_2 g.
\end{aligned}$$

These can be converted into a system of eight first-order equations of the form (1.12). These equations are affine and can be solved by the eigenvalue methods in Chapter 2. Note that if the springs have nonzero natural length, additional affine terms are added to the equations, see §2.1 and Exercise 9.10. ■

Equations (1.12) are an example of a "Hamiltonian system." More generally, let $\{(q_i, p_i) : i = 1, 2, \ldots, n\}$ denote n pairs of variables corresponding to a scalar configuration component, q_i, and its corresponding momentum, p_i; each pair represents a *degree of freedom* of the system. The Hamiltonian function is the total energy of the system

$$H(q, p) = T(p) + V(q),$$

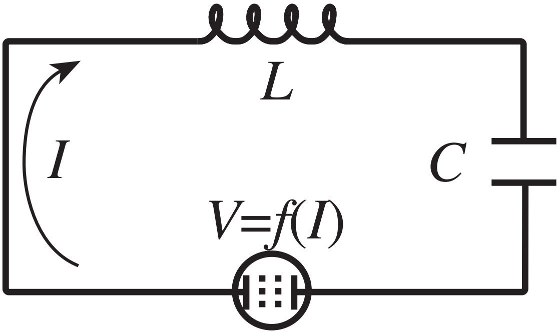

Figure 1.5. *Van der Pol circuit.*

where T is the kinetic energy and V is the potential energy. It is easy to verify that the single function H generates (1.12) if we set $T = \sum_i \frac{|p_i|^2}{2m_i}$ and use the relations

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial q_i}. \tag{1.13}$$

Hamiltonian systems will be used as examples in many sections of this book; the geometry of Hamiltonian dynamics will be studied extensively in Chapter 9.

Oscillating Circuits

Electrical circuits typically combine inductive elements that store magnetic energy, capacitive elements that store electrical energy, resistive elements that dissipate energy, and voltage or current sources. Each circuit element is characterized by a relationship between the current I that flows through it and the voltage V that drops across it. Nonlinearity arises in circuits through elements such as vacuum tubes and solid-state devices such as transistors or operational amplifiers. A circuit with a triode tube, studied by the Dutch electrical engineer Balthazar van der Pol in 1922, gives rise to a famous system that bears his name (Nayfeh and Mook 1979, §3.1.7; van der Pol 1922).

A simplified circuit that also gives van der Pol's model is shown in Figure 1.5. It consists of a single loop containing an inductor, a capacitor, and a vacuum tube. (Here the circuitry driving the tube is omitted.) The voltage drop across an inductor is proportional to the rate of change of the current through it: $V_L = L\dot{I}$. Capacitors are characterized by $I = C\dot{V}_C$, so that the current is proportional to the rate of change of the voltage drop. A vacuum tube has a current-voltage characteristic that can be represented by a function, $V_T = f(I)$, here assumed to be

$$V_T = -RI + NI^3,$$

which means that it acts as a negative resistor ($-RI$) when the current is small but dissipates energy when it is large.

Kirchoff's law gives the equation for the circuit: the sum of the voltage drops around any loop is zero (this is nothing more than energy conservation):

$$L\dot{I} + V_T + V_C = 0.$$

Combining this with the equation for the capacitor gives the system

$$\dot{V}_C = \frac{1}{C} I,$$
$$\dot{I} = -\frac{1}{L} V_C + \frac{R}{L} I - \frac{N}{L} I^3.$$

It is more traditionally written as a second-order equation for the current, obtained by differentiating the current equation and substituting for $\dot{V}_C$:

$$\ddot{I} + \frac{1}{L}\left(3NI^2 - R\right)\dot{I} + \frac{1}{LC} I = 0. \tag{1.14}$$

We will see in Chapter 6 that this equation indeed has oscillatory solutions. Indeed, there is a unique periodically oscillating solution that is an attractor; it is called a *limit cycle*. This equation also exhibits a prototypical *bifurcation*, that is, a qualitative change in solutions with a change in parameters; these will be studied in Chapter 8. The van der Pol oscillator undergoes an *Andronov–Hopf bifurcation* when the negative resistance R crosses zero.

Fluid Mixing

The motion of a fluid can properly be considered a dynamical system; however, its phase space is a function space and has infinitely many dimensions. For example, to specify the state of a fluid, its velocity, $\mathbf{v}$, must be given at every point in the fluid domain—this corresponds to the *Eulerian velocity field*. The simplest fluids obey a set of partial differential equations (PDEs): the Navier–Stokes equations. As we noted in §1.1, the dynamical systems in this book all will be finite dimensional.

There is an interesting case in which a finite dimensional dynamical system is relevant to fluid mechanics: the motion of a small particle in the fluid. In the simplest approximation, the particle will move along with the fluid so that its velocity $\dot{x}$ at a point x, and time t in the fluid must equal the fluid velocity field $\mathbf{v}(x,t)$. Supposing that the Navier–Stokes equations have been solved (a large supposition!) so that $\mathbf{v}$ is known, we then see that the particle obeys the system

$$\dot{x} = \mathbf{v}(x,t). \tag{1.15}$$

For a three-dimensional fluid, $\mathbf{v} \in \mathbb{R}^3$, and the phase space of our system is the domain of the fluid motion. The dynamics represented by (1.15) is called the motion of a *passive scalar* or the Lagrangian dynamics of the fluid.

If the particle is not neutrally buoyant, then its dynamics is influenced by gravity and it cannot be treated as a passive scalar. Similarly, when the particle has significant mass, there will be drag terms in the dynamical equations because the inertia requires a force to cause the particle's velocity to change. Moreover, a finite-size particle with inertia will itself change the fluid flow as the fluid is forced to move around the particle—the dynamics of the particle is no longer "passive."

The passive scalar dynamics (1.15) does apply to the motion of a blob of dye placed in the fluid, providing that it has the same density as the surrounding fluid and that molecular diffusivity is small enough that it is unimportant over the time scale of interest.

An interesting example velocity field, called the "ABC flow," was introduced by Arnold in 1965:

$$\mathbf{v} = (A\sin z + C\cos y, B\sin x + A\cos z, C\sin y + B\cos x)^T. \tag{1.16}$$

This velocity field is periodic in space and is incompressible—it satisfies $\nabla \cdot \mathbf{v} = 0$ and has the so-called Beltrami property: $\mathbf{v} = \nabla \times \mathbf{v}$. Moreover, it is an exact solution of the Navier–Stokes equations for any values of the amplitudes (A, B, C) when an appropriate forcing term is added to counter viscous dissipation; see Exercise 6. When the viscosity is large enough (or more precisely when the ratio of inertial forces to viscous forces—the "Reynolds number"—is small enough) this solution is even a stable solution of the Navier–Stokes equations. The ABC flow has also been used in studies of the dynamo effect: the enhancement of magnetic fields by stretching of the fluid motion. The parameters ABC can also be thought of as representing Arnold—the inventor, Beltrami—for the flow condition, and Childress—who made fundamental contributions to dynamo theory.

Since the ABC velocity field is steady (the Eulerian velocity depends only upon space) the ODE system (1.15) is autonomous. However, the solutions to this set of equations are very complicated, unless two of the parameters are set to zero. Indeed, this system is a prototype *chaotic* system (Dombre et al. 1986). A signature of chaos is that nearby trajectories will often diverge exponentially quickly in time:

$$|x_1(t) - x_2(t)| \sim e^{\lambda t} \, |x_1(0) - x_2(0)| \,. \tag{1.17}$$

Here the exponent, λ, is called the Lyapunov exponent; see Chapter 7. For the ABC flow with $A = B = C = 1$, it is found from numerical studies that $\lambda \approx 0.055$. For example, if the nearby trajectories correspond to points in a blob of dye of linear size 10^{-6}, then by $t \approx 280$, the dye will have spread over a distance of order 2π, becoming well mixed even in the absence of diffusion. On the other hand, the ABC flow also has many regular trajectories; these often cover two-dimensional tori. The complex mixture of regular and chaotic solutions makes the study of systems like this both challenging and fun!

1.5 ▪ Two-Dimensional Dynamics

Just as for the one-dimensional case that we discussed in §1.3, a graphical analysis of motion in the plane is also often possible. Letting $z = (x, y)$ represent a point in the plane, a general two-dimensional ODE is

$$\dot{z} = f(z) = \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} P(x,y) \\ Q(x,y) \end{pmatrix}. \tag{1.18}$$

As for the one-dimensional case, the equilibria, $S = \{(x, y) : P(x, y) = Q(x, y) = 0\}$, are important organizing centers for the motion, and a first step in analyzing any ODE system is to find these. In Chapter 4 and Chapter 6 we will study how the global dynamics is influenced by the local dynamics in the neighborhood of equilibria.

Nullclines

To gain additional information about the equilibria it is also useful to consider the *nullclines*, curves on which a single component of the velocity vanishes,

$$\begin{aligned} N_x &= \{(x, y) : P(x, y) = 0\}, \\ N_y &= \{(x, y) : Q(x, y) = 0\}. \end{aligned} \tag{1.19}$$

Since these sets are defined by a single equation, they generically define curves or collections of curves in the plane. On the set N_x the velocity is strictly vertical, and on

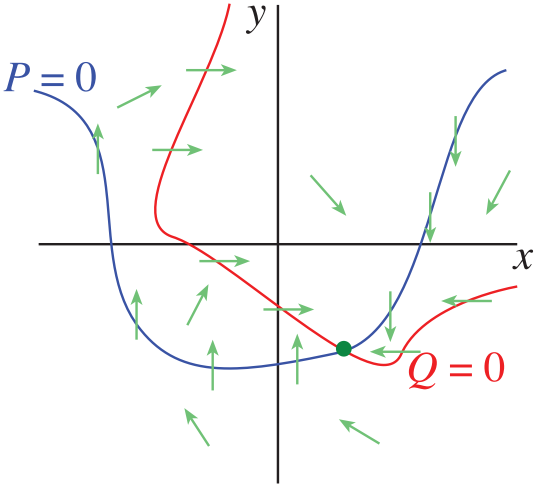

Figure 1.6. *Sketch of nullclines (blue and red curves) and the corresponding vector field. The vector field typically reverses on a nullcline upon passing through an equilibrium (green dot).*

the set N_y it is horizontal. Equilibria correspond to the intersections of the nullclines: $S = N_x \cap N_y$. Inside each region bounded by nullcline curves or extending to infinity, the velocity vector lies in a particular quadrant; see Figure 1.6. It is easiest to see why this is useful by an example.

Example 1.6. The Lotka–Volterra system for the competitive interaction of two species is given by (1.10); rewriting it for variables (x,y) gives

$$\begin{aligned} \dot{x} &= x(a-bx-cy), \\ \dot{y} &= y(d-ex-fy). \end{aligned} \tag{1.20}$$

Since x and y represent populations, they must be nonnegative. Consequently only the first quadrant of the plane is relevant, so the phase space is $M = \{(x,y) : x \geq 0, y \geq 0\}$. Recall that the coefficients (a,b,c,d,e,f) are positive for the biological application. Each nullcline is a union of two lines:

$$N_x = \{x=0\} \cup \left\{ y = \frac{1}{c}(a-bx) \right\}, \quad N_y = \{y=0\} \cup \left\{ y = \frac{1}{f}(d-ex) \right\}.$$

Since N_x includes the y-axis where the velocity is vertical and N_y includes the x-axis where the velocity is horizontal, no orbits can cross the axes. Therefore, orbits that start in M remain in M for all $t \in \mathbb{R}$: it is an *invariant set*; see §4.1.

The set $S = N_x \cap N_y$ typically consists of points, though there are special cases where S contains a line (see Exercise 7). It is important to note that an equilibrium corresponds to the intersection of one of the curves in N_x with one of the curves in N_y. This means, for example, that the intersection of the line $\{(x,y) : x = 0\}$ with $\{(x,y) : y = (a-bx)/c\}$ is not an equilibrium since both curves are in N_x. Since we have assumed that all of the parameters are positive, there are always three equilibria in M: the points $(0,0)$, $(0,d/f)$, $(a/b,0)$. The fourth equilibrium at

$$(x^*, y^*) = \left(\frac{af-cd}{bf-ce}, \frac{bd-ae}{bf-ce} \right)$$

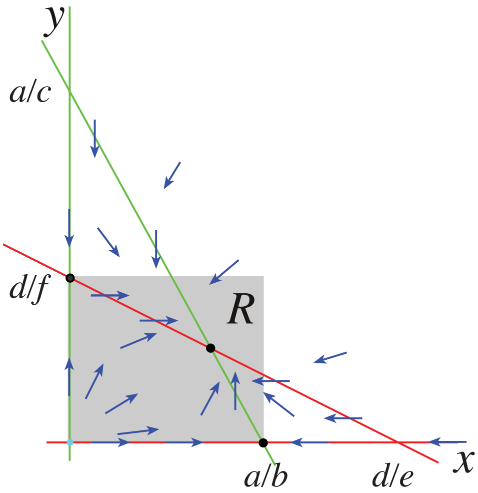

Figure 1.7. *Phase portrait of the Lotka–Volterra system for the case $s = 1$ where there are four equilibria. The closed rectangle R is forward invariant. For this case, the equilibrium at (x^*, y^*) is a global attractor for all orbits in the interior of M.*

is in the interior of M when the terms $af - cd$, $bd - ae$, and $bf - ce$ are nonzero and have the same sign. For the choice

$$s = \operatorname{sgn}(af - cd) = \operatorname{sgn}(bd - ae) = 1, \tag{1.21}$$

it is not hard to see that $bf - ce > 0$ as well, so that $(x^*, y^*) \in \operatorname{int}(M)$, the interior of the phase space. In this case the vector field has the form shown in Figure 1.7. The nullclines divide M into regions that correspond to fixed quadrants of the velocity vector; for this case there are four such regions. Note that both $\dot{x}$ and $\dot{y}$ are negative for large enough values of (x, y). In particular,

$$x > a/b \quad \Rightarrow \quad \dot{x} < 0,$$

so x is monotone decreasing. This implies that all initial conditions to the right of the vertical line $\{(x, y) : x = a/b\}$ move leftward, and if $y > 0$ they will eventually cross this line. Similarly, whenever $y > d/f$, y decreases monotonically. Consequently, the rectangle $R = \{(x, y) : 0 \le x \le a/b,\ 0 \le y \le d/f\}$ is a *forward invariant set*: all orbits that start in R stay in R thereafter. Moreover, every initial condition in $\operatorname{int}(M) \setminus R$, the interior of the phase space that is not in R, eventually must enter R.

So far we have seen that the velocity vector lies in the third quadrant above and to the right of the nullclines, as shown in Figure 1.7. The quadrant of the velocity typically changes upon each crossing of a nullcline. For example, the velocity vector must lie in the first quadrant near the origin since a and $d > 0$; therefore, $(0, 0)$ is a *source*. For the case shown, the two equilibria on the axes are *saddles*—the solutions that begin on the axes are attracted to and eventually limit to the equilibria; however, all points near these equilibria but off the axes eventually move away. The equilibrium at (x^*, y^*) is a *sink*; indeed when $s = 1$, (x^*, y^*) is a *global attractor* for all initial conditions in the interior of the first quadrant (see also Exercises 7 and 8). Confirmation of this qualitative analysis by linearization, and formal definitions for these terms, will be given in Chapter 2 and Chapter 4. ∎

The limiting behavior as $t \to \infty$ for each initial condition in the previous example is very simple—each one is attracted to an equilibrium. One of the goals of our global analysis in Chapter 6 will be to classify which asymptotic behaviors are possible and which actually do occur.

Phase Curves

It is sometimes possible to find the solutions of (1.18) as curves in the phase plane by ignoring their time dependence. The idea is that if an orbit is locally the graph of a function, $y = Y(x)$, then since $\dot{y} = \frac{dY}{dx}\dot{x}$ along a trajectory, the function Y obeys the differential equation

$$\frac{dY}{dx} = \frac{\dot{y}}{\dot{x}} = \frac{Q(x,Y)}{P(x,Y)} = F(x,Y). \tag{1.22}$$

Note that this equation is a single, first-order ODE for the function $Y(x)$; usually it is nonautonomous since the new vector field $F(x,Y)$ depends on the new independent variable, x.

Example 1.7. The system

$$\begin{aligned} \dot{x} &= e^{x+y}(x+y), \\ \dot{y} &= e^{x+y}(x-y) \end{aligned} \tag{1.23}$$

is not obviously explicitly solvable for $(x(t), y(t))$. However, the equation for the phase curves, (1.22), is relatively simple:

$$\frac{dy}{dx} = \frac{x-y}{x+y}.$$

Since this ODE is nonautonomous, it cannot be solved using (1.6); however, it does fall into a classical case first treated by Leibniz in 1691, that of *homogeneous* ODEs. Such equations can be solved explicitly using a variable transformation trick (see, e.g., (Ince 1956)): define a new variable $z = y/x$ so that

$$\frac{dz}{dx} = \frac{1}{x}\frac{dy}{dx} - \frac{y}{x^2} = \frac{1}{x}\left(\frac{1-z}{1+z} - z\right) = -\frac{(z+1)^2-2}{x(1+z)}.$$

Since the vector field for this equation is a product of a function of z and a function of x, it is *separable* and can be solved explicitly. Generally a system of the form $dz/dx = F(z)G(x)$ has a quadrature solution of the form

$$\int_{z_o}^{z} \frac{dz}{F(z)} = \int_{x_o}^{x} G(x)dx. \tag{1.24}$$

For our system we obtain, after some algebra,

$$y = -x \pm \sqrt{2x^2 + c}. \tag{1.25}$$

There are two branches to this solution, indicating that the assumption that $y = Y(x)$ is a graph fails—indeed it does whenever $y = -x$. However, the orbits can be obtained by squaring (1.25):

$$(y+x)^2 - 2x^2 = c.$$

Consequently, the orbits correspond to a family of hyperbolas. Our solution, however, has given us no information about the time dependence of the trajectories. ■

The ODE (1.22) does not make sense when P vanishes; it can, however, be viewed as a *differential form*:[5]

$$\alpha = -Q(x,y)dx + P(x,y)dy. \tag{1.26}$$

Along an orbit, $dx = Pdt$ and $dy = Qdt$ so that $\alpha(x(t),y(t)) = (-PQ+QP)dt = 0$ for any trajectory. The more general form (1.26) frees us from using the particular parameterization $(x(t),y(t))$ of the trajectory; any curve $C = \{(x(s),y(s)) : s \in \mathbb{R}\}$ for which $\alpha|_C \equiv 0$ is a trajectory. A differential form is *exact* if it is a perfect derivative, $\alpha = dH$; in other words, since $dH = (\partial H/\partial x)dx + (\partial H/\partial y)dy$, then

$$\frac{\partial H}{\partial x} = -Q, \quad \frac{\partial H}{\partial y} = P.$$

In this case the system (1.18) is Hamiltonian, (1.13). Moreover, since

$$\alpha(x(t),y(t)) = \left(\frac{\partial H}{\partial x}\frac{dx}{dt} + \frac{\partial H}{\partial y}\frac{dy}{dt}\right)dt = \frac{dH}{dt}dt = 0,$$

the Hamiltonian is constant along the orbits, and they lie on the energy contours, $H(x,y) = E$.

More generally, the one-form α may be a multiple of a perfect differential:

$$\alpha = F(x,y)dH. \tag{1.27}$$

In this case the system (1.18) has the form

$$\dot{x} = F\frac{\partial H}{\partial y}, \quad \dot{y} = -F\frac{\partial H}{\partial x}.$$

This reduces to the Hamiltonian system (1.13) if we formally define the new time variable,

$$\tau = \int_0^t F(x(s),y(s))ds, \tag{1.28}$$

because $\frac{dx}{d\tau} = \frac{dx}{dt}\frac{1}{F} = \frac{\partial H}{\partial y}$, etc.

Example 1.8. The system (1.23) is easily seen to fall into the case (1.27) with

$$H = \tfrac{1}{2}\left(y^2 - x^2\right) + xy \tag{1.29}$$

and $F = e^{x+y}$. Consequently, the phase curves lie on contours of H. The Hamiltonian dynamics in the new time τ is linear and can easily be solved using the methods of Chapter 2; see Exercise 2.2. The contours of H are shown in Figure 1.8. ■

Although the phase curve equation (1.22) is sometimes useful (see §6.2), it is not of much help in general: for most ODEs, (1.22) will not have solutions in terms of "elementary" or even "special" functions, such as Bessel or elliptic functions. Nevertheless, classical texts on ODEs contain many such "tricks" that work on special classes of systems (Ince 1956). However, even *if* analytical solutions can be obtained, their behavior is often difficult to extract from the often complex formulas.

[5] A differential one-form is a linear combination of the differentials dx_i. These are used extensively in differential geometry.

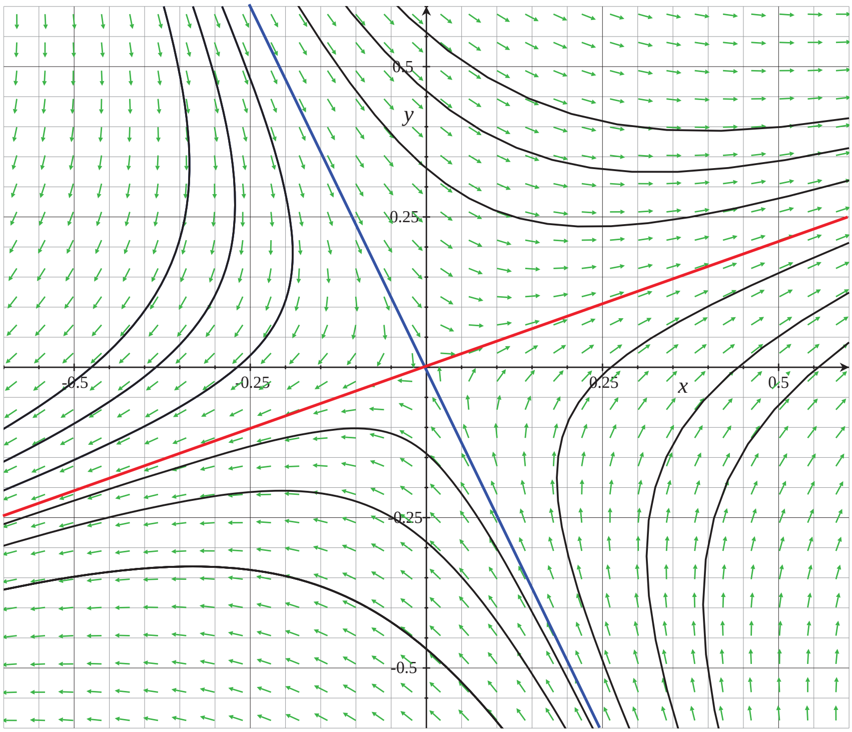

Figure 1.8. *Phase portrait of the flow for* (1.23)*, or equivalently, the contours of the Hamiltonian* (1.29)*. The red line is the unstable manifold of the origin, and the blue is the stable manifold. See Chapters* 2 *and* 4*.*

1.6 ▪ The Lorenz Model

Even though many physical systems are modeled by PDEs—infinite dimensional dynamical systems—there are cases in which the dynamics is sufficiently dissipative that it contracts onto a finite dimensional subspace. Indeed, there may even exist a finite dimensional set to which all solutions are attracted. This set is often a fractal (see §7.3), but in some cases it can be shown to be a subset of a smooth manifold, a so-called inertial manifold (Eden et al. 1994). In this case the long time dynamics could, at least in principle, be studied using an ODE model.

There are other cases in which a finite dimensional approximation of a PDE is appropriate. For example, near the onset of an instability, there are often only finitely many unstable modes (solutions of a spatial eigenvalue problem). The weakly nonlinear dynamics for parameters just beyond the threshold of instability is often very well approximated by a finite dimensional system.

In 1963, Edward Lorenz studied such a model in a famous paper entitled "Deterministic Nonperiodic Flow" (Lorenz 1963). Lorenz was studying a simple model for the weather—a fluid that moves only in two dimensions and is contained in a rectangular box. It is heated from below; the lower boundary, $z = 0$, has temperature T_o; and it is cooled at the top $z = H$ with temperature $T_o - \Delta T$; see Figure 1.9. When the temperature difference is small, the fluid is motionless and the temperature decreases linearly from the bottom to the top of the box—this is the *conducting* state, $T_c = T_o - \Delta T \frac{z}{H}$. The PDEs that model perturbations from this state are called the

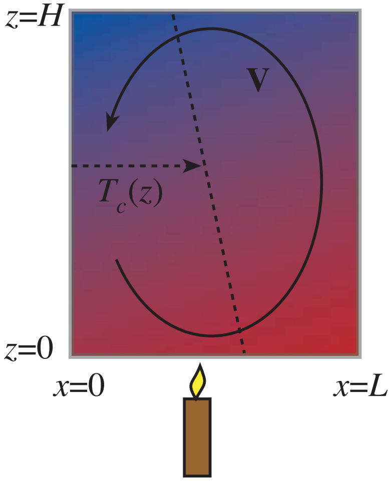

Figure 1.9. *Lorenz fluid model.*

Boussinesq equations:

$$\begin{aligned}\frac{\partial}{\partial t}\nabla^2\psi+(\mathbf{v}\cdot\nabla)\nabla^2\psi&=\nu\nabla^4\psi+g\alpha\frac{\partial\theta}{\partial x},\\ \frac{\partial}{\partial t}\theta+(\mathbf{v}\cdot\nabla)\theta&=\frac{\Delta T}{H}\frac{\partial\psi}{\partial x}+\varkappa\nabla^2\theta.\end{aligned}\tag{1.30}$$

Here the fluid velocity is given by $\mathbf{v}=\nabla\times\hat{y}\psi$, where $\psi(x,z,t)$ is called the stream function—consequently, the velocity is assumed to lie in the vertical xz-plane. The nonlinear terms are represented by the *advective* operator $\mathbf{v}\cdot\nabla$. The perturbation in temperature from the conducting state is represented by $\theta(x,z,t)=T-T_c$. The parameters in the equations are the kinematic viscosity ν, gravitational acceleration g, thermal expansivity α (i.e., the coefficient of thermal expansion), and thermal diffusivity $\varkappa$.

At a critical value of the temperature difference ΔT, the conducting state becomes unstable and the fluid will begin to move. The motion is in the form of a *convection roll*, with hot fluid rising, being cooled, and then falling. Lorenz represented the roll by the spatial forms

$$\begin{aligned}\psi&=A\sin(\pi x/L)\sin(\pi z/H),\\ \theta&=B\cos(\pi x/L)\sin(\pi z/H)-C\sin(2\pi z/H)\end{aligned}\tag{1.31}$$

that depend upon three amplitudes; $A(t)$ represents the fluid velocity, and $B(t)$ and $C(t)$ represent the perturbation of the temperature. The time dependence of these amplitudes can be obtained by substituting the ansatz (1.31) into the Boussinesq equations. This form will not give an exact solution of (1.30) because the advective nonlinear terms will generate spatial structure that is not represented by the assumed three modes. Lorenz applied the idea of *Galerkin truncation* to this system by neglecting all of these additional terms. The result is a system of three ODEs for (A, B, C):

$$\begin{aligned}
\dot{A} &= \frac{\pi g \alpha}{L k^2} B - \nu k^2 A, \\
\dot{B} &= \frac{\pi \Delta T}{LH} A - \frac{\pi^2}{LH} AC - \varkappa k^2 B, \\
\dot{C} &= \frac{\pi^2}{2LH} AB - \frac{4\pi^2 \varkappa}{H^2} C,
\end{aligned} \tag{1.32}$$

where $k^2 = \pi^2(L^{-2} + H^{-2})$ is a squared wavenumber. These ODEs can be scaled to eliminate many of the parameters (see Exercise 9), defining new variables, $x(\tau), y(\tau), z(\tau)$, that are rescaled amplitudes—do not confuse these with the spatial variables of the original PDEs—that depend upon the rescaled time τ to give a simplified set of equations:

$$\begin{aligned}
\dot{x} &= \sigma(y - x), \\
\dot{y} &= r x - x z - y, \\
\dot{z} &= x y - b z.
\end{aligned} \tag{1.33}$$

The new parameters are the Prandtl number $\sigma = \nu/\varkappa$, representing the competition between viscous and thermal diffusions; the Rayleigh number $r = \frac{g\alpha\Delta T}{\varkappa\nu} \frac{\pi^2}{L^2 H k^6}$, representing the applied heat; and a geometric factor $b = \left(\frac{2\pi}{Hk}\right)^2$.

We will return to the Lorenz equations several times in later chapters. In particular the structure of their equilibria and attractors will be investigated in Chapter 4 and their chaotic dynamics in Chapter 7.

1.7 ▪ Quadratic ODEs: The Simplest Chaotic Systems

As we will discuss in Chapter 2, the simplest dynamical systems are linear in their variables; although their dynamics is not all that interesting, linear models do help us understand behavior of more general systems in the neighborhood of equilibria. By contrast, nonlinear ODEs like the Lorenz model and the ABC flow can have amazingly complex solutions. This complexity was discovered by Lorenz in his numerical study of (1.33) and was given the name *chaos* (albeit in a different context) by T.Y. Li and J.A. Yorke in 1975 (Li and Yorke 1975).

Informally, chaos corresponds to aperiodic motion that exhibits “sensitive dependence on initial conditions.” That is, the solutions of two nearby initial states rapidly diverge from one another. Typically the divergence is exponential as in (1.17). A formal definition of chaos will be given in Chapter 7.

As we will see in Chapter 6, chaos cannot occur for one- or two-dimensional ODE systems. Accordingly, three-dimensional systems, like the Lorenz model, are the lowest-dimension, autonomous ODEs that can exhibit chaos. The Lorenz model (1.33) is also remarkable in that the nonlinearity is of particularly simple form—it is contained in just two quadratic terms, xy and xz.

The general quadratic system in three dimensions has 30 terms: each equation can have a constant term, three different linear terms, and six distinct quadratic terms. Clint Sprott set himself the task of finding the simplest such systems as measured by those with the minimal number of terms (Sprott 1994). He looked for chaotic behavior numerically in systems that have a single quadratic term and came up with a list of equations that exhibited chaos for some range of parameter values. A subset of these are listed in Table 1.1 with Sprott’s original labeling.

Table 1.1. *Quadratic, chaotic differential equations.*

Sprott's #	ODE	Reduced Parameters (others set to +1)	Chaotic Parameter Values
B	$\dot{x} = ayz, \quad \dot{y} = bx - cy$ $\dot{z} = d - exy \quad (ae > 0)$	d	$d = 1$
C	$\dot{x} = ayz, \quad \dot{y} = bx - cy$ $\dot{z} = d - ex^2, \quad (abce > 0)$	d	$d = 1$
F	$\dot{x} = ay + bz, \quad \dot{y} = cx + dy$ $\dot{z} = ex^2 - fz$	c, d	$c = -1, d = 0.5$
G	$\dot{x} = ax + bz, \quad \dot{y} = cxz + dy$ $\dot{z} = -ex + fy, \quad (be > 0)$	a, d	$a = 0.4, d = -1$
H	$\dot{x} = ay + bz^2, \quad \dot{y} = cx + dy$ $\dot{z} = ex - fz$	a, d	$a = -1, d = 0.5$
K	$\dot{x} = axy - bz, \quad \dot{y} = cx - dy$ $\dot{z} = ex + fz, \quad (be > 0)$	d, f	$d = 1, f = 0.3$
M	$\dot{x} = -az, \quad \dot{y} = -bx^2 - cy$ $\dot{z} = d + ex + fy$	d, e	$d = e = 1.7$
O	$\dot{x} = ay, \quad \dot{y} = bx - cz$ $\dot{z} = dx + exz + fy$	b, f	$b = 1, f = 2.7$
P	$\dot{x} = ay + bz, \quad \dot{y} = -cx + dy^2$ $\dot{z} = ex + fy, \quad (be > 0)$	a, c	$a = 2.7, c = 1$
Q	$\dot{x} = -az, \quad \dot{y} = bx - cy$ $\dot{z} = dx + ey^2 + fz$	d, f	$d = 3.1, f = 0.5$
S	$\dot{x} = -ax - by, \quad \dot{y} = cx + dz^2$ $\dot{z} = e + fx$	b, e	$b = 4, e = 1$
1	$a\dddot{x} + b\ddot{x} - c\dot{x}^2 + dx = 0$	b	$b = 2.017$
2	$a\dddot{x} + b\dot{x} - cx^2 + d = 0$, $(ab > 0)$	d	$d = 0.025$

These equations have up to six parameters, $a, b, \ldots, f$. However, upon rescaling the variables—like we did for the Lorenz system—the number of relevant parameters can be reduced to one or two; these are called the reduced parameters in the table (see Exercise 10). The values of these reduced parameters for which Sprott observed chaotic behavior are listed in the last column.

You, the reader, are encouraged to adopt one of Sprott's systems as your very own. Throughout this text a number of exercises will refer to this system. You can also apply many of the techniques that will be covered in later chapters to the study of your system. Many of these systems have not been completely analyzed and you may discover new phenomena in your study!

1.8 ▪ Exercises

1. In population dynamics, *depensation* or the *Allee effect* (Allee et al. 1949) corresponds to the reduction in birth rate when a population is small due to the difficulty of finding mates and the harmful effects of inbreeding. A simple model to account for this that generalizes the logistic model (1.7) is

$$\dot{N} = -rN\left(1 - \frac{N}{E}\right)\left(1 - \frac{N}{K}\right),$$

where $0 < E < K$.

(a) Discuss the biological meaning of the variable $N(t)$ and the parameters r, E, and K.

(b) Analyze this system using the methods of §1.3, assuming $r, E, K > 0$.

2. "Habitat conversion from forests to agriculture and then to degraded land is the single biggest factor in the present biological diversity crisis" (Dobson, Bradshaw, and Baker 1997). Let F be the area covered by forest, A the area devoted to agriculture, U the unused land area, and P the human population. A simple model for habitat conversion is

$$\begin{aligned}
\dot{F} &= sU - dPF, \\
\dot{A} &= dPF + bU - aA, \\
\dot{U} &= aA - (b+s)U, \\
\dot{P} &= rP\left(1 - \frac{h}{A}P\right).
\end{aligned} \tag{1.34}$$

(a) Interpret the constants s, d, b, a, and h in the model. In particular, what is the assumed carrying capacity of this environment? What is the interpretation of the nonlinear term dPF? Why is it reasonable to include the area U in the model?

(b) So that this model makes sense, the total land area, T, must be constant. Demonstrate that this is the case for (1.34). Reduce the model to three equations using the fact that T is constant.

(c) Find the equilibrium solution(s) for this model for a given total area T.

3. The Michaelis–Menten mechanism describes the catalysis of a reaction by an enzyme (Michaelis and Menten 1913). The chemical notation for this reaction is

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E + P.$$

Here the enzyme E combines with the substrate S to make an intermediate complex, ES, that is converted into the product P, releasing the enzyme for another reaction. The notation $A \xrightarrow{k} B$ refers to the elementary system $\dot{b} = ka$, $\dot{a} = -ka$, where b and a are the concentrations of species A and B, and k is the rate constant. A binary reaction, such as $A + B \xrightarrow{k} C$, corresponds to the nonlinear system $\dot{c} = -\dot{a} = -\dot{b} = kab$. Note that these elementary reactions have conservation laws that reflect the conversion of one species into another. For example, in the latter case $c(t) + a(t) =$ constant and $c(t) + b(t) =$ constant.

(a) Convert the Michaelis–Menten reaction into a system of four ODEs for the concentrations e, s, c, and p of the enzyme, substrate, complex, and product, respectively. Each arrow in the reaction diagram above refers to an elementary reaction that adds to the rates.

(b) There are two conservation laws for your system. Assuming that the initial product, $p(0)$, and complex, $c(0)$, concentrations are zero, these two laws can be thought of as conservation of enzyme, $e(0) = e_o$, and substrate, $s(0) = s_o$. Use these two laws to eliminate $p(t)$ and $e(t)$ from your four equations, leaving a system of two ODEs.

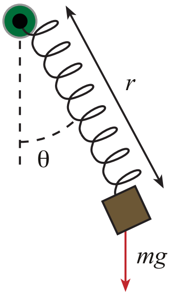

Figure 1.10. *Spring-pendulum of Exercise 5.*

(c) Define new variables $\tau = k_1 e_o t$, $S = s/K_s$, $C = c/e_o$, where $K_s = (k_{-1} + k_2)/k_1$, and rescale the two equations. Show that they can be written

$$\frac{dS}{d\tau} = -S + (1 - \eta + S)C,$$
$$\varepsilon\frac{dC}{d\tau} = S - (1 + S)C$$

with the dimensionless parameters $\varepsilon = e_o/K_s$ and $\eta = k_2/(k_{-1} + k_2)$.

(d) Often the parameter $\varepsilon \ll 1$, which indicates that the complex evolves much more rapidly than the substrate. Consider the limit $\varepsilon = 0$, and reduce your system to a single equation for S. The saturating nonlinearity in this ODE is typical of catalytic reactions.

4. A system of point masses that are coupled by harmonic springs is defined by the equations

$$m_i \ddot{x}_i = -k_i(x_i - x_{i+1}) - k_{i-1}(x_i - x_{i-1}), \quad i = 0, \ldots, n-1,$$

where $x \in \mathbb{R}$, $x_n \equiv x_o$, $x_{-1} \equiv x_{n-1}$, and $k_{-1} \equiv k_{n-1}$.

(a) Describe the physical system that these equations model.

(b) Rewrite the system of n second-order equations as a system of $2n$ first-order equations.

(c) Write the system in (b) as a matrix differential equation (see §2.1).

5. The planar spring-pendulum is modeled by the set of equations

$$\begin{aligned} m\ddot{r} &= mr\dot{\theta}^2 + mg\cos\theta - k(r - L), \\ r^2\ddot{\theta} &= -2r\dot{r}\dot{\theta} - gr\sin\theta. \end{aligned} \tag{1.35}$$

(a) Describe the physical system (e.g., Figure 1.10) that these equations model and explain each term in the equations.

(b) Define the "angular momentum" by $p_\theta = mr^2\dot{\theta}$ and the radial momentum by $p_r = m\dot{r}$. Rewrite the spring-pendulum system as a set of four first-order ODEs for $x = (r, \theta, p_r, p_\theta)$.

(c) Find the equilibrium solution(s), x_{eq}, of the equations, i.e., those solutions for which x is constant.

6. Consider the ABC vector field (1.16).

   (a) Show that (1.16) is incompressible: $\nabla \cdot \mathbf{v} = 0$.

   (b) Show that (1.16) satisfies the Beltrami property: $\mathbf{v} = \nabla \times \mathbf{v}$.

   (c) Show that (1.16) is a solution of the Euler equation

   $$\frac{\partial}{\partial t}\mathbf{v} + \mathbf{v} \cdot \nabla \mathbf{v} = -\nabla P$$

   for some suitable pressure P. The simplest way to do this is to use the vector identity $\mathbf{v} \cdot \nabla \mathbf{v} = {}^1\!/_2 \nabla(\mathbf{v} \cdot \mathbf{v}) - \mathbf{v} \times \nabla \times \mathbf{v}$.

   (d) Show that (1.16) is a solution of the Navier–Stokes equations

   $$\frac{\partial}{\partial t}\mathbf{v} + \mathbf{v} \cdot \nabla \mathbf{v} = \nu \nabla^2 \mathbf{v} + F$$

   for some suitable choice of forcing field F.

7. The Lotka–Volterra system (1.20) has a number of possible phase portraits depending upon parameters. To investigative these it is first convenient to eliminate as many parameters as possible.

   (a) Rescale time and the variables x and y using the scaling transformations

   $$x = \alpha \xi, \; y = \beta \eta, \text{ and } t = \delta \tau$$

   to obtain the differential equations for the new variables $(\xi(\tau), \eta(\tau))$. Show that the parameters (α, β, δ) can be selected to obtain the simplified model

   $$\begin{aligned} \dot{\xi} &= \xi(1 - \xi - C\eta), \\ \dot{\eta} &= D\eta(1 - E\xi - \eta), \end{aligned}$$

   where $C, D, E > 0$.

   (b) Show there are five distinct possibilities for the qualitative dynamics depending upon the values of C and E. Sketch the phase portraits for each case.

   (c) Find the set of initial conditions in each case that are asymptotic to each of the equilibria.

8. The principle of competitive exclusion states that if two species occupy the same ecological niche, then one of them will become extinct. For the Lotka–Volterra model (1.20), being in the same "niche" means that $c/b = f/e$, for this implies that the competitive effect of y on x is relatively the same as that of y on itself. (This is the same as $CE = 1$ for the scaling in Exercise 7.) Prove the exclusion principle for the Lotka–Volterra model in this case (with one exceptional value).

9. Derive the Lorenz model (1.33) from the Boussinesq equations (1.30).

   (a) Substitute (1.31) into (1.30) and collect terms with common spatial dependence. Truncate by neglecting all terms that do not depend upon space in the same way as the terms in (1.31) to obtain the three ODEs (1.32).

(b) Define $x = c_1 A, y = c_2 B$, $z = c_3 C$, and $\tau = c_4 t$ to obtain the differential equations for $x(\tau)$, $y(\tau)$, and $z(\tau)$. Choose the constant scaling factors c_i so that the equations simplify to obtain the Lorenz model (1.33).

10. Adopt one of Sprott's quadratic systems from Table 1.1 as your very own ODE model.[6] This model will be referred to in the exercises in each chapter.

(a) From your variables (x, y, z) and t define a new set of variables (ξ, η, ζ) and τ using a general scaling transformation

$$x = \alpha\xi,\ y = \beta\eta,\ z = \gamma\zeta, \text{ and } t = \delta\tau$$

to find a set of differential equations for $(\xi(\tau), \eta(\tau), \zeta(\tau))$ that have the minimum number of parameters. Note that the chain rule gives $\frac{dx}{dt} = \frac{\alpha}{\delta}\frac{d\xi}{d\tau}$, etc. You will need to solve four nonlinear equations to obtain $(\alpha, \beta, \gamma, \delta)$ in terms of $(a, b, c, \ldots)$ so that all the parameters in your ODEs for (ξ, η, ζ) are 1 except for those listed as "reduced parameters" in the table (keep the same signs in the equations). The nonreduced parameters should be assumed to be nonzero, and in some cases (noted in the table) they may have to be assumed to have a certain sign. Note that the "reduced parameters" will be different from the original ones, e.g., $d \to \hat{d}$.

(b) For your reduced system of ODEs (which you can write as x, y, z again, and drop the "hats" on the reduced parameters) find all the equilibria, i.e., real-valued points (x, y, z) such that $\dot{x} = \dot{y} = \dot{z} = 0$. Is the number of equilibria constant as the (reduced) parameters vary? Do the equilibria ever collide? Discuss.

[6] If your system is a single third-order equation, first rewrite it as a system of three first-order equations.

Chapter 2

Linear Systems

> *...instead of the great number of precepts of which logic is composed, I believed that the four following would prove perfectly sufficient for me ... never to accept anything for true which I did not clearly know to be such ... divide each of the difficulties under examination into as many parts as possible ... conduct my thoughts in such order that, by commencing with objects the simplest and easiest to know, I might ascend by little and little, and, as it were, step by step, to the knowledge of the more complex ... and the last, ... make enumerations so complete, and reviews so general, that I might be assured that nothing was omitted.* (René Descartes, *Discourse on the Method of Rightly Conducting the Reason, and Seeking Truth in the Sciences*, 1637)

In this chapter we will review and extend the standard techniques for solving linear systems of ordinary differential equations (ODEs). Linear differential equations are primarily important because their behavior determines the stability of orbits of more general, nonlinear systems.

Much of the material on linear systems is included in elementary courses on differential equations, so our presentation will be brief. We will, however, introduce some crucial stability concepts that will be useful in more general contexts, and we will pause to consider a couple of more advanced topics, such as the splitting of a matrix into its diagonalizable (semisimple) and nilpotent parts, as well as the treatment of linear, time-periodic systems (Floquet theory). The former will be essential to our study of bifurcations, and the latter to our study of the stability of periodic orbits.

2.1 ▪ Matrix ODEs

The simplest differential equations are linear; they arise as models for systems in which the response is proportional to the input. Such systems include harmonic springs, simple electric circuits, population models, and many others. Formally, a function f is linear when it satisfies the conditions of

- ▷ *linear superposition*: $f(x+y) = f(x) + f(y)$ for each x, y in its domain, and
- ▷ *linear scaling*: $f(cx) = cf(x)$ for each scalar constant c.

Keep in mind that these conditions are typically only satisfied approximately in physical situations: a spring will become nonlinear if it is stretched sufficiently, and the population growth rate will change if competition for resources is important—recall the logistic model (1.7).

The fundamental significance of linearity is that the phase space is naturally a vector space—a set closed under the operations of addition and scalar multiplication. Since the phase space variables are typically physical quantities, they are real variables; thus, it is natural to assume that the phase space is $\mathbb{R}^n$. When f is a vector field on $\mathbb{R}^n$, we say it is a function with domain and range $\mathbb{R}^n$, i.e., $f : \mathbb{R}^n \to \mathbb{R}^n$. For f to be linear, the superposition and scaling conditions imply that its ith component must have the form $f_i(x) = \sum_{j=1}^{n} a_{ij} x_j$ for a set of $n \times n$ constants a_{ij}. In other words, $A = (a_{ij})$ is an $n \times n$ matrix and the vector field is given in matrix notation as[7]

$$f(x) = Ax, \; x \in \mathbb{R}^n.$$

The resulting differential equation is

$$\frac{dx}{dt} = Ax. \tag{2.1}$$

Since x is real-valued, A is assumed to be real as well.

Example 2.1 (Harmonic Oscillators). A spring can be modeled by a linear force law: $F = -k(x - L)$, where L is the equilibrium length of the spring and k is the spring constant. Newton's law for the motion of the spring is $m\ddot{x} = F = -k(x - L)$. This is a second-order ODE, but it is not linear: it is *affine* because of the constant force kL. However, it can be transformed into a linear one by subtracting the equilibrium solution $x^* = L$. Let $\xi = x - x^*$ represent the deviation from equilibrium. Then ξ obeys the equation $\ddot{\xi} = -\frac{k}{m}\xi$. This can be written in the standard form (1.4) as a system of first-order ODEs by letting $\dot{\xi} = \eta$, so that $\dot{\eta} = \ddot{\xi} = -k\xi/m$. In matrix form, this system of two equations becomes

$$\frac{d}{dt}\begin{pmatrix} \xi \\ \eta \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\frac{k}{m} & 0 \end{pmatrix}\begin{pmatrix} \xi \\ \eta \end{pmatrix}. \quad \blacksquare$$

Eigenvalues and Eigenvectors

The standard solution technique for linear ODEs utilizes the eigenvalues and eigenvectors of the matrix A. Recall that an eigenvector, v, is a nonzero solution to the equation

$$Av = \lambda v \tag{2.2}$$

for an eigenvalue λ. This equation has a solution only when the matrix $A - \lambda I$ is singular or equivalently when the characteristic polynomial

$$p(\lambda) \equiv \det(\lambda I - A) = 0. \tag{2.3}$$

Since (2.2) is a homogeneous equation, if v is an eigenvector, then so is any nonzero multiple, cv, for $c \in \mathbb{R} \setminus \{0\}$. As a consequence, one is free to choose the length of the eigenvector to be any convenient, nonzero value.

The characteristic polynomial (2.3) is an nth-order polynomial and so it has n zeros, λ_i. Some of the zeros may be identical, but these should be counted with their

[7] We will not distinguish vectors or matrices with boldface type, preferring to define variables, such as x, to be elements of a particular space, e.g., $x \in \mathbb{R}^n$. This becomes particularly apropos when the phase space is not a vector space, for then vector notation would not be appropriate.

▷ *algebraic multiplicity*: If a polynomial can be written $p(r) = (r - \lambda)^k q(r)$, with $q(\lambda) \neq 0$, then λ is a root with algebraic multiplicity k.

An eigenvalue whose algebraic multiplicity is larger than one is called a *multiple eigenvalue*. The fundamental theorem of algebra states that an nth degree polynomial has exactly n zeros when they are counted with their algebraic multiplicity.

Each eigenvector corresponds to a simple solution of the ODE: assume that $x(t) = c(t)v$ for $c : \mathbb{R} \to \mathbb{R}$ and substitute this into (2.1) to obtain

$$\dot{c}v = cAv = c\lambda v,$$

which when $v \neq 0$ implies that $\dot{c} = \lambda c$ since the eigenvector is constant. The general solution of this scalar ODE is $c(t) = e^{\lambda t} c_o$ for an arbitrary constant c_o. Therefore, the vector

$$x(t) = c_o e^{\lambda t} v \tag{2.4}$$

is a solution to (2.1). Geometrically, (2.4) corresponds to a straight-line solution (when λ is real): $x(t)$ is a vector along v whose length changes exponentially with time.

Example 2.2. Consider the 2×2 system

$$\dot{x} = \begin{pmatrix} -8 & -5 \\ 10 & 7 \end{pmatrix} x. \tag{2.5}$$

The characteristic polynomial is $p(\lambda) = \lambda^2 + \lambda - 6 = (\lambda - 2)(\lambda + 3)$, so there are two eigenvalues, each with algebraic multiplicity one, $\lambda_1 = 2$ and $\lambda_2 = -3$. The eigenvector equations (2.2) are

$$(A - 2I)v_1 = \begin{pmatrix} -10 & -5 \\ 10 & 5 \end{pmatrix} v_1 = 0 \Rightarrow v_1 = \begin{pmatrix} 1 \\ -2 \end{pmatrix},$$

$$(A + 3I)v_2 = \begin{pmatrix} -5 & -5 \\ 10 & 10 \end{pmatrix} v_2 = 0 \Rightarrow v_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

This gives the two solutions

$$x_1 = c_1 e^{2t} \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \quad x_2 = c_2 e^{-3t} \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \quad \blacksquare$$

Because the vector field in the ODE (2.1) is linear, it obeys the linear superposition principle; hence any linear combination of solutions is a solution: indeed, if x_1 and x_2 solve (2.1), then so does $y = c_1 x_1 + c_2 x_2$ for any constants c_1 and c_2 since

$$\dot{y} = c_1 \dot{x}_1 + c_2 \dot{x}_2 = c_1 A x_1 + c_2 A x_2 = A(c_1 x_1 + c_2 x_2) = Ay.$$

This implies that the set of solutions of (2.1) is a vector space: it is closed under the operations of linear superposition and linear scaling. As a consequence, if there are k different eigenvector solutions of the form (2.4), then there is a more general solution of the form

$$x(t) = \sum_{i=1}^{k} c_i e^{\lambda_i t} v_i \tag{2.6}$$

for any values of the constants c_i.

The case that there are n "different" eigenvectors is the optimal one, for then the sum (2.6) has $k = n$ terms with n arbitrary constants c_i. Each eigenvector provides a distinct piece of information if together they span the phase space $\mathbb{R}^n$. The *span* of a set of vectors is the set of points that can be reached by linear combinations of the vectors:

$$\operatorname{span}\{v_1, v_2, \dots, v_n\} \equiv \left\{ w = \sum_{i=1}^{n} c_i v_i : c \in \mathbb{R}^n \right\}. \tag{2.7}$$

If the span of the eigenvectors is $\mathbb{R}^n$, then A is said to have a *complete* set of eigenvectors. An equivalent statement is that the eigenvectors are *linearly independent*,[8] which means that the $n \times n$ matrix whose columns are given by the eigenvectors

$$P = [v_1, v_2, \dots, v_n] \tag{2.8}$$

is nonsingular (i.e., $\det P \neq 0$ so P^{-1} exists).

Example 2.3. For the system (2.5) each eigenvalue has an algebraic multiplicity of one. Moreover, the two eigenvectors are independent since

$$\det[v_1, v_2] = \det\begin{pmatrix} 1 & 1 \\ -2 & -1 \end{pmatrix} = 1.$$

Superposition of the two solutions yields the more general solution

$$x(t) = \begin{pmatrix} e^{2t}c_1 + e^{-3t}c_2 \\ -2e^{2t}c_1 - e^{-3t}c_2 \end{pmatrix} = \begin{pmatrix} e^{2t} & e^{-3t} \\ -2e^{2t} & -e^{-3t} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}. \qquad \blacksquare$$

Matrices that have multiple eigenvalues (algebraic multiplicity larger than one) are unusual in the set of all matrices, but they will arise especially in Chapter 8 for the study of bifurcation theory. Such an eigenvalue may have more than one eigenvector, though it need not. This number is called the

> ▷ *geometric multiplicity*: An eigenvalue λ has geometric multiplicity k if it has k linearly independent eigenvectors v_i, i.e., $(A - \lambda I)v_i = 0$ and $\dim(\operatorname{span}\{v_1, v_2, \dots, v_k\}) = k$.

Recall that the column space or *range* of a matrix is defined to be the span of its column vectors: if $B = [b_1, b_2, \dots, b_n]$ is a matrix with column vectors b_i, then

$$\operatorname{rng}(B) = \operatorname{span}\{b_1, b_2, \dots, b_n\}. \tag{2.9}$$

The *rank* of B is the dimension of its range:

$$\operatorname{rank}(B) \equiv \dim(\operatorname{rng}(B)). \tag{2.10}$$

Accordingly, the geometric multiplicity of λ is $\operatorname{rank}([v_1, v_2, \dots, v_k])$.

Alternatively, since the eigenvector is a solution of a homogeneous equation $(A - \lambda I)v = Bv = 0$, it is appropriate to consider the null space or *kernel*,

$$\ker(B) \equiv \{v \in \mathbb{R}^n : Bv = 0\}. \tag{2.11}$$

[8] Here we are speaking of eigenvectors—not generalized eigenvectors; see §2.6.

Consequently, each eigenvector is an element of $\ker(A-\lambda I)$. The dimension of the kernel is called the *nullity* of a matrix

$$\text{nullity}(B) \equiv \dim(\ker(B)). \tag{2.12}$$

Consequently, the geometric multiplicity of λ is $\text{nullity}(A-\lambda I)$. The fundamental theorem of linear algebra implies that

$$\text{nullity}(B) + \text{rank}(B) = n \tag{2.13}$$

when B has n columns.

Example 2.4. The matrix

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

is upper triangular, so the eigenvalues can be read directly from the diagonal, $\lambda \in \{1,1,2\}$. The algebraic multiplicity of $\lambda = 1$ is two, and the rank of

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

is two since there are two independent vectors in the columns, so its nullity is one. In fact, the complete family of solutions to $(A-I)v = 0$ corresponds to the vectors $v = (c,0,0)^T$ for any $c \in \mathbb{R}$. These vectors all lie along the x-axis and define a one-dimensional space, $\ker(A-I)$; consequently, the geometric multiplicity of $\lambda = 1$ is one. ∎

A basic theorem of linear algebra states that the geometric multiplicity of λ is at most its algebraic multiplicity. As we will see in §2.6, there are n independent eigenvectors when the geometric multiplicity of each eigenvalue λ is equal to its algebraic multiplicity; otherwise there is a *deficiency* of eigenvectors.

Diagonalization

When the matrix of eigenvectors, $P = [v_1, v_2, \ldots, v_n]$, is nonsingular (i.e., when there is no deficiency), the eigenvectors can be used to diagonalize A. To see how this happens, suppose first that we let A act on the $n \times 2$ matrix $[v_1, v_2]$:

$$A[v_1, v_2] = [Av_1, Av_2] = [\lambda_1 v_1, \lambda_2 v_2] = [v_1, v_2]\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}.$$

Note that the last form must be written with the matrix of eigenvalues on the right side of the matrix of eigenvectors. Generalizing this to n eigenvectors gives

$$AP = P\Lambda, \tag{2.14}$$

where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Multiplying by P^{-1} on the left then gives

$$P^{-1}AP = \Lambda. \tag{2.15}$$

In this case, we say that A is *diagonalizable* or *semisimple*. A transformation $A \to P^{-1}AP$ is called a *similarity*. In other words, a semisimple matrix is one that can be diagonalized by a similarity transformation.

When A is diagonalizable, the general solution to the ODE can be obtained by transforming to eigenvector coordinates: let $y = P^{-1}x$; then

$$\frac{dy}{dt} = P^{-1}\frac{dx}{dt} = P^{-1}Ax = P^{-1}APy = \Lambda y.$$

This implies that in the new coordinates, the equations decouple, $\dot{y}_i = \lambda_i y_i$, and have the general solutions $y_i(t) = c_i e^{\lambda_i t}$. In vector notation we can write $y = e^{t\Lambda}c$, where c is the column vector of coefficients c_i, and we define the symbol $e^{t\Lambda}$ as the diagonal matrix[9]

$$e^{t\Lambda} \equiv \mathrm{diag}\left(e^{\lambda_1 t}, e^{\lambda_2 t}, \ldots, e^{\lambda_n t}\right). \tag{2.16}$$

For the moment, this exponential symbol is defined only for this diagonal case (2.16); in §2.3, we will generalize the concept to arbitrary matrices.

Using this notation, a solution to (2.1) is

$$x(t;c) = Py = Pe^{t\Lambda}c.$$

Here, as in §1.2, we emphasize that x depends on the parameters c by adding them to its arguments; thus now $x : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$. To solve the initial value problem, $x(0) = x_o$, note that when $t = 0$ our solution reduces to $x_o = Pc$. This equation is solvable for c since P is nonsingular, so the solution becomes

$$x(t;x_o) = Pe^{t\Lambda}P^{-1}x_o. \tag{2.17}$$

Since this solution is valid for each and every choice of initial condition, x_o, we are justified, according to the definition in §1.2, in calling (2.17) the general solution to (2.1).

Example 2.5 (Symbolic Methods). All computer algebra programs have commands for diagonalizing and exponentiating matrices. Although the reader is encouraged to acquire the skills to manipulate matrices by hand, it quickly becomes tedious to do so when the dimension exceeds three. Even the computation of a 3 × 3 determinant involves so many signs that your author has to do the calculation several times to even hope to get the right answer! Some simple commands to manipulate matrices and compute their exponentials are given in the appendix. ■

A consequence of the eigenvector-eigenvalue analysis is that linear ODEs are essentially trivial[10]—that is, the solution procedure reduces to linear algebra. However, it is worth spending a little more time worrying about two things:

- What if some of the eigenvalues are complex (see §2.5)?
- What if the set of eigenvectors is not complete (see §2.6)?

We first pause, however, to consider the geometry of the phase portraits for two-dimensional systems.

[9] Note that we put the t on the left of Λ as it is a scalar that multiplies every element of Λ.

[10] "Trivial" is a technical term often used simply to indicate one's superiority to one's fellow beings. Use it with care!

2.2 ▪ Two-Dimensional Linear Systems

The properties of the eigenvalues of the arbitrary real, 2 × 2 matrix, $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, can be easily obtained in general. Its eigenvalues are roots of the characteristic polynomial

$$\begin{aligned} p(\lambda) &= \lambda^2 - \tau\lambda + \delta = 0, \\ \tau &\equiv \operatorname{tr}(A) = a + d, \\ \delta &\equiv \det(A) = ad - bc. \end{aligned} \tag{2.18}$$

Thus, the eigenvalues depend only on the values of τ, the *trace* of A, and δ, the *determinant* of A. The roots of p are

$$\lambda_{\pm} = \frac{\tau \pm \sqrt{\Delta}}{2}, \quad \Delta \equiv \tau^2 - 4\delta = (a-d)^2 + 4bc. \tag{2.19}$$

Here, Δ is the *discriminant* of p. There are five different eigenvalue regions in the (τ, δ)-plane, as shown in Figure 2.1. The insets in the figure show complex λ-planes and the dots correspond to the two eigenvalues. The eigenvalues are real when $\Delta > 0$ or equivalently below the parabola $\delta = \tau^2/4$. In the upper half-plane $\Delta < \tau^2$, so that the real part of the eigenvalues has the sign of τ. When $\Delta > 0$ in the first quadrant, both eigenvalues are real and positive, and when $\Delta > 0$ in the second quadrant, both eigenvalues are negative. These cases are called *nodes*. In the lower half-plane $\delta < 0$ so that $\Delta > \tau^2$ and the two eigenvalues have opposite signs; this case is called a *saddle*. Finally, above the parabola $\Delta < 0$ so that the eigenvalues are complex and are conjugates of one another; these cases are called *foci*. The real part of the eigenvalues is positive or negative depending upon the sign of τ; when there is an eigenvalue with positive real part the system is *unstable* (stability is formally defined in §2.7).

Whenever $\Delta \neq 0$, there are two eigenvectors, $v_{\pm}$, corresponding to the two eigenvalues, $\lambda_{\pm}$. Provided that $\lambda \neq a$, Gaussian elimination (elementary row operations) reduces the eigenvector equation to

$$(A - \lambda I)v = \begin{pmatrix} a-\lambda & b \\ c & d-\lambda \end{pmatrix} v \sim \begin{pmatrix} a-\lambda & b \\ 0 & p(\lambda)/(a-\lambda) \end{pmatrix} v = 0,$$

which has rank one when $p(\lambda) = 0$. On the other hand, if $\lambda = a$, but $\lambda \neq d$, then the first row could be eliminated by a similar row operation. Note that the case $\lambda = a = d$ is impossible when $\Delta \neq 0$ since then $\lambda = a \pm \tfrac{1}{2}\sqrt{\Delta}$. Consequently, when $\Delta \neq 0$ each eigenvalue has exactly one eigenvector. Moreover, it is easy to verify that the two eigenvectors are linearly independent.

When $\Delta \neq 0$ the matrix $P = [v_+, v_-]$ is nonsingular, and according to (2.6) and (2.17), the general solution is of the form

$$x(t) = c_+ e^{\lambda_+ t} v_+ + c_- e^{\lambda_- t} v_-. \tag{2.20}$$

The five regions in the (τ, δ) parameter space correspond to geometrically distinct types of motion—to distinct "phase portraits." These can be easily understood using (2.20).

(A) *Unstable node*: $\lambda_+ > \lambda_- > 0$. In this case there are special "straight-line" solutions corresponding to $c_+ = 0$ or to $c_- = 0$. For these cases $x(t)$ grows exponentially with t along the ray through the origin defined by the respective eigenvector. The sign of the nonzero c determines whether the solution moves in the direction of v or —v. Since there are unbounded solutions, this case is called *unstable*, as we will discuss

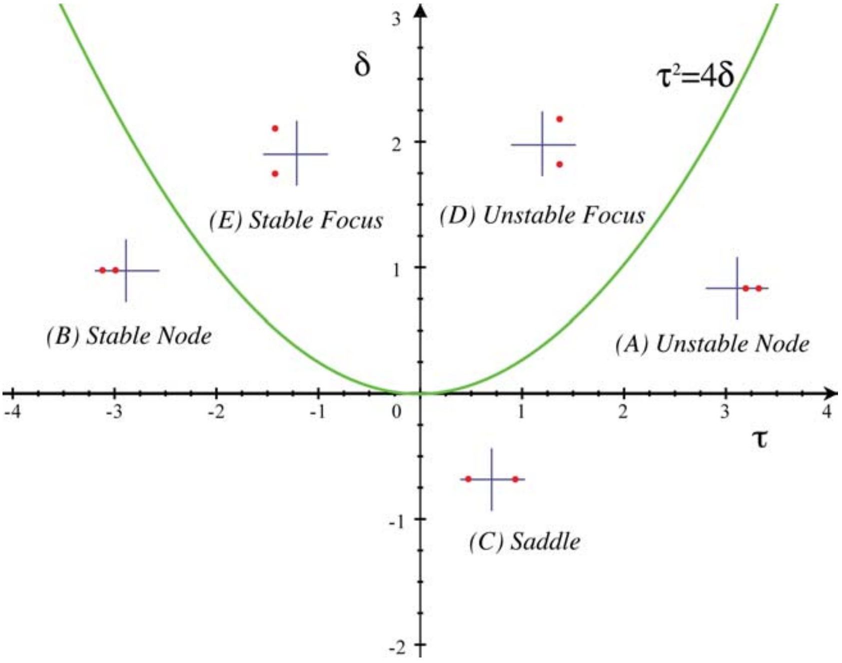

Figure 2.1. *Classification of the eigenvalues for a* 2×2 *linear system in the parameter space of the trace,* τ*, and determinant,* δ.

further in §2.7. When both $c_\pm \neq 0$ the solution is a curve, as shown in Figure 2.2. As t increases the λ_+ solution dominates, so the curves are asymptotically parallel to v_+. By contrast, when $t \to -\infty$, both terms approach zero, but $e^{\lambda_+ t} \to 0$ much more quickly than $e^{\lambda_- t}$, so the solution curve approaches a ray defined by the vector v_-.

(B) *Stable node*: $\lambda_- < \lambda_+ < 0$. The geometry is essentially the same as the previous case, but the arrows will be reversed since the solutions asymptotically approach the origin as $t \to \infty$. Since every solution is bounded as $t \to \infty$, this case is called *stable*. When t is large, the exponential $e^{\lambda_- t}$ is much smaller than $e^{\lambda_+ t}$, so the solution curves are asymptotic to v_+ near the origin, opposite to the previous case. Furthermore, as $t \to -\infty$ the "$-$" exponent dominates, and the curves are asymptotic to v_-. Thus, the phase portrait is just like that in Figure 2.2 with both the arrows and the vector labels $v_\pm$ reversed.

(C) *Saddle*: $\lambda_- < 0 < \lambda_+$. The straight-line solution $c_+ e^{\lambda_+ t} v_+$ moves away from the origin with increasing t, and the solution $c_- e^{\lambda_- t} v_-$ asymptotically approaches the origin. Because there are unbounded solutions, this case is called unstable—even though the special "$-$" direction corresponds to solutions that approach the origin. More general solutions are asymptotic to the v_+ solution as t increases and to v_- as t decreases, as shown in Figure 2.3. If $\lambda_- = -\lambda_+$, the solution curves are hyperbolas with $v_\pm$ as asymptotes, so this case is sometimes called "hyperbolic;"[11] more generally, the solution curves are qualitatively similar to hyperbolas.

(D) *Unstable focus*: $\lambda_\pm = \alpha \pm i\beta$, $\alpha > 0$, $\beta > 0$. Since we assume that the matrix A is real, whenever the eigenvalues are complex they are complex conjugates. This also follows explicitly from the formula (2.19) since $\alpha = {}^1\!/_2\tau$ and $\beta = {}^1\!/_2\sqrt{|\Delta|}$. It

[11] We reserve the term *hyperbolic* for the more general situation that $\mathrm{Re}(\lambda_i) \neq 0$; see §2.7.

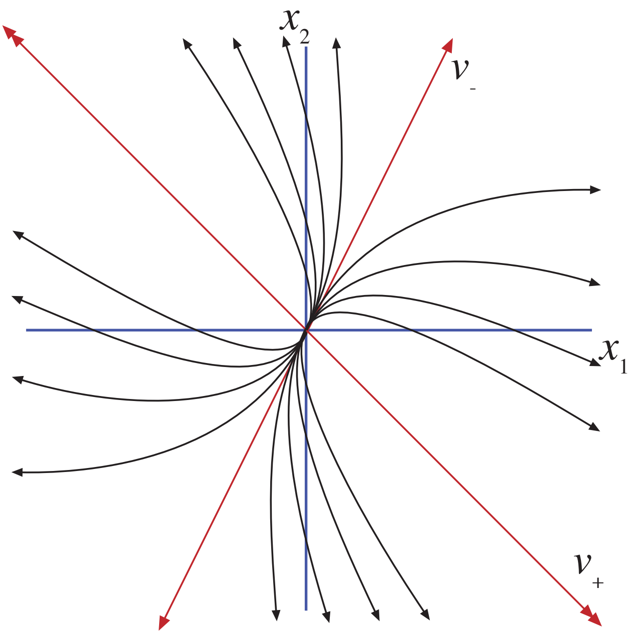

Figure 2.2. *Phase portrait of an unstable node with $v_+ = (1,-1)^T$, and $v_- = (1,2)^T$ and $\lambda_+ = 2\lambda_-$. The arrows denote the direction of motion.*

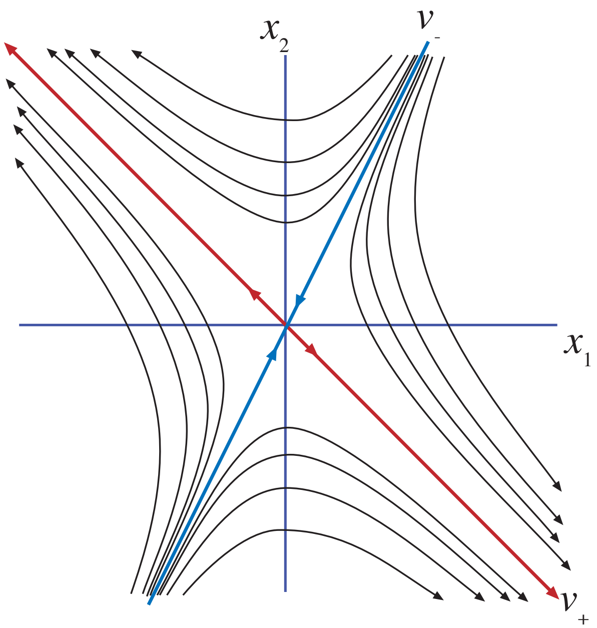

Figure 2.3. *Phase portrait of a saddle with $v_+ = (1,-1)^T$, and $v_- = (1,2)^T$ and $\lambda_- = -2\lambda_+$. The arrows denote the direction of motion.*

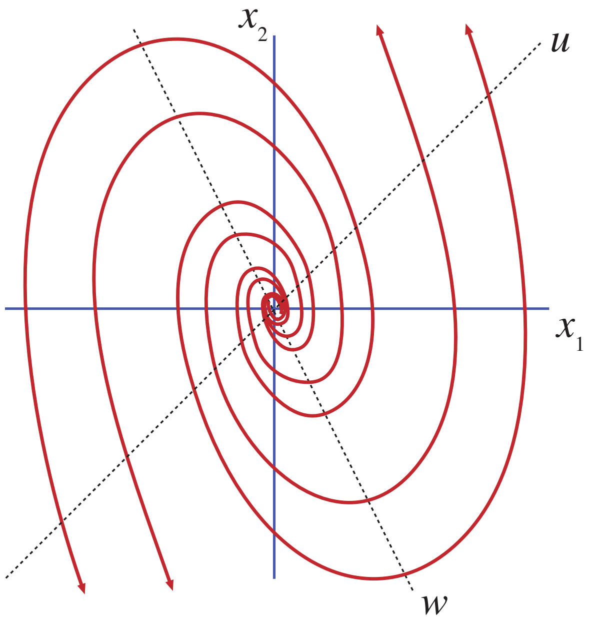

Figure 2.4. *Phase portrait of an unstable focus with* $u = (1,1)^T$, *and* $w = (1,-2)^T$ *and* $\alpha = 0.3\beta$. *Here the motion is counterclockwise since* $\det[u,w] < 0$.

also follows from (2.2) in this case that the eigenvectors are conjugates, $v_\pm = u \pm iw$. Finally, so that the solution is real, we must also assume that $c_\pm = ½(g \pm ih)$ as well. In this case, simple algebra using Euler's formula,

$$e^{(\alpha+i\beta)t} = e^{\alpha t}(\cos\beta t + i\sin\beta t), \tag{2.21}$$

can be used to rewrite (2.20) in the explicitly real form

$$\begin{aligned} x(t) &= \tfrac{1}{2}(g+ih)e^{\alpha t}(\cos\beta t + i\sin\beta t)(u+iw) + \tfrac{1}{2}(g-ih)e^{\alpha t}(\cos\beta t - i\sin\beta t)(u-iw), \\ &= e^{\alpha t}(g\cos\beta t - h\sin\beta t)u - e^{\alpha t}(g\sin\beta t + h\cos\beta t)w. \end{aligned}$$

To unpack this solution, note that it can be written as the product of several terms. Letting $P = [u,w]$ be a matrix with columns u and w, we have

$$x(t) = e^{\alpha t}P\begin{pmatrix} \cos\beta t & \sin\beta t \\ -\sin\beta t & \cos\beta t \end{pmatrix}\begin{pmatrix} g \\ -h \end{pmatrix}. \tag{2.22}$$

The motion consists of a clockwise rotation by angle βt applied to the vector $(g,-h)^T$, generating a circle. This is followed by the application of the matrix P, transforming the circle to an ellipse. Finally, the first coefficient corresponds to an exponentially growing amplitude. Consequently, the motion is an expanding elliptical spiral, as shown in Figure 2.4. Since multiplication by the matrix P preserves orientation, when $\det(P) > 0$ the motion is clockwise; otherwise, it is counterclockwise.

(E) *Stable focus*: $\lambda_\pm = \alpha \pm i\beta$, $\alpha < 0$, $\beta > 0$. Here the motion is still governed by (2.22); however, in this case the orbits spiral inward, approaching the origin as $t \to \infty$.

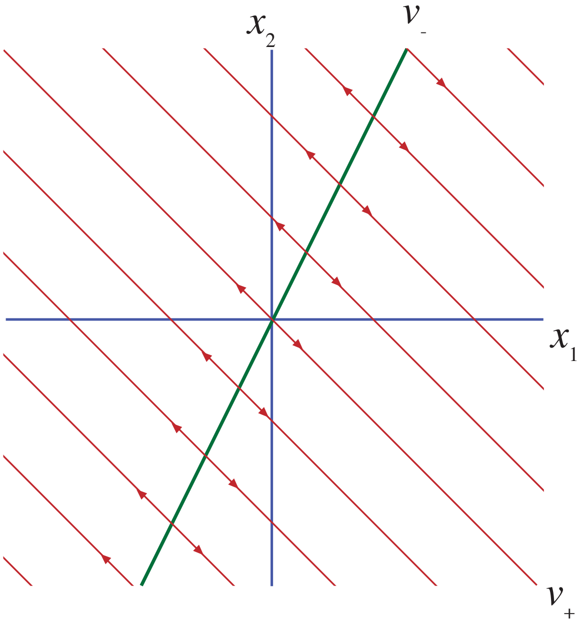

Figure 2.5. *Unstable line of equilibria for* $\lambda_- = 0$, $\lambda_+ > 0$.

The boundaries between the five regions in Figure 2.1 correspond to special categories. If $\delta = 0$, then one of the eigenvalues is zero. In this case we say that the equilibrium is *degenerate* or *nonisolated*. Indeed, the straight-line solution corresponding to the zero eigenvalue is an equilibrium for any value of the constant c. A third boundary case corresponds to complex eigenvalues, but $\tau = 0$.

(a) *Unstable degenerate equilibrium*: $\lambda_- = 0$, $\lambda_+ > 0$. This corresponds to the positive τ-axis in Figure 2.1, i.e., the boundary between the unstable node and saddle regions. The set of solutions $x(t) = c_- v_-$ is a line of equilibria. Solutions that begin off this line (with $c_+ \neq 0$) move to infinity along straight-line trajectories parallel to v_+; see Figure 2.5.

(b) *Stable degenerate equilibria*: $\lambda_+ = 0$, $\lambda_- < 0$. This corresponds to the negative τ-axis in Figure 2.1, i.e., the boundary between the stable node and saddle regions. The line of equilibria $x(t) = c_+ v_+$ exponentially attracts all other solutions along lines parallel to v_-.

(c) *Center*: $\lambda_\pm = \pm i\beta$. This case corresponds to the positive δ-axis in Figure 2.1. Here (2.22) still applies, but since $\alpha = 0$, the motion is confined to ellipses.

The eigenvalues have algebraic multiplicity two when $\Delta = 0$, corresponding to the parabola in Figure 2.1. There are two eigenvectors when the nullity of $A - \lambda I$ is two; this can happen only when $A - \lambda I = 0$, so that A is diagonal and a multiple of the identity. More generally, nullity$(A - \lambda I) = 1$. Thus, when the eigenvalues are equal, it is typical that the geometric multiplicity is smaller than the algebraic multiplicity. In this case there is only one eigenvector and it provides a solution of the form $x(t) = ce^{\lambda t} v$ that involves only a single arbitrary constant. Therefore, this solution cannot be the general solution; this case will be treated in the coming sections.

2.3 ▪ Exponentials of Operators

As we saw in (2.16) and (2.17), when A is semisimple the solution to the matrix ODE $\dot{x} = Ax$ involves the exponential of Λ, the diagonal matrix of the eigenvalues of A. In the next few sections, we will develop methods to compute e^{tA} the for general matrices and show how the exponential provides the general solution to (2.1). To begin, we discuss the more general case of the exponential of a linear operator.

An operator T on a vector space E maps a vector $v \in E$ into another vector $w = T(v) \in E$. An operator is linear if it satisfies the superposition and scaling properties of §2.1. If E has dimension n, and the vectors $\{e_1, e_2, \ldots, e_n\}$ are a basis for E, then a linear operator on E can be represented by matrix A, by setting $T(e_j) = \sum_{i=1}^{n} a_{ij} e_i$ so that the theory of linear operators reduces to matrix algebra.[12] However, the more general notation is useful since the results in this section apply more generally to operators on complete, infinite-dimensional vector spaces.

Suppose $x \in E$ is a vector and that there is some notion of length or *norm* of x, denoted by $|x|$. In this book, E will usually be Euclidean space and the norm will be the ordinary Euclidean length. Given such a notion of length, a norm for a linear operator T on E can also be defined as:

$$||T|| = \sup_{|x|>0} \frac{|T(x)|}{|x|} = \sup_{|x|=1} |T(x)|, \tag{2.23}$$

where sup denotes the supremum, which is the least upper bound. The $||\cdot||$ notation is used to distinguish this operator norm from the vector space norm $|\cdot|$. An operator for which $||T|| < \infty$ is *bounded*.

Example 2.6. Suppose $T(x) = Ax = \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix} x$. Let $x = \begin{pmatrix} a \\ b \end{pmatrix}$ and use the Euclidean norm so that $|x| = \sqrt{a^2 + b^2}$. According to (2.23), $||T||$ can be computed by maximizing the function $f(a,b) = |T(x)|^2 = (2a+b)^2 + b^2$, subject to the constraint $|x| = 1$. One way to do this is to use Lagrange multipliers: find the extrema of the function $F = f - \lambda\left(|x|^2 - 1\right)$. To do this, differentiate with respect to a and b to obtain

$$\frac{\partial F}{\partial a} = 4(2a+b) - 2\lambda a = 0,$$

$$\frac{\partial F}{\partial b} = 2(2a+b) + 2b - 2\lambda b = 0.$$

This can be written as a homogeneous linear system

$$\begin{pmatrix} 8-2\lambda & 4 \\ 4 & 4-2\lambda \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{2.24}$$

which has a nonzero solution only when its determinant vanishes. This implies that $\lambda = 3 \pm \sqrt{5}$. Solving the system above and normalizing the solution gives

$$(a,b) = \frac{1}{\sqrt{10 \mp 2\sqrt{5}}} \left(-2, 1 \mp \sqrt{5}\right) \Rightarrow |T(x)|^2 = 3 \pm \sqrt{5}.$$

[12] Note that the action of T on the jth basis vector has a_{ij} as its ith component. Using a_{ij} instead of a_{ji} is natural since then the action of T on a general vector $v = \sum_{j=1}^{n} c_j e_j$ is $T(v) = \sum_{i=1}^{n} e_i \sum_{j=1}^{n} a_{ij} c_j$. Thus T is represented by the matrix $A = (a_{ij})$ acting on the vector c of components of v.

The larger value corresponds to the plus sign, which yields $||T|| = \sqrt{3+\sqrt{5}}$.

The value of the operator norm does depend upon the norm used for the vector space. In the current example, the sup-norm for x, $|x|_\infty \equiv \max_i |x_i|$ would give $||T||_\infty = 3$. In the next chapter the sup-norm will be used often, as it simplifies some of the analysis. ■

A more instructive way to compute the operator norm for the finite dimensional case is to use the equivalence between linear operators and matrices; then (in the Euclidean norm)

$$|T(x)|^2 = x^T A^T A x = x^T S x$$

is a quadratic form in the symmetric matrix $S = A^T A$ formed from the product of the transpose of A with A. As is well known, any symmetric matrix can be diagonalized by an orthogonal transformation, i.e., there is a matrix O such that $O^{-1} = O^T$ and $S = O^T \Lambda O$ with Λ diagonal. The eigenvalues of S are the elements of $\Lambda = \text{diag}(r_1^2, r_2^2, \ldots, r_n^2)$; the r_i^2 are all nonnegative since S is positive semidefinite ($x^T S x \geq 0$). The nonnegative square roots of these elements, $r_i \geq 0$, are called the *singular values* of the original matrix T. The transformation $x = O^T y$ can be used to simplify the expression for $|T(x)|^2$:

$$|T(x)|^2 = (O^T y)^T S (O^T y) = y^T O (O^T \Lambda O) O^T y = y^T \Lambda y = \sum_{i=1}^{n} r_i^2 y_i^2.$$

Under the constraint that $|x|^2 = |y|^2 = 1$, the largest value this can take is the maximum of the squared singular values. Consequently,

$$||T|| = \max_{i=1,\ldots,n} r_i.$$

Thus every $n \times n$ matrix corresponds to a bounded linear operator.

Example 2.7 (continued). Note that for the example above, $A^T A = S = \begin{pmatrix} 4 & 2 \\ 2 & 2 \end{pmatrix}$, which is not accidentally $1/2$ of the matrix involved in the linear system (2.24). Consequently the squared singular values are $r_\pm^2 = 3 \pm \sqrt{5}$, so that $||T|| = \sqrt{3+\sqrt{5}}$, as before. ■

The exponential of an operator T is formally defined by the power series

$$e^T \equiv \sum_{k=0}^{\infty} \frac{T^k}{k!}. \tag{2.25}$$

If this series converges, it defines a linear operator as well. It is not hard to show that convergence follows whenever T is bounded.

Lemma 2.8. *If T is a bounded linear operator, then e^T is as well.*

Proof. Choose an arbitrary $x \in E$ and consider the value of $e^T(x)$. By definition (2.25) this is a series whose terms are elements of E. The norm of this series is bounded by the sum of the norms of each term. By the definition of the operator norm, for any x,

$$|T(x)| \leq ||T|| \; |x|,$$

$$\left|T^k(x)\right| = \left|T\left(T^{k-1}(x)\right)\right| \leq ||T|| \left|T^{k-1}(x)\right| \leq \cdots \leq ||T||^k \, |x|.$$

Consequently, each of the terms in the series $e^T(x)$ can be bounded by

$$\left|\frac{T^k(x)}{k!}\right| \le \frac{||T||^k}{k!}|x| = M_k.$$

The series of real numbers

$$\sum_{k=0}^{\infty} M_k = \sum_{k=0}^{\infty} \frac{||T||^k}{k!}|x| = e^{||T||}|x|$$

converges for any finite value of $||T||$. By the Weierstrass M-test $\left|e^T(x)\right| \le e^{||T||}|x|$ and the series for $e^T(x)$ converges uniformly in x. Moreover $\left\|e^T\right\| \le e^{||T||}$, so the exponential is a bounded operator. □

Since, as we have seen, the norm of an $n \times n$ matrix is its maximum singular value, then every matrix corresponds to a bounded linear operator. Thus Lemma 2.8 implies the following.

Corollary 2.9. *The exponential of every linear operator on $\mathbb{R}^n$ is a bounded linear operator.*

The following properties of the exponential operator $e^T : E \to E$ are easily verified from the definition (2.25):

(i) $e^0 = I$.
(ii) $\left(e^T\right)^{-1} = e^{-T}$ (term-by-term multiplication of the series for $e^T e^{-T}$).
(iii) If A and B are commuting linear operators, i.e., $AB - BA = 0$, then $e^{A+B} = e^A e^B$ (see Exercise 6).
(iv) If B is nonsingular, then $e^{BAB^{-1}} = Be^A B^{-1}$ (factors of $B^{-1}B$ cancel in each term of the sum (2.25)).
(v) If $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, then $e^\Lambda = \operatorname{diag}(e^{\lambda_1}, e^{\lambda_2}, \ldots, e^{\lambda_n})$ (since $\Lambda^k = \operatorname{diag}(\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k)$).
(vi) If v is an eigenvector of T with eigenvalue λ, then $e^T v = e^\lambda v$.

As noted above, for every linear operator on $\mathbb{R}^n$, there is an associated matrix. Since any matrix A commutes with itself, as well with any multiple of itself, rule (iii) implies that if $x(\tau) = e^{\tau A}x_o$, then

$$e^{tA}x(\tau) = e^{tA}e^{\tau A}x_o = e^{(t+\tau)A}x_o = x(t+\tau),$$

so that flowing forward for a time τ and then a time t is equivalent to flowing forward for a time $t + \tau$. This property will be shown to hold more generally for autonomous ODEs in Chapter 4. Rule (iii) is not true when the matrices A and B do not commute.

Example 2.10 (Baker–Campbell–Hausdorff Theorem). Suppose we attempt to define a matrix C by

$$e^C = e^A e^B.$$

If the commutator

$$[A,B] = AB - BA \tag{2.26}$$

is zero, then $C=A+B$ by property (iii). The remarkable Baker–Campbell–Hausdorff theorem implies more generally that if the norms of A and B are small enough, then C exists and can be computed in terms of commutators of A and B. To compute the first few terms in this expression, expand the exponentials in a power series:

$$\begin{aligned} e^A e^B &= \left(I+A+\frac{1}{2}A^2+\frac{1}{6}A^3+\cdots\right)\left(I+B+\frac{1}{2}B^2+\frac{1}{6}B^3+\cdots\right) \\ &= I+A+B+\frac{1}{2}\left(A^2+2AB+B^2\right)+\frac{1}{6}\left(A^3+3AB^2+3A^2B+B^3\right)+\cdots . \end{aligned} \tag{2.27}$$

The matrix C can be computed term by term by its exponential expansion $e^C=I+C+\frac{1}{2}C^2+\cdots$. If we set set $C=A+B+D+E+\cdots$, then both series have the lowest-order term I and linear terms $A+B$. To construct the next few terms of C, assume that D is quadratic and E is cubic in the matrices A and B; then

$$\begin{aligned} e^C &= I+A+B+\left[D+\frac{1}{2}(A+B)^2\right] \\ &\quad +\left[E+\frac{1}{2}((A+B)D+D(A+B))+\frac{1}{6}(A+B)^3\right]+\cdots . \end{aligned} \tag{2.28}$$

Comparing the quadratic terms in (2.27) and (2.28) gives

$$D=\frac{1}{2}\left(A^2+2AB+B^2\right)-\frac{1}{2}(A+B)^2=\frac{1}{2}[A,B].$$

The cubic terms become

$$\begin{aligned} E &= \frac{1}{6}\left(A^3+3AB^2+3A^2B+B^3\right)-\frac{1}{2}(A+B)D-\frac{1}{2}D(A+B)-\frac{1}{6}(A+B)^3 \\ &= \frac{1}{12}\left(AB^2-2BAB+B^2A+A^2B-2ABA+BA^2\right) \\ &= \frac{1}{12}[A,[A,B]]-\frac{1}{12}[B,[A,B]]. \end{aligned}$$

Thus we see that (at least through the first few terms) apart from the linear terms, the matrix C can be expressed solely in terms of commutators of the matrices A and B:

$$C=A+B+\frac{1}{2}[A,B]+\frac{1}{12}[A,[A,B]]-\frac{1}{12}[B,[A,B]]+\cdots .$$

An explicit although rather complicated formula for the coefficients of C was first obtained by the Russian mathematician Eugene Dynkin in 1947 (Hall 2003).

This theorem will be especially important to us for the study of nonautonomous linear systems in §2.8. ■

In some special cases, the definition (2.25) and the rules (i)-(vi) can be used to directly compute the exponential.

Example 2.11 (Nilpotent Matrices). If N is a nilpotent matrix, i.e., if there is a $k\geq 0$ such that $N^k=0$, then the exponential series terminates after a finite number of terms. This property allows a simple computation of the exponential for some operators. For example, consider

$$A=\begin{pmatrix} a & b \\ 0 & a \end{pmatrix}=\begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}+\begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix}=S+N.$$

Now $[S,N]=\left(\begin{smallmatrix} 0 & ab \\ 0 & 0 \end{smallmatrix}\right)-\left(\begin{smallmatrix} 0 & ba \\ 0 & 0 \end{smallmatrix}\right)=0$, so that by property (iii) $e^A=e^Se^N$; moreover, since $N^2=0$, and S is diagonal,

$$e^A=\begin{pmatrix} e^a & 0 \\ 0 & e^a \end{pmatrix}(I+N)=e^a\begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix}. \quad ■$$

Example 2.12 (Roots of the Identity). If the matrix A is a root of the identity, then the series separates into a set of simple subseries. For example, the matrix

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

has powers $\sigma^2 = -I$, $\sigma^3 = -\sigma$, and $\sigma^4 = I$, so its exponential is given by

$$e^{t\sigma} = I \sum_{m=0}^{\infty} \frac{(-1)^m}{(2m)!} t^{2m} + \sigma \sum_{m=0}^{\infty} \frac{(-1)^m}{(2m+1)!} t^{2m+1} = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} \tag{2.29}$$

since the two power series in t define the cosine and sine functions, respectively. ■

Example 2.13. If a matrix can be written as a sum of two commuting matrices whose exponentials can be computed, then exponentiation is easy. For example,

$$B = \begin{pmatrix} a & b \\ -b & a \end{pmatrix} = aI + b\sigma. \tag{2.30}$$

Note by (2.26) that $[aI, b\sigma] = 0$; indeed, any matrix commutes with a multiple of the identity. Consequently (2.29) gives

$$e^{tB} = e^{taI} e^{tb\sigma} = e^{at} \begin{pmatrix} \cos(bt) & \sin(bt) \\ -\sin(bt) & \cos(bt) \end{pmatrix}. \quad ■ \tag{2.31}$$

2.4 ▪ Fundamental Solution Theorem

Theorem 2.14. *Let A be an $n \times n$ matrix. Then the initial value problem*

$$\dot{x} = Ax, \quad x(0) = x_o, \tag{2.32}$$

has the unique solution $x(t) = e^{tA} x_o$.

Proof. We first demonstrate that the proposed solution works. To compute the derivative, use its basic definition as a limit with the series (2.25) and the fact that a matrix commutes with a multiple of itself to obtain

$$\frac{d}{dt} e^{tA} = \lim_{h \to 0} \frac{e^{(t+h)A} - e^{tA}}{h} = \lim_{h \to 0} \frac{e^{hA} - I}{h} e^{tA} = \lim_{h \to 0} \left(\frac{1}{h} \sum_{n=1}^{\infty} \frac{(hA)^n}{n!} \right) e^{tA}$$

$$= \lim_{h \to 0} \left(\frac{hA}{h} + \sum_{n=2}^{\infty} h^{n-1} \frac{A^n}{n!} \right) e^{tA} = \lim_{h \to 0} \left(A + h \sum_{j=0}^{\infty} h^j \frac{A^{j+2}}{(j+2)!} \right) e^{tA} = Ae^{tA}.$$

Here, the last equality holds because the series in the last expression converges to a linear operator, $T_{A,h}$, and $hT_{A,h} \to 0$ as $h \to 0$. Since $\frac{d}{dt}\left(e^{tA} x_o\right) = Ae^{tA} x_o = Ax$, it is certainly a solution.

To show that the solution is unique, suppose that $y(t)$ is another solution. Then differentiation and the product rule imply

$$\frac{d}{dt}\left(e^{-tA} y(t)\right) = -Ae^{-tA} y(t) + e^{-tA} A y(t) = \left[-Ae^{-tA} + e^{-tA} A\right] y(t) = 0.$$

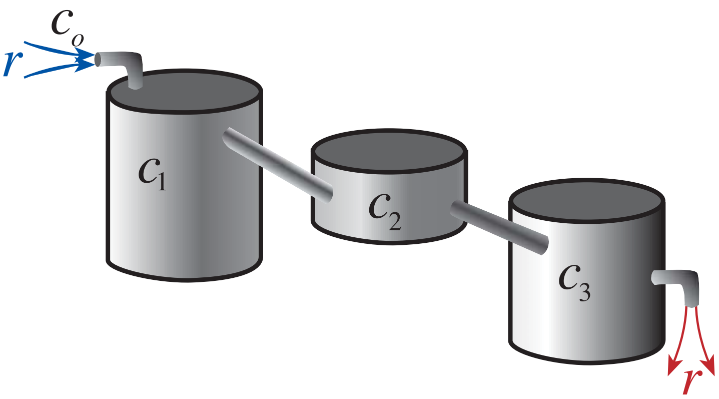

Figure 2.6. *Flow compartment model.*

The term in brackets is zero because the matrices A and e^{-tA} commute. Therefore, $e^{-tA}y(t) = y_o$ a constant, so $y(t) = e^{tA}y_o$ by property (ii). Moreover, $y_o = x_o$, since this solution must satisfy the initial value problem. □

We have reduced the problem of solving linear systems to that of finding the exponential of the matrix A. If A has a complete set of eigenvectors, the exponential is easily obtained by diagonalization.

Example 2.15 (Compartmental Mixing). A chemical mixer consists of three tanks sequentially connected by pipes; see Figure 2.6. A solution of salt with concentration c_o kg/liter flows into the first tank at a flow rate of r liters/sec. The fluid is well mixed in this tank—by an impeller—to have uniform concentration $c_1(t)$; it flows out to the second tank at the same flow rate, r. This continues in the second tank with concentration c_2 flowing to the third with concentration c_3. Finally, the fluid leaves the third tank at the same flow rate r. Since the flow rates are equal, the total volume of fluid in each tank is constant in time; call these values V_i. Each tank begins at $t = 0$ with zero salt concentration.

The ODE model that governs the concentrations is constructed by computing the rate of mass flow into and out of each tank. For example, a mass of rc_o kg/sec flows into the first tank and rc_1 flows out. The complete model is the system

$$\frac{d}{dt}V_1c_1 = r(c_o - c_1), \frac{d}{dt}V_2c_2 = r(c_1 - c_2), \frac{d}{dt}V_3c_3 = r(c_2 - c_3).$$

This system is affine because of the constant term rc_o in the first equation. As usual, we can eliminate this by subtracting the equilibrium solution $c_1^* = c_2^* = c_3^* = c_o$, i.e., defining new dynamical variables $x_i = c_i - c_o$; then we obtain a linear initial value problem of the form (2.32) with initial condition $x(0) = (-c_o, -c_o - c_o)^T$ and matrix

$$A = \begin{pmatrix} -\alpha & 0 & 0 \\ \beta & -\beta & 0 \\ 0 & \gamma & -\gamma \end{pmatrix},$$

where $\alpha = r/V_1$, $\beta = r/V_2$, and $\gamma = r/V_3$. This matrix has eigenvalues $-\alpha$, $-\beta$, and $-\gamma$; note that this implies that the equilibrium is a stable node—all solutions limit to this constant solution as $t \to \infty$. For numerical simplicity, let us assume that the fluid volumes are $V_1 = 1$, $V_2 = 1/3$, and $V_3 = 1/2$ liters, so that the eigenvalues are $\lambda = -r$, $-3r$, and $-2r$. To compute the exponential of A we use properties (iv) and (v) of the exponential. A short calculation gives the matrix of eigenvectors

$$P = \begin{pmatrix} 2 & 0 & 0 \\ 3 & 1 & 0 \\ 6 & -2 & 1 \end{pmatrix}.$$

The matrix exponential $e^{tA} = Pe^{t\Lambda}P^{-1}$ is then

$$e^{tA} = \begin{pmatrix} e^{-rt} & 0 & 0 \\ \frac{3}{2}\left(e^{-rt} - e^{-3rt}\right) & e^{-3rt} & 0 \\ 3\left(e^{-rt} - 2e^{-2rt} + e^{-3rt}\right) & 2\left(e^{-2rt} - e^{-3rt}\right) & e^{-2rt} \end{pmatrix}.$$

As a check, note that when $t = 0$ this reduces to the identity. Finally, multiplying by the initial vector $x(0)$ and adding back the equilibrium gives the solution $c(t) = x(t) + c_o$,

$$c(t) = c_o \begin{pmatrix} 1 - e^{-rt} \\ 1 - \frac{1}{2}\left(3e^{-rt} - e^{-3rt}\right) \\ 1 - 3e^{-rt} + 3e^{-2rt} - e^{-3rt} \end{pmatrix}.$$

Consequently, the solution approaches the equilibrium; for large t the deviation from the equilibrium state is along the slowest decaying eigenvector and is approximately $-{}^1\!/\!_2 v_1 e^{-rt}$. ∎

We have shown that the vector $e^{tA}x_o$ is the unique solution to (2.32). Now consider a set of initial conditions $x_{jo} = v_j$, $j = 1, 2, \ldots, n$, with arbitrary initial vectors v_j. Since the corresponding solutions are the vectors $x_j(t) = e^{tA}v_j$, we can put the initial conditions into a matrix $\Psi_o = [v_1, v_2, \ldots, v_n]$ and the solutions into a matrix $\Psi(t) = [x_1(t), x_2(t), \ldots, x_n(t)]$ to demonstrate that $\Psi(t)$ is the solution of a matrix differential equation:

Theorem 2.16. *The matrix initial value problem*

$$\frac{d}{dt}\Psi = A\Psi, \quad \Psi(0) = \Psi_o, \tag{2.33}$$

has the unique solution $\Psi(t) = e^{tA}\Psi_o$.

In particular, when $\Psi_o = I$, the solution to (2.33) is $\Phi(t) = e^{tA}$; this is called the *(principal) fundamental matrix solution*. We will return to it in §2.8.

We now return to the problem of how to compute the exponential of a matrix for the general case. As we will see in §2.6, when a matrix is deficient, it can be written as the sum of a *semisimple* matrix and a *nilpotent* matrix that commute. This will make the computation of the exponential possible in general. First, however, we pause to consider the complex case.

2.5 ▪ Complex Eigenvalues

We saw in §2.2 that when the eigenvalues are complex it is possible to use complex eigenvectors and Euler's formula (2.21) for the complex exponential to compute the solutions. However, if the dynamical system is real, then values of e^{tA} must be real as well, and it seems strange to have complex values for the intermediate results. As we will see, this can be avoided.

First, note that if the matrix A is real, then so are the coefficients of the characteristic polynomial $p(\lambda) = \det(\lambda I - A)$. Therefore, if $p(\lambda)$ has a complex root $\lambda = a + ib$, then its conjugate $\bar{\lambda} = a - ib$ is also a root. Moreover, if $Av = \lambda v$, then $A\bar{v} = \overline{\lambda v} = \bar{\lambda}\bar{v}$. Therefore, the corresponding eigenvectors are also complex conjugates.

Example 2.17. For the matrix $A = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, the eigenvalues are $\lambda = \pm i$ and the eigenvectors are $v = \left(\begin{smallmatrix} 1 \\ \pm i \end{smallmatrix}\right)$. Choosing $P = \left(\begin{smallmatrix} 1 & 1 \\ i & -i \end{smallmatrix}\right)$ and using (2.15) and property (iv) of Corollary 2.9 gives

$$e^{tA} = Pe^{t\Lambda}P^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix}\begin{pmatrix} e^{it} & 0 \\ 0 & e^{-it} \end{pmatrix}\begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix} = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix},$$

which is the same as the real matrix, (2.29), obtained using the infinite series. ∎

Suppose that the $n \times n$ real matrix A has a complex eigenvector v and eigenvalue λ. These can be written in terms of their real and imaginary parts as

$$\lambda = a + ib, \quad v = u + iw.$$

Since $Av = \lambda v = (au - bw) + i(aw + bu)$ and A is real, then

$$Au = au - bw, \quad Aw = bu + aw.$$

If we let $P = [u, w]$ be the $n \times 2$ matrix with real columns u and w, then these two equations can be combined to obtain

$$AP = P\begin{pmatrix} a & b \\ -b & a \end{pmatrix}, \tag{2.34}$$

giving a real "normal form" that is not diagonal but relatively simple. We computed the exponential of this 2×2 block in (2.31).

Example 2.18. Consider the 2×2 system

$$A = \begin{pmatrix} 0 & -2 \\ 1 & 2 \end{pmatrix}, \quad p(\lambda) = \lambda^2 - 2\lambda + 2.$$

The eigenvalues are $\lambda = 1 \pm i$, and corresponding eigenvectors are $v = (-1 \pm i, 1)^T$. Using the real and imaginary parts of v we use

$$P = [u, w] = \begin{pmatrix} -1 & 1 \\ 1 & 0 \end{pmatrix}, \quad P^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$$

to obtain

$$P^{-1}AP = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix},$$

$$e^{tA} = Pe^{t}\begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}P^{-1} = e^{t}\begin{pmatrix} \cos t - \sin t & -2\sin t \\ \sin t & \cos t + \sin t \end{pmatrix}.$$ ∎

In general, suppose that there are k real eigenvalues and $2m$ complex ones. Assuming that the set of vectors $\{v_1, v_2, \ldots, v_k, u_1, w_1, \ldots, u_m, w_m\}$ is complete, then the matrix

$$P = [v_1, v_2, \ldots, v_k, u_1, w_1, \ldots, u_m, w_m]$$

is nonsingular, and our result implies that

$$P^{-1}AP = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & \vdots & & 0 \\ \vdots & & \ddots & 0 & & \vdots \\ 0 & \cdots & 0 & B_1 & & 0 \\ \vdots & & \cdots & 0 & \ddots & \vdots \\ 0 & 0 & \cdots & & 0 & B_m \end{pmatrix}$$

is block diagonal with 1×1 blocks λ_j and 2×2 blocks B_j of the form (2.30). The matrix $P^{-1}AP$ can be written as the sum of commuting matrices

$$P^{-1}AP = \Lambda_1 + \Lambda_2 + \cdots + \Lambda_k + C_1 + \cdots + C_m,$$

where

$$\Lambda_i = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \lambda_i & \vdots \\ 0 & \cdots & 0 \end{pmatrix}, \quad C_j = \begin{pmatrix} 0 & & \cdots & & 0 \\ \vdots & a_j & & b_j & \vdots \\ & -b_j & & a_j & \\ 0 & & \cdots & & 0 \end{pmatrix}$$

for $i = 1, \ldots, k$ and $j = 1, \ldots, m$. This means that the solution of the differential equation with complex eigenvalues can be written in terms of the 1×1 matrices $e^{\lambda_i t}$ and the 2×2 blocks (2.31). Finally the exponential of A is of the form

$$e^{tA} = P \begin{pmatrix} e^{\lambda_1 t} & 0 & \cdots & \\ 0 & \ddots & 0 & \cdots \\ \vdots & 0 & e^{tB_1} & \\ & \vdots & 0 & \ddots \end{pmatrix} P^{-1}.$$

With this construction we can now straightforwardly compute the exponential of any matrix that is diagonalizable. In the next section we will consider the more general case.

2.6 ▪ Multiple Eigenvalues

Recall that for an operator $T : E \to E$ on a complex vector space E, an eigenvector with eigenvalue λ is defined as a nonzero solution to (2.2), i.e., the eigenvector v is an element of the null space or kernel, (2.11), of the operator $T - \lambda I$. For the case of multiple eigenvalues—when the algebraic multiplicity is larger than one—it is not sufficient to consider this null space; instead, one must consider the

▷ *generalized eigenspace*: Suppose λ_j is an eigenvalue of a linear operator T with algebraic multiplicity n_j. The generalized eigenspace of λ_j is

$$E_j \equiv \ker\left[\left(T - \lambda_j I\right)^{n_j}\right]. \tag{2.35}$$

One reason that the generalized eigenspace is important dynamically is because it is an

> ▷ *invariant subspace*: A space E is invariant under an operator T if for every $v \in E$, it follows that $T(v) \in E$.

Lemma 2.19 (Invariance). *Each of the generalized eigenspaces of a linear operator T is invariant under T. That is, if E_j is a generalized eigenspace, then $T : E_j \to E_j$.*

Proof. Suppose that $v \in E_j$ so that $\left(T - \lambda_j I\right)^{n_j} v = 0$. To show that $Tv \in E$, compute

$$(T-\lambda_j I)^{n_j} Tv = (T-\lambda_j I)^{n_j-1} T(T-\lambda_j I)v = T(T-\lambda_j I)^{n_j} v = 0,$$

since the operator T commutes with itself and with I. Therefore, whenever $v \in E_j$, $Tv \in E_j$, and so the operator T leaves E_j invariant. □

Just as an eigenvector is a nonzero solution to $(T - \lambda I)v = 0$, we define a

> ▷ *generalized eigenvector*: A nonzero solution to $(T - \lambda_j I)^{n_j} v = 0$, where n_j is the algebraic multiplicity of λ_j, is a generalized eigenvector of T.

It turns out that each generalized eigenspace E_j has dimension equal to n_j, and for the finite-dimensional case, the space spanned by the collection of all of the generalized eigenspaces is the full space.

Theorem 2.20 (Primary Decomposition). *Let T be a linear operator on a finite-dimensional, complex vector space E, with distinct eigenvalues $\lambda_1, \dots, \lambda_r$, and let E_j be the generalized eigenspace of T with eigenvalue λ_j. Then $\dim(E_j)$ is the algebraic multiplicity of λ_j and the generalized eigenvectors span E, i.e., $E = E_1 \oplus E_2 \oplus \cdots \oplus E_r$.*

Consequently the (complex) generalized eigenvectors $\left\{v_1, v_2, \dots, v_{n_j}\right\}$ form a basis for the generalized eigenspace E_j. This theorem is proved in most texts on linear algebra (Hirsch and Smale 1974, Appendix III; Olver and Shakiban 2006, §8.6; Strang 1988, Appendix B).

Generalized eigenvectors are not uniquely defined by their definition: there are infinitely many possible basis choices for the generalized eigenspace (2.35).

Example 2.21. Consider the matrix

$$A = \begin{pmatrix} 6 & 2 & 1 \\ -7 & -3 & -1 \\ -11 & -7 & 0 \end{pmatrix}, \quad p(\lambda) = \lambda^3 - 3\lambda^2 + 4 = (\lambda-2)^2(\lambda+1).$$

From the characteristic polynomial, we see that A has a double eigenvalue $\lambda_1 = 2$. The geometric multiplicity of λ_1 is one since

$$(A-2I)v = 0 \Rightarrow v = c(-1,1,2)^T, \quad c \in \mathbb{R},$$

is a one-dimensional set, so there is only one eigenvector. To find the generalized eigenspace, first compute

$$(A-2I)^2 = \begin{pmatrix} -9 & -9 & 0 \\ 18 & 18 & 0 \\ 27 & 27 & 0 \end{pmatrix}.$$

Since the first two columns of this matrix are the same and the last is zero, its rank is one, so its nullity is two, and there is a two-dimensional space, E_1, of generalized eigenvectors. Indeed, the general solution to $(A-2I)^2v=0$ is $v=(a,-a,b)^T$, which has two arbitrary constants. As generalized eigenvectors we could choose, for example, $(a,b)=(1,0)$ to obtain $v_1=(1,-1,0)^T$ and $(a,b)=(0,1)$ to obtain $v_2=(0,0,1)^T$; moreover, any two linearly independent sets of values of a and b can be used to construct the basis. Note that the eigenvector is also an element of E_1; it is given by the choice $(a,b)=(-c,2c)$.

The eigenvalue $\lambda_2=-1$ has multiplicity one, and its eigenspace is spanned by the eigenvector $v_3=(-1,2,3)^T$. ∎

Semisimple-Nilpotent Decomposition

The decomposition theorem, Theorem 2.20, leads directly to a strategy for finding the exponential of a matrix A using a basis of generalized eigenvectors. If we denote these vectors by $v_1,\dots,v_n$, where, say, $v_1,\dots,v_{n_1}$ give a basis for E_1,and so forth, then the primary decomposition theorem implies that the matrix $P=[v_1,\dots,v_n]$ is nonsingular. As usual, let

$$\Lambda=\operatorname{diag}(\lambda_1,\dots,\lambda_n),$$

and then define a matrix

$$S=P\Lambda P^{-1}. \tag{2.36}$$

This means that $SP=P\Lambda$, or, equivalently, $Sv_i=\lambda_i v_i$. Accordingly, S is *diagonalizable by a similarity transformation* or is

> ▷ *semisimple*: A matrix S is semisimple if there is a (possibly complex) nonsingular matrix P such that $P^{-1}SP=\Lambda$ is diagonal.

The matrix S captures, in some sense, the eigenvalues of A. What is left over? We claim that decomposing $A=S+N$ gives a remainder, N, which is

> ▷ *nilpotent*: A matrix N is nilpotent with nilpotency k if $N^k=0$ but $N^{k-1}\neq 0$.

It is not too hard to show that the maximum nilpotency of an $n\times n$ matrix is n (see Exercise 9).

We have previously seen by example in §2.3 that it is not hard to compute the exponential of a nilpotent matrix. So that the semisimple-nilpotent decomposition is useful for finding the exponential of A, it is important that N commutes with S.

Lemma 2.22. *Let $N\equiv A-S$, where $S=P\Lambda P^{-1}$. Then N commutes with S and is nilpotent with order at most the maximum of the algebraic multiplicities of the eigenvalues of A.*

Proof. Using the definition (2.26), note first that $[S,N]=[S,A-S]=[S,A]$. For any $v\in E_j$, since $Sv=\lambda_j v$,

$$[S,A]v=SAv-A\lambda_j v=(S-\lambda_j I)Av=0,$$

where $Av\in E_j$ because E_j is invariant. Now by Theorem 2.20, E is a direct sum of the E_j and any vector w can be written as a linear combination of $v_j\in E_j$; therefore,

$[S,A]w = 0$. Since this is true for an arbitrary vector, then $[S,A] = 0$, and so $[S,N] = 0$.

To see that N is nilpotent, suppose the maximum algebraic multiplicity of the eigenvalues is m; then for any $v \in E_j$, since $[S,A] = 0$,

$$\begin{aligned} N^m v &= (A-S)^m\, v = (A-S)^{m-1}(Av - \lambda_j v) \\ &= \left(A - \lambda_j I\right)(A-S)^{m-1}\, v = \cdots \\ &= (A - \lambda_j I)^m v = 0. \end{aligned} \tag{2.37}$$

By Theorem 2.20, this relation holds for any $v \in E$; thus, $N^m = 0$. □

Note that the order of N could be less than m; for example, if A is semisimple itself, then $N = 0$, independent of the multiplicities of the eigenvalues.

The semisimple-nilpotent decomposition of a matrix is unique.

Theorem 2.23. *A matrix A on a complex vector space E has a unique decomposition, $A = S + N$, where S is semisimple, N is nilpotent, and $[S,N] = 0$.*

Proof. We have already constructed one such decomposition. Suppose that it is not unique: let $A = \hat{S} + \hat{N}$ be another such decomposition. Recall that S leaves the generalized eigenspaces of A invariant. Suppose that $v \in E_j$. Since $[A,\hat{S}] = 0$, $\left(A - \lambda_j I\right)^{n_j} \hat{S}v = \hat{S}\left(A - \lambda_j I\right)^{n_j} v = 0$; therefore, $\hat{S}$ also leaves E_j invariant. Furthermore, since v is an eigenvector of S, $[S,\hat{S}]v = (S - \lambda_j I)\hat{S}v = 0$; by Theorem 2.20, this is true for all $w \in E$, and consequently $[S,\hat{S}] = 0$. This immediately also implies $[N,\hat{N}] = 0$. Now consider the difference

$$\hat{S} - S = (A - \hat{N}) - (A - N) = N - \hat{N}.$$

Since $\hat{S}$ and S commute and are each semisimple, so is their difference (see Exercise 12). Similarly, since N and $\hat{N}$ commute and are each nilpotent, so is their difference; indeed, let m be the maximum of the nilpotencies of N and $\hat{N}$; then

$$\left(N - \hat{N}\right)^{2m} = \sum_{k=0}^{2m} (-1)^k \left(\begin{array}{c} 2m \\ k \end{array} \right) N^k \hat{N}^{2m-k} = 0,$$

where $\binom{2m}{k}$ is the binomial coefficient. Each term in the sum vanishes since at least one of the matrices is raised to a power greater than or equal to m. Consequently we have shown that $\hat{S} - S$ is diagonalizable and nilpotent. The only such matrix is identically zero, since the only diagonal, nilpotent matrix is 0 itself, and $P0P^{-1} = 0$ for any nonsingular P. Therefore, $\hat{S} = S$ and $\hat{N} = N$. □

A perhaps surprising implication of Theorem 2.23 is that the matrices S and N do not depend upon the ordering of the eigenvalues/eigenvectors, nor on the choice basis for E_j!

The Exponential

The semisimple-nilpotent decomposition leads to a compact and relatively computable formula for the exponential. Letting $A = S + N$, where $S = P\Lambda P^{-1}$, since N is nilpotent,

$$e^{tA} = e^{tS}e^{tN} = Pe^{t\Lambda}P^{-1}\left(\sum_{j=0}^{n-1}\frac{(tN)^j}{j!}\right). \tag{2.38}$$

Here the finite sum for N terminates at the nth term, since necessarily $N^n = 0$ (see Exercise 9).

Unfortunately, computing this general expression still can be labor intensive, as we will see from some examples.

Example 2.24. To complete the classification of the qualitatively distinct cases for the 2×2 matrices that we began in §2.2, consider a matrix on the parabola $\tau^2 = 4\delta$ of Figure 2.1. Writing this equation as $(a-d)^2 = -4bc$ and assuming that $b = \alpha^2 \geq 0$ and $c = -\beta^2 \leq 0$ gives a matrix of the form

$$A = \begin{pmatrix} \lambda+\alpha\beta & \alpha^2 \\ -\beta^2 & \lambda-\alpha\beta \end{pmatrix}. \tag{2.39}$$

It has a single eigenvalue λ with multiplicity two, and since $(A-\lambda I)^2 = 0$, the generalized eigenspace for λ is $E_1 = \mathbb{R}^2$. Therefore, a suitable choice for P is I, and $S = \operatorname{diag}(\lambda,\lambda)$. In this case $N = A - \lambda I$, and $N^2 = 0$, so that

$$e^{tA} = e^{\lambda t}(I + tN).$$

Consequently, the general solution of the ODE is

$$x(t) = e^{\lambda t}\begin{pmatrix} (1+t\alpha\beta)x_1(0) + t\alpha^2 x_2(0) \\ -t\beta^2 x_1(0) + (1-t\alpha\beta)x_2(0) \end{pmatrix}.$$

When $\alpha = \beta = 0$, A has two eigenvectors, and $N = 0$. The general solution is simply $x(t) = e^{\lambda t}x(0)$, so that every solution moves along a ray through the origin. This case is called a *proper node*.

If α and β are not both zero, A has only one eigenvector, $v = (\alpha, -\beta)^T$. Note that $Nv = 0$, so that if $x(0) = cv$, then the solution is $x(t) = e^{\lambda t}x(0)$, a straight-line solution. Every other solution is asymptotic to the form

$$x(t) \to te^{\lambda t}(\beta x_1(0) + \alpha x_2(0))v, \; t \to \pm\infty.$$

Therefore, all solutions are asymptotic to the eigenvector v. This case, shown in Figure 2.7, is called an *improper node* because infinitely many solutions approach the origin along a single direction. ∎

Example 2.25 (Multiplicity n). The previous example is a special case of a single eigenvalue of multiplicity n. When this is true, $E = E_1$, so that every vector is in E_1. Consequently, we are free to choose the v_i so that $P = I$, which gives $S = \lambda I$. The associated nilpotent matrix is

$$N = A - \lambda I.$$

Since $\ker(A-\lambda I)^n = E$, then $N^n = 0$ and $[S,N] = 0$. Amazingly, we have written $A = S + N$,where S is semisimple (in fact diagonal) and N is nilpotent, and we did not even need to find the eigenvectors! The exponential then follows easily:

$$e^{tA} = e^{\lambda t}\left(I + tN + \frac{t^2}{2}N^2 + \cdots + \frac{t^{n-1}}{(n-1)!}N^{n-1}\right).$$

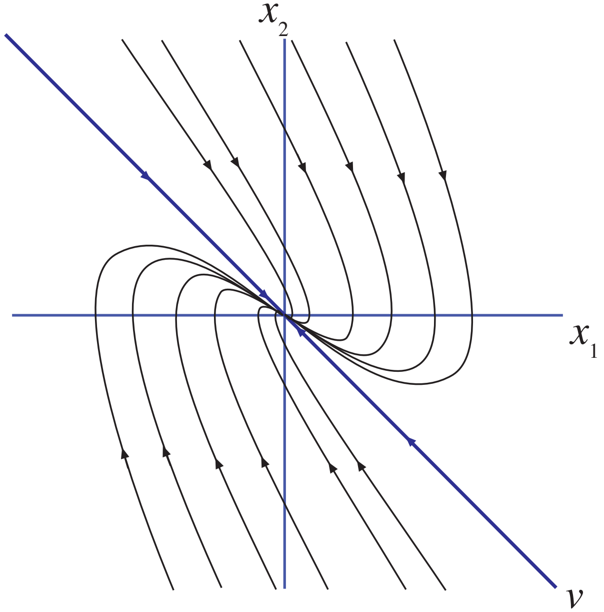

Figure 2.7. *Phase portrait of the stable improper node* (2.39) *with* $\lambda < 0$ *and* $\alpha = \beta > 0$.

A simple case for this would be an upper triangular matrix with a single eigenvalue λ, such as

$$A = \begin{pmatrix} \lambda & 1 & 1 \\ 0 & \lambda & 2 \\ 0 & 0 & \lambda \end{pmatrix}, \quad S = \lambda I, \quad N = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

In this case

$$N^2 = \begin{pmatrix} 0 & 0 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad N^3 = 0,$$

and the exponential becomes

$$e^{tA} = e^{\lambda t} \begin{pmatrix} 1 & t & t+t^2 \\ 0 & 1 & 2t \\ 0 & 0 & 1 \end{pmatrix}. \quad \blacksquare$$

Example 2.26. Consider the multiplicity-two case

$$A = \begin{pmatrix} -1 & 1 & -2 \\ 0 & -1 & 4 \\ 0 & 0 & 1 \end{pmatrix}, \quad p(\lambda) = (\lambda+1)^2(\lambda-1) = 0. \tag{2.40}$$

The eigenspace for $\lambda_3 = 1$ is obtained by solving

$$(A-I)v_3 = \begin{pmatrix} -2 & 1 & -2 \\ 0 & -2 & 4 \\ 0 & 0 & 0 \end{pmatrix} v_3 = 0, \text{ thus } v_3 = \begin{pmatrix} 0 \\ 2 \\ 1 \end{pmatrix}.$$

To obtain the generalized eigenspace, for $\lambda_1 = \lambda_2 = -1$, solve

$$(A+I)^2 v = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 8 \\ 0 & 0 & 4 \end{pmatrix} v = 0, \text{ thus } v = \begin{pmatrix} a \\ b \\ 0 \end{pmatrix},$$

for arbitrary constants a and b. The space E_1 is spanned by $v_1 = (1,0,0)^T$ and $v_2 = (0,1,0)^T$. Setting $P = [v_1, v_2, v_3]$ gives

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}, \; P^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{pmatrix},$$

$$S = P\Lambda P^{-1} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 4 \\ 0 & 0 & 1 \end{pmatrix}, \; N = A - S = \begin{pmatrix} 0 & 1 & -2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Since $N^2 = 0$, the final answer is

$$e^{tA} = Pe^{t\Lambda}P^{-1}e^{tN} = P \begin{pmatrix} e^{-t} & 0 & 0 \\ 0 & e^{-t} & 0 \\ 0 & 0 & e^t \end{pmatrix} P^{-1}(I + tN),$$

$$= \begin{pmatrix} e^{-t} & 0 & 0 \\ 0 & e^{-t} & -2e^{-t}+2e^t \\ 0 & 0 & e^t \end{pmatrix} \begin{pmatrix} 1 & t & -2t \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} e^{-t} & te^{-t} & -2te^{-t} \\ 0 & e^{-t} & -2e^{-t}+2e^t \\ 0 & 0 & e^t \end{pmatrix}. \quad \blacksquare$$

Alternative Methods

As was noted by (Moler and Loan 1978), there are at least 19 different algorithms for computing the matrix exponential, some less useful than others, at least for numerical computations. Here is a 20th way that appears to be quite useful (Harris, Fillmore, and Smith 2001). Denoting the characteristic polynomial of A by $p(\lambda)$, recall that the Cayley–Hamilton theorem (Olver et al 2006; Strang 1988) states that

$$p(A) = 0. \tag{2.41}$$

Moreover, using $\frac{d^n}{dt^n} e^{tA} = A^n e^{tA}$ to replace the powers of A in each term of the polynomial p by derivatives implies that

$$p\left(\frac{d}{dt}\right) e^{tA} = 0,$$

so that every component of e^{tA} solves the nth order *scalar* ODE $p\left(\frac{d}{dt}\right) u(t) = 0$. This ODE has a fundamental set of n solutions, call them $\varphi_j(t)$, $j = 0, 1, \ldots, n-1$, i.e., the solutions such that

$$p\left(\frac{d}{dt}\right) \varphi_j(t) = 0, \;\; \frac{d^i}{dt^i} \varphi_j(0) = \delta_{ij}$$

for $i = 0, 1, \ldots, n-1$. Here δ_{ij} is the Kronecker delta

$$\delta_{ii} = 1 \text{ and } \delta_{ij} = 0 \text{ if } i \neq j. \tag{2.42}$$

Consequently, $\varphi_0(0) = 1$, and $\dot{\varphi}_1(0) = 1$, etc. Accordingly, any solution can be written as a linear combination of the fundamental solutions:

$$e^{tA} = \varphi_0(t)F_0 + \varphi_1(t)F_1 + \cdots + \varphi_{n-1}(t)F_{n-1},$$

where the F_i are constant matrices. It is easily seen by differentiating this expression i times and setting $t = 0$ that

$$\frac{d^i}{dt^i}\, e^{tA}\Big|_{t=0} = A^i = \sum_{j=0}^{n-1} \frac{d^i}{dt^i} \varphi_j(0)F_j = F_i.$$

This gives the expression

$$e^{tA} = \varphi_0(t)I + \varphi_1(t)A + \varphi_2(t)A^2 + \cdots + \varphi_{n-1}(t)A^{n-1}. \tag{2.43}$$

This form could have been anticipated from the series expression for e^{tA}. Indeed, the Cayley–Hamilton theorem implies that every power A^k for $k \geq n$ can be expressed as a linear combination of the matrices $I, A, \ldots, A^{n-1}$. The form (2.43) follows directly, although a little more work is needed to identify the coefficients φ_i as the fundamental solutions.

Example 2.27. Let $A = \begin{pmatrix} 1 & 2 \\ 4 & -1 \end{pmatrix}$ so that $p(\lambda) = \lambda^2 - 9$. The solutions to $p\left(\frac{d}{dt}\right)u(t) = \ddot{u} - 9u = 0$ are linear combinations of $u_\pm(t) = e^{\pm 3t}$, and a bit of algebra gives the fundamental solutions:

$$\varphi_0 = \tfrac{1}{2}\left(e^{3t} + e^{-3t}\right), \quad \varphi_1 = \tfrac{1}{6}\left(e^{3t} - e^{-3t}\right).$$

Therefore,

$$e^{tA} = \varphi_0 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \varphi_1 \begin{pmatrix} 1 & 2 \\ 4 & -1 \end{pmatrix} = \frac{1}{3}\begin{pmatrix} 2e^{3t} + e^{-3t} & e^{3t} - e^{-3t} \\ 2e^{3t} - 2e^{-3t} & e^{3t} + 2e^{-3t} \end{pmatrix}. \quad \blacksquare$$

2.7 ▪ Linear Stability

There are several definitions of stability of dynamical systems. We will discuss the most useful one—Lyapunov stability—in Chapter 4. For now, think of stability as being related to the idea that solutions are bounded as $t \to \infty$. For example, the sign of λ governs the long-time behavior of the solution of the single differential equation $\dot{x} = \lambda x$. If $\lambda > 0$, the solution is unbounded, while if $\lambda \leq 0$, it is bounded (for positive time).

More generally, the solution of $\dot{x} = Ax$ is $x(t) = e^{tA}x_o$, and each element of the exponential matrix is a sum of terms that are multiplied by exponentials of the eigenvalues, $e^{\lambda t}$. This means the spectrum determines whether there are exponentially growing or decaying terms. This leads to the definition of

> ▷ *spectral stability*: A linear system is spectrally stable if none of its eigenvalues has a positive real part.

The sign of the real part of the eigenvalue distinguishes the subspaces on which the solutions have growing or decaying behavior. Denote the (complex) generalized eigenvectors by $v_j = u_j + iw_j$. Then

▷ $E^u = \text{span}\left\{u_j, w_j : \text{Re}(\lambda_j) > 0\right\}$ is the *unstable* subspace,

▷ $E^c = \text{span}\left\{u_j, w_j : \text{Re}(\lambda_j) = 0\right\}$ is the *center* subspace, and

▷ $E^s = \text{span}\left\{u_j, w_j : \text{Re}(\lambda_j) < 0\right\}$ is the *stable* subspace.

Note that by Theorem 2.20, $E = E^u \oplus E^c \oplus E^s$. Moreover, since each of the generalized eigenspaces is invariant, so are the stable, center, and unstable subspaces.

Consequently, we can describe the evolution in each subspace by constructing a "restriction," say, $A|_{E^u}$ of A. For example, if $P = [v_1, v_2, \ldots, v_k]$ is the $n \times k$ matrix formed from a basis for E^u, then every vector x in E^u has a unique expansion in this basis, i.e.,

$$x = \sum_{j=1}^{k} c_j v_j = Pc \in E^u.$$

Since E^u is an invariant subspace, then each column of the matrix AP is in E^u and has such an expansion: the jth column can be written $(AP)_j = \sum_{i=1}^{k} v_i u_{ij}$. Collecting these columns uniquely defines $U = \left(u_{ij}\right) = A|_{E^u}$ as the $k \times k$ matrix that solves

$$AP = PU.$$

This can be solved for U by multiplying each side by P^T and noting that the Gram matrix $P^T P$ is nonsingular to give

$$U = (P^T P)^{-1}(P^T A P).$$

The dynamical evolution of x can be determined by allowing the coefficients c_i to depend on time. Then

$$\dot{x} = P\dot{c} = APc = PUc.$$

Uniqueness of the basis representation then implies that $\dot{c} = Uc$. Thus, U represents the dynamics in the subspace E^u. A similar representation could be obtained in any invariant subspace.

Example 2.28. Consider again the example (2.40). The eigenvalue $\lambda_3 = 1$ has eigenvector $v_3 = (0, 2, 1)^T$. Consequently the matrix $U = A|_{E^u}$ is the 1×1 matrix defined by the equation

$$Av_3 = 1v_3 = v_3 U,$$

and $U = (1)$. The dynamics restricted to this subspace is simply $\dot{c}_3 = 1c_3$. The stable subspace with eigenvalue $\lambda_1 = -1$ has basis $v_1 = (1, 0, 0)^T$ and $v_2 = (0, 1, 0)^T$, so that the stable matrix $S = A|_{E^s}$ is the 2×2 matrix defined by

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

which gives $U = \left(\begin{smallmatrix} -1 & 1 \\ 0 & -1 \end{smallmatrix}\right)$. The dynamics in this subspace is therefore

$$\begin{pmatrix} \dot{c}_1 \\ \dot{c}_2 \end{pmatrix} = \begin{pmatrix} -c_1 + c_2 \\ -c_2 \end{pmatrix},$$

and c_1, c_2 are simply the x_1 and x_2 components. ∎

A system with no center subspace is

▷ *hyperbolic*: A linear system is hyperbolic if all its eigenvalues have nonzero real parts.

The importance of hyperbolic systems stems from their simple behavior under perturbation. Imagine choosing a matrix at random. Only rarely would a matrix that has any pure imaginary eigenvalues occur; in some sense, the set of such matrices occurs with probability zero. More precisely, one says that hyperbolic systems are *generic*. By contrast, the dynamical consequences of perturbing a system with a center subspace are much more complicated, as we will see in Chapter 8.

Showing that a system is spectrally stable or not is relatively easy, since the eigenvalues can be computed by solving the characteristic polynomial. (The Routh–Hurwitz theorem gives a stability criterion; see Exercise 11.) However, this does not tell the whole story: when the nilpotent part of A is nonzero, then e^{tA} contains terms that are powers of t multiplied by exponentials, that is, terms of the form $t^k e^{\lambda t}$. Note that if the real part of λ is negative, then this function is still bounded for any $k \geq 0$ and that it asymptotically approaches zero as $t \to \infty$. Indeed, it is not hard to see that every initial condition in the stable subspace asymptotically approaches the origin.

Lemma 2.29. *If A is an $n \times n$ matrix, then there are constants $K \geq 1$ and $\alpha > 0$ such that*

$$\left| e^{tA} x_o \right| \leq K e^{-\alpha t} \left| x_o \right|, \; t \geq 0, \tag{2.44}$$

for any $x_o \in E^s$, the stable space of A. Consequently, $e^{tA} x_o \to 0$ as $t \to \infty$.

Proof. According to (2.38), the solution is of the form

$$e^{tA} x_o = P e^{t\Lambda} P^{-1} \left(\sum_{j=0}^{n-1} \frac{(tN)^j}{j!} \right) x_o .$$

Since $x_o \in E^s$, and this is an invariant subspace of dimension n_s, we need only consider the matrix $e^{tA}\big|_{E^s}$. Each element of this matrix will be a linear combination of terms from the stable eigenvectors, i.e., of the form $t^k e^{a_j t} e^{i b_j t}$, where $\lambda_j = a_j + i b_j$, $j = 1, 2, \ldots, n_s$, $a_j < 0$, and $k < n_s$. Indeed, according to (2.37) the restriction of N to a generalized eigenspace has nilpotency at most n_k, the algebraic multiplicity of λ_k. Consequently the maximum power of t in any term will be $n_k - 1$. More explicitly, a general element of the exponential restricted to the stable subspace must have the form

$$\left(e^{tA}\big|_{E^s} \right)_{lm} = \sum_{j=1}^{n_s} \sum_{k=0}^{n_j - 1} t^k e^{a_j t} \left(c_{jklm} \cos(b_j t) + d_{jklm} \sin(b_j t) \right)$$

for some set of coefficients c_{jklm}, and d_{jklm}. Choose an $\alpha > 0$ such that $a_j < -\alpha < 0$. Then there is a K such that $t^{n_s} e^{(\alpha + a_j)t} \sqrt{c^2_{jklm} + d^2_{jklm}} < K/n_s^2$ for all $j, k \in [1, n_s]$, $l, m \in [1, n]$, and $t \geq 0$. Consequently, each term in the sum has the bound $\frac{K}{n_s^2} e^{-\alpha t}$. This directly implies the result. □

In this lemma, we did not work very hard to find an optimal value for K. It can be shown (with much more work), that with the selection of a new norm that is adapted to A, the constant K can be chosen to be equal to one (see, e.g., (Chicone 1999, Theorem 2.34; Robinson 1999, Theorem 5.1)).

If there is an eigenvalue λ with zero real part (i.e., the center subspace is not empty), terms of the form $t^k e^{\lambda t}$ grow with t when $k > 0$; therefore, when there are eigenvalues with zero real part, stability is affected by the nilpotent part of A.

A stronger concept than spectral stability is one that would guarantee that all solutions are bounded. If all solutions are bounded, then a system is linearly stable:

> ▷ *linear stability*: A linear system is linearly stable if all its solutions are bounded as $t \to \infty$.

As we argued above, any initial condition in the stable subspace, $x_o \in E^s$, has a bounded solution for $t > 0$. Similarly, any initial condition in the unstable subspace, $x_o \in E^u$, has an unbounded solution as $t \to \infty$. Solutions in the center space can be bounded, but in general, when the multiplicity of an eigenvalue in this subspace is larger than one, they are not.

The strongest concept for stability of linear systems is

> ▷ *asymptotic linear stability*: A linear system is asymptotically linearly stable if all of its solutions approach 0 as $t \to \infty$.

This occurs whenever $E = E^s$.

Theorem 2.30 (Asymptotic Linear Stability). $\lim_{t\to\infty} e^{tA} x_o = 0$ *for all* x_o *if and only if all eigenvalues of* A *have negative real part.*

Proof. If all the eigenvalues have negative real part, then Lemma 2.29 implies that $\lim_{t\to\infty} e^{tA} x_o = 0$. Conversely, if there is an eigenvalue with positive real part, then there is an initial condition in the eigenspace corresponding to this eigenvalue, so that the solution grows exponentially without bound. Finally, if there is an eigenvalue with zero real part, then solutions in this subspace have terms of the form $t^j e^{iIm(\lambda_k)t}$ and do not go to zero. □

Similarly, when all the eigenvalues have positive real part, the solution goes asymptotically to zero as $t \to -\infty$.

Example 2.31. Consider the system with matrix

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

which has characteristic polynomial $p(\lambda) = \lambda^3 + 2\lambda^2$. Hence the eigenvalues are $\lambda = -2$ with multiplicity 1 and $\lambda = 0$ with multiplicity 2. Since there are no eigenvalues with positive real part, the system is spectrally stable. It is not hyperbolic, since there are two zero eigenvalues. To find the stable subspace we must solve for the eigenvector

$$(A + 2I)v = \begin{pmatrix} 0 & -1 & -2 \\ -2 & 0 & -2 \\ 2 & 1 & 4 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = 0.$$

This implies that $v_1 = (1, 2 - 1)^T$, or any nonzero multiple of this. Consequently,

$$E^s = \operatorname{span}(v_1) = \left\{ \begin{pmatrix} c \\ 2c \\ -c \end{pmatrix} : c \in \mathbb{R} \right\}.$$

Theorem 2.20 implies that E^c is the complement of E^s, since the generalized eigenvectors span $\mathbb{R}^3$. To demonstrate this, find generalized eigenvectors by solving

$$(A-0I)^2 v = \begin{pmatrix} 2 & 2 & 2 \\ 4 & 4 & 4 \\ -2 & -2 & -2 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = 0.$$

This is equivalent to the single equation $a+b+c=0$, so that there are two arbitrary constants in v (we knew this already since $\dim(E^c)=2$). One representation of the solution is $v = av_2 + bv_3$, where $v_2 = (1,0,-1)^T$ and $v_3 = (0,1,-1)^T$. Consequently,

$$E^c = \operatorname{span}(v_2, v_3) = \left\{ \begin{pmatrix} a \\ b \\ -a-b \end{pmatrix} : a, b \in \mathbb{R} \right\}.$$

Finally we ask, is the system linearly stable? For this to be the case, the nilpotent part of A must vanish, or alternatively there must be two independent eigenvectors corresponding to $\lambda = 0$. The eigenvalue problem $(A-0I)v = 0$ has only a single solution, $v = (1,0,-1)^T$. Since the nilpotent part is nonzero our system is not linearly stable. This is confirmed by finding

$$\begin{aligned} S = P\Lambda P^{-1} &= \begin{pmatrix} 1 & 1 & 0 \\ 2 & 0 & 1 \\ -1 & -1 & -1 \end{pmatrix} \begin{pmatrix} -2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \frac{1}{2} \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -2 & 0 & -2 \end{pmatrix} \\ &= \begin{pmatrix} -1 & -1 & -1 \\ -2 & -2 & -2 \\ 1 & 1 & 1 \end{pmatrix}, \end{aligned}$$

giving a nilpotent part

$$N = A - S = \begin{pmatrix} -1 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix},$$

which is easily seen to satisfy $N^2 = 0$. Finally, the exponential is

$$e^{tA} = Pe^{t\Lambda}P^{-1}(I+tN) = \frac{1}{2} \begin{pmatrix} e^{-2t}+1-2t & e^{-2t}-1 & e^{-2t}-1-2t \\ 2e^{-2t}-2 & 2e^{-2t} & 2e^{-2t}-2 \\ -e^{-2t}+1+2t & -e^{-2t}+1 & -e^{-2t}+3+2t \end{pmatrix},$$

confirming that this system is unstable since there are terms that grow linearly in time. In particular, if $x_0 = (1,0,0)^T$, then $x(t) \to 2t(-1,0,1)^T \to \infty$. Note that not all solutions are unbounded. For example, if $x_0 = (0,1,0)^T$, then $x(t) \to (-1,0,1)^T$. Nevertheless, a single unbounded solution is enough to declare the system unstable. ∎

2.8 ▪ Nonautonomous Linear Systems and Floquet Theory

A linear physical system that is externally forced can often be modeled by the affine set of ODEs,

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

Such differential equations can be easily solved using the "integrating factor" method; see Exercise 17. It is considerably more difficult to solve a linear system when the matrix A depends upon time,

$$\dot{x} = A(t)x, \; x(t_o) = x_o. \tag{2.45}$$

Nonautonomous equations like these can arise in mechanical systems if the forcing changes the effective spring constants; for example, a person pumping his legs on a swing will change the effective length of the pendulum and thereby modulate the coefficient g/l that governs the linear oscillation frequency. Equations of the form (2.45) also occur as the linearization of the dynamics about a periodic orbit of period T. In this case the matrix A is a periodic function of time, $A(t+T) = A(t)$. Gaston Floquet developed the theory of the solutions of such systems in the 1880s (Chicone 1999, §2.4; Floquet 1883; Yakubovitch and Starzhinskii 1975, Chapter 5).

To solve (2.45), it is convenient to consider a matrix differential equation of the form (2.33), replacing the vector $x(t)$ by a matrix. The general solution is most conveniently represented in terms of the principal *fundamental matrix* solution, which is the solution $\Phi(t, t_o)$ of the matrix initial value problem

$$\frac{d}{dt}\Phi = A(t)\Phi, \quad \Phi(t_o, t_o) = I. \tag{2.46}$$

Here we have added a second argument to Φ to indicate that the initial condition is applied at time t_o. As for the autonomous case, the solution of the original system with initial value $x(t_o) = x_o$ is simply given by $x(t) = \Phi(t, t_o)x_o$. Thus, if we can find $\Phi(t, t_o)$, we also have the general solution to (2.45). We will ignore for the moment the more delicate question of the existence and uniqueness of Φ; this will follow more generally from Theorem 3.24, requiring only that $A(t)$ be a continuous function of time. Uniqueness implies that the fundamental matrix solution obeys the relation

$$\Phi(t, r) = \Phi(t, s)\Phi(s, r) \tag{2.47}$$

for all $t, s, r \in \mathbb{R}$.

When A is constant $\Phi(t, t_o) = e^{(t-t_o)A}$, and we proved in §2.4 that this is the unique solution. However, this formula no longer works for the time-dependent case, and more importantly, the "obvious" generalization

$$\Phi(t, t_o) = \exp\left(\int_{t_o}^{t} A(s)ds\right) \quad \text{(incorrect!)} \tag{2.48}$$

is usually wrong since the matrix $A(s_1)$ does not generally commute with $A(s_2)$ when $s_1 \neq s_2$ (see Exercises 18–19). Moreover, as the following example shows, the eigenvalues of the matrix $A(t)$ at a fixed value of time may have nothing to do with the properties of the solution of (2.45).

Example 2.32. Here is an example that points out the pitfalls of looking at the eigenvalues of $A(t)$ (Markus and Yamabe 1960). Consider the time-dependent matrix

$$A(t) = \begin{pmatrix} -1 + \alpha\cos^2 t & 1 - \alpha\cos t \sin t \\ -1 - \alpha\cos t \sin t & -1 + \alpha\sin^2 t \end{pmatrix}.$$

It is easy to see that the eigenvalues of this matrix are independent of time because $\mathrm{tr}(A) = \alpha - 2$, and $\det(A) = 2 - \alpha$, so

$$\lambda = \tfrac{1}{2}\left(\alpha - 2 \pm \sqrt{\alpha^2 - 4}\right).$$

When $\alpha < 2$, the eigenvalues indicate that this system may be stable. However, the differential equation $\dot{x} = A(t)x$ has two simple explicit solutions, as can be easily verified by substitution:

$$x_1(t) = \begin{pmatrix} \cos t \\ -\sin t \end{pmatrix} e^{(\alpha-1)t}, \qquad x_2(t) = \begin{pmatrix} \sin t \\ \cos t \end{pmatrix} e^{-t}. \tag{2.49}$$

Therefore, when $\alpha > 1$ the first solution is unbounded and thus the system is unstable. Consequently, for the range $1 < \alpha < 2$ the system is unstable, even though the eigenvalues of $A(t)$ would suggest that it should be stable. This example shows that the eigenvalues of a nonautonomous matrix do not generally determine the stability of the corresponding ODE. ■

For the case that A is a periodic matrix, an important quantity is the value of the fundamental matrix at one period; it is called the

▷ *monodromy matrix*, $M \equiv \Phi(T, 0)$.

Given the initial condition $x(0) = x_o$, then $x(T) = Mx_o$. To continue this solution past T requires finding the solution of the initial value problem

$$\dot{x} = A(t)x, \quad x(T) = Mx_o.$$

Define a new time variable $\tau = t - T$, and use $A(\tau + T) = A(\tau)$ to see that this is the same as the initial value problem (2.45), with x_o replaced by Mx_o, so its solution is $\Phi(\tau, 0)Mx_o$. This implies

$$x(2T) = M^2 x_o.$$

In consequence, to get the long-time behavior of any solution, we merely need to compute M^n.

The eigenvalues of M are called the *Floquet multipliers*. Suppose x_o is an eigenvector of M with eigenvalue μ; then

$$x(nT) = \mu^n x_o = e^{n \ln \mu} x_o.$$

The exponent $\ln \mu$ is called a *Floquet exponent*; it is a special case of the Lyapunov exponent that we will meet in Chapter 7.

Example 2.33. Continuing the previous example, note that the matrix $A(t)$ is periodic with period $T = \pi$. Moreover, the two solutions (2.49) are linearly independent, and since $x_1(0) = (1, 0)^T$ and $x_2(0) = (0, 1)^T$, the fundamental solution is $\Phi(t, 0) = [x_1(t), x_2(t)]$. Evaluating this at $t = \pi$ gives the monodromy matrix

$$M = \Phi(\pi, 0) = \begin{pmatrix} -e^{\pi(\alpha-1)} & 0 \\ 0 & -e^{-\pi} \end{pmatrix},$$

showing that the Floquet multipliers are $\mu_1 = \text{-e}^{\pi(\alpha-1)}$ and $\mu_2 = \text{-e}^{-\pi}$. Note that when $\alpha > 1$, there is one Floquet multiplier with magnitude larger than one and one with magnitude smaller than one. ■

In general, the monodromy matrix M is nonsingular. In fact, there is a simple equation for the evolution of the determinant of Φ that holds even when $A(t)$ is not periodic. This theorem generalizes the standard result by Abel for the "Wronskian" of a second-order ODE.

Theorem 2.34 (Abel). *The determinant of the fundamental matrix is*

$$\det(\Phi(t,t_o)) = \exp\int_{t_o}^{t} \operatorname{tr}(A(s))ds. \tag{2.50}$$

Note that $\operatorname{tr}(A(s))$ *is a scalar, so the exponential is the ordinary, scalar exponential.*

Proof. Our goal is to obtain a simple ODE for $\det(\Phi)$. The derivative of the determinant of Φ can be computed using the cofactor formula. Recall that the cofactor, c_{ij}, is $(-1)^{i+j}$ times the determinant of the $(n-1)\times(n-1)$ matrix obtained by omitting the ith row and the jth column from Φ. Multiplying c_{ij} by Φ_{ij} and summing over j, i.e., summing along the ith row, gives

$$\det(\Phi) = \sum_{j=1}^{n} c_{ij}\Phi_{ij}.$$

This formula is true for any choice of row i. If instead we multiply c_{ij} by Φ_{kj}, and then sum over j, then this is equivalent to computing the determinant of the matrix with the ith row replaced by the kth row. Since the resulting matrix has two equal rows, its determinant is zero. This generalization of the cofactor formula can be written as

$$\det(\Phi)\ \delta_{ik} = \sum_{j=1}^{n} c_{ij}\Phi_{kj}, \tag{2.51}$$

where δ_{ij} is the Kronnecker delta (2.42). Equivalently, (2.51) can be written in matrix notation as $\det(\Phi)I = C\Phi^T$. Finally, note that the only term in $\det(\Phi)$ that contains a specific element Φ_{ij} is the term $c_{ij}\Phi_{ij}$, so that

$$\frac{\partial}{\partial \Phi_{ij}} \det(\Phi) = c_{ij}. \tag{2.52}$$

Using (2.46), (2.51), (2.52), and the chain rule, the time derivative of the fundamental matrix is

$$\begin{aligned}\frac{d}{dt}\det(\Phi(t)) &= \sum_{i,j=1}^{n} c_{ij}(t)\frac{d}{dt}\Phi_{ij}(t) = \sum_{i,j,k=1}^{n} c_{ij}(t)a_{ik}(t)\Phi_{kj}(t)\\ &= \sum_{i,k=1}^{n} a_{ik}(t)\left(\sum_{j=1}^{n} c_{ij}(t)\Phi_{kj}(t)\right) = \sum_{i,k=1}^{n} a_{ik}(t)\delta_{ik}\det(\Phi(t)).\end{aligned}$$

Simplifying yields

$$\frac{d}{dt}\det(\Phi(t)) = \left(\sum_{i,k=1}^{n} \delta_{ik}a_{ik}(t)\right)\det(\Phi(t)) = \operatorname{tr}(A(t))\ \det(\Phi(t)).$$

This scalar differential equation for the determinant of Φ can be easily integrated to time t to obtain the promised (2.50). □

Since $\det(\Phi(T,0)) = \det(M)$, M is nonsingular. Consequently, all the Floquet multipliers are nonzero and the Floquet exponents are well defined. Abel's theorem will be used in §4.11 and in §7.2 to aid the study of the stability of periodic and aperiodic orbits.

In addition to the Floquet exponents, $\ln \mu_j$, it is also convenient to define the logarithm of the Floquet matrix, $\ln M$, itself. However, it is not obvious that the logarithm of a general matrix is always well defined, as is the case for the exponential. Since the MacLaurin series defined $\exp(M)$, it would be reasonable to use a similar series for the logarithm,

$$\ln(1-x) = -\sum_{j=1}^{\infty} \frac{x^j}{j}; \tag{2.53}$$

however, this converges only for $|x| < 1$. Since $\ln M = \ln(I-(I-M))$, we assume the series definition can be used only for $||I-M|| < 1$. How can we define $\ln M$ in general?

Lemma 2.35. *Any nonsingular matrix A has a (possibly complex) logarithm*

$$\ln A = P\ln(\Lambda)P^{-1} - \sum_{j=1}^{n-1} \frac{(-S^{-1}N)^j}{j},$$

where $A = S+N$, $S = P\Lambda P^{-1}$ is semisimple, N is nilpotent, Λ is the diagonal matrix of eigenvalues, and P is the matrix of generalized eigenvectors of A.

Proof. The semisimple-nilpotent decomposition, Theorem 2.23, gives $A = S+N$, where S is semisimple, N is nilpotent, and $[S,N]=0$. Since A is assumed nonsingular, S is also nonsingular since its eigenvalues are the same as those of A.

Consider first the case of a semisimple, nonsingular matrix S. By definition there exists a diagonalizing transformation P such that $P^{-1}SP = \Lambda$, where Λ is diagonal and has all entries nonzero but is possibly complex. Defining $\ln\Lambda \equiv \operatorname{diag}(\ln\Lambda_{ii})$, then $e^{\ln\Lambda} = \Lambda$, and

$$S = Pe^{\ln\Lambda}P^{-1} = \exp\left(P\ln\Lambda P^{-1}\right), \tag{2.54}$$

so that $\ln S \equiv P\ln\Lambda P^{-1}$. Hence $\ln S$ exists for any nonsingular, semisimple S.

Now suppose that N is any nilpotent matrix. We claim that $\ln(I+N)$ exists. Indeed, using the series (2.53) formally (ignoring convergence), define a matrix B by

$$B = -\sum_{j=1}^{\infty} \frac{(-N)^j}{j} = -\sum_{j=1}^{n-1} \frac{(-N)^j}{j}. \tag{2.55}$$

This is more than a formal definition, however, because, when N is nilpotent, only finitely many terms in this series are nonzero; consequently, (2.55) converges for any N. Moreover we claim that $e^B = I+N$. Formal manipulation of the power series gives

$$e^B = \sum_{k=0}^{\infty} \frac{1}{k!}\left(-\sum_{j=1}^{\infty} \frac{(-N)^j}{j}\right)^k = I+N$$

because this is true for scalar values, and $\left[N^j, N^k\right] = 0$ for any integers j and k. Moreover these series converge because the exponential series converges for any bounded linear operator, and the inner series has only finitely many nonzero terms. In conclusion, $B = \ln(I+N)$ is given by (2.55) for any nilpotent N.

Finally, consider the general case:

$$A = S + N = S(I + S^{-1}N).$$

Note that since N is nilpotent and $[S,N] = 0$, then $S^{-1}N$ is also nilpotent: if $N^k = 0$, then $\left(S^{-1}N\right)^k = S^{-k}N^k = 0$. Therefore, both terms, S and $(I + S^{-1}N)$, have logarithms. By analogy with the property $\ln(ab) = \ln a + \ln b$, we claim that $\ln A$ is given by

$$B = \ln S + \ln(I + S^{-1}N),$$

where the first term is given by (2.54) and the second by (2.55) with $N \to S^{-1}N$. Note that $\left[S, I + S^{-1}N\right] = 0$, and so by their definitions, $\left[\ln S, \ln(I + S^{-1}N)\right] = 0$ as well. This implies that

$$e^B = e^{\ln S + \ln(I+S^{-1}N)} = e^{\ln S} e^{\ln(I+S^{-1}N)} = S(I + S^{-1}N) = A,$$

as claimed. □

Although $\ln A$ exists, it is not unique. Indeed, just as for a scalar, where the exponential of $\ln(a) + 2n\pi i$ is independent of $n \in \mathbb{Z}$, the eigenvalues of $\ln A$ are unique only up to addition of $2n\pi i$ (see Exercise 13d).

The definition of $\ln M$ can be used to obtain a nice form for the solutions to a periodic linear system.

Theorem 2.36 (Floquet 1883). *Let M be the monodromy matrix for a T-periodic linear system $\dot{x} = A(t)x$ and $TB = \ln M$ its logarithm. Then there exists a T-periodic matrix $\mathscr{P}$ such that the fundamental matrix solution is*

$$\Phi(t,0) = \mathscr{P}(t)e^{tB}. \tag{2.56}$$

Proof. Let $\Psi(t) = \Phi(t + T, 0)$. Since $A(t)$ is periodic, then $\frac{d}{dt}\Psi = A(t + T)\Psi = A(t)\Psi$, with $\Psi(0) = M$. Now since Φ is the fundamental matrix solution, every solution $x(t)$ is of the form $\Phi(t,0)x(0)$; accordingly $\Psi(t) = \Phi(t,0)M$, and

$$\Phi(t + T, 0) = \Phi(t,0)M = \Phi(t,0)e^{TB}.$$

Since e^{tB} is nonsingular, define $\mathscr{P}(t) \equiv \Phi(t,0)e^{-tB}$ so that

$$\mathscr{P}(t + T) = \Phi(t + T, 0)e^{-(t+T)B} = \Phi(t,0)e^{TB}e^{-(t+T)B} = \mathscr{P}(t).$$

Therefore, $\mathscr{P}$ is T-periodic. □

As usual, it is not always satisfactory to write the solution of a real linear system in terms of complex functions. However, at the expense of doubling the period, a real form can be found, as follows.

Theorem 2.37. *Let Φ be the fundamental matrix solution for the time T-periodic linear system* (2.45). *Then there exist a real $2T$-periodic matrix $\mathscr{Q}$ and real matrix R such that*

$$\Phi(t,0) = \mathscr{Q}(t)e^{tR}.$$

Proof. In Exercise 21, you will show that for any nonsingular matrix M, there exists a real matrix R such that $M^2 = e^{2TR}$. Define $\mathscr{Q}(t) = \Phi(t,0)e^{-tR}$, and then

$$\mathscr{Q}(t + 2T) = \Phi(t + 2T, 0)e^{-2TR}e^{-tR} = \Phi(t,0)M^2M^{-2}e^{-tR} = \mathscr{Q}(t).$$

Therefore, $\mathcal{Q}$ is $2T$-periodic. □

In fact, one need only extend the period to $2T$ when M has negative real multipliers (see Exercise 21). These, as we will see later in Chapter 8, typically arise near a "period-doubling bifurcation."

2.9 ▪ Exercises

You should do these problems by hand; however, feel free to use a computer to check your answers if that is possible.

1. Near an equilibrium an ODE can be simplified by expanding the equations to first order in the deviations of the variables from their equilibrium values. The resulting system is linear. Formally for $\dot{x} = f(x)$, set $x = x_{eq} + \delta x$, and use $f(x_{eq}) = 0$ to find

$$\delta\dot{x} = f(x_{eq} + \delta x) \approx f(x_{eq}) + \frac{\partial f}{\partial x}(x_{eq})\delta x + \cdots \approx A\delta x.$$

Here you must remember that x is a vector, and so the matrix A has elements $a_{ij} = \partial f_i \big/ \partial x_j$. Carry out this expansion for the equilibria you found in Exercise 1.2 and compute the 4×4 matrix A for each case.

2. Find the general solution to the two-dimensional linear system for the Hamiltonian (1.29) and show that the phase portrait given in Figure 1.8 is correct.

3. Show that if T is a bounded linear operator and is invertible, then

$$\left\|T^{-1}\right\| \geq \frac{1}{\|T\|}.$$

4. Suppose T is a bounded linear operator on X that leaves a complete, vector subspace $E \subset X$ invariant (i.e., whenever $v \in E$ then $T(v) \in E$). Show that e^T also leaves E invariant. (For the definition of complete, normed space, see Sec. 3.2.)

5. In this problem we will prove the following lemma.

Lemma 2.38. *A linear operator T is bounded if and only if it is continuous.*

(a) Recall that continuity means that if $x_n \to x$, then $T(x_n) \to T(x)$. First show that linearity implies that if T is continuous at $x = 0$, then it is continuous everywhere. (*Hint*: Consider a sequence $x_n \to 0$ and then use superposition to find the limit of $T(x_n + y)$.)

(b) Suppose T is bounded; then show that $x_n \to 0$ implies that $|T(x_n)| \to 0$. Argue that this implies T is continuous.

(c) Suppose T is not bounded; then show that it is not continuous at $x = 0$. (*Hint*: Argue that there is sequence x_n such that $|T(x_n)| > n\,|x_n|$. Now let $y_n = x_n / n\,|x_n|$). Argue that you have proved that if T is continuous, it is bounded.

6. Here we will prove the next lemma by two methods.

Lemma 2.39. $e^{tA}e^{tB} = e^{t(A+B)}$ *for all* $t \in \mathbb{R}$ *if and only if* $[A,B] = 0$.

(a) Using the series definition of the exponential, expand the product on the left and group like powers of t. Use the binomial theorem $(x+y)^n = \sum_{j=0}^{n}\binom{n}{j}x^j y^{n-j}$ to identify the result with the series for the exponential on the right. This proves the "if" part.

(b) An alternative, more elegant, method is based on the fundamental solution theorem, Theorem 2.16. First, show that if $[A,B]=0$, then the matrix function $F(t) = Be^{tA}$ satisfies the same initial value problem as the function $G(t) = e^{tA}B$. Use uniqueness to conclude that $F = G$.

(c) Now let $\Phi(t) = e^{tA}e^{tB}$ and find the differential equation for Φ. Using the commutation relation in (b), show that it solves the same initial value problem as $e^{t(A+B)}$. Again use uniqueness to obtain the "if" part.

(d) To prove the "only if" part, assume that $e^{t(A+B)} = e^{tA}e^{tB}$ and argue that differentiation with respect to time gives $F(t) = G(t)$. By differentiating again, finally show that $[A,B]=0$.

7. Find all possible values of a, b, c, and d for which the 2×2 matrix $\begin{pmatrix} a & b \\ c & d\end{pmatrix}$ is

(a) semisimple,

(b) nilpotent.

8. Prove that if A and B are *similar* matrices (i.e., $B = P^{-1}AP$ for some nonsingular matrix P), then they have the same eigenvalues, and these have the same multiplicities.

9. Here we will prove, without relying on Theorem 2.20, that the maximum nilpotency for an $n \times n$ matrix is n.

(a) First show that if N is nilpotent, then all of its eigenvalues are zero. (*Hint*: Consider $N^j v$ where v is an eigenvector.)

(b) Use the Cayley–Hamilton theorem (2.41) to show that if all the eigenvalues of a matrix are zero, then it is nilpotent with nilpotency at most n.

(c) Construct examples of 3×3 nilpotent matrices with nilpotencies 1, 2, 3.

10. Classify the dynamics of the following ODEs $\dot{x} = Ax$ using the categories of §2.2 and §2.6. Sketch the phase portraits.

(a) $A = \begin{pmatrix} 1 & 3 \\ 2 & -1\end{pmatrix}$, (b) $A = \begin{pmatrix} 4 & 2 \\ -3 & 1\end{pmatrix}$, (c) $A = \begin{pmatrix} 0 & 2 \\ -1 & 2\end{pmatrix}$, (d) $A = \begin{pmatrix} 2 & 1 \\ -1 & 0\end{pmatrix}$,

(e) $A = \begin{pmatrix} 4 & -2 \\ 2 & -1\end{pmatrix}$, (f) $A = \begin{pmatrix} 1 & -2 \\ 1 & 4\end{pmatrix}$, (g) $A = \begin{pmatrix} 2 & 1 \\ 1 & 1\end{pmatrix}$.

11. The Routh–Hurwitz criterion determines whether the roots of a polynomial have all negative real parts and hence is a test for asymptotic stability. Here we consider just the three-dimensional case, the cubic generalization of (2.18):

$$p(\lambda) = \lambda^3 - \tau\lambda^2 + \sigma\lambda - \delta.$$

Show that all the roots of p have negative real parts, $\mathrm{Re}(\lambda_i) < 0$, if and only if $\tau < 0$ and $\tau\sigma < \delta < 0$. (*Hints*: Use the symmetric polynomials $\tau = \lambda_1 + \lambda_2 + \lambda_3, \sigma = \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3$, and $\delta = \lambda_1\lambda_2\lambda_3$; the value $p(0)$; and of the critical points, λ_c where $p'(\lambda_c) = 0$. Consider separately the cases of all real eigenvalues and of a complex pair.)

12. The following lemma is useful for the proof of the uniqueness of the decomposition into semisimple and nilpotent matrices in Theorem 2.23.

    Lemma 2.40. *If A and B are semisimple matrices, then there is a matrix P that simultaneously diagonalizes both A and B if and only if $[A,B] = 0$.*

    (a) Prove the "only if" part of the lemma. (*Hint*: Assume that there is such a P and consider the quantity $P^{-1}ABP$.)

    (b) Prove the "if" part of the lemma. (*Hint*: Assume that $[A,B] = 0$, and that $v_i,\ i = 1,\ldots,n_k$, are a basis for the eigenspace E_k of A with eigenvalue λ_k. Show that $Bv_i = \sum_{j=1}^{n_k} c_{ij}v_j$. Find a suitable linear combination $w_i = \sum_{j=1}^{n_k} d_{ij}v_j$ so that w_i is simultaneously an eigenvector of both A and B.)

    (c) Prove that if A and B are commuting semisimple matrices, then $A+B$ and AB are semisimple.

13. Compute e^A for each of the following matrices:

    (a) $\begin{pmatrix} 2 & 0 \\ 0 & -1 \end{pmatrix}$, (b) $\begin{pmatrix} 2 & 3 \\ 0 & 1 \end{pmatrix}$,

    (c) $\begin{pmatrix} 2 & -1 & -1 \\ -1 & 0 & -1 \\ 1 & 3 & 4 \end{pmatrix}$, (d) $\begin{pmatrix} \ln\left(\frac{16}{27}\right) + 8\pi i & 6\ln\left(\frac{2}{3}\right) + 12\pi i \\ 2\ln\left(\frac{3}{2}\right) - 4\pi i & \ln\left(\frac{81}{8}\right) - 6\pi i \end{pmatrix}$.

    (*Hint*: The eigenvalues of (d) are $\ln 2 + 2\pi i$ and $\ln 3$.)

    (e) Explain why the result of (d) is related to the fact that $e^{2\pi i} = 1$.

14. Solve the initial value problem $\frac{dx}{dt} = Ax, \quad x(0) = x_o$ with

    $$A = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 2 & -4 \\ 1 & 4 & 2 \end{pmatrix}$$

    and $x_o = (1,1,0)^T$.

15. Compute e^{tA} for the matrices

    (a) $\begin{pmatrix} -1 & -2 \\ 4 & 3 \end{pmatrix}$, (b) $\begin{pmatrix} 5 & -2 \\ 2 & 1 \end{pmatrix}$, (c) $\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$, (d) $\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 3 & 1 \end{pmatrix}$

    (e) $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}$, (f) $\begin{pmatrix} -2 & -1 & 1 \\ 0 & -2 & 2 \\ 0 & 0 & -2 \end{pmatrix}$, (g) $\begin{pmatrix} 2 & 0 & 1 \\ 1 & 2 & -2 \\ -1 & 0 & 2 \end{pmatrix}$.

16. Find the stable, unstable, and center subspaces of the linear systems defined by matrices

$$\text{(a)} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad \text{(b)} \begin{pmatrix} 2 & 1 \\ 0 & -4 \end{pmatrix}, \quad \text{(c)} \begin{pmatrix} -2 & 3 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}.$$

17. A forced, linear system can often be modeled by an equation of the form

$$\dot{x} = Ax + f(t), \; x(0) = x_o. \tag{2.57}$$

(a) One way to solve this system is to move the Ax term to the left of the equation and multiply both sides by the "integrating factor" e^{-tA}, and realize that the left-hand side is a total derivative. Using this, find the general solution to (2.57). What assumptions on $f(t)$ are required?

(b) If $f(t) = b$ is a constant and A is nonsingular, then the integral in your solution can be done explicitly. Compare this solution with that obtained by the method of subtracting the equilibrium $x^* = A^{-1}b$ from x.

(c) Suppose that $f(t) = b$ is a constant, and $b \in \text{rng}(A)$, but A is singular. Can you simplify the solution you found in (a) to that of the form in (b)?

(d) Discuss the case that $b \notin \text{rng}(A)$.

18. Consider the general nonautonomous linear matrix ODE

$$\frac{d}{dt}\Phi = A(t)\Phi, \quad \Phi(0,0) = I. \tag{2.58}$$

(a) An obvious guess for the solution is the exponential (2.48). Expand this exponential in a series, keeping terms to second order (quadratic terms in A). Substitute the result into the ODE and show that it is generally not correct to this order.

(b) Show that the problem terms you computed in (a) will vanish if $[A(s), A(t)] = 0$ for all $t, s \in \mathbb{R}$.

(c) Indeed, supposing that $[A(s), A(t)] = 0$, show that (2.48) is a solution to *all* orders in the exponential series.

19. Consider a special case of the ODE (2.58) with

$$A(t) = \begin{pmatrix} 1 & t \\ 0 & -1 \end{pmatrix}.$$

(a) Show that the commutator $[A(s), A(t)] \neq 0$ when $t \neq s$. Thus the solution should not be given by the exponential (2.48).

(b) Compute the exponential of the matrix $B(t) = \int_0^t A(s)ds$ explicitly, and show that it *does not* solve the ODE (2.58). (*Hint*: It is easy to find eigenvectors and eigenvalues of B for each t.)

(c) Find the true solution Φ to (2.58) for this case by first finding the general solution to $\dot{x} = A(t)x$. (*Hint*: It is easy to solve for the second component, $x_2(t)$.)

20. Compute a logarithm of the matrices

$$\text{(a)}\begin{pmatrix}1/2 & 5/4\\ 5 & 1/2\end{pmatrix},\quad \text{(b)}\begin{pmatrix}2 & 1\\ 0 & 2\end{pmatrix},\quad \text{(c)}\begin{pmatrix}-2 & 3\\ 0 & -2\end{pmatrix},\quad \text{(d)}\begin{pmatrix}-5 & -8\\ 2 & 3\end{pmatrix}.$$

If these matrices were monodromy matrices for a periodically time-dependent linear system, classify the stability of the system.

21. Although any nonsingular matrix, A, has a logarithm, it is possible that all values of $\ln A$ are complex. In this problem you will prove that A^2 has a real logarithm.

   (a) Show that if A has all real eigenvalues, then A^2 has positive eigenvalues. Use this to prove that $\ln(A^2)$ can be taken to be real.

   (b) Show that if A has a complex eigenvalue $\lambda = re^{-i\theta}$, with multiplicity one, then it is similar to a block diagonal matrix with a 2×2 block $B = r\begin{pmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{pmatrix}$ on the diagonal. Show that this matrix has a real logarithm, $\ln B = \ln r\, I + \begin{pmatrix}0 & \theta\\ -\theta & 0\end{pmatrix}$.

   (c) Finally, suppose that A has complex eigenvalues of multiplicity larger than one. Show that the semisimple part of A can still be put in a real form with 2×2 blocks as in (b).

   (d) Putting together (a), (b), and (c), prove that for any nonsingular matrix A, A^2 has a real logarithm.

22. Prove that if A is nonsingular, then $\det(A) = e^{\operatorname{tr}(\ln A)}$.

23. Consider your adopted quadratic ODEs (recall Exercise 1.10) in their reduced form (i.e., set all "nonessential parameters" to +1—keep the signs as given in the original equation). Call the reduced variables $\xi = (x, y, z)^T$ for simplicity.

   (a) Choose one of the equilibria, ξ^*, of your system. Define a new dynamical vector $\delta\xi = \xi - \xi^*$, and find the differential equations for $\delta\xi$.

   (b) Linearize the equations for $\delta\xi$ by dropping all the nonlinear terms. You will obtain a linear system $\delta\dot{\xi} = A\delta\xi$.

   (c) For the "chaotic value" of the parameters, classify the stability of your system by finding the eigenvalues of the matrix A and the spaces E^s, E^u, and E^c (perhaps numerically if the cubic, characteristic polynomial is not easily factored).

Chapter 3
Existence and Uniqueness

> *An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.* (Pierre-Simon Laplace, *Essai philosophique sur les probabilities,* 1814)

The goal of this chapter is to prove the fundamental theorems of existence and uniqueness for solutions of ordinary differential equations (ODEs). As Laplace most eloquently stated, if one knows precisely the initial condition for the system of ODEs that describe the dynamics of a closed universe, it is possible—in principle—to construct the solution. The analysis in this chapter will also lead to a review of some fundamental mathematical machinery, such as the contraction-mapping theorem. We will find this theorem of use in many more exotic locales in later chapters.

The hypotheses of the existence theorem reveal some surprising requirements on the vector field for the solution of an ODE to exist and be unique. The theorem also makes clear that solutions of differential equations need not exist for all time, but only over limited intervals, even when the vector field is perfectly well behaved.

3.1 ▪ Set and Topological Preliminaries

Some of the basic notions from topology are essential in the study of dynamical systems, so we pause for a moment to collect some notation and recall a few of the ideas from set theory and topology that will be needed. Some common mathematical notation will be often used:

▹ $\mathbb{R}$ is the real line, and $\mathbb{R}^+ = \{x \in \mathbb{R} : x > 0\}$.[13]

▹ $\mathbb{R}^n$ is n-dimensional Euclidean space.

▹ $\mathbb{Z}$ is the set of all integers.

[13] The notation $\{a : b\}$ means the set of all a such that b holds. So, for example, $\{x \in \mathbb{R} : |x| < 1\}$ is the set of all real numbers between minus one and plus one.

- ▷ $\mathbb{N}$ is the set of natural numbers (the nonnegative integers and zero).

The Euclidean norm is denoted by $|x|$. A solid ball of radius r around a point x_o is the *closed* set

$$B_r(x_o) = \{x \in \mathbb{R}^n : |x - x_o| \leq r\}. \tag{3.1}$$

We will be dealing primarily with differential equations on $\mathbb{R}^n$. The slightly more general case of "manifolds" is based on this analysis, since a manifold is a space that, locally, looks like Euclidean space.[14] Some common manifolds are

- ▷ $\mathbb{S}^d = \left\{(x_1, x_2, \ldots, x_{d+1}) : x_1^2 + x_2^2 + \cdots + x_{d+1}^2 = 1\right\}$ is the d-dimensional sphere; it is the boundary of a unit ball in $d+1$ dimensions;
- ▷ $\mathbb{T}^d$ is the d-dimensional torus; and
- ▷ $\mathbb{S}^1 = \mathbb{T}^1$ is the circle.

Note that the "common sphere" embedded in three-dimensional space is denoted $\mathbb{S}^2$, the two-sphere, since it is a two-dimensional set. Additional notations include

- ▷ $\in$, an *element* of a set;
- ▷ $\subset$, a *subset*;
- ▷ $\cap$, *intersection*;
- ▷ $\cup$, *union*; and
- ▷ $A \setminus B = \{x \in A : x \notin B\}$, the *relative complement* of B in A.

For example, $3 \in \{5, 3, 2\}$, $\{0, 1, 2\} \cap \{2, 1\} = \{1, 2\}$, $\bigcap_{j>0} \{n \in \mathbb{N} : n < j\} = \{0\}$, $\bigcup_{j=3}^{10} \{n \in \mathbb{N} : n < j\} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, and $\{1, 2, 3\} \setminus \{3, 4, 6\} = \{1, 2\}$. The qualifier symbols are denoted

- ▷ $\exists$, meaning *there exists*, and
- ▷ $\forall$, meaning *for all*.

A topological space is characterized by a collection of *open sets*. For Euclidean space the basic open sets are the open balls, $\{x : |x - x_o| < r\}$. By definition, a union of any number of open sets is declared open, as is the intersection of any finite number of open sets. Similarly, the basic closed sets are the closed balls $B_r(x_o)$. By definition, the intersection of any number of closed sets is closed, as well as the union of finitely many closed sets. The word *neighborhood* is used to denote some arbitrary set that encloses a designated point:

- ▷ *neighborhood*: N is a neighborhood of a point x if N contains an open set containing x.

Note that a neighborhood can be open or closed, but it must contain some open set. This excludes calling the set $\{x\}$ a neighborhood of x; however, for any $r > 0$, the closed ball $B_r(x)$ is a neighborhood of x. Often, we think of neighborhoods as being "small" sets in some sense, but this is not a requirement.

[14] Manifolds will be discussed more completely in Chapter 5.

Convergence

Sequences are ordered lists; for example, $S = \left\{ s_j \in \mathbb{R}^K : j \in \mathbb{N} \right\}$. A sequence is *convergent* if it approaches a fixed value, s^*, i.e., if $\left| s_j - s^* \right| \to 0$ as $j \to \infty$. Formally, we say that the sequence S converges if for every $\varepsilon > 0$ there is an $N(\varepsilon)$ such that whenever $n \geq N(\varepsilon)$, then $|s_n - s^*| < \varepsilon$.

More generally a point x_o is called a *limit point* of the sequence x_j if there is a subsequence $\left\{ s_{k_i} : k_i \in \mathbb{N},\ k_j \to \infty \text{ as } j \to \infty \right\}$ that converges to x^*. For example, the sequence $\left\{ (-1)^j : j \in \mathbb{N} \right\}$ has both 1 and -1 as limit points. With this notion we can formally define a

▷ *closed set*: A set S is closed if it includes all of its limit points; that is, if s^* is a limit point of some sequence in S, then $s^* \in S$.

The *closure* of a set S, denoted $\overline{S}$, is the union of the set and the limit points of every sequence in S. The interior of S, denoted $\text{int}(S)$, is the set of points in S that have a neighborhood that is also in S.

The boundary of a set S is denoted ∂S; it is the set of points that are in $\overline{S}$ but not in $\text{int}(S)$, i.e., $\partial S = \overline{S} \setminus \text{int}(S)$. Consequently $\partial B_1(0) = \mathbb{S}^{n-1}$ is the unit sphere. A set is *bounded* if it is contained in some ball $B_r(0)$; otherwise, it is *unbounded.* A set that is both closed and bounded is called a

▷ *compact set*: A closed and bounded set in a finite-dimensional space is compact.

One of the basic theorems of topology states that every compact set, $C \subset \mathbb{R}^n$, can be covered by a finite number of balls: $C \subset \bigcup_{i=1}^{N} B_{r_i}(x_i)$.[15] Another important result relates compactness to convergent subsequences.

Theorem 3.1 (Bolzano–Weierstrass). *Suppose every element of a sequence is contained in a compact set. Then the sequence has at least one limit point.*

Uniform Convergence

If a sequence depends upon parameters—the elements of the sequence are functions, say, $f_n(x)$—then there is another notion of convergence that is important, that of

▷ *uniform convergence*: A sequence $\{f_n(x) : n \in \mathbb{N}, x \in E\}$ converges uniformly if for every $\varepsilon > 0$ there is an $N(\varepsilon)$ that can be chosen independently of x, such that whenever $n \geq N(\varepsilon)$, then $|f_n(x) - f^*(x)| < \varepsilon$ for all $x \in E$.

A sequence of continuous functions that converges need not converge to a continuous function; however, if the convergence is uniform, then the limit is continuous. Recall that a continuous function $f \in C^0(E)$ is one for which for every $x \in E$ and every $\varepsilon > 0$, there is a $\delta(\varepsilon, x)$ such that $|f(y) - f(x)| < \varepsilon$ whenever $|y - x| < \delta$. Here we allowed the distance δ to depend on both the accuracy ε and the choice of point x.

[15] Indeed, this is usually taken as the more general definition of compact: a set for which *every* open cover has a finite subcover.

Lemma 3.2. *The limit of a uniformly convergent sequence of continuous functions is continuous.*

Proof. Let $u(x)$ denote the limit of $u_n(x)$; we must show that there is a $\delta(\varepsilon, x)$ such that $|u(y)-u(x)| < \varepsilon$ whenever $|y-x| < \delta$. Insert four new terms that sum to zero into this norm:

$$\begin{aligned} |u(y)-u(x)| &= |u(y)-u_n(y)+u_n(y)-u_n(x)+u_n(x)-u(x)| \\ &\leq |u(y)-u_n(y)|+|u_n(y)-u_n(x)|+|u_n(x)-u(x)|. \end{aligned}$$

Since by assumption u_n converges uniformly, then for any $x \in E$ and any $\varepsilon/3$ there is an N such that $|u_n(x)-u(x)| < \varepsilon/3$ whenever $n \geq N$. Moreover, since u_n is continuous for any fixed n, there is a $\delta(\varepsilon, x)$ such that $|u_n(x)-u_n(y)| < \varepsilon/3$ for each $y \in \text{int}(B_\delta(x))$. As a consequence,

$$|u(y)-u(x)| < \frac{\varepsilon}{3}+\frac{\varepsilon}{3}+\frac{\varepsilon}{3} = \varepsilon,$$

so u is continuous. □

There is also a uniform version of continuity:

▷ *uniform continuity*: A function f is uniformly continuous on E if for every $x \in E$ and every $\varepsilon > 0$, there is a $\delta(\varepsilon)$, independent of x, such that $|f(y)-f(x)| < \varepsilon$ whenever $|y-x| < \delta$.

It is not too hard to show that when E is a compact set, then every continuous function on E is also uniformly continuous (see Exercise 2).

A generalization of Lemma 3.2 is easily obtained: if each of the elements of a uniformly convergent sequence is uniformly continuous, then the limit is also uniformly continuous.

3.2 ▪ Function Space Preliminaries

A function $f : D \to R$ is a map from its domain D to its range R; that is, given any point $x \in D$, there is a unique point $y \in R$, denoted $y = f(x)$. In our applications the domain is often a subset of Euclidean space, $E \subset \mathbb{R}^n$, and the range is $\mathbb{R}^n$; in this case, $f : E \to \mathbb{R}^n$ is given by n components $f_i(x_1, x_2, \ldots, x_n), i = 1, 2, \ldots, n$. The set of functions denoted $C(E)$ or $C^0(E)$ consists of those functions on the domain E whose components are continuous. Colloquially we say "f is C^0" if it is a member of this set. If it is necessary to distinguish different ranges, the set of continuous functions from D to R is denoted $C^0(D,R)$; the second argument is often omitted if it is obvious. When E is an open subset of $\mathbb{R}^n$, a function $f : E \to \mathbb{R}^n$ is differentiable at a point $x \in E$ if there exists an $n \times n$ matrix $Df(x)$ such that

$$\lim_{|h|\to 0} \frac{1}{|h|}|f(x+h)-f(x)-Df(x)h| = 0 \tag{3.2}$$

When it exists, this matrix is unique and is called the *Jacobian matrix*

$$Df(x) \equiv \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}. \tag{3.3}$$

Conversely, Df exists at x when all its partial derivatives, $\partial f_i / \partial x_j$, exist *and* are continuous in a neighborhood of x. We say f is $C^1(E)$—continuously differentiable—if the elements of $Df(x)$ are continuous on the open set E. Colloquially we will say that f is *smooth* when it is a C^1 function of its arguments.

Spaces of functions, like $C(E)$ and $C^1(E)$, are examples of infinite dimensional linear spaces, or *vector spaces*. Just as for ordinary vectors (recall §2.1), linearity means that whenever f and $g \in C(E)$, then so is $c_1 f + c_2 g$ for any (real) scalars c_1 and c_2. Much of our theoretical analysis will depend upon convergence properties of sequences of functions in some such space. To talk about convergence it is necessary to define a *norm* on the space; such norms will be denoted by $||f||$ to distinguish them from the finite dimensional Euclidean norm $|x|$. We already met one such norm, the operator norm, in (2.23). For continuous functions, the *supremum* or *sup-norm*, defined by

$$||f|| \equiv \sup_{x \in E} |f(x)| \tag{3.4}$$

will often be used. For example, if $E = \mathbb{R}$, and $f = \tanh(x)$, then $||f|| = 1$. Other norms include the L_p norms,

$$||f||_p = \left(\int_E |f(x)|^p \, dx \right)^{1/p},$$

but these will not have much application in this book. This formula becomes the sup-norm in the limit $p \to \infty$, which is why the sup-norm is also called the L_∞ norm and is often denoted $||f||_\infty$.

Metric Spaces

A normed space is an example of a metric space. A metric is a distance function $\rho(f,g)$ that takes as arguments two elements of the space and returns a real number, the "distance" between f and g. A metric must satisfy the three properties

1. $\rho(f,g) \geq 0$, and $\rho(f,g) = 0$ only when $f \equiv g$ (positivity),
2. $\rho(f,g) = \rho(g,f)$ (symmetry), and
3. $\rho(f,h) \leq \rho(f,g) + \rho(g,h)$ (triangle inequality).

Associated with any norm $||f||$ is a metric defined by $\rho(f,g) = ||f - g||$. Therefore, a normed vector space is also a metric space; however, metric spaces need not be vector spaces, since in a metric space there is not necessarily a linear structure.

A sequence of functions f_n that are elements of a metric space X is said to *converge* to f^* if $\rho(f_n, f^*) \to 0$ as $n \to \infty$. Since the distance $\rho(f_n, f^*)$ is simply a number, the usual definition of limit can be used for this convergence. Note that the norm (3.4) bounds the Euclidean distance: if we use

$$\rho(f,g) = \|f-g\|_\infty, \text{ then } |f(x)-g(x)| \le \rho(f,g).$$

Thus, convergence of a sequence of functions f_n in norm implies that the sequence of points $f_n(x)$ converges *uniformly*.

Another notion often used to discuss convergence is that of

> ▷ *Cauchy sequence*: Given a metric space X with metric ρ, a sequence $f_n \in X$ is Cauchy if, for every $\varepsilon > 0$ there is an $N(\varepsilon)$ such that $\rho(f_n, f_m) < \varepsilon$ whenever $m, n \ge N(\varepsilon)$.

Informally, a Cauchy sequence satisfies

$$\rho(f_n, f_m) \to 0 \text{ as } m, n \to \infty,$$

where m and n approach infinity independently. One advantage of this idea is that the value of the limit of a sequence need not be known in order to check if it is Cauchy.

It is easy to see that every convergent sequence is a Cauchy sequence. However, it is not necessarily true that every Cauchy sequence converges.

Example 3.3. Consider the sequence of functions $f_n(x) = \sin(nx)/n \in C[0,\pi]$, the continuous functions on the interval $[0,\pi]$. This sequence converges to $f^* = 0$ in the sup norm because

$$\|f_n - 0\| = \frac{1}{n} \to 0.$$

The sequence is also Cauchy because

$$\|f_m - f_n\| \le \frac{1}{n} + \frac{1}{m} \le \frac{2}{N} < \frac{3}{N} \quad \forall m, n \ge N.$$

Thus for any $\varepsilon > 0$, we may choose $N(\varepsilon) = 3/\varepsilon$ so that the difference is smaller than ε. ■

Example 3.4. Consider the sequence $f_n = \sum_{j=1}^{n} \frac{x^j}{j}$ of functions in $C(-1,1)$. Assuming that $m > n$, then

$$\|f_m - f_n\| = \left\| \sum_{j=n+1}^{m} \frac{x^j}{j} \right\| = \sum_{j=n+1}^{m} \frac{1}{j} \ge \int_n^m \frac{dy}{y+1} = \ln\left(\frac{m+1}{n+1}\right),$$

since the supremum of $|x^j|$ on $(-1,1)$ is 1. This does not go to zero for m and n arbitrarily large but otherwise independent. For example, selecting $m = 2N+1$ and $n = N$ gives a difference larger than $\ln 2$. Consequently, the sequence is not Cauchy.

Note that for any fixed $x \in (-1,1)$ this sequence converges to the function $-\ln(1-x)$; however, it does not converge uniformly since the number of terms needed to obtain an accuracy ε depends upon x. Thus in the sense of the L_∞ norm, the sequence does not converge on $C(-1,1)$. ■

A space X that is nicely behaved with respect to Cauchy sequences is called a

▷ *complete space*: A metric space X is complete if every Cauchy sequence in X converges to an element of X.

For the case of linear spaces a complete space is called a

▷ *Banach space*: A complete normed linear space is a Banach space.

Some spaces, like the interval $[a,b]$ with the Euclidean norm, are complete, and some, like an open interval, are not. The space $C(E)$ with the L_∞ norm is complete when E is compact.

Theorem 3.5 (Completeness of $C^0(E)$). *The space $C^0(E,\mathbb{R}^n)$ with the sup-norm (3.5), is complete when E is compact.*

Proof. Suppose that $f_n \in C^0(E,\mathbb{R}^n)$, $n \in \mathbb{N}$, is a Cauchy sequence. We must show that it converges to a function $f^* \in C^0(E)$. We first claim that for any fixed $x_o \in E$, $f_n(x_o)$ converges as a sequence in $\mathbb{R}^n$. Indeed since the sequence f_n is Cauchy, for all $\varepsilon > 0$, there is an $N(\varepsilon)$ such that whenever $n,m \geq N$, $|f_n(x_o)-f_m(x_o)| \leq ||f_n - f_m|| < \varepsilon$; therefore, the sequence $f_n(x_o) \in \mathbb{R}^n$ is Cauchy. Any Cauchy sequence in $\mathbb{R}^n$ is bounded, and thus contained in a compact set. Consequently, Theorem 3.1 implies there is a convergent subsequence, $f_{n_k}(x_o) \to f^*(x_o)$ as $n_k \to \infty$. In fact the entire sequence converges to the same point because whenever $n \geq N(\varepsilon)$,

$$|f_n(x_o)-f^*(x_o)| \leq |f_n(x_o)-f_{n_k}(x_o)| + |f_{n_k}(x_o)-f^*(x_o)| < 2\varepsilon,$$

since for a subsequence $n_k \geq n$. Moreover, since N depends only on ε, the convergence is uniform.

Define the function $f^* : E \to \mathbb{R}^n$ to be the limit obtained as above for each $x \in E$. We must show that f^* is continuous. Since E is compact, each f_n is uniformly continuous (recall Exercise 2). Thus for any $\varepsilon > 0$, there is a $\delta(\varepsilon)$ so that $|f_n(x) - f_n(y)| < \varepsilon$ for any $x, y \in E$ such that $|x-y| < \delta$. Choosing $N(\varepsilon)$ as before, whenever $n \geq N$, we have

$$|f^*(x)-f^*(y)| \leq |f^*(x)-f_n(x)| + |f_n(x)-f_n(y)| + |f_n(y)-f(y)| < 5\varepsilon.$$

Since this is true for any $\hat{\varepsilon} = 5\varepsilon > 0$, f^* is (uniformly) continuous. □

Completeness of a space depends on the choice of norm. The next example shows that the space of continuous functions is not complete in the L_2-norm.

Example 3.6. Let $f_n \in C[-1,1]$ be the sequence

$$f_n = \begin{cases} 1, & x \leq 0, \\ \dfrac{1}{1+nx}, & x > 0. \end{cases} \tag{3.5}$$

With the L_2-norm, this sequence limits to the function $f = \begin{cases} 1, & x \leq 0 \\ 0, & x > 0 \end{cases}$ because

$$||f_n - f||_2 = \left(\int_0^1 \frac{dx}{(1+nx)^2} \right)^{1/2} = \frac{1}{\sqrt{1+n}} \underset{n\to\infty}{\to} 0.$$

Note that the limit, however, is not in $C[-1,1]$. In the L_2-norm, the sequence is also a Cauchy sequence:

$$||f_n - f_m||_2^2 = \int_0^1 \left(\frac{1}{1+nx} - \frac{1}{1+mx}\right)^2 dx \le \int_0^1 \left[\left(\frac{1}{1+nx}\right)^2 + \left(\frac{1}{1+mx}\right)^2\right] dx$$
$$= \frac{1}{1+n} + \frac{1}{1+m} \le \frac{2}{N},$$

for any $n, m \ge N$—of course every convergent sequence is Cauchy. As a consequence, the L_2-norm is not complete on the space $C[-1,1]$. ■

Example 3.7. Now consider the sequence (3.5) with the sup-norm. In this case the sequence does not converge to f, since

$$||f_n - f|| = \max\left\{ |1-1|, \sup_{x\in(0,1]} \left|\frac{1}{1+nx}\right| \right\} = \max\{0,1\} = 1.$$

Accordingly, the very definition of convergence can depend upon the choice of norm. Moreover, this sequence is not Cauchy in the sup-norm:

$$||f_n - f_m|| = \sup_{x\in[0,1]} \left| \frac{1}{1+nx} - \frac{1}{1+mx} \right| = \sup_{x\in[0,1]} \left| \frac{m-n}{(1+nx)(1+mx)} x \right|.$$

Differentiation of this expression shows that its maximum occurs at $x = (mn)^{-1/2}$ and has the value $||f_n - f_m|| = \left|\frac{\sqrt{m}-\sqrt{n}}{\sqrt{m}+\sqrt{n}}\right|$ that does not approach zero for all $m, n \ge N \to \infty$. For example, $||f_{4N} - f_N|| = \frac{1}{3}$. This proves that the sequence is not Cauchy. ■

Since complete spaces are so important, it is worthwhile noting that given one such space we can construct more of them by taking subsets, as in the next lemma.

Lemma 3.8. *A closed subset of a complete metric space is complete.*

Proof. To see this, first note that if $f_j \in Y \subset X$ is a Cauchy sequence on a complete space X, then $f_j \to f^* \in X$. Moreover, since f is a limit point of the sequence f_j, and a closed set Y includes all of its limit points, then $f \in Y$. □

The issues that we have discussed are rather subtle and worthy of a second look—see Exercises 1-3.

Contraction Maps

We have already used the concept of an operator, or map, $T : X \to X$, from a metric space to itself in Chapter 2: an $n \times n$ matrix is a map from $\mathbb{R}^n$ to itself. We will have many more occasions to use maps in our study of dynamical systems, including the proof of the existence and uniqueness theorem in §3.3. This proof will rely heavily on what is perhaps the most important theorem in all of analysis, Stefan Banach's 1922 fixed-point theorem.

Theorem 3.9 (Contraction Mapping). *Let $T : X \to X$ be a map on a complete metric space X. The map T is a contraction if there exists a constant $c < 1$ such that for all*

$f,g \in X$,

$$\rho(T(f),T(g)) \leq c\rho(f,g). \tag{3.6}$$

In this case T has a unique fixed point, $f^* = T(f^*) \in X$.

Proof. The result will be obtained iteratively. Choose an arbitrary $f_0 \in X$. Define the sequence $f_{n+1} = T(f_n)$. We wish to show that f_n is a Cauchy sequence. Applying (3.6) repeatedly yields

$$\rho(f_{n+1},f_n) = \rho\left(T(f_n),T(f_{n-1})\right) \leq c\rho(f_n,f_{n-1}) \leq c^2\rho(f_{n-1},f_{n-2}) \leq \cdots \leq c^n\rho(f_1,f_0).$$

Therefore, for any integers $m > n$, the triangle inequality implies that

$$\rho(f_m,f_n) \leq \sum_{i=n}^{m-1}\rho(f_{i+1},f_i) \leq \sum_{i=n}^{m-1}c^i\rho(f_1,f_0) = \frac{1-c^{m-n}}{1-c}c^n\rho(f_1,f_0) \leq Kc^n,$$

where $K = \rho(f_1,f_0)/(1-c)$. Since $c < 1$, then for any $\varepsilon > 0$ there is an N such that for all $m,n \geq N$, $\rho(f_m,f_n) \leq Kc^N < \varepsilon$. This implies that the sequence f_n is Cauchy and, since X is complete, that the sequence converges.

The limit, f^*, is a fixed point of T. Indeed, suppose that N is large enough so that $\rho(f_n,f^*) < \varepsilon$ for all $n > N$, then

$$\begin{aligned}\rho(T(f^*),f^*) &\leq \rho(T(f^*),f_{n+1}) + \rho(f_{n+1},f^*)\\ &= \rho(T(f^*),T(f_n)) + \rho(f_{n+1},f^*) < (c+1)\varepsilon.\end{aligned}$$

Because this is true for any $\varepsilon > 0$, the distance is zero and $T(f^*) = f^*$.

Finally, we show that the fixed point is unique. Suppose to the contrary that there are two fixed points $f \neq g$. Then, $\rho(f,g) = \rho(T(f),T(g)) \leq c\rho(f,g)$. Since $c < 1$, this is impossible unless $\rho(f,g) = 0$, but this contradicts the assumption $f \neq g$; thus, the fixed point is unique. □

Example 3.10. Consider the space $C^0(\mathbb{S})$ of continuous functions on the circle with circumference one, i.e., continuous functions that are periodic with period one: $f(x+1) = f(x)$. For any $f \in C^0(\mathbb{S})$ define the operator

$$T(f)(x) = \tfrac{1}{2}f(2x).$$

Note that $T(f) \in C^0(\mathbb{S})$, and, using the sup-norm, that $||T(f) - T(g)|| = 1/2||f - g||$; therefore, T is a contraction map on $C^0(\mathbb{S})$. What is its fixed point? According to the theorem, any initial function will converge to the fixed point under iteration. For example, let $f_0(x) = \sin(2\pi x)$. Then $f_1(x) = 1/2\sin(4\pi x)$, and after n steps, $f_n = \frac{1}{2^n}\sin(2^{n+1}\pi x)$. A previous example showed that this sequence converges to $f^* = 0$ in the sup-norm. By Theorem 3.9, $f^* = 0$ is the unique fixed point. ■

Example 3.11. As a slightly more interesting example, consider the same function space but let

$$T(f)(x) = \cos(2\pi x) + \tfrac{1}{2}f(2x). \tag{3.7}$$

Note that T decreases the sup-norm by a factor of $1/2$ as before, so it is still contracting. For example, the sequence starting with the function $f_0(x) = \sin(2\pi x)$ is

$$f_1(x) = \cos(2\pi x) + \tfrac{1}{2}\sin(4\pi x),$$

$$f_2(x) = \cos(2\pi x) + \tfrac{1}{2}\cos(4\pi x) + \tfrac{1}{4}\sin(8\pi x),$$

$$f_j(x) = \sum_{n=0}^{j-1}\frac{\cos(2^{n+1}\pi x)}{2^n} + \frac{1}{2^j}\sin(2^{j+1}\pi x).$$

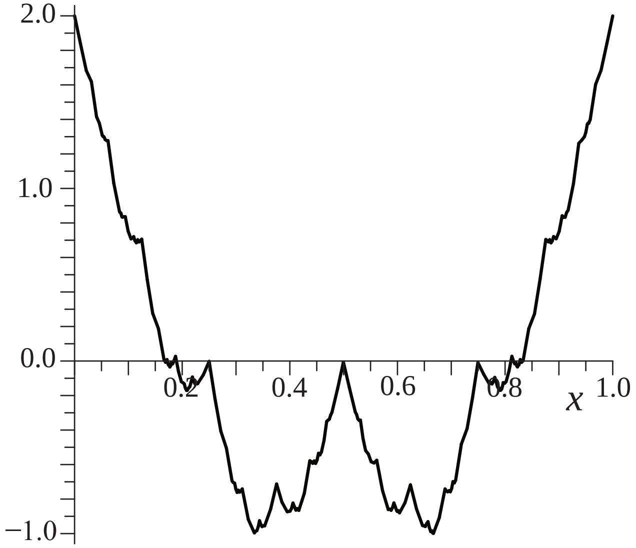

Figure 3.1. *The fixed point of the operator* (3.7).

The last term goes to zero in the sup-norm, and by the contraction-mapping theorem, the result is guaranteed to be unique and continuous. The fixed point is not an elementary function but is easy to graph; see Figure 3.1; it is an example of a Weierstrass function (Falconer 1990). ■

Lipschitz Functions

Another ingredient that we will need in the existence and uniqueness theorem is a notion that is stronger than continuity but slightly less stringent than differentiability:

> ▷ *Lipschitz*: Suppose (X, ρ_X) and (Y, ρ_Y) are metric spaces with the indicated distance functions. A function $f : X \to Y$ is Lipschitz if for all $x_1, x_2 \in X$, there is a K such that
>
> $$\rho_Y(f(x_1), f(x_2)) \leq K\rho_X(x_1, x_2). \tag{3.8}$$
>
> The smallest such K is called the Lipschitz constant for f on X.

For example if $X = Y = \mathbb{R}$ and $\rho_X = \rho_Y$ is the Euclidean metric, then when f is Lipschitz, the slope of the chord connecting $(x, f(x))$ and $(y, f(y))$ is at most K in absolute value. More generally the graph of a Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}^n$ must be contained in the cone $\{(\xi, \eta) : |\eta - f(x)| \leq K|\xi - x|\}$, for each vertex $(x, f(x))$.

The Lipschitz property implies more than continuity, but less than differentiability.

Lemma 3.12. *A Lipschitz function is uniformly continuous.*

Proof. For each ε, set $\delta = \varepsilon/K$. Then whenever $\rho_X(x_1, x_2) < \delta$, (3.8) implies that $\rho_Y(f(x_1), f(x_2)) < \varepsilon$. Consequently f is continuous at x_1, say, and moreover, is uniformly so because δ is chosen independently of x_1. □

If the space X is unbounded, then the assumption that f is Lipschitz is very strong. For example, $f = x^2$ is not Lipschitz on $\mathbb{R}$, even though it is Lipschitz on every bounded interval (a,b). A weaker notion is

> ▷ *locally Lipschitz*: $f : X \to Y$ is locally Lipschitz if for every $x \in X$, there is a neighborhood $N(x)$ such that f is Lipschitz on $N(x)$.

Note that for a locally Lipschitz function, the constant K can vary with the point and indeed may become arbitrarily large. Nevertheless, the restriction of any such function to a compact set is (globally) Lipchitz.

Lemma 3.13. *Suppose that $f : X \to Y$ is locally Lipschitz and $A \subset X$ is compact. Then f is Lipschitz on A.*

Proof. By assumption, for each $x_j \in A$, there is a ball $B_{r_j}(x_j)$ on which f is Lipschitz with constant K_j. Since A is compact, there is a finite collection of these balls—even if we decrease their radii by a factor of two—that covers A, i.e., $A \subset \bigcup_{j=1}^{n} B_{r_j/2}(x_j)$. Since n is finite, there exist positive constants $K = \max_j K_j$ and $\delta = \frac{1}{2}\min_j r_j$. To show f is Lipschitz on A, consider two points $\xi, x \in A$.

First, if $\rho_X(\xi,x) \le \delta$ then there is j such that $\xi, x \in B_{r_j}(x_j)$; indeed there is a j for which $x \in B_{r_j/2}(x_j)$ since these balls cover A, and the triangle inequality then implies

$$\rho_X(\xi,x_j) \le \rho_X(\xi,x) + \rho_X(x,x_j) \le \delta + \frac{r_j}{2} + \le r_j.$$

In this case, we have $\rho_Y(f(\xi), f(x)) \le K\rho_X(\xi,x)$, by the local Lipschitz assumption.

On the other hand, if $\rho_X(\xi,x) > \delta$, then we argue as follows. Since A is compact and f is continuous, there is an M such that $\rho_Y(f(\xi), f(x)) \le M$. Setting $\hat{K} = \max(K, M/\delta)$, we now have $\rho_Y(f(\xi), f(x)) \le \hat{K}\delta \le \hat{K}\rho_X(\xi,x)$. Thus f is Lipschitz on A with constant $\hat{K}$. □

When a function is continuously differentiable on an open set in $\mathbb{R}^n$, it is locally Lipschitz. This is a simple consequence of the following lemma.

Lemma 3.14. *Suppose that $A \subset \mathbb{R}^n$ is compact and convex and $f \in C^1(A,\mathbb{R}^n)$. Then f is Lipschitz with constant $K = \max_{x\in A} ||Df||$.*

Proof. Since A is convex, the points on a line between any two points $x, y \in A$, are also in A. Accordingly, $\xi(s) = x + s(y-x) \in A$ when $0 \le s \le 1$. Therefore

$$f(y) - f(x) = \int_0^1 \frac{d}{ds}(f(\xi(s)))ds = \int_0^1 Df(\xi(s))(y-x)ds,$$

which amounts to a mean value theorem. Since A is compact and the norm of the Jacobian $||Df||$ is continuous, it has a maximum value K, as defined in the lemma. Thus

$$|f(y) - f(x)| \le \int_0^1 ||Df(\xi(s))||\, |y-x|\, ds \le K|y-x|. \tag{3.9}$$

This is exactly the promised condition. □

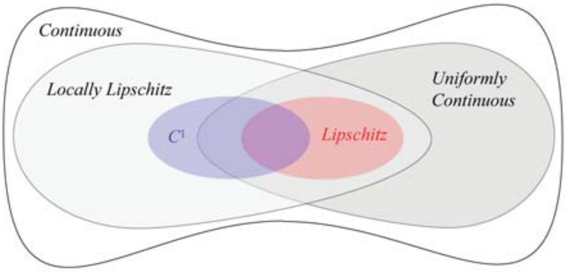

Figure 3.2. *Venn diagram relating continuity, differentiability and the Lipschitz property.*

Corollary 3.15. *If $E \subset \mathbb{R}^n$ is open and $f \in C^1(E,\mathbb{R}^n)$, then f is locally Lipschitz.*

Proof. For any $x \in E$, there is an r such that $B_r(x) \subset E$. Since $B_r(x)$ is compact and convex, then Lemma 3.14 applies. $\square$

Some of the relationships between continuous, Lipschitz, and smooth functions are summarized in Figure 3.2.

Example 3.16. The function $f(x) = |x|$ is continuous and Lipschitz on $\mathbb{R}$. It is obviously C^1 on $\mathbb{R}^+$ and $\mathbb{R}^-$, and if x and y have the same sign, then $|f(x)-f(y)| = |x-y|$. So the only thing to be checked is the Lipschitz condition when the points have the opposite sign. Although this is obvious geometrically, let us be formal: let $x > 0 > y$; then $|f(x)-f(y)| = ||x|-|y|| \le x + |y| = |x-y|$. So f is Lipschitz with $K = 1$.

However, the function $f(x) = \sqrt{x}$ is not Lipschitz on $[0,1]$ even though it is uniformly continuous. For example, choosing $x = 4\varepsilon$, $y = \varepsilon$, we then have

$$|f(x)-f(y)| = \sqrt{x} - \sqrt{y} = \sqrt{\varepsilon} = \frac{\varepsilon}{\sqrt{\varepsilon}} = \frac{1}{\sqrt{\varepsilon}}\frac{4\varepsilon - \varepsilon}{3} = \frac{1}{3\sqrt{\varepsilon}}|x-y|,$$

so that as ε becomes small, the needed value of $K \to \infty$. $\blacksquare$

All these formal definitions have been given to provide us with the tools to prove that solutions to certain ODEs exist and, if the initial values are given, are unique. We are finally ready to begin this analysis.

3.3 ▪ Existence and Uniqueness Theorem

Before we can begin to study properties of the solutions of differential equations, we must discover if there *are* solutions in the first place: do solutions exist? The foundation of the theory of differential equations is the theorem proved by the French analyst Charles Emile Picard in 1890 and the Finnish topologist Ernst Leonard Lindelöf in 1894 that guarantees the existence of solutions for the initial value problem

$$\dot{x} = f(x), \;\; x(t_o) = x_o \tag{3.10}$$

for a solution $x : \mathbb{R} \to \mathbb{R}^n$ and vector field $f : \mathbb{R}^n \to \mathbb{R}^n$. We were able to avoid this discussion in Chapter 2 because linear differential equations can be solved explicitly. Since this is not the case for more general ODEs, we must now be more careful.

The main tool that we will use in developing the theory is the reformation of the differential equation as an integral equation. Formally integrating the ODE in (3.10) with respect to t yields

$$x(t) = x_o + \int_{t_0}^{t} f(x(\tau))d\tau. \tag{3.11}$$

This equation, while correct, is actually not a "solution" for $x(t)$ since in order to find x, the integral on the right-hand side must be computed—but this requires knowing x.

Begin by imagining that (3.11) can be solved to find a function

$$x : J \to \mathbb{R}^n$$

on some time interval $J = [t_o - a, t_o + a]$. Since the integral in (3.11) does not require differentiability of $x(\tau)$, we will only assume that it is continuous. However, given that such a solution $x(t)$ to (3.11) exists, then it is actually a solution to the ODE (3.10).

Lemma 3.17. *Suppose $f \in C^k(E, \mathbb{R}^n)$ for $k \geq 0$ and $x \in C^0(J, E)$ is a solution of* (3.11)*. Then $x \in C^{k+1}(J, E)$ and is a solution to* (3.10)*.*

Proof. First note that if x solves (3.11), then $x(t_o) = x_o$. Since $x \in C^0(J)$, the integrand $f(x(\tau))$ is also continuous and so the right-hand side of (3.11), being the integral of a continuous function, is C^1; consequently, the left-hand side of (3.11), $x(t)$, is also differentiable. By the fundamental theorem of calculus, the derivative of the right-hand side is precisely $f(x(t))$; thus, $\dot{x} = f(x)$. Now we use induction to show that $x \in C^{k+1}$. Suppose that $x \in C^j(J)$ for some $0 \leq j \leq k$. If follows that $f(x(\tau)) \in C^j$ and the right-hand side of (3.11) is C^{j+1}, so $x \in C^{j+1}(J)$. □

Equation (3.11) can be viewed as an operator acting on functions $u(t)$

$$T(u) = x_o + \int_{t_o}^{t} f(u(\tau))d\tau. \tag{3.12}$$

So that (3.12) is well defined, u must be chosen from some suitable function space, for example, continuous functions. Since every solution to (3.10) obeys (3.11), Lemma 3.17 implies that a continuous function $x(t)$ solves our initial value problem *if and only if* it is a fixed point of T:

$$x^* = T(x^*).$$

This leads to a strategy called *Picard iteration*. Starting with a test function, $u_0(t)$, apply T to obtain what we hope is a "better" approximation, $u_1 = T(u_0)$. Repeatedly applying this operator generates a sequence

$$u_j(t) = T(u_{j-1})(t), \; j = 1, 2, \ldots, \tag{3.13}$$

that, with any luck, will converge to a fixed point, $u_j \to x^*$. Moreover if the limit is continuous, then Lemma 3.17 implies it is C^1 and consequently a solution of the ODE.

Example 3.18 (Picard Iteration). Consider the one-dimensional, linear, initial value problem $\dot{x} = rx$, $x(0) = x_o$ for $x \in \mathbb{R}$. The operator (3.12) becomes

$$T(u) = x_o + r \int_0^t u(s)ds.$$

Choose some more or less arbitrary starting function, say, the constant function $u_0(t) = x_o$. The first approximation is then $u_1(t) = T(u_0)(t) = x_o + r x_o t$, and then

$$u_2(t) = x_o + r\int_0^t x_o(1+rs)\,ds = x_o\left(1 + rt + \frac{r^2}{2}t^2\right),$$

$$u_3(t) = x_o + r\int_0^t x_o\left(1 + rs + \frac{r^2}{2}s^2\right)ds = x_o\left(1 + rt + \frac{r^2}{2}t^2 + \frac{r^3}{6}t^3\right).$$

It is clear that this sequence generates the power series for the well-known solution $x_o e^{rt}$. More interestingly, even if another initial guess is used we still find the same solution. For example, choosing $u_0(t) = t$, then

$$u_1(t) = x_o + r\int_0^t s\,ds = \left(x_o + \frac{r}{2}t^2\right),$$

$$u_3(t) = x_o + r\int_0^t \left(x_o + \frac{r}{2}s^2\right)ds = x_o(1+rt) + \frac{r^2}{6}t^3.$$

After each iteration, the "bad" term—the last term in these expressions—moves to a higher power, leaving behind the series for the exponential. ■

Of course, we have not shown that Picard iteration converges in general—and indeed it is possible to choose sufficiently badly behaved initial functions u_0 so that this iteration does not converge. To prove an existence theorem, it is necessary to restrict the class of allowed functions to a set on which the vector field in (3.10) is bounded; that is, functions for which the speed $|f(u(t))|$ is bounded. To use the contraction-mapping theorem it is essential that this subset be a complete set. Luckily, as Lemma 3.8 implies, any closed subset of a complete space is still complete. Thus we need only to choose a closed subset of C^0 to be assured of completeness.

Once we do this, we find that the contraction-mapping theorem is perfectly suited for proving the existence of a fixed point of the operator (3.12) and hence of solutions to the initial value problem (3.10).

Theorem 3.19 (Picard–Lindelöf Existence and Uniqueness). *Suppose that for $x_o \in \mathbb{R}^n$, there is a $b > 0$ such that $f : B_b(x_o) \to \mathbb{R}^n$ is Lipschitz with constant K. Then the initial value problem* (3.10) *has a unique solution, $x(t)$ for $t \in J = [t_o - a, t_o + a]$, provided that*

$$a = b/M \text{ where } M = \max_{x \in B_b(x_o)} |f(x)|. \tag{3.14}$$

Note that a and M both depend upon the values of b and x_o. Since this is such an important theorem, we will give three separate proofs!

Proof (using contraction). For the first proof will use the contraction-mapping theorem. This proof does not quite give the "optimal" bound on a, but it has the advantage of being elegant.

To begin, define a complete metric space on which the contraction map is to operate. This will consist of all continuous functions $x(t)$ that do not leave the ball $B_b(x_o)$ during the time interval J:

$$V = C^0(J, B_b(x_o)). \tag{3.15}$$

This is a closed set since the range $B_b(x_o)$ is closed. If we use the sup-norm metric (3.4), then V is a closed subset of the complete space $C^0(J, \mathbb{R}^n)$ and hence it is complete by Lemma 3.8.

Since f is Lipschitz on $B_b(x_o)$, it is continuous; therefore, the integral of $f(x(t))$, for any $x(t) \in V$, is a continuous function of t. We will show the operator T defined by (3.12) maps V into itself and is a contraction. This would imply, using the contraction-mapping theorem, that T has a unique fixed point. By Lemma 3.17 any such fixed point is a solution to (3.10) and conversely every solution to (3.10) is a fixed point of T.

When $x \in V$, then $T(x)$ is automatically continuous since f is continuous, but we must show that $T(x) \in V$. Now, since $f \in C^0(B_b(x_o))$, and $B_b(x_o)$ is a closed subset, then f is bounded on $B_b(x_o)$, so that M can be defined as in (3.14). If $t_o \le t \le t_o + a$,[16]

$$|T(x)(t) - x_o| \le \int_{t_0}^{t} |f(x(\tau))|\, d\tau \le M\,|t - t_0| \le Ma;$$

the final inequality holds also when $t_o - a \le t \le t_o$. The right-hand side can be no larger than b, so we must have $a \le b/M$. In this case $T(x)(t) \in B_b(x_o)$, so that $T(x) \in V$. To show that T is a contraction mapping, consider two functions x and $y \in V$. Then, because f is Lipschitz,

$$\begin{aligned} |T(x)(t) - T(y)(t)| &\le \int_{t_0}^{t} |f(x(\tau)) - f(y(\tau))|\, d\tau \\ &\le K \int_{t_0}^{t} |x(\tau) - y(\tau)|\, d\tau \le Ka\,||x - y|| \end{aligned}$$

when $t \in J$. As a consequence, $||T(x) - T(y)|| \le c\,||x - y||$, where $c = Ka < 1$ providing $a < 1/K$. Consequently T is a contraction and has a unique fixed point provided that

$$a \le b/M \text{ and } a < 1/K. \tag{3.16}$$

By Lemma 3.17, the fixed point of T is the unique solution to the initial value problem (3.10) over the time interval J. □

The only deficiency of this first proof is the extra restriction on a in (3.16). For the second proof the contraction-mapping theorem will not be used, but iteration of T will still be the strategy. However, in this case, a special initial function is chosen. There are two advantages: the time interval J does not have the "artificial" restriction (3.16), and everything can be done explicitly, without appealing to the completeness of V. A disadvantage is that the proof is much longer.

Proof (using Picard). Define the operator T (3.12), the bound M (3.14), and the space V (3.15) as before. Let

$$u_0 \equiv x_o \text{ and } u_j \equiv T(u_{j-1}). \tag{3.17}$$

We will show that $u_j \in V$ by induction. To begin, the function u_0 is obviously in V. Now suppose that $u_{j-1} \in V$, then $u_j \in C^0(J)$, since it is defined by the integral. Moreover, the curve $u_j(t)$ is contained in the cone with vertex at (t_o, x_o) and slope M, as sketched in Figure 3.3, because

$$\left|u_j - x_o\right| \le \int_{t_0}^{t} \left|f(u_{j-1}(\tau))\right| d\tau \le M\,|t - t_0|. \tag{3.18}$$

[16] Note that $|\int_{t_o}^{t} f(t)dt| \le \int_{t_o}^{t} |f(t)|\,dt$ when $t \ge t_o$ but that the second integral should be reversed when $t \le t_o$. In many of the inequalities below, we will usually assume the former, but the end results will be valid in either case.

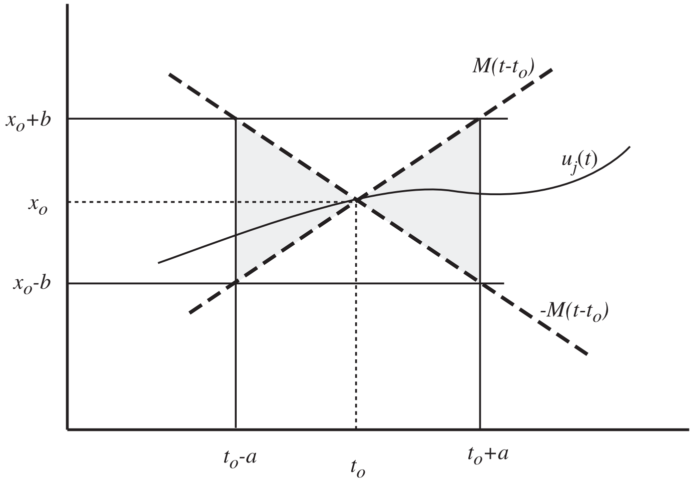

Figure 3.3. *Cone containing the solution to the Picard iteration* (3.17).

Thus $u_j(t) \in B_b(x_o)$, providing $a \le b/M$ as before. In consequence, each of the functions $u_j \in V$.

We want to show that the sequence u_j is convergent. Define

$$\Delta_j(t) = \left| u_j(t) - u_{j-1}(t) \right|.$$

The result (3.18) implies that $\Delta_1 \le M|t - t_o|$. Using this and the Lipschitz property of f gives a recursive bound on the Δ_j:

$$\Delta_{j+1} \le \int_{t_o}^{t} \left| f(u_j(\tau)) - f(u_{j-1}(\tau)) \right| d\tau \le K \int_{t_o}^{t} \Delta_j(\tau) d\tau.$$

Explicitly, the iteration of this recursion gives $\Delta_2 \le \tfrac{1}{2} MK|t - t_o|^2$, and hence

$$\Delta_j \le \frac{M}{K} \frac{(K|t - t_o|)^j}{j!} \le \frac{M}{K} \frac{(Ka)^j}{j!}.$$

To show that the sequence $u_n \to u$ as $n \to \infty$, write

$$u_n = x_o + \sum_{j=1}^{n} \left(u_j - u_{j-1} \right). \tag{3.19}$$

If this series converges as $n \to \infty$, then the sequence u_n does as well. Convergence of the series follows from the Weierstrass M-test. Since $\left| u_j - u_{j-1} \right| = \Delta_j$, and

$$\sum_{j=1}^{\infty} \Delta_j \le \sum_{j=1}^{\infty} \frac{M}{K} \frac{(Ka)^j}{j!} \le \frac{M}{K} \left(e^{Ka} - 1 \right),$$

then the Weierstrass M-test implies that the series (3.19) converges uniformly. Thus the sequence u_n is uniformly convergent, and Lemma 3.2 implies that the limit, $u(t)$, is continuous.

Finally, the limiting function $u(t)$ is a fixed point of T, as can be seen from

$$\begin{aligned}|u(t)-T(u)(t)| &\le \left|u(t)-u_n(t)+T(u_{n-1})(t)-T(u)(t)\right| \\ &\le |u(t)-u_n(t)|+K\int_{t_0}^{t}\left|u_{n-1}(s)-u(s)\right|ds.\end{aligned}$$

Since the sequence u_{n-1} converges, for any $\varepsilon > 0$ there is an N such that $\left|u(t)-u_{n-1}(t)\right| <$ ε for all $n \ge N$. Using this in the equation above gives

$$|u(t)-T(u)(t)| < \varepsilon(1+Ka).$$

Since this is true for any $\varepsilon > 0$, then $u = T(u)$ and is therefore a solution of (3.10).

It remains to show that $u(t)$ is unique.[17] Suppose $x(t) \in V$ is any solution, $T(x) = x$; then

$$|x(t)-x_o| = \left|\int_{t_o}^{t} f(x(s))ds\right| \le M\,|t-t_o|. \tag{3.20}$$

The implication is that $x \in V$ providing $t \in J$ and $a \le b/M$, as before. Now we show that x must be the same as $u(t)$ by showing that $\left|x-u_j\right| \to 0$. Using the same inductive procedure as before, and the inequality (3.20) with $x_o = u_0$, implies

$$\begin{aligned}\left|x(t)-u_j(t)\right| &\le \int_{t_0}^{t}\left|f(x(s))-f(u_{j-1}(s))\right|ds \\ &\le K\int_{t_0}^{t}\left|x(s)-u_{j-1}(s)\right|ds \le \frac{M}{K}\frac{[K\,|t-t_0|]^{j+1}}{(j+1)!}.\end{aligned}$$

Since this bound approaches zero as $j \to \infty$, then $u_j \to x$. However, $u_j \to u$ as well; therefore, $u = x$. □

Since the contraction mapping theorem yields a much more compact proof, it would be nice if it could be modified to yield the same result, that $a = b/M$. One way to accomplish this is to use a slightly different norm on V.

Proof (using Bielecki). We still define the space V (3.15) so that $x(t)$ must remain in $B_b(x_o)$, but now define a new norm, the Bielecki norm (Bielecki 1956), given by

$$||f||_L = \sup_{t\in J} e^{-L|t-t_o|}\,|f(t)|,$$

for some positive constant L. The continuous functions, $C^0(J, B_b(x_o))$, with this norm are also a complete space. Repeating the contraction mapping proof with this norm gives $a = b/M$ provided that $L \ge K$. We leave the completion of this proof as an exercise (see Exercise 8). □

Example 3.20. Existence can be proved when $f \in C^0$ without the additional Lipschitz assumption (Coddington and Levinson 1955, Theorem 1.1.2); however, for uniqueness the Lipschitz condition is needed. For example, consider the one-dimensional equation

$$\dot{x} = f(x) = |x|^{\alpha} \tag{3.21}$$

[17] This also follows easily from Grönwall's lemma—see §3.4 and Exercise 10.

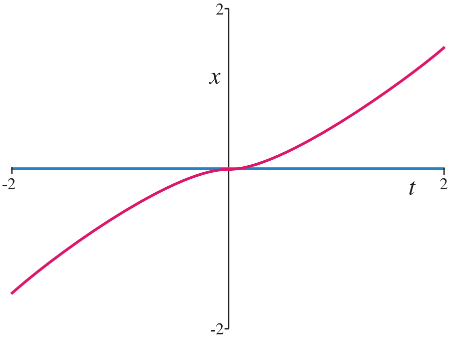

Figure 3.4. *Solutions of* $\dot{x} = |x|^{1/3}$, $x(0) = 0$.

for $0 < \alpha < 1$. Although f is continuous, it is not Lipschitz around $x = 0$, because there is no finite K for which $|x|^\alpha < K|x|$ for all $|x| < 1$, since that would require $K > |x|^{\alpha-1}$, but the right-hand side is unbounded. Moreover, there are at least *two* solutions to the initial value problem with $x(0) = x_o = 0$. First $x = 0$ is a solution, as can be seen by simply substituting it into the ODE. A second solution can be obtained by separation of variables as in (1.6) and carefully considering the signs (still assuming $1 - \alpha > 0$):

$$x(t) = \operatorname{sgn}(t)\big((1-\alpha)|t|\big)^{\frac{1}{1-\alpha}}.$$

The solution for $\alpha = 1/3$ is shown in Figure 3.4. Note that this solution satisfies the necessary condition that $\dot{x} \geq 0$ for all t. There are *infinitely* many other solutions of (3.21) that obey $x_o = 0$ as well—we leave it as a challenge to the reader to find them!
∎

Example 3.21. An equation such as (3.21) might seem artificial, but it is an approximate model for a physical system. Consider a mass slowly sliding on a ramp whose height is given by $y = H(x)$.[18] The motion is given by Newton's equations with gravity, drag, and normal forces. Denote the magnitude of the drag by F_d and of the force normal to the ramp by F_n. Let θ be the angle of the surface below the horizontal, v the speed, and $v(\cos\theta, -\sin\theta)$ the velocity; see Figure 3.5. Then the drag force is $-F_d(\cos\theta, -\sin\theta)$, and the Newtonian equations are

$$\begin{aligned} m\ddot{x} &= -F_d\cos\theta + F_n\sin\theta, \\ m\ddot{y} &= F_d\sin\theta + F_n\cos\theta - mg. \end{aligned}$$

For the heavily damped case, the inertial terms (those on the left-hand sides) can be neglected. Solving for the normal force in the y equation and substituting it into the x equation yields

$$0 \approx -F_d\cos\theta + (-F_d\sin\theta + mg)\tan\theta \;\Rightarrow\; F_d = mg\sin\theta.$$

We will assume that $F_d = \gamma v$; this is valid if the Coulomb friction between the mass and the ramp can be neglected and either the mass is embedded in a low Reynolds

[18] The equations of motion for the system without drag are most easily obtained by using a Lagrangian (see, for example, (9.29):

- $L(x,\dot{x}) = {}^1\!/\!_2 m(\dot{x}^2 + \dot{y}^2) - mgH(x) = {}^1\!/\!_2 m(1 + H'(x)^2)\dot{x}^2 - mgH(x)$.
- The Euler–Lagrange equations are $\left(1 + H'(x)^2\right)\ddot{x} + H'H''\dot{x}^2 = -gH'$.

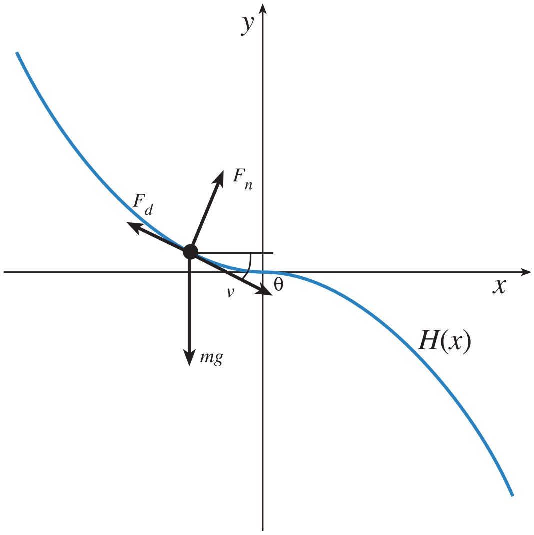

Figure 3.5. *Force diagram and function $H(x)$ from (3.22) for $\alpha = 0.8$.*

number fluid flow or the ramp is lubricated. Using $v = \sqrt{\dot{x}^2 + \dot{y}^2} = \dot{x}\sqrt{1+H'^2}$, where $H' = -\tan\theta$, gives the ODE

$$\dot{x} = -\frac{kH'(x)}{1+(H'(x))^2}$$

for a constant $k = mg/\gamma$. This has the form of (3.21) if we set

$$H'(x) = \frac{1}{2|x|^\alpha}\left(\sqrt{1-4x^{2\alpha}} - 1\right), \tag{3.22}$$

which limits to $-|x|^\alpha$ for small x; consequently, $H \approx \frac{-\operatorname{sgn}(x)}{1+\alpha}|x|^{1+\alpha}$. ■

Example 3.22. Differentiability of f is not required for the existence and uniqueness theorem to work. For example, if $f = 1 - |x|$, then f is Lipschitz on $\mathbb{R}$ with $K = 1$ (i.e., f is contained in a cone with slope 1), and the unique solution for $x_o = 0$ is

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

Note that this is C^1 at $t = 0$, as Theorem 3.19 guarantees. ■

The Picard–Lindelöf theorem only guarantees that the solution exists over the interval, $|t - t_o| \le a = b/M$. Since b appears in the numerator of $a(b, x_o)$, it appears that the interval of existence grows with b, and so if f is well behaved, then one should choose b very large. However, M is a function of b as well, so the choice of the optimal value for b is a little more complicated.

Example 3.23. Here we explore the maximal domain of existence implied by the Picard–Lindelöf theorem for the problem

$$\dot{x} = x^2, \quad x(0) = x_o. \tag{3.23}$$

For simplicity let us assume that $x_o > 0$. In the ball $B_b(x_o)$, $|f(x)| \leq (x_o + b)^2 = M$. According to Theorem 3.19, the solution can be proved to exist for

$$|t| \leq a = \frac{b}{(x_o + b)^2}.$$

To get the largest interval, compute the maximum of this function over possible choices of b; it is easy to see that this occurs when $b = x_o$, giving the maximal interval

$$|t| \leq \frac{1}{4x_o}. \tag{3.24}$$

Using separation of variables (recall (1.6)), the general solution to (3.23) is easily found:

$$x(t) = \frac{x_o}{1 - t x_o}, \tag{3.25}$$

which has an interval of existence $t \in (-\infty, 1/x_o)$. Note that this interval is asymmetric and is also larger than (3.24). Nevertheless, the fact that the actual interval of existence does not extend to $+\infty$ shows that the bound on this interval in Theorem 3.19 is not an artificial result of the methodology used in the proof. ∎

As shown by this example, it is important to keep in mind that a solution of a perfectly well-behaved nonlinear ODE need not exist for all time. This behavior is to be contrasted with that of the linear equations studied in Chapter 2 whose solutions, $e^{tA}x_o$, do exist for all time.

Recall from §1.2 that a nonautonomous equation $\dot{x} = f(t,x)$ can be converted into an autonomous one by adding t to the list of variables. Hence the Picard–Lindelöf theorem applies—providing f is Lipschitz in t as well as x. It is sometimes useful to have a special existence theorem for nonautonomous equations for which that assumption can be relaxed.

Theorem 3.24 (Nonautonomous Existence and Uniqueness). *Suppose $f : J \times B_b(x_o) \to \mathbb{R}^n$ is a uniformly Lipschitz function of x with constant K, and a continuous function of t on $J = [t_o - c, t_o + c]$. Then the initial value problem*

$$\dot{x} = f(t,x), \quad x(t_o) = x_o \tag{3.26}$$

has a unique solution for $t \in [t_o - a, t_o + a]$ with $a = \min(c, b/M)$, where

$$M = \max_{\substack{x \in B_b(x_o) \\ t \in J}} |f(t,x)|.$$

The assumption of "uniformly" Lipschitz means that K can be taken to be independent of t. The proof is left as an exercise (see Exercise 9). It can also be shown that continuity in t is not necessary for existence and uniqueness; see Exercise 10.

Example 3.25. When the elements of a matrix A are uniformly continuous functions of time on an interval J, the vector field of the nonautonomous linear equation $\dot{x} = A(t)x$ is Lipschitz in x with constant $K = \sup_{t \in J} \|A(t)\|$. Thus by Theorem 3.24, an initial value problem for this system has a unique solution. Moreover, it can be shown that the existence interval can be extended to J, see Exercise 12. This result was used in §2.8 in the development of Floquet theory. ∎

3.4 ▪ Dependence on Initial Conditions and Parameters

In this section we will discuss how a solution of an ODE depends on the choice of initial condition as well as on parameters in the vector field f. To do this, we need to add some notation to the solution to indicate its dependence on the initial value:

$$\dot{x} = f(x),\ x(0) = y \quad \Rightarrow \quad x(t) = u(t;y). \tag{3.27}$$

We use the semicolon to separate the primary argument of u from its implicit, secondary dependence on y. Using this notation, the initial condition becomes $u(0;y) = y$. Below, we will show that when $f \in C^1$, then $u \in C^1$ as a function of y. This permits the definition of the linearization of the flow about the solution using the Jacobian matrix (recall (3.3)):

$$\Phi(t;y) \equiv D_y u \equiv \frac{\partial u}{\partial y}. \tag{3.28}$$

Note that since $u(0;y) = y$, then $\Phi(0;y) = I$. If in addition $u \in C^2$, then the chain rule yields

$$\begin{aligned} \frac{d}{dt}\frac{\partial}{\partial y}u(t;y) &= \frac{\partial}{\partial y}\dot{u}(t;y) = \frac{\partial}{\partial y}f(u(t;y)) = Df(u(t;y))\frac{\partial}{\partial y}u(t;y), \\ \frac{d}{dt}\Phi &= Df(u(t;y))\Phi. \end{aligned} \tag{3.29}$$

This nonautonomous linear differential equation is the *linearization* or *variational equation*. We discussed a similar linear, matrix equation in (2.46), when we studied Floquet theory.

First we will show that it makes sense to think of u as a function of initial condition. To do that we must first show that solutions with nearby initial conditions can be defined on a common interval of time.

Lemma 3.26 (Neighborhood Existence). *Suppose that for a given $x_o \in \mathbb{R}^n$, there is a $b > 0$ such that $f : B_b(x_o) \to \mathbb{R}^n$ satisfies a Lipschitz condition with constant K, and that $M = \max_{x\in B_b(x_o)} |f(x)|$. Then the family of solutions $u(t;y)$ of (3.27) exists and is unique for each $y \in B_{b/2}(x_o)$ and $t \in [-a,a]$ providing $a < \min\{1/K, b/(2M)\}$.*

Proof. As in the contraction mapping proof of Theorem 3.19, define the closed set V of continuous functions (3.15), where $J = [-a,a]$, since we have set $t_o = 0$. We now label the operator T by the specific initial condition y:

$$T_y(u) = y + \int_0^t f(u(\tau))d\tau. \tag{3.30}$$

If $y \in B_{b/2}(x_o)$, and $u \in V$, then $T_y(u) \in V$ providing $a \le b/(2M)$, because

$$\left|T_y(u) - x_o\right| \le |y - x_o| + \int_0^t |f(u(s))|\,d\tau \le \frac{b}{2} + Ma \le b.$$

Moreover T_y is a contraction on V, as before, providing $a < 1/K$. In conclusion, for each $y \in B_{b/2}(x_o)$, T_y has a unique, continuous fixed point $u(t;y)$ that is a solution of the ODE for $t \in J$. □

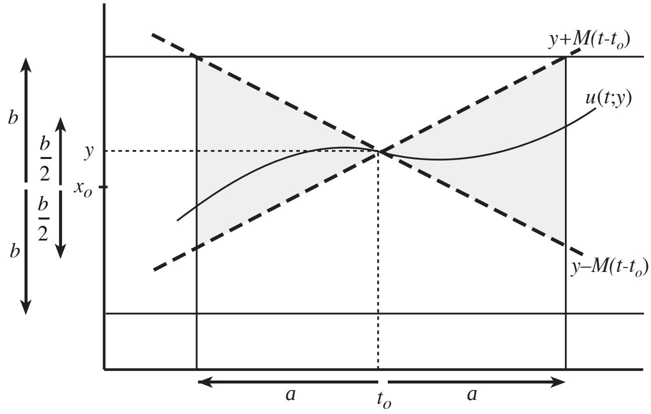

Figure 3.6. *Existence of solutions for initial conditions in a neighborhood of radius b about x_o requires using a smaller ball.*

Note that the initial conditions can be varied only over a ball with *half* the radius of the ball where f is assumed to be nice and that the solution can be shown to exist only for *half* of the time. This is because all the solutions must stay in B_b for all $|t| < a$; see Figure 3.6. We could adjust these factors of $1/2$, increasing one at the expense of decreasing the other. Finally, as before, the requirement that $a < 1/K$ could be eliminated with a little more work.

Example 3.27. Consider the initial value problem (3.23) taking as the central point $x_o = 0$ so that $f : B_b(0) \to \mathbb{R}$. The Lipschitz constant on this domain is $K = 2b$ and $|f|$ is bounded by $M = b^2$. The theorem then guarantees that a unique solution exists for $|y| < b/2$, providing $a < \min\left\{(2b)^{-1}, b\big/(2b^2)\right\} = (2b)^{-1}$. Note that the actual solution (3.25) for an initial condition $y \in B_{b/2}(0)$ blows up a time $t = 1/y$, so the shortest time occurs when $y = b/2$. Thus the true solution exists at least four times longer than the theorem gives us. ■

So far we have seen that the solution $u(t;y)$ exists for a range of initial conditions and is C^1 in t whenever the vector field f is Lipschitz. Our goal now is to discuss the smoothness of the dependence of $u(t;y)$ on y. For example, we will see that when the vector field is Lipschitz, u is a Lipschitz function of y.

The main tool used to prove this is a lemma about *differential inequalities*. Some care must be exercised here. For example, suppose that $f < g$; does it follow that $\dot{f} < \dot{g}$? A simple counterexample shows this is not true: $f(t) = \cos 3t$ and $g(t) = 2$. The converse statement is also not true: for example, if $f(t) = \sin t$ and $g(t) = 2t$, then indeed $\dot{f}(t) = \cos t < \dot{g}(t) = 2$, but note that $f > g$ when $t < 0$. In contrast, note that if $\dot{f} \le \dot{g}$, it follows that f increases less rapidly than g, so that $f(t) - f(t_o) \le g(t) - g(t_o)$ *provided* $t \ge t_o$. It is important, of course, that we assume that both $f, g \in C^1$ for this to work. This simple idea leads to the lemma proved by Thomas Grönwall in 1919.

Lemma 3.28 (Grönwall). *Suppose $g, k : [0,a] \to \mathbb{R}$ are continuous, $a > 0$, $k(t) \ge 0$,* *and g obeys the inequality*

$$g(t) \le G(t) \equiv c + \int_0^t k(s)g(s)ds \tag{3.31}$$

for all $0 \le t \le a$. *Then for all* $t \in [0,a]$,

$$g(t) \le c e^{\int_0^t k(s)ds}. \tag{3.32}$$

Proof. Since g and k are continuous, then G is C^1 and $G(0) = c$. Differentiation of G from (3.31) gives

$$\dot{G}(t) = k(t)g(t) \le k(t)G(t);$$

consequently, $\dot{G} - kG \le 0$. Multiplying by the positive "integrating factor" $e^{-\int_0^t k(s)ds}$ gives

$$e^{-\int_0^t k(s)ds}(\dot{G}(t) - kG) = \frac{d}{dt}\left(G(t)e^{-\int_0^t k(s)ds}\right) \le 0.$$

Integrating this inequality finally implies

$$G(t)e^{-\int_0^t k(s)ds} \le G(0) \;\Rightarrow\; G(t) \le c e^{\int_0^t k(s)ds}.$$

Since $g \le G$, we obtain (3.32). □

A similar lemma holds when c is allowed to be a function of time—see Exercise 11.

Grönwall's inequality makes the proof of our desired theorem very easy.

Theorem 3.29 (Lipschitz Dependence on Initial Conditions). *Let* $x_o \in \mathbb{R}^n$, *and suppose there is a* $b > 0$ *such that* $f : B_b(x_o) \to \mathbb{R}^n$ *is Lipschitz with constant K and that* $J = [-a,a]$ *is the common interval of existence for solutions* $u : J \times B_{b/2}(x_o) \to B_b(x_o)$. *Then* $u(t;y)$ *is uniformly Lipschitz in y with Lipschitz constant* e^{Ka}.

Proof. Suppose $u(t;y)$ and $u(t;z)$ are two solutions starting in $B_{b/2}(x_o)$. They have a common interval of existence J. When $t \in [0,a]$, the integral form (3.11) implies that

$$\begin{aligned} |u(t;y) - u(t;z)| &\le |y - z| + \int_0^t |f(u(\tau;y)) - f(u(\tau;z))| d\tau \\ &\le |y - z| + K\int_0^t |u(\tau;y) - u(\tau;z)| d\tau. \end{aligned}$$

This is precisely Grönwall's form (3.31) with $c = |y - z|$, and $k(t) = K$, so (3.32) becomes

$$|u(t;y) - u(t;z)| \le |y - z| e^{Kt}. \tag{3.33}$$

A similar inequality holds for $t \in [-a,0]$, giving our result. □

A slightly different proof is sketched in Exercise 10.

We can use this Lipschitz dependence of $u(t;y)$ on y to prove that when f is C^1, then u is also C^1 in y. The proof of this result requires a bit more work than the previous one.

Theorem 3.30 (Smooth Dependence on Initial Conditions). *Suppose $f : E \to \mathbb{R}^n$ is C^1 on an open set E. Then there is an $a > 0$ such that the solution $u(t;y)$ of (3.27) is a C^1 function of y for $t \in J = [-a,a]$.*

Proof. Since f is C^1 on an open set, it is locally Lipschitz by Corollary 3.15. Hence, for any initial condition $x_o \in E$ and any subset $B_b(x_o) \subset E$, f is Lipschitz on $B_b(x_o)$ with constant $K(x_o, b)$. By Lemma 3.26, there is a unique solution $u(t;y)$ for all $y \in B_{b/2}(x_o)$ on a common interval J. As in (3.29), define the fundamental matrix Φ to be the solution of the initial value problem

$$\frac{d}{dt}\Phi = Df(u(t;y))\Phi, \;\; \Phi(0;y) = I.$$

Just as we argued in §3.3, Φ exists by Theorem 3.24. Indeed, since $u \in C^1$ as a function of t and $Df(x)$ is a continuous function of x, the matrix $A(t) = Df(u(t;y))$ is a continuous function of t. Thus, there exist unique solutions to $\dot{x} = A(t)x, x(0) = \hat{e}_i$ for each of the unit vectors $\hat{e}_i$ on the interval J. These solutions define the columns of Φ.

Now suppose $|h| \le b/2$ and consider

$$g(t) \equiv |u(t;y+h) - u(t;y) - \Phi(t;y)h|.$$

We can insert the integral form (3.11) into each term in g to obtain

$$g(t) = \left| \int_0^t [f(u(\tau;y+h)) - f(u(\tau;y)) - Df(u(\tau;y))\Phi(\tau;y)h]\, d\tau \right|, \tag{3.34}$$

where we have simplified using $h = \Phi(0;y)h$. The goal is to show that $g \to 0$ as $h \to 0$ faster than $|h|$,[19] as this would imply that Φ is the derivative of $u(t;y)$.

The C^1 function $f : B_b(x_o) \to \mathbb{R}^n$ is uniformly C^1 by extending the argument sketched in Exercise 2. Thus Taylor's theorem implies

$$f(w) = f(u) + Df(u)(w-u) + R(u,w)$$

such that the remainder, R, is small, i.e., for any $\varepsilon > 0$, there is a $\delta(\varepsilon)$ such that if $u, w \in B_b(x_o)$ then

$$|R(u,w)| \le \varepsilon\, |w-u| \text{ when } |w-u| < \delta. \tag{3.35}$$

Consequently, using the operator norm (2.23) on the domain $B_b(x_o)$ we have

$$|f(w) - f(u)| \le \|Df\|\, |w-u| + |R|.$$

Using this in (3.34) gives for any $t \in [0,a]$,

$$\begin{aligned} g(t) &\le \int_0^t |f(u(\tau;y+h)) - f(u(\tau;y)) - Df(u(\tau;y))\Phi(\tau;y)h|\, d\tau \\ &\le \int_0^t \|Df\|\, |u(\tau;y+h) - u(\tau;y) - \Phi(\tau;y)h|\, d\tau + \int_0^t |R(u(\tau;y), u(\tau;y+h))|\, d\tau. \end{aligned}$$

[19] That is, we want to show that $g = o(h)$. See §4.4 for a definition of the "little oh" notation.

Now we use the Lipschitz bound (3.33) in the form $|u(\tau;y+h)-u(\tau;y)| \le he^{Ka}$ and the bound (3.35) to obtain

$$g(t) \le \int_0^t ||Df||\, g(\tau)d\tau + \varepsilon\, |h|\, e^{Ka} a, \tag{3.36}$$

providing $|h| \le r = \delta(\varepsilon)e^{-Ka}$. This restriction implies that for each $\varepsilon > 0$ we have a ball $B_r(y)$ of acceptable initial conditions for (3.36) but that $|h|$ can be arbitrarily small for any $\varepsilon > 0$. Equation (3.36) is again of the form of Grönwall's inequality (3.31). Since the Lipschitz constant K bounds $||Df||$ according to (3.9), we have

$$g(t) \le \varepsilon\, |h|\, a e^{Ka} e^{Kt}.$$

As this is true for any $\varepsilon > 0$, then $g(t)/|h| \to 0$ as $h \to 0$. This implies, recalling (3.2), that $u \in C^1$ as promised, and that its derivative is indeed Φ. □

It is interesting that the proof of Theorem 3.30 implies that when $f \in C^1$, the matrix $D_y u = \Phi$ solves the variational ODE (3.29) without the assumption that $u \in C^2$ that was used at the beginning of this section.

As a final result, suppose that the vector field depends continuously upon some parameters μ—for example, the n-body gravitational equations depend upon the masses of each body and the universal gravitational constant. We will show that the solution also depends continuously on μ. This result is related to the concept of *structural stability*: properties of the solutions should not change dramatically if the parameters of a system are varied. Such considerations are important in modeling since typically the values of parameters in the vector field will be uncertain.

Theorem 3.31 (Continuous Dependence on Parameters). *Suppose $f : B_b(x_o) \times B_r(\mu_o) \to \mathbb{R}^n$ has uniformly Lipschitz dependence on $x \in B_b(x_o)$ and is a uniformly continuous function of parameters $\mu \in B_r(\mu_o)$. Then the ODE $\dot{x} = f(x;\mu)$ has a unique solution $u(t;y,\mu)$ for $y \in B_{b/2}(x_o)$ that is a uniformly continuous function of μ for t in some interval J.*

Proof. Use the same idea as the Lipschitz dependence on initial conditions result, but now choose two solutions $u(t;y,\mu)$ and $u(t;y,\nu)$ with $\mu,\nu \in B_r(\mu_o)$. The usual arguments imply that these have a common interval of existence J. Moreover (suppressing the dependence upon the initial condition),

$$|u(t;\mu)-u(t;\nu)| \le \int_0^t |f(u(\tau;\mu);\mu)-f(u(\tau;\nu);\nu)|\,d\tau.$$

Write

$$\begin{aligned} f(u(\tau;\mu);\mu)-f(u(\tau;\nu);\nu) = {} & f(u(\tau;\mu);\mu)-f(u(\tau;\mu);\nu) \\ & + f(u(\tau;\mu);\nu)-f(u(\tau;\nu);\nu). \end{aligned}$$

Since f is uniformly continuous in μ, for any $\varepsilon > 0$ there is a $\delta(\varepsilon)$ such that whenever $|\mu-\nu| < \delta$, then $|f(x;\mu)-f(x;\nu)| < \varepsilon$. Using this gives

$$\begin{aligned} |u(t;\mu)-u(t;\nu)| &\le \varepsilon \int_0^t d\tau + \int_0^t |f(u(\tau;\mu);\nu)-f(u(\tau;\nu);\nu)|\,d\tau \\ &\le \varepsilon a + K\int_0^t |u(\tau;\mu)-u(\tau;\nu)|\,d\tau, \end{aligned}$$

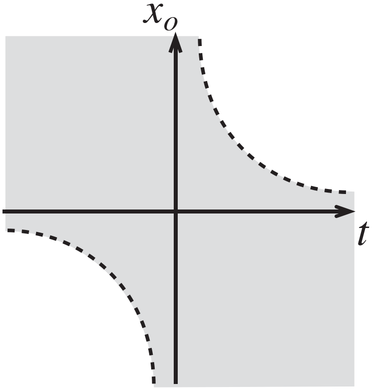

Figure 3.7. *Shaded region is the domain of existence for* (3.23).

which gives, by Grönwall's lemma (3.31), $|u(t;\mu)-u(t;\nu)| \le \varepsilon a e^{Ka}$ for any $\varepsilon > 0$ and $t \in J$. □

3.5 ▪ Maximal Interval of Existence

The existence theorem implies that when f is locally Lipschitz at a point x_o, the solution can be found on a closed interval $J = [t_o - a, t_o + a]$. Since the estimates used to obtain J are certainly not optimal, the true solution typically exists over a much larger interval. The largest such interval will be called the

> ▷ *maximal interval of existence*: The maximal interval of existence, $J(t_o, x_o)$, is the largest interval of time that includes t_o for which the solution, $x(t)$, to the initial value problem (3.10) exists.

If the solution can be found explicitly, then we can compute the maximal interval and find the maximal domain of existence in space–time.

Example 3.32. For the initial value problem (3.23), f is locally Lipschitz on $\mathbb{R}$ and the existence and uniqueness theorem applies for any x_o. The solution was given in (3.25) and exists for the maximal interval $(-\infty, x_o^{-1})$ if $x_o > 0$, for $(-\infty, \infty)$ if $x_o = 0$, and (x_o^{-1}, ∞) if $x_o < 0$. Note for each x_o that the interval is open but that it depends upon initial conditions: the domain of existence is an open subset of $\mathbb{R}^2$, as sketched in Figure 3.7. Moreover, as t approaches the boundary of the domain, $t \to x_o^{-1}$, we have $x(t) \to \infty$. ■

As indicated by the example, we now show that the maximal interval is indeed always open. Then we will show that if this interval is bounded, the solution must leave the domain of definition of the vector field f as it approaches the bounded endpoint of J.

Theorem 3.33 (Maximal Interval of Existence). *Let E be an open set and $f : E \to \mathbb{R}^n$ be locally Lipschitz. Then there is a maximal, open interval $J = (\alpha, \beta)$ containing t_o such that the initial value problem $\dot{x} = f(x)$, $x(t_o) = x_o$, has a unique solution $x : J \to E$.*

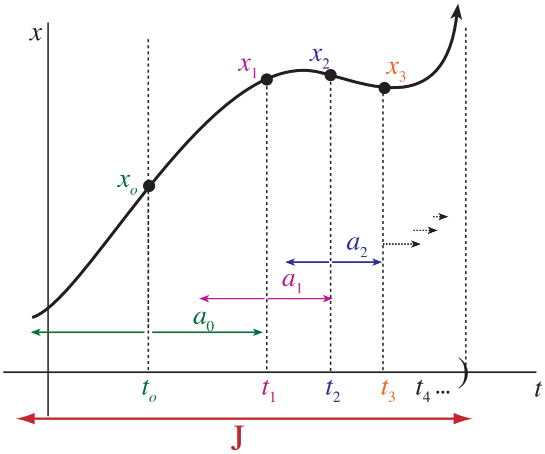

Figure 3.8. *Maximal interval of existence is constructed by repeatedly applying the existence theorem.*

Proof. For the purposes of this proof, denote the local solution to the initial value problem (3.10) by $x(t) = u(t; t_o, x_o)$. Theorem 3.19 guarantees that in each closed ball $B_{b_o}(x_o) \subset E$ there is a solution on an interval $J_0 = [t_o - a_0, t_o + a_0]$. Indeed, the theorem implies that $u(t; t_o, x_o) \in B_{b_0}(x_o) \subset E$ and is C^1; therefore,

$$\lim_{t \to t_o + a_0} u(t; t_o, x_o) = x_1 \in B_{b_0}(x_o),$$

and $x_1 \in E$ since E is open. Apply Theorem 3.19 again for the initial value problem with $x(t_1) = x_1$ on another ball $B_{b_1}(x_1) \subset E$ to find a new solution $u(t; t_1, x_1)$ on an interval $J_1 = [t_1 - a_1, t_1 + a_1]$ around $t_1 = t_o + a_0$. Note that $J_0 \cap J_1$ is not empty, and uniqueness implies that $u(t; t_o, x_o) = u(t; t_1, x_1)$ on their common interval of definition, $J_0 \cap J_1$.

In this way, as sketched in Figure 3.8, the solution can be extended to obtain a unique solution on a larger interval. Let J be the union of all such intervals and $x(t)$ be the unique solution just constructed on J.

The interval J must be open. Suppose to the contrary that J has a closed endpoint, for example, that $J = (\alpha, \beta]$. Then as before $x(\beta) \in E$, and so the solution can be extended to a larger interval; therefore, J is open. □

Example 3.34. Consider one final time our favorite example, $\dot{x} = x^2, x(0) = x_o > 0$ with solution (3.25). Recall that our computation for the existence and uniqueness theorem gave an interval of existence (3.24) with $a_0 = \frac{1}{4x_o}$ using the choice $b = x_o$, so that $J_0 = [-a_0, a_0]$. To apply Theorem 3.33 it would be necessary to calculate $x_1 = x(a_0)$; in general, this is impossible. In our case, however, the solution is known, and $x_1 = 4x_o/3$. Starting over with this value as the initial condition, $x(t_1) = x_1$, existence is guaranteed for at least an interval $J_2 = [t_1 - a_1, t_1 + a_1]$ with $a_1 = \frac{1}{4x_1} = \frac{3}{16x_o}$. Then $t_2 = t_1 + a_1 = \frac{1}{4x_o}\left(1 + \frac{3}{4}\right) = \frac{1}{x_o}\left(1 - (\frac{3}{4})^2\right)$, so that $x_2 = x(t_2) = \left(\frac{4}{3}\right)^2 x_o$. Continuing in this way for n steps gives existence up to a time

$$t_n = \frac{1}{4x_o}\left(1 + \frac{3}{4} + \cdots \left(\frac{3}{4}\right)^{n-1}\right) = \frac{1}{x_o}\left(1 - \left(\frac{3}{4}\right)^n\right).$$

This sequence converges to $t_\infty = 1/x_o$. Note, however, that existence is guaranteed only up to times equal to t_n for finite n, and in the limit the solution exists in the open interval with upper limit t_∞. If the same game is played for decreasing t, then the solution $x(t)$ becomes smaller in size, so that the successive intervals of existence do not decrease, but rather get larger. This is why the solution exists for all negative time. ■

Theorem 3.35 (Unboundedness). *Suppose E is an open set and $f : E \to \mathbb{R}^n$ is locally Lipschitz. Let $J = (\alpha, \beta)$ be the maximal interval of existence for* (3.10). *If β is finite, then for any compact set $K \subset E$ there is a $t \in [t_o, \beta)$ such that $x(t) \notin K$. Similarly, if α is finite, then for any compact set $K \subset E$ there is a $t \in (\alpha, t_o]$ such that $x(t) \notin K$.*

Proof. Consider the case that β is finite. Suppose the theorem were false. Then there would be a compact set K such that $x(t) \in K$ for all $t \in [t_o, \beta)$. Since f is continuous and K is compact, f is bounded on K: denote the bound as usual by $M = \max_{x \in K} |f(x)|$. The integral equation (3.11) implies that for any $t_1 \leq t_2 < \beta$, $|x(t_1) - x(t_2)| \leq M |t_1 - t_2|$. This means that if $t_j \in [t_o, \beta)$ is a sequence such that $t_j \to \beta$, then $x(t_j)$ is a Cauchy sequence. Since K is a closed subset of $\mathbb{R}^n$, every Cauchy sequence converges. Moreover, if $t_j \to \beta$ and $\tau_j \to \beta$ are any two such sequences, then $|x(t_j) - x(\tau_j)| \to 0$, so the limit

$$\lim_{t \to \beta} x(t) = x_1$$

exists. Consequently, if we define $x(\beta) = x_1$, then the function $x(t)$ is continuous on $[t_o, \beta]$. We can now apply Theorem 3.19 again using the initial condition $x(\beta) = x_1$ to show that there is a solution of (3.10) in some interval around β. By uniqueness, this is the same solution as $x(t)$ on the intersection of their existence intervals. Consequently, β is not the upper limit of the interval of existence, and we have reached a contradiction. The proof for α finite is similar. □

Corollary 3.36. *If β is finite, then either $\lim_{t \to \beta} x(t)$ does not exist or $\lim_{t \to \beta} x(t) \in \partial E$.*

Proof. Since β is finite, Theorem 3.35 implies that x leaves every compact set contained in E. If the limit exists, then it cannot be in E, since then the solution could be extended as before. However, since every point on $x(t)$ is in E for $t < \beta$, this means that $\lim_{t \to \beta} x(t) \in \partial E$. □

Example 3.37. Consider the initial value problem on $\mathbb{R}$:

$$\dot{x} = f(x) = \frac{1}{x}, \quad x(0) = x_o > 0.$$

Now the function f is well defined only for $x \neq 0$, so the entire real line cannot be used as the space E. Instead we have to choose the positive or negative half-line. Since $x_o > 0$, set $E = \mathbb{R}^+$. The solution to this problem is $x(t) = \sqrt{x_o^2 + 2t}$ and is contained in E for the maximal interval $J = (-x_o^2/2, \infty)$. In this case x does approach a limit at the lower endpoint of J:

$$\lim_{t \to -x_o^2/2} x(t) = 0 \in \partial E. \quad ■$$

Example 3.38. Consider the system on $\mathbb{R}^2$ defined by

$$\dot{x} = \frac{1}{1-y}, \quad \dot{y} = y,$$

with initial conditions $x(0) = x_o$ and $y(0) = y_o$. The differential equation is locally Lipschitz on any subset of the plane that does not include the line $y = 1$. A solution is found by first solving the equation for y to obtain $y(t) = y_o e^t$ and then substituting this into the x equation to give a separable equation in x and t. Solving this gives

$$x(t) = x_o + \ln\left|\frac{1-y_o}{e^{-t} - y_o}\right|$$

so that the solution to the system is

$$u(t; x_o, y_o) = \left(x_o + \ln\left|\frac{1-y_o}{e^{-t} - y_o}\right|, \; y_o e^t \right).$$

When $y_o > 1$, the solution is defined on the interval $t \in (-\ln y_o, \infty)$, and when $0 < y_o < 1$, it is defined on the interval $t \in (-\infty, -\ln y_o)$. Note that

$$u(t; x_o, y_o) \to (\infty, 1) \quad \text{as} \quad t \to -\ln y_o,$$

so the solution does not approach a limit at this endpoint of J. However, when $y_o \leq 0$, then the solution is defined on $(-\infty, \infty)$. ■

3.6 ▪ Exercises

1. Determine whether the following sequences are Cauchy in the space $C^0(\mathbb{R}, \mathbb{R})$ with the sup-norm:

   (a) $f_n(x) = \sin(2\pi n x)$,
   (b) $f_n = \tan(2\pi x/n)$,
   (c) $f_n(x) = \frac{1}{x^2+n^2}$,
   (d) $f_n(x) = \frac{nx}{1+(nx)^2}$.

   If the sequence is Cauchy, find its limit.

2. Show that if $f \in C^0(E)$ and E is compact, then f is uniformly continuous. (*Hint*: Use the fact that every compact set can be covered by a finite number of balls, $B_{\delta_i}(x_i)$. Argue that you can choose the balls so that for each $y \in B_{2\delta_i}(x_i)$, $|f(y) - f(x_i)| < \varepsilon/2$. Set δ to the minimum radius of these balls. Now prove that for any y and z, if $|z - y| < \delta$, then $|f(z) - f(y)| < \varepsilon$.)

3. Consider the sequence

$$f_n(x) = \begin{cases} 0 & x \leq 0 \\ nx & 0 < x < \frac{1}{n} \\ 1 & \frac{1}{n} \leq x \end{cases},$$

   for $n = 1, 2, \ldots$.

(a) Show that $f_n \in C^0[-1,1]$. However, argue that if the sequence f_n converges, it must converge to a discontinuous function.

(b) Is this sequence Cauchy in the sup-norm?

(c) Is this sequence Cauchy in the L_1 norm?

(d) What can you conclude about the completeness of $C^0[-1,1]$ with these two norms from this example?

4. Consider the operator

$$T(f) = \sin(2\pi x) + \lambda \int_{-1}^{1} \frac{f(y)}{1+(x-y)^2} dy$$

on the space of functions $C^0[-1,1]$ equipped with the sup-norm (3.4).

(a) Show that if $f \in C^0[-1,1]$, then so is $T(f)$.

(b) Find a $\lambda_o > 0$ such that if $|\lambda| < \lambda_o$, then $T(f)$ is a contraction mapping, and if $|\lambda| > \lambda_o$, then it is not. (*Hint*: To show the second part it is sufficient to find a pair of functions, f, g, for which $\rho(T(f), T(g)) > \rho(f, g)$.)

(c) Investigate the fixed point numerically. Start with $f(x) = 0$, and then try several other initial states. Try values of λ both smaller and larger than λ_o. (*Hint*: Numerical integration may be necessary.)

5. Another example for which an ODE like (3.21) is appropriate is known as Torricelli's law. Consider a cylinder of cross-sectional area A with a small hole of area $a \ll A$ in its bottom. Suppose that the cylinder is filled with water up to a height $h(0) = h_o$. The goal is to determine the height of the fluid as it leaks out of the hole.

Bernoulli's principle, essentially conservation of energy for an inviscid, incompressible fluid, states that

$$\tfrac{1}{2}\rho v^2 + \rho g z + p = const.$$

namely the sum of the kinetic, potential, and pressure energy densities is constant. At the top of the water, $p = p_a$, atmospheric pressure. At the hole this is also the case. Assume that the velocity of the water surface and of the leaking jet are both in the z direction.

(a) Since $a \ll A$, argue that the velocity v_h of the top of the water column, $z = h(t)$, is much less than the velocity v_0 at the hole, $z = 0$. Neglecting v_h, obtain a relation between $v_0(t)$ and $h(t)$.

(b) Equating the rate that the volume of fluid in the container changes with the rate that it streams out the hole, obtain an ODE for h.

(c) Find the general solution for your ODE.

(d) Discuss the physical meaning of the non-uniqueness of solutions to your ODE: what happens when the fluid completely drains from the container?

6. Show that the initial value problem

$$\dot{x} = \cos(t)\,|x|^{1/2}, \quad x(0) = 0,$$

has at least two different solutions. Sketch them in the (x,t)-plane. Why is the solution not unique?

7. Consider the initial value problem

$$\dot{x} = x^3, \quad x(0) = a.$$

   (a) Using Picard iteration (3.13) with $u_0(t) = 0$, find the first three successive approximations $u_1(t), u_2(t), u_3(t)$ to the solution.

   (b) Find the exact solution of this problem and expand it in a Taylor series about $t = 0$. Show that the first few terms of this series agree with the Picard iterates.

   (c) How does the number of correct terms grow with iteration?

8. Complete the third proof of Theorem 3.19 using the Bielecki norm and the contraction-mapping theorem.

9. Here we will prove Theorem 3.24. Suppose $f : J \times B_b(x_o) \to \mathbb{R}^n$, where $J = [t_o - c, t_o + c]$ is uniformly Lipschitz in $B_b(x_o)$ and C^0 in J. Thus there exists a constant K such that $|f(t,x) - f(t,y)| \leq K|x - y|$ for all $t \in J$ and $x, y \in B_b(x_o)$. Prove that there is an $a > 0$ such that solutions of the nonautonomous initial value problem (3.26) exist and are unique for $t \in [t_o - a, t_o + a]$.

10. Here is an alternative proof that solutions are unique and have Lipschitz dependence upon initial conditions when f is Lipschitz. Suppose that $u, v : J \to B_b(x_o)$ are two solutions of the ODE

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

where $f : J \times B_b(x_o) \to \mathbb{R}^n$ has a uniformly Lipschitz dependence on x with constant K. We make no assumptions about the dependence of f on t. Define $\varphi(t) = |u(t) - v(t)|^2$.

   (a) Use the inner product $\langle u, v \rangle = \sum_{i=1}^n u_i v_i$ and the Schwarz inequality, $|\langle u, v \rangle| \leq |u|\,|v|$ for vectors in $\mathbb{R}^n$, to find an ordinary differential inequality for φ, i.e., an equation of the form $\dot{\varphi}(t) \leq F(t, \varphi)$.

   (b) Using this inequality show $\frac{d}{dt}\left(e^{-2Kt}\varphi(t)\right) \leq 0$. Therefore, if $t > t_o$, show that

$$|u(t) - v(t)| \leq e^{K(t - t_o)}\,|u(t_o) - v(t_o)|.$$

   Conclude that the solution is unique and that two nearby solutions deviate at most exponentially in time.

11. Suppose that $g(t)$ obeys the inequality

$$g(t) \leq c(t) + \int_0^t k(s)g(s)ds,$$

where g and k obey the hypotheses of Lemma 3.28, and suppose that $c \in C^1(J)$ and is nondecreasing, $\dot{c} \geq 0$. Prove that $g(t) \leq c(t)e^{\int_0^t k(s)ds}$.

12. Consider the linear initial value problem $\dot{x} = A(t)x,\ x(0) = x_o$, where the matrix A is a continuous function of time on an interval (α, β) with $\alpha < 0 < \beta$. Your goal is to prove that the maximal interval of existence for this system contains the interval (α, β).

(a) Start by assuming that the maximal interval has a right-hand endpoint $b < \beta$. Argue that $||A|| \leq M$ on $[0, b]$.

(b) Use the integral form (3.11) to show that

$$|x(t)| \leq |x_o| + M \int_0^t |x(s)| ds$$

for any $t \in [0, b]$.

(c) Conclude from Grönwall's inequality (3.31) that $|x(t)|$ is bounded on $[0, b]$. Finally, extend Theorem 3.35 to contradict the assumption $b < \beta$.

(d) What can you conclude if A is continuous on $\mathbb{R}$?

13. Find the explicit solution and the maximal interval of existence for the initial value problems

(a) $\dot{x} = t x^3, \quad x(0) = x_o,$

(b) $\dot{x} = -x^2 \cos(t), \quad x(\pi/2) = x_o,$

(c) $\dot{x} = \frac{x^2}{\sqrt{t}}, \quad x(1) = x_o.$

Note that the maximal interval depends upon x_o, is open and must contain t_o. Plot the intervals in the (t, x_o) plane.

14. Consider the initial value problem

$$\begin{aligned} \dot{x} &= y/z \\ \dot{y} &= -x/z \quad , \quad (x, y, z) = (1, 0, 1) \text{ at } t = 1. \\ \dot{z} &= 1 \end{aligned}$$

(a) Convert this system to cylindrical coordinates (r, θ, ζ), where $r^2 = x^2 + y^2$, $\theta = \arctan(y/x)$, and $\zeta = z$. Find the initial conditions in the new coordinate system.

(b) Solve the new system and show that its solution exists in the maximal interval $J = (0, \infty)$.

(c) Apply Theorem 3.19 to the new system and determine the maximal interval guaranteed by the theorem.

15. Consider your adopted quadratic equations (recall Exercise 1.10) in their reduced form (i.e., set all the "nonessential parameters" to +1—keep the signs as given in the original equation). Call the reduced variables (x, y, z) for simplicity. Consider the set of solutions that start at the origin at $t = 0$ and stay in the ball $B_b(0) \subset \mathbb{R}^3$. Find a value a such that the existence and uniqueness theorem guarantees your system has a unique solution for a time interval $[-a, a]$. What is the maximal interval that you can obtain by varying b?

Chapter 4
Dynamical Systems

Science, as well as history, has its past to show—a past indeed, much larger; but its immensity is dynamic, not divine. (James Martineau)

So far, our approach to the study of dynamics has been completely traditional: we concentrated on some simple, solvable systems—especially linear systems—and we proved that more general, nonlinear systems actually have solutions. By contrast, the theory of "dynamical systems" is more concerned with qualitative properties. In this chapter we will seek to develop a classification of the qualitative properties of dynamics and to understand asymptotic behavior—what happens as $t \to \infty$. The first part of this study concerns the trajectories of a dynamical system in a local neighborhood. The goals are to classify equilibria by their stability, invariant manifolds, and topological type. This information will be used in later chapters to understand bifurcations and global dynamics.

4.1 ▪ Definitions

Behold the rule we follow, and the only one we can follow: when a phenomenon appears to us as the cause of another, we regard it as anterior. It is therefore by cause that we define time. (Henri Poincaré, 1914)

According to the *Encyclopedia Britannica*, dynamics is the "branch of physical science that is concerned with the motion of material objects in relation to the physical factors that affect them: force, mass, momentum, energy." Since Newton showed that mechanical systems are governed by differential equations, these do indeed provide good examples of dynamics. However, a more general definition is

> ▷ *dynamical system*: An *evolution rule* that defines a trajectory as a function of a single parameter (*time*) on a set of *states* (the *phase space*) is a dynamical system.

Dynamical systems are therefore categorized according to properties of their phase space, of their evolution rule, and of time itself. In this book, we consider systems with a continuous phase space, M, that is typically $\mathbb{R}^n$ or a more general space called a "manifold" such as the cylinder or torus.[20] Systems with a discrete phase space include the heads–tails model of a coin toss and "cellular automata" (Wolfram 1983). We will also primarily study systems with a continuous time variable, $t \in \mathbb{R}$. Systems with a discrete time variable are called "mappings" (Alligood, Sauer, and Yorke 1997; Devaney 1986).

The evolution rule can be deterministic or stochastic. A system is *deterministic* if for each state in the phase space there is a unique consequent, i.e., the evolution rule is a function taking a given state to a unique, subsequent state. Systems that are nondeterministic are called *stochastic*; a standard example is the idealized coin toss. For this case, the phase space is finite, consisting of the two states, heads and tails, and time is discrete, taking only the values at which the coin is examined. The evolution rule states that a head or a tail is equally likely at the next toss, independent of the current state of the coin.

When the evolution rule is deterministic, then for each time, t, it is a mapping from the phase space to the phase space,

$$\varphi_t : M \to M, \tag{4.1}$$

so that $x(t) = \varphi_t(x_o)$ denotes the position of the system at time t that started at x_o. Here we assume that t takes values in some allowed range and that the initial value of time is zero, so that $\varphi_0(x_o) = x_o$.

Every dynamical system has *orbits* or *trajectories*; namely, the sequence of states that follow from or lead to a given initial state. The forward orbit is the set of subsequent states

$$\Gamma_x^+ \equiv \{\varphi_t(x) : t \geq 0\}. \tag{4.2}$$

Similarly, the *preorbit*, Γ_x^-, is the set of sequences of states that lead, according to the evolution rule, to the initial state. When the function φ_t is one to one, then $\Gamma_x^- = \{\varphi_t(x) : t \leq 0\}$; otherwise, it is possible that several prior points could lead to the same x. Finally, the full *orbit* of a point x, Γ_x, is simply the union of the forward and preorbits of x.

The simplest orbit is an *equilibrium*, where the orbit is a single point: $\Gamma_x = \{x\}$. A point x is on a *periodic orbit* if there is a $T > 0$ such that

$$\varphi_T(x) = x, \tag{4.3}$$

that is, for which it returns to x. More generally orbits can be quasiperiodic, aperiodic, or chaotic; we will discuss these in later sections.

An orbit is a special case of an

▷ *invariant set*: A set Λ is invariant under a rule φ_t if $\varphi_t(\Lambda) = \Lambda$ for all t; that is, for each $x \in \Lambda$, $\varphi_t(x) \in \Lambda$ for any t.

Thus for each point x in an invariant set Λ, the entire orbit of x must be in Λ as well. Just as we define a forward orbit, we can also define a

▷ *forward invariant set*: A set Λ is forward invariant if $\varphi_t(\Lambda) \subset \Lambda$ for all $t > 0$.

[20] For our purposes, it is sufficient to think of a manifold simply as a smooth, multidimensional surface embedded in $\mathbb{R}^n$; see §5.5. More formal definitions are given in courses on differential geometry.

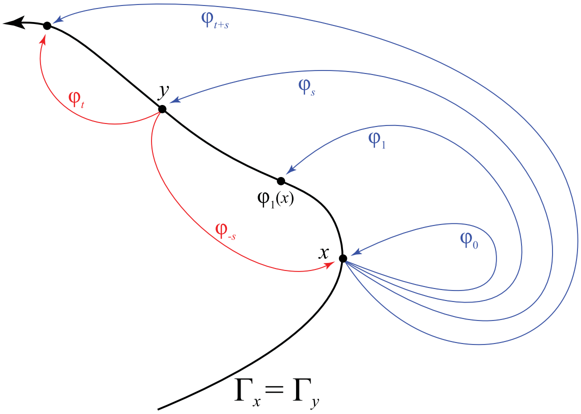

Figure 4.1. *Illustration of the group property of a flow,* $\varphi_t(y) = \varphi_t(\varphi_s(x)) = \varphi_{t+s}(x)$.

4.2 ▪ Flows

In §3.4 the solution of the initial value problem (3.27) with initial condition y was denoted by $u(t;y)$, and it was shown that when the vector field is C^1, then u is a C^1 function of both t and y. Beginning with this section, we will let

$$u(t;y) \to \varphi_t(y),$$

as in (4.1), so that the evolution rule is now thought of as a map from the phase space to itself that is parameterized by time. To emphasize this change of point-of-view, we define a class of evolution rules without reference to ordinary differential equations (ODEs):

> ▷ *flow*: Suppose the phase space for a dynamical system is a manifold M. A *complete flow* $\varphi_t(x)$ is a one-parameter, differentiable mapping $\varphi : \mathbb{R} \times M \to M$, such that
>
> (a) $\varphi_0(x) = x$, and
>
> (b) for all t and $s \in \mathbb{R}$,
>
> $$\varphi_t \circ \varphi_s = \varphi_{t+s}, \tag{4.4}$$
>
> where the composition symbol, ∘, means $\varphi_t \circ \varphi_s(x) \equiv \varphi_t(\varphi_s(x))$ for each $x \in M$.

For each fixed x, $\varphi_t(x)$ defines a curve in M as t varies over $\mathbb{R}$—the orbit Γ_x. Property (b) is known as the *group property*, since it implies that under the operation of composition, the family of maps $\{\varphi_t : t \in \mathbb{R}\}$ is an additive group (see Figure 4.1). For example, the group property for $s = -t$ implies $\varphi_t \circ \varphi_{-t} = \varphi_0 = id$ (here id is the "identity" function, $id(x) = x$), hence φ_t is an invertible function of x for each t, and moreover

$$(\varphi_t)^{-1} = \varphi_{-t}.$$

Consequently, for each t the flow φ_t is a one-to-one and onto map on M: it is a *bijection*. The group property also implies that two distinct trajectories cannot cross: if two trajectories ever touch, say, at a point $y = \varphi_t(x) = \varphi_s(z)$, then (4.4) implies that $\varphi_{t+r}(x) = \varphi_{s+r}(z)$ for all $r \in \mathbb{R}$, and the trajectories coincide.

Example 4.1. If φ is a flow and $\gamma = \Gamma_x$ is a periodic orbit, then the group property and (4.3) imply that

$$\varphi_{T+s}(x) = \varphi_s(x),$$

and so $\varphi_{2T}(x) = x$ and indeed $\varphi_{kT}(x) = x$ for any integer k. If T is the minimum positive value for which $\varphi_T(x) = x$, it is called the *period* of γ. It is also easy to see from the group property that if y is any other point on γ, then it has the same period as x: $\varphi_T(y) = y$.

Consequently, when γ is periodic, it is a closed loop; it can be viewed as an embedding of the circle $\mathbb{S}^1$ into the phase space, $\gamma : \mathbb{S}^1 \to \mathbb{R}^n$. ■

A flow is *complete* when it is defined for all t, so that the group property applies for all time. Usually, when we use the term "flow" without any qualification we mean a complete flow. Note that the group property implies that $x(t) = \varphi_{t-s}(\varphi_s(x_o)) = \varphi_{t-s}(x(s))$ for any time s along the trajectory. Therefore, $x(s)$ can also be viewed as the "initial condition" for the trajectory $x(t)$, but one that is imposed at time s.

Since a flow is differentiable, it has an associated ODE, or more precisely a

> ▷ *vector field*: A vector field is a function $f : M \to \mathbb{R}^n$ that defines a vector $v = f(x)$ at each point x in n-dimensional the phase space M.

The vector field associated with a flow is defined by

$$f(x) = \left.\frac{d}{dt}\varphi_t(x)\right|_{t=0}. \tag{4.5}$$

This vector field is interesting because the flow is a solution of the differential equation $\dot{x} = f(x)$, as we show next.

Lemma 4.2. *If $\varphi_t(x)$ is a flow, then it is a solution of the initial value problem*

$$\frac{d}{dt}\varphi_t(x_o) = f(\varphi_t(x_o)), \quad \varphi_0(x_o) = x_o,$$

for the vector field defined in (4.5).

Proof. Let $x(t) = \varphi_t(x_o)$. Differentiating and using the group property yields

$$\frac{dx}{dt} = \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left[\varphi_{t+\varepsilon}(x_o) - \varphi_t(x_o)\right] = \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left[\varphi_\varepsilon(x(t)) - \varphi_0(x(t))\right] = f(x(t)).$$

Therefore, the flow is the solution of the differential equation $\dot{x} = f(x)$. □

When the flow is complete, the solutions to this differential equation exist for all time: their maximal interval of existence is $(-\infty, \infty)$.

Example 4.3. The function $\varphi_t(x) = xe^{\lambda t}$ is a smooth map $\varphi : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$, and can be seen to satisfy the flow properties (a) and (b). Differentiation gives $\frac{d}{dt}\varphi_t(x) = \lambda\varphi_t(x)$, so that the vector field associated with φ_t is simply $f(x) = \lambda x$. Of course, φ_t is the general solution of the ODE $\dot{x} = \lambda x$. ■

Example 4.4. Consider the function $\varphi_t : \mathbb{R}^2 \to \mathbb{R}^2$ defined by

$$\varphi_t(x) = \begin{pmatrix} \varphi_{1t}(x) \\ \varphi_{2t}(x) \end{pmatrix} = \begin{pmatrix} x_1 e^{-t} \\ x_2 e^{x_1(e^{-t}-1)} \end{pmatrix}.$$

This function is clearly defined for all $(x_1,x_2)\in\mathbb{R}^2$ and $t\in\mathbb{R}$, and it is C^1 on this domain. To see that it satisfies the flow properties note first that $\varphi_0(x)=x$ and that

$$\varphi_t(\varphi_s(x))=\begin{pmatrix} \varphi_{1s}(x)e^{-t} \\ \varphi_{2s}(x)e^{\varphi_{1s}(x)(e^{-t}-1)} \end{pmatrix}=\begin{pmatrix} x_1e^{-(s+t)} \\ x_2e^{x_1(e^{-s}-1)}e^{x_1e^{-s}(e^{-t}-1)} \end{pmatrix}=\varphi_{s+t}(x).$$

Thus $\varphi_t(x)$ is a flow. The vector field (4.5) associated with this flow is given by differentiation:

$$\left.\frac{d}{dt}\varphi_t(x)\right|_{t=0}=\begin{pmatrix} -x_1e^{-t} \\ -x_2x_1e^{-t}e^{x_1(e^{-t}-1)} \end{pmatrix}_{t=0}=\begin{pmatrix} -x_1 \\ -x_1x_2 \end{pmatrix}=f(x).$$

Note that $f(x)$ is itself C^1 on $\mathbb{R}^2$. ▮

Not every differential equation defines a complete flow, because, as we saw in §3.5, the solutions do not necessarily exist for all time. However, if they do, then the flow is complete.

Lemma 4.5. *Let E be an open subset of $\mathbb{R}^n$, and $f:E\to\mathbb{R}^n$ a C^1 vector field such that the initial value problem $\dot{x}=f(x), x(0)=x_o$, has a solution $u(t;x_o)\in E$ that exists for all $t\in\mathbb{R}$ and all $x_o\in E$. Then $\varphi_t(x_o)\equiv u(t;x_o)$ is a complete flow.*

Proof. Theorem 3.30 implies that $u(t;x_o)$ is a differentiable function of both t and x_o. Moreover, the solution is unique in any interval in which it exists. To identify the solution as a flow, the group property must be demonstrated. Choose an $s\in\mathbb{R}$ and define $x_1=u(s,x_o)$. The initial value problem starting at x_1 has a solution that, by uniqueness, is given by the same function $u(t;x_1)$. However, uniqueness also implies that this new solution must follow the original solution; therefore,

$$u(s+t;x_o)=u(t;x_1)=u(t;u(s;x_o)).$$

This is the group property (4.4). □

4.3 ▪ Global Existence of Solutions

Theorem 3.19 (existence and uniqueness) implies that if a vector field $f:E\to\mathbb{R}^n$ is Lipschitz, then the initial value problem

$$\dot{x}=f(x), \quad x(0)=x_o, \tag{4.6}$$

has a unique solution for t within a maximal, open interval $J=(\alpha,\beta)$ (recall Theorem 3.33). As we have noted, when f is C^1, such a solution defines a flow, although the flow is not complete when either α or β is not infinite. Recall that a *complete* flow must obey the group property (4.4) for all t and $s\in\mathbb{R}$, and so the interval of existence must be all of $\mathbb{R}$. This makes the discussion of the global properties of the solutions of ODEs somewhat problematic.

There are several ways in which this problem can be obviated. For example, we will prove next that whenever the vector field f is bounded, the solutions exist for all time.

Theorem 4.6 (Bounded Global Existence). *If $f:\mathbb{R}^n\to\mathbb{R}^n$ is locally Lipschitz and bounded, then the solution of* (4.6) *exists for all $t\in\mathbb{R}$.*

Proof. Since f is locally Lipschitz, a solution $x(t) = u(t; x_o)$ exists on some maximal, open interval (α, β). By assumption, there is an m such that $|f(x)| \leq m$. The integral equation (3.11) then gives the inequality (for $t > 0$)

$$|x(t) - x_o| \leq \int_0^t |f(x(s))|\, ds \leq mt.$$

If β were finite, then this inequality implies that $x(t)$ is contained in the compact set $\{x : |x - x_o| \leq m\beta\}$; however, this contradicts Theorem 3.35 (unboundedness). Consequently, β is not the maximal value, and indeed there is no finite upper limit for the interval of existence. Similarly, it can be argued that α cannot be finite and therefore that the solution exists for all t. □

For example, the flow of the vector field $f(x) = \operatorname{sech}(x)$ on $\mathbb{R}$ is complete. Unfortunately, as shown in §3.5, the flow of an unbounded vector field such as $f(x) = x^2$ is not typically complete. Nevertheless, it is possible to show that any such flow is *equivalent* to a complete flow.

Theorem 4.7. *If $f(x)$ is locally Lipschitz on $\mathbb{R}^n$, then* (4.6) *is equivalent to*

$$\frac{dy}{d\tau} = F(y) = \frac{f(y)}{1 + |f(y)|}$$

upon reparameterizing time. The solutions exist for all $\tau \in \mathbb{R}$ since F is Lipschitz and bounded.

The use of the term "equivalence" for changing the definition of the time variable will be discussed more in §4.7.

Proof. The original equation has a solution $x(t)$ in some maximal interval (α, β). Define $y(\tau(t)) = x(t)$ using the new time variable

$$\tau = \int_0^t (1 + |f(x(s))|)\, ds. \tag{4.7}$$

Since $d\tau/dt = 1 + |f(x(t))| > 0$, the transformation (4.7) is strictly monotone increasing, so it defines a one-to-one mapping τ. Moreover, the differential equation for $y(\tau)$ is

$$\frac{dy}{d\tau} = \frac{dx}{dt}\frac{dt}{d\tau} = \frac{f(x)}{1 + |f(x)|} = F(y(\tau)). \tag{4.8}$$

Using the identity $(ab - cd) = \frac{1}{2}[(a - c)(b + d) + (b - d)(a + c)]$, it is not too hard to show that the new vector field F is locally Lipschitz:

$$\begin{aligned}
|F(y) - F(x)| &= \frac{|f(x)(1 + |f(y)|) - f(y)(1 + |f(x)|)|}{(1 + |f(x)|)(1 + |f(y)|)} \\
&= \frac{1}{2}\frac{|(f(x) - f(y))(2 + |f(x)| + |f(y)|) + (|f(y)| - |f(x)|)(f(x) + f(y))|}{(1 + |f(x)|)(1 + |f(y)|)} \\
&\leq |f(x) - f(y)| \frac{1 + |f(x)| + |f(y)|}{(1 + |f(x)|)(1 + |f(y)|)}.
\end{aligned}$$

Since the ratio above is bounded by one, F has the same Lipschitz constant as f. Moreover, as the new vector field F is bounded, Theorem 4.6 implies that the solutions of (4.8) exist for all time. If α (β) is finite, Theorem 3.35 implies that solution $x(t)$ must be unbounded as $t \to \alpha$ ($t \to \beta$); consequently, the transformation (4.7) maps J onto the infinite interval $(-\infty, \infty)$. □

Global existence also can be proved for vector fields that are globally Lipschitz.

Theorem 4.8 (Lipschitz Global Existence). *Suppose that $f(x)$ is globally Lipschitz on $E = \mathbb{R}^n$. Then the solution to (4.6) exists for all time.*

Proof. Beginning just as in the proof of Theorem 4.6, we obtain from the integral equation (3.11) the inequality

$$|x(t) - x_o| \le \int_0^t |f(x(s))| ds \le \int_0^t (|f(x(s)) - f(x_o)| + |f(x_o)|) ds$$

for any $t \in [0, \beta)$. The first term in the integral can be bounded using the global Lipschitz constant, K, for f. Suppose that β is finite; then

$$|x(t) - x_o| \le \beta |f(x_o)| + K \int_0^t |x(s) - x_o| ds,$$

which by the Grönwall inequality (3.31) implies that $|x(t) - x_o| \le \beta |f(x_o)| e^{Kt}$. Hence, when $0 \le t < \beta$, $x(t)$ is contained in the compact set $\{x : |x - x_o| \le \beta |f(x_o)| e^{K\beta}\}$. However, by Theorem 3.35 (unboundedness) this is impossible, so β cannot be finite. A similar argument shows that α is not finite. □

In some cases, a system of ODEs has a singularity that gives rise to a finite interval of existence. However, we can also often use the idea of rescaling time in this case to obtain a set of equations with global solutions.

Example 4.9. Consider two point masses interacting through mutual gravitational forces, and suppose that the velocity of the particles is tangent to the line connecting their masses. Choose a reference frame fixed on one mass, and let the origin correspond to the position of this mass. Denoting the position of the second particle by $x \in \mathbb{R}$, Newton's equations for this system are then

$$\dot{x} = v, \quad \dot{v} = -\frac{K}{x^2} \operatorname{sgn}(x), \tag{4.9}$$

where $K = G(m_1 + m_2)$. This is a Hamiltonian system—recall (1.12)—on the two-dimensional phase space of position, x, and velocity, v, with energy $H = {}^1\!/\!_2 v^2 - K/|x|$. However, we must restrict our attention to the set where $x \neq 0$ to avoid a singularity in the equation; consequently, the interval of existence is finite when a collision occurs (for example, when $H < 0$). In 1920 Levi-Civita developed a transformation that *regularizes* this collision singularity (Siegel and Moser 1971). By analogy with (4.7), he defines a new time by

$$\tau = \int_0^t \frac{ds}{x(s)}.$$

To simplify the equations, Levi–Civita also defines new dynamical variables (u, w) using the transformation

$$\begin{aligned} x &= u^2 \\ v &= 2\frac{w}{u} \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} u &= \sqrt{x} \\ w &= \frac{1}{2} v\sqrt{x}, \end{aligned}$$

which is well defined for $x > 0$. Substituting these transformations into the system (4.9) gives

$$\begin{aligned} \frac{du}{d\tau} &= \frac{1}{2} v\sqrt{x} = w, \\ \frac{dw}{d\tau} &= \frac{w^2}{u} - \frac{K}{2u} = \frac{1}{2} H u, \end{aligned} \tag{4.10}$$

where $H = (2w^2 - K)\big/u^2$ is the energy in the new coordinates. Since H is a constant, this system is effectively linear and its solutions are very simple; recall (2.20). Note that this linear system is defined for all (u, w) and has a global interval of existence. When $H < 0$, the solutions to (4.10) are oscillatory, and u changes sign; the negative values of u correspond to fictitious imaginary positions of the masses.

It is much more complicated to regularize the collision of more than two point masses. The three-body collision was studied by (McGehee 1974), but the behavior near a simultaneous collision of more than three bodies is still an unresolved question. ■

With these results, the concept of “flow” can be used to represent dynamics in most situations of interest—though with a possible reparameterization of time.

4.4 ▪ Linearization

The simplest orbit of a dynamical system is one that does not move, an

▷ *equilibrium*: A point x^* is an equilibrium of $\dot{x} = f(x)$ if $f(x^*) = 0$.

Some authors use the term “critical point” or “singular point” in place of equilibrium. Neither of these is, in my opinion, good terminology, as they seem to imply that something critical or singular happens at equilibria, when in fact an equilibrium is not critical or singular at all! It is simply a place where there is no motion. Moreover, it is standard to use the term “critical point” for a point where the derivative of a function vanishes.

Example 4.10. If the ODE is a *gradient system* $\dot{x} = \nabla V(x)$, then equilibria occur at critical points of the “potential” V. Therefore, in this case the terminology “critical point” is appropriate for the equilibria. The dynamics of a gradient system can be visualized by drawing the contours of the potential, since the velocity is perpendicular to surfaces of constant V. ■

When $f(x)$ is C^1, it is reasonable to hope that the motion in the neighborhood of an equilibrium can be studied by a Taylor series expansion of the ODE about x^*. To do this, substitute $x(t) = x^* + \delta x(t)$ into the ODE (4.6) using $f(x^*) = 0$ to obtain

$$\begin{aligned} \frac{d}{dt}(x^* + \delta x) = \frac{d}{dt}\delta x &= f(x^* + \delta x) = f(x^*) + Df(x^*)\delta x + o(\delta x), \\ \frac{d}{dt}\delta x &= Df(x^*)\delta x + o(\delta x), \end{aligned} \tag{4.11}$$

by Taylor's theorem. Here the notation (pronounced "little *oh* of δx") means

> ▷ $g(x) = o(f(x))$ as $x \to a$ if for all $\varepsilon > 0$ there is a neighborhood $N(\varepsilon)$ of a such that $|g(x)| < \varepsilon\,|f(x)|$ for all $x \in N(\varepsilon)$.

Recall from §3.1 that a neighborhood of a point a is any set that contains an open set containing a. A similar notation is the "big oh" symbol, which means

> ▷ $g(x) = O(f(x))$ as $x \to a$ if there is a neighborhood N of a and a $c \geq 0$ such that $|g(x)| < c\,|f(x)|$ for all $x \in N$.

When $f \in C^2$, then Taylor's theorem implies that the remainder term in (4.11) is actually $O(\delta x^2)$.

If we simply discard the $o(\delta x)$ terms in (4.11), we obtain an ODE called the

> ▷ *linearization*: If $f \in C^1(E)$, then the linearization of $\dot{x} = f(x)$ at the equilibrium $x^* \in E$ is the differential equation
>
> $$\dot{y} = Df(x^*)y. \tag{4.12}$$

No justification, other than the desire for simplicity, has been given for neglecting the higher-order terms in (4.11); nevertheless, (4.12) does give a faithful local representation for the motion in some cases. Note that $Df(x^*) = A$ is a constant matrix and so all our techniques from Chapter 2 for solving linear systems apply. In particular, the general solution is $\Phi(t,0)y_0$ where $\Phi(t,0)$ is the fundamental matrix (2.46).

In §2.7 the solutions of linear ODEs were classified by their generalized eigenspaces according to the sign of the real part of the eigenvalues, resulting in the decomposition $E = E^u \oplus E^s \oplus E^c$ into the direct sum of unstable, stable, and center eigenspaces. We can now use this decomposition to classify the behavior "near" an equilibrium. We first generalize the notion of hyperbolic linear systems in §2.7 to general equilibria:

> ▷ *hyperbolic*: an equilibrium x^* of a C^1 vector field f is hyperbolic if none of the eigenvalues of $Df(x^*)$ have zero real part, or equivalently when E^c is trivial.

Hyperbolic equilibria fall into three classes:

> ▷ *sink*: an equilibrium is a sink if all of the eigenvalues of $Df(x^*)$ have negative real parts (are in the left half of the complex plane), or equivalently when $E = E^s$;
>
> ▷ *source*: an equilibrium is a source if all of the eigenvalues of $Df(x^*)$ have positive real parts, or equivalently when $E = E^u$; and
>
> ▷ *saddle*: an equilibrium is a saddle if it is hyperbolic, but not a sink or a source, equivalently when $E = E^s \oplus E^u$.

Recall that in §2.2 an equilibrium was called a stable node when its eigenvalues are real and negative and an unstable node when they are real and positive. The classification into sink and source above includes these cases but also allows the eigenvalues to be complex. When some or all of the eigenvalues of $Df(x^*)$ are complex, we can indicate this by adding some additional modifiers to the classification:

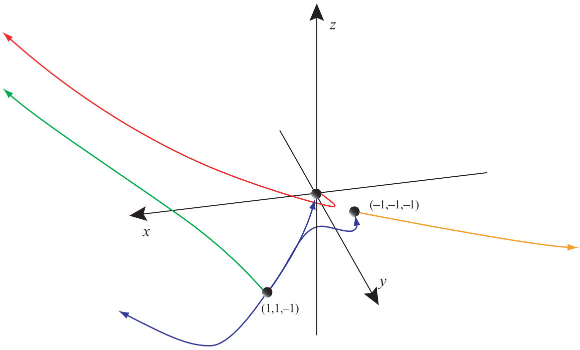

Figure 4.2. *Several orbits of the system* (4.13) *near its three equilibria.*

- ▷ *focus*: there is a subspace with complex eigenvalues with nonzero real part, or
- ▷ *center*: there is a subspace with purely imaginary eigenvalues.

For example, a four-dimensional saddle with two pairs of eigenvalues $\lambda_{1,2} = 1 \pm 2i$ and $\lambda_{3,4} = -2 \pm 4i$ is called a saddle-focus. There are many varieties of foci, depending upon the number of complex eigenvalues. If we wish to be more precise in the classification, we can specify the dimension of each of the invariant subspaces.

Example 4.11. Consider the set of ODEs on $\mathbb{R}^3$:

$$\begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{pmatrix} = f(x,y,z) = \begin{pmatrix} x - y \\ z + y^2 \\ x + yz \end{pmatrix}. \tag{4.13}$$

Solving the three equations $f(x,y,z) = 0$ gives three equilibria, $(0,0,0)$, $(1,1,-1)$, and $(-1,-1,-1)$. The Jacobian of the vector field at a general point is

$$Df = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 2y & 1 \\ 1 & z & y \end{pmatrix}.$$

The characteristic polynomial of this matrix is

$$p(\lambda) = \det(\lambda I - Df) = \lambda^3 - (3y+1)\lambda^2 - (z - 3y - 2y^2)\lambda + 1 + z - 2y^2.$$

Perhaps the hardest part of linear stability analysis is to find the roots of $p(\lambda)$. The critical points and critical values of p can be used to determine the relevant information even without explicitly finding the eigenvalues (recall Exercise 2.11). For example, a cubic polynomial always has one real root; however, it has three real roots only if it has two real critical points, two values c_i such that $p'(c_i) = 0$, and if the signs of p at the two critical points are opposite, so $p(c_1)p(c_2) < 0$.

The first equilibrium $(0,0,0)$ of (4.13) has the characteristic polynomial

$$p(\lambda) = \lambda^3 - \lambda^2 + 1.$$

Since $p'(c) = 3c^2 - 2c$, there are critical points at $c_1 = 0$ and $c_2 = 2/3$, where $p(c_i) > 0$. Thus, there is only one real root. Since $p(0) = 1$, the real root, λ_1, is negative; and since $p(-1) = -1$, then $-1 < \lambda_1 < 0$. A numerical solution shows that $\lambda_1 \approx -0.7548$.

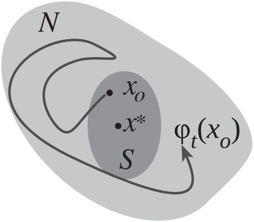

Figure 4.3. *Lyapunov stability.*

The remaining roots must be complex, $\lambda_{2,3} = \alpha \pm i\beta$. The sum of the eigenvalues is $\mathrm{tr}(Df) = 1 = \lambda_1 + 2\alpha$, so that $\alpha = \frac{1}{2}(1-\lambda_1) > \frac{1}{2}$. Numerically, $\alpha \approx 0.8774$. As a consequence, the origin is a hyperbolic saddle. Since one pair of eigenvalues is complex, it can be called a saddle-focus. Here, E^u is two-dimensional, and E^s is one-dimensional.

The second equilibrium, $(1,1,-1)$, has the characteristic polynomial

$$p(\lambda) = \lambda^3 - 4\lambda^2 + 6\lambda - 2.$$

The critical points of p are complex, so p has only one real root. Since $p(0) < 0$ and $p(1) > 0$, then $0 < \lambda_1 < 1$. Moreover, since $Re(\lambda_{2,3}) = \alpha = \frac{1}{2}(\mathrm{tr}(Df) - \lambda_1) = \frac{1}{2}(4-\lambda_1) > 0$, this point is a source-focus and has a three-dimensional unstable space.

Finally, the equilibrium $(-1,-1,-1)$, has characteristic polynomial

$$p(\lambda) = \lambda^3 + 2\lambda^2 - 2,$$

which has critical points at $c_1 = 0$ and $c_2 = -4/3$, where $p(c_i) < 0$, so again there is a single real root, $0 < \lambda_1 < 1$. So $\alpha = \frac{1}{2}(\mathrm{tr}(Df) - \lambda_1) = \frac{1}{2}(-2 - \lambda_1) < 0$. Thus, this point is a saddle-focus with a two-dimensional stable space and a one-dimensional unstable space. Some orbits of this system are shown in Figure 4.2. ∎

One of the major questions that we will soon address is, "To what extent does the solution of the full system *look like* the solution of the linear system?" Moreover, what is meant by *look like*? A partial answer to this will be provided by the Hartman–Grobman theorem in §4.8.

4.5 ▪ Stability

In §2.7 we said a system is *linearly stable* if it has bounded forward orbits; in other words, each orbit stays a bounded distance from the equilibrium at the origin. In that section we also defined the concepts of *spectral stability* and *asymptotic linear stability*. For nonlinear systems, these definitions are deficient: simply being bounded does not characterize the long time dynamics. A better definition of stability refers to orbits that are close: an equilibrium is stable if orbits that start "nearby" stay "nearby." Aleksandr Lyapunov (pronounced lēah·pū′·nof) (1857–1918) formalized this idea in 1892:

> ▹ *Lyapunov stability*: An equilibrium x^* of a flow φ_t is (Lyapunov) stable if for every neighborhood N of x^* there is a neighborhood $S \subset N$ such that if $x \in S$, then $\varphi_t(x) \in N$ for all $t \geq 0$.

This construction is sketched in Figure 4.3. An equilibrium that is not stable is called *unstable*.

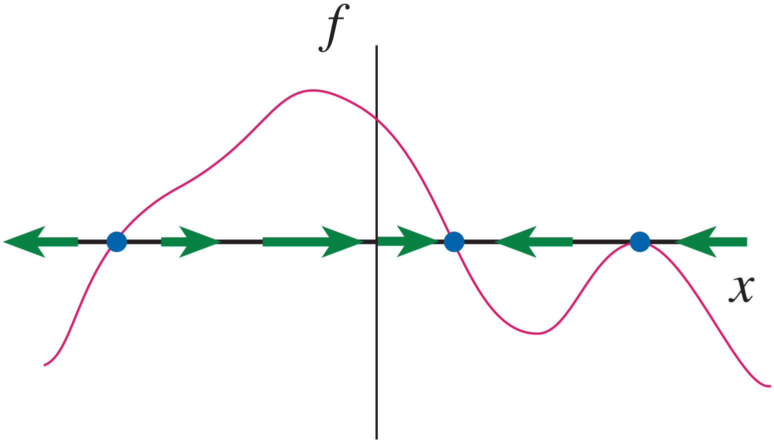

Figure 4.4. *Illustration of the three types of equilibria for a one-dimensional ODE. The left equilibrium is a source, the middle a sink, and the right is semistable.*

For a metric space, Lyapunov stability is equivalent to the assertion that for every $\varepsilon > 0$ there is a $\delta > 0$ such that whenever $x \in B_\delta(x^*)$, we have $\varphi_t(x) \in B_\varepsilon(x^*)$ for all $t \geq 0$; recall (3.1). Whenever the word "stability" is used without qualification, it should be taken to mean "Lyapunov stability."

For a one-dimensional ODE, the stability of an equilibrium, x^*, is easily investigated by examining the graph of the function f near x^*, as we discussed in §1.3. For example, if there is a $\delta > 0$ such that $f(x) < 0$ for $x \in (x^*, x^* + \delta)$ and $f(x) > 0$ for $x \in (x^* - \delta, x^*)$, then x^* is Lyapunov stable, since all points in the interval $(x^* - \delta, x^* + \delta)$ move toward x^* monotonically. This is illustrated by the middle equilibrium in Figure 4.4. Generalizing the terminology from the linear case, such a point is a *sink*. By contrast, if the signs of f are reversed, then the flow moves locally away from the equilibrium and x^* is unstable, and it is called a *source* (e.g., the leftmost equilibrium in Figure 4.4). If x^* is a zero and f has the same sign on both sides, then the point is often somewhat misleadingly called *semistable*—even though by Lyapunov's definition it is really unstable! This case corresponds to the rightmost equilibrium in Figure 4.4. If $f(x) = 0$ on an interval about x^*, then there is an interval of equilibria, and each equilibrium in the interior of this interval is stable.

These notions of sink, source, and semistable equilibria are topological: they follow without any assumptions on the smoothness of f. When $f \in C^1(\mathbb{R})$, however, these stability properties are related to hyperbolicity. For example, when $Df(x^*) \neq 0$, the equilibrium is hyperbolic; it is stable when $Df(x^*) < 0$ and unstable when $Df(x^*) > 0$.

Example 4.12. The logistic ODE (1.7), $\dot{x} = rx(1-x)$, has an unstable equilibrium $x^* = 0$ when $r > 0$, because $Df(0) = r$, and a stable one at $x^* = 1$ where $Df(1) = -r$. Moreover, every initial condition in the interval $(0, \infty)$ moves monotonically toward 1. Indeed, for any $\varepsilon > 0$, choose any $\delta \in (0, \min(\varepsilon, 1))$ and $x \in [1-\delta, 1+\delta]$; then $|\varphi_t(x) - 1| < \delta < \varepsilon$. Hence $x^* = 1$ is Lyapunov stable. ∎

Example 4.13. $f(x) = x^2 - x\cos x$. This function, shown in Figure 4.5, has precisely two zeros, $x_0 = 0$, and $x_1 = \cos(x_1) \approx 0.739085$. The solution $x(t)$ is monotone increasing if $x < x_0$ or $x > x_1$, and monotone decreasing in the interval (x_0, x_1). Accordingly, $x^* = 0$ is a stable equilibrium, while $x^* = x_1$ is unstable. ∎

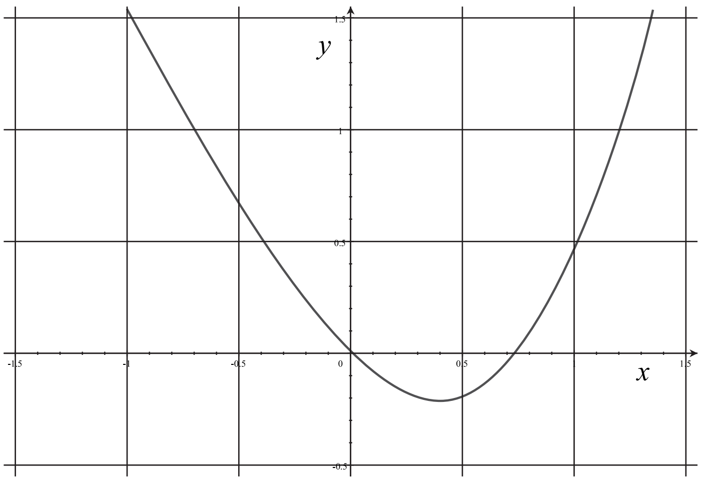

Figure 4.5. *Graph of* $f(x) = x^2 - x\cos x$.

A nonhyperbolic equilibrium, one for which $Df(x^*) = 0$, can be either stable or unstable. For example, the point $x = 0$ for $\dot{x} = x^2$ is semistable but not Lyapunov stable, even though all points starting with negative initial conditions asymptotically approach the origin. The problem is that there is no neighborhood containing the origin for which points stay close.

Example 4.14. Suppose$f \in C^1(\mathbb{R})$ and $Df(0) = 0$. There are four typical cases:

(a) $f(x) = -x^3$, here graphical analysis implies $x = 0$ is stable, a *sink*;

(b) $f(x) = +x^3$, unstable, a *source*;

(c) $f(x) = \pm x^2$, *semistable*; and

(d) $f(x) \equiv 0$, infinitely many equilibria. ■

This monotonic motion toward or away from an equilibrium is specific to one-dimensional systems; higher-dimensional systems can exhibit oscillation. Moreover, even in the linear case, the distinction between the two neighborhoods S and N is needed because the eigenvectors of a matrix are not typically orthogonal.

Example 4.15. A matrix is *normal* if it commutes with its adjoint: $[A^*, A] = 0$, where $A^* = \bar{A}^T$ is the conjugate transpose of A. It is not hard to see that the eigenspaces of a normal matrix are orthogonal. The dynamics of a stable linear system with a nonnormal matrix can exhibit a surprising temporary growth. Consider, for example,

$$\dot{x} = \begin{pmatrix} -1 & 10 \\ 0 & -2 \end{pmatrix} x \Rightarrow x(t) = c_1 e^{-t} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + c_2 e^{-2t} \begin{pmatrix} -10 \\ 1 \end{pmatrix}. \tag{4.14}$$

The general solution shows that every initial condition is attracted to the origin, so the origin should be stable. However, points that start in the disk of radius δ about the origin can leave, at least for a while. For example, setting $c_1 = 9$, $c_2 = 1$, then $x(0) = (-1, 1)$. However, the second eigenvector quickly decays, leaving a large horizontal component. Consequently, the orbit can move away from the origin for some time, as shown in Figure 4.6.

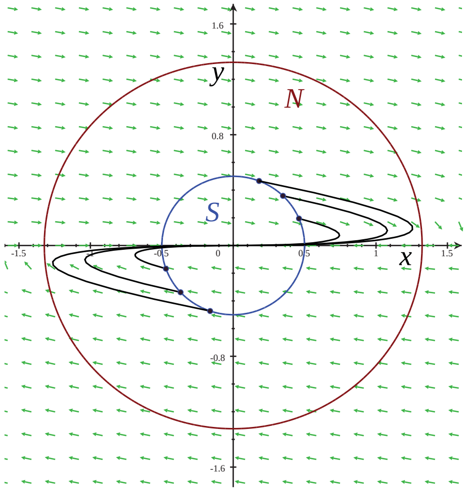

Figure 4.6. *Orbits of the system* (4.14) *that start in a neighborhood S never leave N.*

However, we can easily obtain a crude bound on $|x(t)|$, given that $|x(0)|^2 = (c_1 - 10c_2)^2 + c_2^2 \le \delta^2$. This implies that both $|c_2| \le \delta$ and $|c_1| \le 11\delta$ so that

$$\begin{aligned} |x(t)| &\le \left|c_1 e^{-t} - 10c_2 e^{-2t}\right| + \left|c_2 e^{-2t}\right| \le |c_1|\, e^{-t} + 11\, |c_2|\, e^{-2t} \\ &\le 22\delta = \varepsilon. \end{aligned} \tag{4.15}$$

So, if we choose $\delta = \varepsilon/22$ we are guaranteed that every point that starts in the δ ball remains in the ε ball. ■

A more stringent version of stability is the property of

▷ *asymptotic stability*: An equilibrium x^* is *asymptotically stable* if it is stable *and* there is a neighborhood N of x^* such that every point in N approaches x^* as $t \to \infty$.

An asymptotically stable equilibrium is also called an *attracting equilibrium*. This is the simplest case of the concept called an *attractor*; see §4.10. Note that by this definition, an attractor must attract a neighborhood.

Example 4.16. We showed that the origin is a stable equilibrium of (4.14). Moreover, the inequality (4.15) implies that every point is asymptotic to the origin, so it is asymptotically stable as well. ■

There are ODEs that have equilibria with a neighborhood that eventually attracts all nearby points but which is nevertheless *not* Lyapunov stable. In this case, nearby points may move a large distance from the equilibrium. A physical model ODE system is often derived to be valid only in some neighborhood of an equilibrium; consequently, when orbits move far from the equilibrium the model may no longer be valid and it would not be appropriate to rely on the eventual return to define asymptotic stability.

Example 4.17. Consider the system

$$\begin{aligned} \dot{r} &= r(1-r), \\ \dot{\theta} &= \sin^2(\theta/2), \end{aligned} \tag{4.16}$$

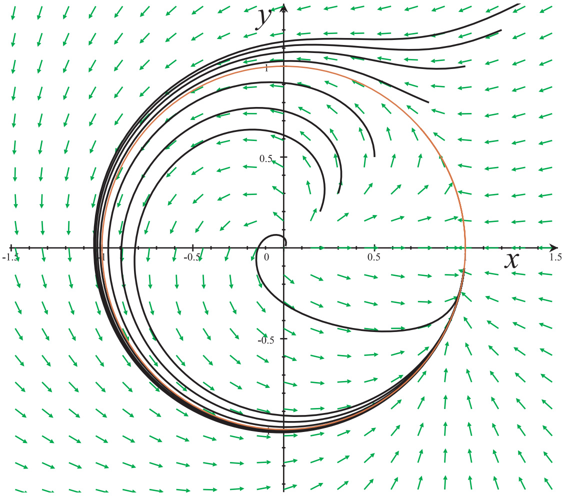

Figure 4.7. *Phase space of the example* (4.16).

where (r, θ) are polar coordinates in the plane. As shown in Figure 4.7, there are two equilibria, the origin and $(1, 0)$. The origin is unstable; indeed the r dynamics is decoupled from the θ dynamics, and graphical analysis immediately shows that every $r > 0$ is asymptotic to $r = 1$. Similarly the θ equation is uncoupled and since $\sin^2(\theta/2) \geq 0$, the point $\theta = 0$ is "semistable." However, since θ is a periodic coordinate, even the points with $\theta = \delta > 0$, which move away from the equilibrium point, will eventually return to $\theta = 0$. Therefore, every initial condition in $\mathbb{R}^2$ except the origin is attracted to the point $(1, 0)$. However, this point is not Lyapunov stable since for any $\varepsilon < 2$, there are nearby points—for example $(1, \delta)$—that leave the ball of radius ε about the equilibrium. ■

Example 4.18 (Vinograd). A more complicated example of this behavior was given in (Vinograd 1957):

$$\dot{x} = \frac{x^2(y - x) + y^5}{r^2(1 + r^4)}, \quad \dot{y} = \frac{y^2(y - 2x)}{r^2(1 + r^4)}, \tag{4.17}$$

where r is the polar radius, $r^2 = x^2 + y^2$. To analyze this system, first note that the origin is the only equilibrium, since $\dot{y} = 0$ implies either $y = 0$ or $y = 2x$. In the latter case if $\dot{x} = 0$ as well, then

$$x^3 + 32x^5 = 0 \Rightarrow x = 0 \text{ or } x^2 = -1/32.$$

So the only real solution is $x = y = 0$. Note that $\dot{y}|_{y=0} = 0$, so the line $y = 0$ is invariant. On this line x is governed by $\dot{x}|_{y=0} = -x/(1 + x^4)$; therefore, since $\operatorname{sgn}(\dot{x}) = -\operatorname{sgn}(x)$ and $\dot{x} \neq 0$ unless $x = 0$, $x(t)$ monotonically moves toward the origin, so that the origin attracts all points on this line. It is much harder to show that every point in the plane approaches the origin as $t \to \infty$, but a numerical solution (shown in Figure 4.8) indicates that this is so. More interestingly, the picture indicates that many orbits in any δ-ball leave the ball $B_{1/2}(0)$ no matter how small δ is chosen. In fact, it seems that there is a family of *homoclinic loops* from the origin, i.e., orbits that leave

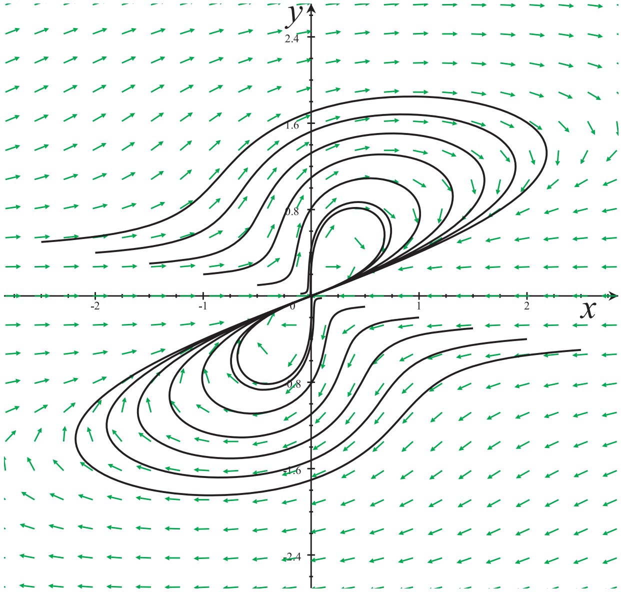

Figure 4.8. *Phase plane of the Vinograd example* (4.17).

the origin and go a finite distance away before returning as $t \to \infty$ (see §5.2). These loops are what prevent the origin from being an attractor. The behavior of this system near the origin is studied in §6.2. ∎

When f is C^1, the local behavior near an equilibrium is often governed by the linearization, (4.12). For example, asymptotic linear stability is sufficient to imply asymptotic stability of the equilibrium for the nonlinear system, if it is differentiable. The main point is that in this case we can extract the nonlinear part of f near x^* by writing

$$f(x) = Df(x^*)(x - x^*) + g(x - x^*).$$

The assumption that f is C^1 is sufficient to guarantee that the remainder term is small, i.e., that $g(\delta x) = o(\delta x)$. This follows from the definition of the derivative, (3.2), taking $h = \delta x_j \hat{e}_j$,

$$0 = \left[\lim_{\delta x_j \to 0} \frac{f_i(x^* + \delta x_j \hat{e}_j) - f_i(x^*)}{\delta x_j} - (Df)_{ij}(x^*) \right] = \lim_{\delta x_j \to 0} \frac{g_i(\delta x_j \hat{e}_j)}{\delta x_j}.$$

Note that if $f(x)$ is C^2, then $g(\delta x) = O(\delta x^2)$, by the Taylor remainder theorem. However, we will not need this additional assumption to prove the desired result.[21]

Theorem 4.19 (Asymptotic Linear Stability implies Asymptotic Stability). *Let $f : E \to \mathbb{R}^n$ be C^1 and have an equilibrium x^* such that all the eigenvalues of $Df(x^*)$ have real parts less than zero. Then x^* is asymptotically stable.*

[21] This theorem also follows from either the Hartman–Grobman or the stable manifold theorem; see §4.8 and Chapter 5.

Proof. Rewrite the differential equations using $y = x - x^*$, defining $A = Df(x^*)$, and $g(y) \equiv f(x^* + y) - Ay$, to obtain

$$\dot{y} = Ay + g(y), \quad y(0) = y_o. \tag{4.18}$$

Variation of parameters can be used to obtain an integral equation for the solution. Let $y = e^{tA}\eta(t)$, and substitute this into the ODE (4.18) to obtain $\dot{\eta} = e^{-tA} g(y(t))$. Formally integrating this equation and substituting again for y gives the integral equation:

$$y(t) = e^{tA} y_o + \int_0^t e^{(t-s)A} g(y(s)) ds. \tag{4.19}$$

By assumption, there is an α such that if λ is any eigenvalue of A, then $\mathrm{Re}(\lambda) < -\alpha < 0$. The estimate (2.44) in §2.7 implies that there is a $K \geq 1$ such that for any vector v

$$\left| e^{tA} v \right| \leq K e^{-\alpha t} |v|, \quad t \geq 0. \tag{4.20}$$

Since f is C^1, then $g(y) = o(y)$, so, for any $\varepsilon > 0$ there is a $\delta > 0$ such that if $|y| < K\delta$, then $|g(y)| < \varepsilon |y|$.

Now assume that $|y_o| < \delta$. Since the solution to (4.18) is continuous there is a $\tau > 0$ such that for all $t \in [0, \tau)$, $|y(t)| < K\delta$. Therefore for any $t \in [0, \tau)$, (4.19) using (4.20) gives

$$|y(t)| \leq K\delta e^{-\alpha t} + K\varepsilon \int_0^t e^{-\alpha(t-s)} |y(s)| ds.$$

Now let $\xi(t) = e^{\alpha t} |y(t)|$, and use Grönwall's Lemma 3.28 to obtain

$$\xi(t) \leq K\delta + K\varepsilon \int_0^t \xi(s) ds \quad \Rightarrow \quad \xi(t) \leq K\delta e^{K\varepsilon t} \quad \Rightarrow \quad |y(t)| \leq K\delta e^{-(\alpha - K\varepsilon)t}.$$

Hence, providing $\varepsilon < \alpha/K$, then $|y|$ stays bounded below $K\delta$ for all $t \in [0, \tau)$.

If $\tau < \infty$ is the smallest time for which the inequality $|y(t)| < K\delta$ does not hold, then we must have $|y(\tau)| = K\delta$. However, since $\alpha - K\varepsilon < 0$, the above inequality implies that $\lim_{t \to \tau} |y(t)| < K\delta$. Therefore $\tau = \infty$, and $|y(t)| \to 0$ as $t \to \infty$.

In conclusion, if S is the ball of radius δ, then N is the ball of radius $K\delta$. □

Example 4.20. The origin is an equilibrium of the system

$$\begin{aligned} \dot{x} &= -x - y - r^2, \\ \dot{y} &= x - y + r^2, \end{aligned}$$

where r is the polar radius. The origin is a stable focus since

$$Df(0,0) = \begin{pmatrix} -1 & -1 \\ 1 & -1 \end{pmatrix}$$

has eigenvalues $\lambda = -1 \pm i$. To show that adding nonlinear terms does not change the topological character, we want to construct an attracting neighborhood of the origin. To study this system, it is easier to use the differential equation for r.[22] Noting that $2r\dot{r} = 2x\dot{x} + 2y\dot{y}$

$$\dot{r} = \frac{1}{r}\left(x\left(-x - y - r^2\right) + y(x - y + r^2)\right) = r(-1 + y - x).$$

[22] We will find this technique extremely useful in our study of the global structure of flows in the plane in Chapter 6.

Since $-r \le x, y \le r$, then $y - x \le 2r$. If $r < 0.5$, then $-1 + y - x < 0$, and so $\dot{r} < 0$ at any point in the open disk of radius ½. This implies that the origin is asymptotically stable, because r is monotonically decreasing. Note that there is another equilibrium point at $(-1, 0)$. This equilibrium has eigenvalues $\lambda = \pm\sqrt{2}$ and is therefore a saddle. Orbits near the saddle can go to infinity. ∎

4.6 ▪ Lyapunov Functions

Lyapunov devised another technique that can potentially show that an equilibrium is stable—the construction of what is now called a "Lyapunov function." An advantage of this method is that it can sometimes prove stability of a nonhyperbolic equilibrium; a disadvantage is that there is no straightforward construction of Lyapunov functions.

Lyapunov functions are nonnegative functions that decrease in time along the orbits of a dynamical system:

> ▷ *Lyapunov function*: A continuous function $L : \mathbb{R}^n \to \mathbb{R}$ is a (*strong*) Lyapunov function for an equilibrium x^* of a flow φ_t on $\mathbb{R}^n$ if $L(x^*) = 0$ and there is an open neighborhood U of x^* such that for all $x \in U \setminus \{x^*\}$, $L(x) > 0$ and
> $$L(\varphi_t(x)) < L(x), \quad \forall\, t > 0. \tag{4.21}$$

The function L is a *weak* Lyapunov function if (4.21) is replaced by $L(\varphi_t(x)) \le L(x)$.

When L is C^1, (4.21) can be guaranteed by requiring that $dL/dt < 0$. This can be computed using the chain rule:

$$\frac{dL}{dt} = \nabla L(x) \cdot f(x). \tag{4.22}$$

Consequently, in the smooth case, L is a Lyapunov function when its gradient points in a direction opposed to that of the vector field f.

If such a nonincreasing function can be found, the equilibrium is stable.

Theorem 4.21 (Lyapunov Functions). *Let x^* be an equilibrium point of a flow $\varphi_t(x)$. If L is a weak Lyapunov function for x^*, then x^* is stable. If L is a strong Lyapunov function, then x^* is asymptotically stable.*

Proof. First we prove stability. We can assume that $x^* = 0$ without loss of generality. Choose any ε small enough so that $B_\varepsilon(0) \subset U$ and define $m = \min\{L(x) : |x| = \varepsilon\}$, as in Figure 4.9. The constant m exists because $\partial B_\varepsilon(0)$ is compact and, since L is positive definite, $m > 0$. Since L is continuous and $L(x^*) = 0$, then L decreases as $x \to 0$ and so there exists a $\delta < \varepsilon$ such that $L(x) \le m$ for all $x \in B_\delta(0)$. Since L is nonincreasing along orbits then $L(\varphi_t(x)) \le m$ for all $x \in B_\delta(0)$. Therefore, since L does not increase beyond the minimum on $|x| = \varepsilon$, $\varphi_t(x) \in B_\varepsilon(0)$. Consequently, the origin is stable.

Now we prove asymptotic stability. If $x \in B_\delta(0)$, then $\varphi_t(x) \in B_\varepsilon(0)$ for all positive time. Since $B_\varepsilon(0)$ is compact, the Bolzano–Weierstrass theorem, Theorem 3.1, implies that for any sequence $t_i \to \infty$, the sequence $\varphi_{t_i}(x)$ must have limit points. Suppose one of these limit points is not the origin, i.e., there is a sequence of times $t_n \to \infty$ such that $\varphi_{t_n}(x) \to z \ne 0$. By continuity $L(\varphi_{t_n}(x)) \to L(z)$, and since L is strictly decreasing, the sequence of values must decrease monotonically with n:

$$L(\varphi_{t_n}(x)) > L(\varphi_{t_{n+1}}(x)) > \cdots > L(z). \tag{4.23}$$

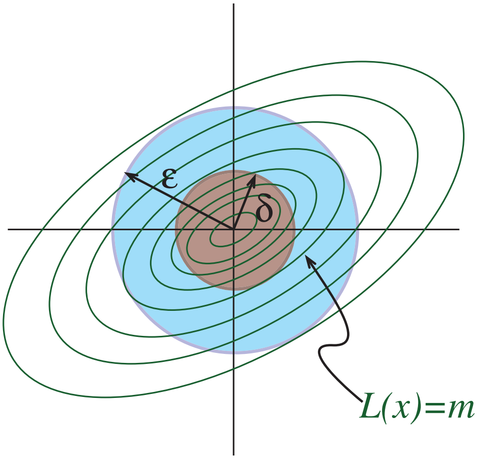

Figure 4.9. *Contours of a Lyapunov function for a stable equilibrium.*

Now consider the orbit, $\varphi_s(z)$ of the limit point z. Again, since z is not the equilibrium, $L(\varphi_s(z)) < L(z)$ for any positive s, and hence by continuity $L(\varphi_{t_n+s}(x)) \to L(\varphi_s(z)) < L(z)$. This implies for large enough n that, since $x(t_n)$ is arbitrarily close to z, $L(\varphi_{t_n+s}(x)) < L(z)$. Finally, given s, construct a subsequence t_m of t_n such that for each m, n, $t_m > t_n + s$, but then

$$L(\varphi_{t_m}(x)) < L(\varphi_{t_n+s}(x)) < L(z);$$

however, this contradicts the inequality (4.23). In conclusion, all limit points of the flow are at $x = 0$, and so the origin is asymptotically stable. □

Example 4.22. Any linear system $\dot{x} = Ax$ that is asymptotically stable has a strong Lyapunov function. Indeed, we assert that one such function has the form $L = x^T Sx$, where S is a symmetric matrix. In this case, the derivative of L is negative, provided that

$$\frac{dL}{dt} = \dot{x}^T Sx + x^T S\dot{x} = x^T(A^T S + SA)x < 0 \quad \forall\, x \neq 0.$$

One way to solve this is to require that S obey

$$A^T S + SA = -I \tag{4.24}$$

for in this case $dL/dt = -|x|^2$. Equation (4.24) is sometimes called the "Lyapunov equation." We claim that (4.24) has a solution whenever all the eigenvalues of A have negative real parts (see Exercise 9). This result also provides an alternative proof of the asymptotic linear stability theorem in §2.7. ■

In general, finding a Lyapunov function for a nonlinear system is a matter of guessing. However, when the equilibrium is asymptotically stable, a Lyapunov function is guaranteed to exist.

Theorem 4.23. *If x^* is an asymptotically stable equilibrium that attracts a neighborhood U, then the function*

$$L(x) = \int_0^\infty e^{-s} \sup_{t \geq s} |\varphi_t(x) - x^*|\, ds \tag{4.25}$$

is a strong Lyapunov function on U.

Proof. By assumption every orbit in U is bounded; this implies the function λ defined by

$$\lambda(x) = \sup_{t\geq 0} |\varphi_t(x) - x^*|$$

for $x \in U$ is continuous. Indeed, asymptotic stability implies that for any $\rho > 0$ there is a time $T(\rho)$ such that for any $x \in U$, $|\varphi_t(x) - x^*| < \rho$ whenever $t > T(\rho)$. As a consequence, the supremum in the definition need only be taken over a finite interval of time. Moreover, since $\varphi_t(x)$ is continuous for any fixed time, the norm $|\varphi_t(x) - x^*|$ is also continuous. To show that λ is also continuous, take any $x, y \in U \setminus B_\rho(x^*)$; then

$$\begin{aligned} |\lambda(x) - \lambda(y)| &= \left| \sup_{0\leq t\leq T(\rho)} |\varphi_t(x) - x^*| - \sup_{0\leq t\leq T(\rho)} |\varphi_t(y) - x^*| \right| \\ &\leq \left| \sup_{0\leq t\leq T(\rho)} (|\varphi_t(x) - x^*| - |\varphi_t(y) - x^*|) \right| \\ &\leq \sup_{0\leq t\leq T(\rho)} |\varphi_t(x) - \varphi_t(y)|. \end{aligned}$$

Since $\varphi_t(x)$ is continuous and the interval $[0, T(\rho)]$ is compact, for any $\varepsilon > 0$ there is a $\delta > 0$ such that whenever $|x - y| < \delta$, then $|\varphi_t(x) - \varphi_t(y)| < \varepsilon$. In this case, $|\lambda(x) - \lambda(y)| < \varepsilon$ which implies continuity.

Notice also that $\lambda(x^*) = 0$, and otherwise that $\lambda(x) > 0$, so it satisfies two of the properties that are needed to be a strong Lyapunov function. Moreover, $\lambda(\varphi_t(x)) \leq \lambda(x)$ when $t \geq 0$, because

$$\begin{aligned} \lambda(\varphi_t(x)) &= \sup_{s\geq 0} |\varphi_s(\varphi_t(x)) - x^*| = \sup_{s\geq 0} |\varphi_{s+t}(x) - x^*| \\ &= \sup_{s\geq t} |\varphi_s(x) - x^*|, \end{aligned}$$

and the last expression is definitely not larger than $\lambda(x)$. Consequently, λ is a weak Lyapunov function. We now show that (4.25) is a strong Lyapunov function. Note that for any $t > 0$,

$$L(\varphi_t(x)) = \int_0^\infty e^{-s} \lambda(\varphi_{s+t}(x)) ds \leq \int_0^\infty e^{-s} \lambda(\varphi_s(x)) ds = L(x).$$

If the two sides of this inequality were equal, then $\lambda(\varphi_{t+s}(x)) = \lambda(\varphi_s(x))$ for all $s > 0$. However, this is impossible since if we set $t = (n—1)s$, then we would have $\lambda(\varphi_{ns}(x)) = \lambda(\varphi_s(x))$ for all n. This cannot happen since if $x \neq x^*$, $\lambda(\varphi_s(x)) \neq 0$, but $\varphi_{ns}(x) \to x^*$, so that $\lambda(\varphi_{ns}(x)) \to 0$. □

Although this theorem guarantees that a strong Lyapunov function exists for an asymptotically stable equilibrium, it is not possible to construct it in general unless the flow can be obtained analytically—in which case there is no reason to find L! However, there are cases in which it is not hard to find a Lyapunov function and for which stability is not obvious (see Exercise 4.23).

Example 4.24. The Lorenz system, (1.33), is

$$\begin{aligned} \dot{x} &= \sigma(y - x), \\ \dot{y} &= rx - y - xz, \\ \dot{z} &= xy - bz, \end{aligned} \tag{4.26}$$

where we assume, as in the physical model, that the parameters r, σ, and b are positive. The equilibrium at the origin has linear stability determined by the Jacobian

$$Df(0) = \begin{pmatrix} -\sigma & \sigma & 0 \\ r & -1 & 0 \\ 0 & 0 & -b \end{pmatrix}.$$

The z direction corresponds to an eigenvector with eigenvalue $\lambda = -b$ and is therefore always attracting for $b > 0$. The other two eigenvalues are determined by

$$\lambda^2 + (\sigma+1)\lambda + \sigma(1-r) = 0.$$

This implies that for the x-y dynamics, the origin is attracting when $r < 1$ but is a saddle when $r > 1$. Consequently in three dimensions, the origin is asymptotically stable when $r < 1$ and is unstable when $r > 1$. Linear analysis cannot tell us what happens when $r = 1$.

We now attempt to construct a Lyapunov function. Beginning with a general quadratic in (x, y, z), one can fairly quickly see that the function

$$L = \frac{1}{2}\left(\frac{x^2}{\sigma} + y^2 + z^2\right)$$

will work. Differentiation yields

$$\begin{aligned} \frac{dL}{dt} &= yx - x^2 + ryx - y^2 - xyz + zxy - bz^2 \\ &= (r+1)xy - \left(x^2 + y^2 + bz^2\right) \\ &= -\left(x - \frac{r+1}{2}y\right)^2 - \left(1 - \frac{(r+1)^2}{4}\right)y^2 - bz^2, \end{aligned}$$

where we completed the square on the first two terms to get the third line. Therefore, when $r < 1$ and $b > 0$, this is negative definite, confirming again that the origin is asymptotically stable. Interestingly, this analysis applies for *any* values of (x, y, z), so that the origin is globally asymptotically stable.

When $r = 1$, $dL/dt = 0$ on the line $Z = \{(x,y,z) : x = y, z = 0\}$. This means that L is not a strong Lyapunov function. However, the following argument will imply that since this set is not invariant (because $dz/dt|_Z \neq 0$), the origin is asymptotically stable in this case as well! ■

As in the previous example, it is sometimes possible to conclude that the equilibrium is asymptotically stable for the case that L is a weak Lyapunov function, provided that we know something about the dynamics on the set where $dL/dt = 0$.

Theorem 4.25 (LaSalle's Invariance Principle). *Suppose x^* is an equilibrium for $\varphi_t(x)$ and suppose that L is a weak Lyapunov function on some compact, forward-invariant neighborhood U of x^*. Let $Z = \{x \in U : dL/dt = 0\}$ be the set where L is not decreasing. Then if $\{x^*\}$ is the largest forward invariant subset of Z, it is asymptotically stable and attracts every point in U.*

Proof. For any $x \in U$, suppose z is a limit point of the trajectory $x(t) \in U$. Then $L(\varphi_s(z)) = L(z)$ for all $s > 0$, since if $L(\varphi_s(z)) < L(z)$ we would have a contradiction with the inequalities in (4.23). Consequently, $\varphi_s(z) \in Z$ for all $s > 0$, so that the orbit of z must be a forward invariant subset of Z, and therefore, by assumption, $z = x^*$. □

Example 4.26. A slightly more realistic model than the logistic equation (1.7) adds "delay," modeling the fact that the gestation period is nonzero, and so the competition that affects current births is in the past. One type of delay is to introduce a second variable y that represents the population at an earlier era. The model then becomes

$$\begin{aligned} \dot{x} &= rx(1-y), \\ \dot{y} &= b(x-y). \end{aligned}$$

Note that at equilibrium $y = x$ and so $x = 0$ or $x = 1$ as for (1.7). Our goal is to show that the point $(1,1)$ is the limit of all initial conditions in the positive quadrant. First note that the positive quadrant is forward invariant. To leave it, the orbit would have to pass through the x- or y-axis. When $x = 0$, $\dot{x} = 0$, so this is an invariant line. When $y = 0$, then $\dot{y} \geq 0$, so the orbit cannot cross to negative y.

We next transform to coordinates centered at the equilibrium of interest. Let $(\xi, \eta) = (x-1, y-1)$ so that

$$\begin{aligned} \dot{\xi} &= -r\eta(1+\xi), \\ \dot{\eta} &= b(\xi - \eta). \end{aligned}$$

Note that $(0,0)$ is a linearly stable equilibrium for this equation when b and r are positive since then $\operatorname{tr}(Df(0,0)) = -b < 0$ and $\det(Df(0,0)) = rb > 0$ (recall §2.2). A simple quadratic function will not work as a Lyapunov function for this system, nor will any polynomial of finite order. However, after some guesswork—see (MacDonald 1978)—a Lyapunov function can be found:

$$L(\xi, \eta) = \xi - \ln(1+\xi) + \frac{r}{2b}\eta^2.$$

Note that $L(0,0) = 0$, and that since $\xi - \ln(1+\xi) \geq 0$ when $\xi > -1$, then L is otherwise positive. Furthermore, differentiation gives

$$\frac{dL}{dt} = -r\eta(\xi+1)\left(1 - \frac{1}{\xi+1}\right) + r\eta(\xi - \eta) = -r\eta^2.$$

Accordingly, L is strictly decreasing except on the set $Z = \{(\xi,\eta), \eta = 0, \xi > -1\}$. However, the equations of motion imply that the only invariant point in Z is the origin since $\dot{\eta}|_Z = b\xi \neq 0$ otherwise. Therefore, according to LaSalle's invariance principle, $(0,0)$ attracts the orbits of all initial conditions with $\xi > -1$. Equivalently, in the original coordinates, the point $(1,1)$ is the forward limit of all points in the right half-plane. ■

Hamiltonian systems—recall §1.4—often have Lyapunov functions. Suppose that $H : \mathbb{R}^2 \to \mathbb{R}$, and consider the Hamiltonian system

$$\dot{x} = \frac{\partial H}{\partial y}, \quad \dot{y} = -\frac{\partial H}{\partial x}. \tag{4.27}$$

The value of $H(x,y)$ typically represents the "energy" of the system. It is constant along trajectories, because

$$\frac{dH}{dt} = \frac{\partial H}{\partial x}\dot{x} + \frac{\partial H}{\partial y}\dot{y} = \frac{\partial H}{\partial x}\frac{\partial H}{\partial y} - \frac{\partial H}{\partial y}\frac{\partial H}{\partial x} \equiv 0. \tag{4.28}$$

Therefore, if $H(x_o, y_o) = E$, then so does $H(x(t), y(t))$. If (x^*, y^*) is an equilibrium, then the function

$$L(x,y) = H(x,y) - H(x^*, y^*)$$

is zero at the equilibrium and constant along trajectories; consequently, if it can be shown that L is positive in some neighborhood of the equilibrium, then it is a weak Lyapunov function.

Example 4.27. Consider the system

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= x - 3ax^2. \end{aligned} \tag{4.29}$$

These equations have the form (4.27), since if $y = \partial H / \partial y$, then $H(x,y) = {}^1\!/\!_2 y^2 + V(x)$, for an arbitrary function V. Similarly, demanding that $x - 3ax^2 = -\partial H/\partial x$ gives $H(x,y) = T(y) - {}^1\!/\!_2 x^2 + ax^3$, for an arbitrary function T. These two equations are consistent, implying that (4.29) is Hamiltonian and we obtain $H(x,y) = {}^1\!/\!_2(y^2 - x^2) + ax^3$.

The system (4.29) has two equilibria, $(0,0)$ and $(1/3a, 0)$. The first is a saddle, and the second is a center. The Hamiltonian provides a Lyapunov function in a neighborhood of the center. We can see this most easily by shifting coordinates, defining $\xi = x - 1/3a$ to obtain

$$H = 1/2\left(y^2 + \xi^2\right) + a\xi^3 + H(1/3a, 0).$$

Therefore, for ξ small enough, H has contours about $y = \xi = 0$ that are approximately circular. In conclusion, $L = {}^1\!/\!_2\left(y^2 + \xi^2\right) + a\xi^3$ is a weak Lyapunov function, and the equilibrium $(1/3a, 0)$ is a "topological center"—see §6.2. ■

We will discuss more examples of this type in §5.1 (see also Exercise 8).

Although Hamiltonian systems correspond to "conservative" dynamics, engineering systems often have damping.

Example 4.28. Suppose $x \in \mathbb{R}^n$ are coordinates and $y \in \mathbb{R}^n$ are the conjugate momenta, with the Hamiltonian $H(x,y) = {}^1\!/\!_2|y|^2 + V(x)$. Here, $V(x)$, the potential energy, gives rise to the force $F = -\nabla V$. This system is conservative; the simplest model for damping is an additional force proportional to the momentum, which gives the set of equations

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -\nabla V(x) - \gamma y, \end{aligned} \tag{4.30}$$

where γ is the damping coefficient. The "energy" of this system is given by the function $H(x,y)$. If we assume that $\nabla V(0) = 0$, so that the origin is an equilibrium, then the origin is a critical point of H, since

$$\nabla H = (\nabla V(x), y)^T.$$

Moreover, when $D^2V(0)$ is a positive definite matrix, the Hessian matrix,

$$D^2H(0) = \begin{pmatrix} D^2V(0) & 0 \\ 0 & I \end{pmatrix},$$

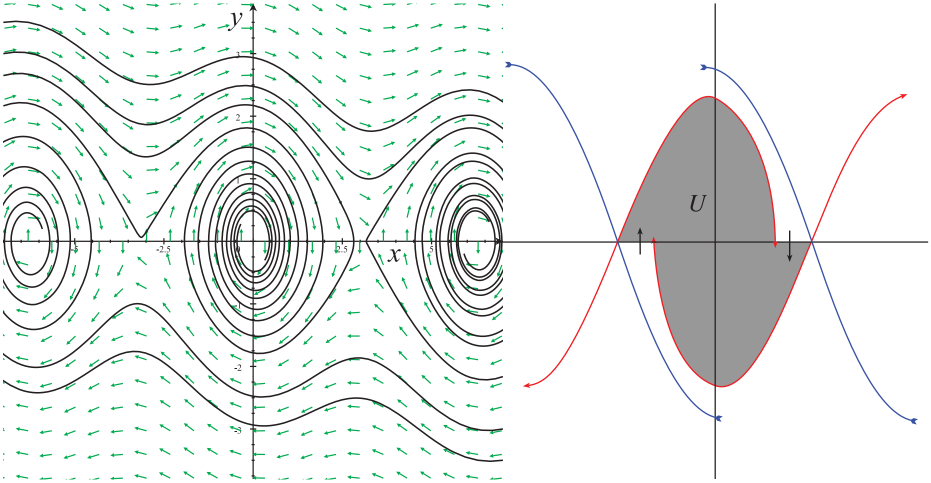

Figure 4.10. *Phase space of the damped pendulum* (4.30) *with* $V(x) = -\cos x$, *and* $\gamma = 0.1$. *V has critical points on the x-axis at $n\pi$. The points* $(2k\pi, 0)$ *are asymptotically stable, while* $((2k+1)\pi, 0)$ *are saddles. On the right is shown a forward invariant region U enclosing the origin. U is bounded by pieces of the unstable manifolds (see* §5.1*) of the saddles at $x = \pm\pi$ and by part of the x-axis. To prove that U exists, we would have to show that the unstable manifolds (see Chapter* 5*) of the saddles first cross the x-axis in the interval* $(-\pi, \pi)$.

is also positive definite so that the origin is a minimum of H. In this case, the contours of H are closed near the origin. Moreover,

$$\frac{dH}{dt} = y \cdot (-\nabla V - \gamma y) + y \cdot \nabla V = -\gamma |y|^2 \leq 0;$$

therefore, the origin is stable.

If 0 is the only critical point of V, then LaSalle's invariance principle implies that the origin is asymptotically stable. The set for which $dH/dt = 0$ is $Z = \{(x,y) : y = 0\}$. Now since $\dot{y}|_Z = -\nabla V(x)$, whenever x is not a critical point of V, then $\dot{y} \neq 0$ on Z. We can conclude that if 0 is the only critical point of V, the only invariant subset of Z is the origin.

The analysis above could be generalized to the case where there are more critical points of V if it could be proved that there exists a neighborhood, U, of the origin—like that depicted in Figure 4.10—that does not include other critical points and that is forward invariant. ∎

4.7 ▪ Topological Conjugacy and Equivalence

An important task in dynamical systems is to determine whether two dynamical systems that seemingly look "different" are actually the same but are just written in different forms. A system that looks complicated may actually be quite simple in a different coordinate system. A classification of equivalent systems will considerably reduce the work to be done, for example, in bifurcation theory (see Chapter 8). Moreover, the study of these equivalence classes leads to notions of sensitivity of dynamics to modification of the system—what is called structural stability.

There are several different notions of equivalence, depending upon the degree of smoothness required for the transformation. The definitions require some notions

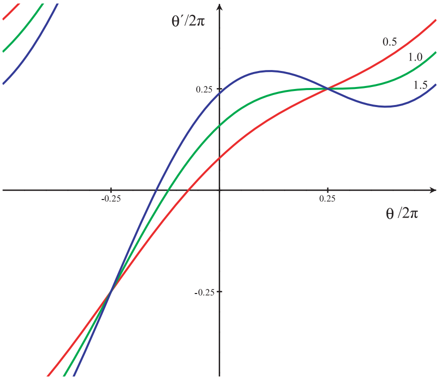

Figure 4.11. *The function* (4.31) *for* $a = 0.5$, 1.0, *and* 1.5. *The last case is not a homeomorphism since the graph is not monotone.*

from basic set theory and topology. Suppose that A and B are two topological spaces (recall §3.1). A map $h : A \to B$ is

- *surjective* or *onto* if for every $b \in B$, there is at least one $a \in A$ such that $h(a) = b$,
- *injective* or *one-to-one* if whenever $h(a) = h(a')$, then $a = a'$, and
- *bijective* if it is both surjective and injective.

Note that a bijective map has an inverse: since for each b there is exactly one a such that $b = h(a)$, the map $h^{-1} : B \to A$ is defined by setting $a = h^{-1}(b)$. Note that h^{-1} is both a left and a right inverse for h: $h(h^{-1}(b)) = b$ and $h^{-1}(h(a)) = a$. These notions are used to define one of the most fundamental concepts in topology:

- *homeomorphism*: A map $h : A \to B$ is a *homeomorphism* if it is continuous, is bijective, and has a continuous inverse.

For example, the map $h : (0, \infty) \to (0, 1)$ defined by $h(x) = 1\big/(1 + x^2)$ is a homeomorphism. Similarly, the map $f : \mathbb{S} \to \mathbb{S}$ defined by

$$f(\theta) = \theta + a \cos\theta \tag{4.31}$$

is a homeomorphism only when $|a| \le 1$, since it is otherwise not one-to-one; see Figure 4.11.[23]

Topology declares that two spaces are equivalent if there is a homeomorphism from one to the other. It is this notion that implies that a mug of coffee and a doughnut are the "same" (though one gives you a buzz from caffeine and the other from sugar). Conversely, if it can be shown that there is no homeomorphism from one space to another, then they are topologically distinct spaces.

It is natural to also define a notion of "smooth" equivalence:

[23] Challenge for the topologically inclined: find an example of a continuous, bijective map between compact sets that is not a homeomorphism. At least one of the spaces must have an exotic topology, because every continuous, bijective map from a compact space to a Hausdorff space is a homeomorphism (Hocking and Young 1961).

▷ *diffeomorphism*: A map $f : A \to B$ is a diffeomorphism if it is a C^1 bijective map with a C^1 inverse.

For example, $f : \mathbb{R} \to \mathbb{R}$, given by $f(x) = x + {}^1\!/_2 \sin x$ is a diffeomorphism, but $f(x) = x^3$ is not because its inverse, $f^{-1}(x) = x^{\frac{1}{3}}$, is not C^1. Note that every diffeomorphism is also a homeomorphism. Recall from §4.2 that a flow is a C^1 bijection from the phase space to itself with a C^1 inverse, and thus the map φ_t for each time t is a diffeomorphism.

With these definitions in our toolbox, we are now prepared to understand the key notion of equivalence of two flows,

▷ *topological conjugacy*: Two flows $\varphi_t : A \to A$ and $\psi_t : B \to B$ are conjugate if there exists a homeomorphism $h : A \to B$ such that for each $x \in A$ and $t \in \mathbb{R}$

$$h(\varphi_t(x)) = \psi_t(h(x)). \tag{4.32}$$

It is clear that for such a homeomorphism to exist, A and B must be topologically equivalent spaces. Often, two systems are simply said to be *conjugate* as a shorthand for topologically conjugate. A diagram that represents (4.32) is

$$\begin{tikzcd} x \arrow[r, "\varphi_t"] \arrow[d, "h"'] & \varphi_t(x) \arrow[d, "h"] \\ y \arrow[r, "\psi_t"] & \psi_t(y) \end{tikzcd} .$$

The two paths in this diagram, $x \xrightarrow{h} y \xrightarrow{\psi_t} \psi_t(y)$ and $x \xrightarrow{\varphi_t} \varphi_t(x) \xrightarrow{h} \psi_t(y)$, which represent the right- and left-hand sides of (4.32), respectively, must give the same result, namely, $\psi_t(h(x))$. We say, in this case, that the "diagram commutes."

Example 4.29. The flow on $\mathbb{R}$ generated by $\dot{x} = -x$ is $\varphi_t(x) = xe^{-t}$. Under the homeomorphism $y = h(x) = x^3$, this is equivalent to the new flow

$$\psi_t(y) = (xe^{-t})^3 = ye^{-3t}.$$

This is the solution of the linear equation $\dot{y} = -3y$. Consequently, these two ODEs are topologically conjugate. ■

Conjugacy implies that each trajectory of ψ corresponds to a trajectory of φ, and vice versa. For example, if x^* is an equilibrium of φ, then since $\varphi_t(x^*) = x^*$ for all $t, \psi_t(h(x^*)) = h(x^*) = y^*$ and so y^* is an equilibrium of ψ. Thus, h provides a one-to-one correspondence between the equilibria of two conjugate flows. Similarly, if $\varphi_t(x_o)$ is a periodic orbit of φ with period T, i.e., $\varphi_{t+T}(x_o) = \varphi_t(x_o)$, then $\psi_t(y_o) = h(\varphi_t(x_o)) = h(\varphi_{t+T}(x_o)) = \psi_{t+T}(y_o)$, so $\psi_t(y_o)$ is also a periodic orbit of ψ with the same period; see Figure 4.12.

Topological conjugacy can be too restrictive a condition because, in addition to the fact that trajectories "look" the same in phase space, (4.32) implies that the curves have identical temporal parameterizations. A slightly more general notion that still captures the shape and direction of the flows as curves in phase space is

▷ *topological equivalence*: Two flows $\varphi_t : A \to A$ and $\psi_t : B \to B$ are equivalent if there exists a homeomorphism $h : A \to B$ that maps the orbits of φ onto the orbits of ψ and preserves the direction of time. That

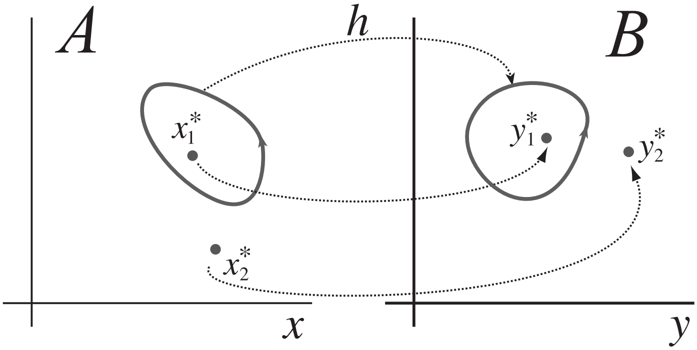

Figure 4.12. *Orbits of conjugate systems must be in a one-to-one correspondence.*

is, there is a surjective map $\tau : A \times \mathbb{R} \to \mathbb{R}$ that is monotone increasing with t and

$$h(\varphi_{\tau(x,t)}(x)) = \psi_t(h(x)). \tag{4.33}$$

Example 4.30. If we temporarily relax the requirement that a flow exist for all time, then

$$\psi_t(y) = \frac{y}{1+ty}$$

is the flow corresponding to the ODE $\dot{y} = -y^2$. For $y \in \mathbb{R}^+$, it exists only on the interval $t \in (-y^{-1}, \infty)$. This flow is equivalent to $\varphi_t(x) = xe^{-t}$ under the transformations $h(x) = x$, and $\tau(x,t) = \ln(1+xt)$, since

$$h(\varphi_{\tau(x,t)}(x)) = xe^{-\ln(1+xt)} = \frac{x}{1+xt} = \psi_t(h(x)).$$

Note that the orbits of ψ are qualitatively the same as those of φ; for example, the point $y = h(0) = 0$ is an equilibrium, and if $y > 0$, then $\psi_t(y) \to 0$ as $t \to \infty$, just as $\varphi_t(x) \to 0$. We used this notion of equivalence in our proof of the theorem in §4.3 that each ODE is equivalent to one with a complete flow. ∎

Two topologically equivalent flows must, in some precise sense, exhibit the same "orbit structure." In particular, for the one-dimensional case, it is quite easy to make a precise statement since the behavior is quite limited.

Theorem 4.31 (One-Dimensional Equivalence). *Two flows φ and ψ in $\mathbb{R}$ are topologically equivalent if and only if their equilibria, ordered on the line, can be put in a one-to-one correspondence, and if and only if the corresponding equilibria have the same topological type (sink, source, or semistable).*

Proof. If a homeomorphism h exists, then to each equilibrium of φ there must be a corresponding equilibrium of ψ and vice versa; thus we can put the equilibria in a one-to-one correspondence. The correspondence is ordered since h is monotone. Conversely, suppose that φ and ψ have corresponding equilibria. We will next explicitly construct h, and show that the flows not only are equivalent but are actually conjugate.[24]

[24] The necessity also follows from the Hartman–Grobman theorem.

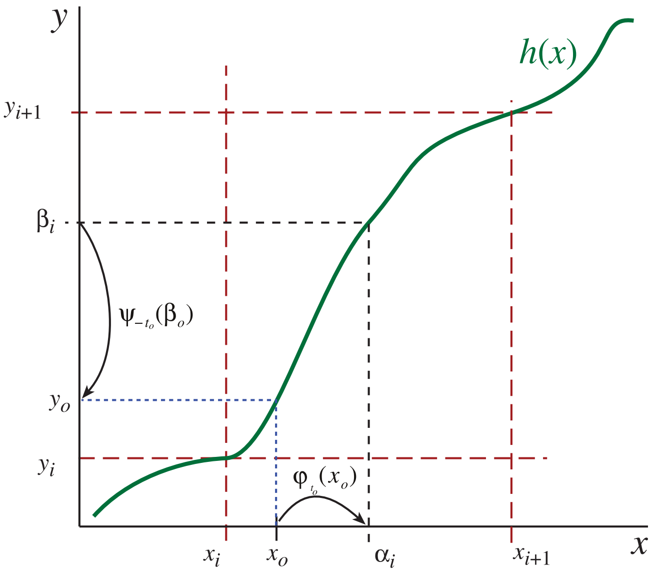

Figure 4.13. *Construction of a homeomorphism for a one-dimensional flow.*

Suppose first, for simplicity, that there are finitely many equilibria. Denote the equilibria of φ by $x_1^* < x_2^* < \cdots < x_n^*$ and of ψ by $y_1^* < y_2^* < \cdots < y_n^*$. It is clear that we must define $h(x_i^*) = y_i^*$. Choose points α_i such that $\alpha_o < x_1^* < \alpha_1 < x_2^* < \cdots < x_n^* < \alpha_n$, and points β_i that are similarly intertwined with y_i^*, as shown in Figure 4.13. We can arbitrarily define $h(\alpha_i) = \beta_i$. To complete the construction of the homeomorphism in an interval between two equilibria $h : (x_i^*, x_{i+1}^*) \to (y_i^*, y_{i+1}^*)$, note that for each $x_o \in (x_i^*, x_{i+1}^*)$, since $\varphi_t(x_o)$ is either monotonically increasing or decreasing with t, there is a *unique* time $t_o \in \mathbb{R}$ such that $\varphi_{t_o}(x_o) = \alpha_i$. As sketched in Figure 4.13, define

$$h(x_o) = y_o = \psi_{-t_o}(\beta_i).$$

This function is a homeomorphism (it is one-to-one since the flow is monotone, and it is continuous and has a continuous inverse since ψ does). Note also that since $\varphi_{t_o-t}(\varphi_t(x_o)) = \alpha_i$ we have

$$h(\varphi_t(x_o)) = \psi_{-(t_o-t)}(\beta_i) = \psi_t\left(\psi_{-t_o}(\beta_i)\right) = \psi_t(h(x_o)),$$

as required. This construction applies in each such interval bounded by two equilibria. We can similarly deal with the two intervals $(-\infty, x_1^*)$ and (x_n^*, ∞). This yields the required homeomorphism on $\mathbb{R}$.

If the number of equilibria is countably infinite, or even uncountably infinite, the analysis is similar. □

Generally, when the dimension of the phase space is larger than one, we must know more than just the number and topological type of the equilibria to determine whether two flows are equivalent; see Exercise 13. We will see such systems in §8.11 when we discuss homoclinic bifurcations.

Sometimes we will not be satisfied by mere topological equivalence—we will want differential properties to be the same. In Example 4.29 we saw that the eigenvalues are not preserved by a topological equivalence (they changed from -1 to -3 at the equilibrium). A notion that does preserve this information is

▷ *diffeomorphic*: Two flows $\varphi_t : A \to A$ and $\psi_t : B \to B$ are diffeomorphic if there is a diffeomorphism h such that $h(\varphi_t(x)) = \psi_t(h(x))$.

We also call two flows *smoothly equivalent* when, in addition to the diffeomorphism h, there is an increasing diffeomorphism $\tau(x,t)$ such that (4.33) is satisfied.

Example 4.32. The map $h : \mathbb{R} \to (-1,1)$ defined by $h(x) = \tanh(x)$ is a diffeomorphism. Applying this to the flow $\varphi_t(x) = xe^{-t}$ gives the new flow $\psi_t = h \circ \varphi_t \circ h^{-1}$, or explicitly

$$\psi_t(y) = \tanh\left(e^{-t}\tanh^{-1}(y)\right).$$

This flow has the vector field

$$\dot{y} = g(y) = \frac{d}{dt}\,\psi_t(y)|_{t=0} = \left(y^2 - 1\right)\tanh^{-1}(y).$$

This ODE has only one equilibrium, $y = 0$, in the interval $(-1,1)$; since $Dg(0) = -1$, it is stable just like $x = 0$ is for the flow φ. The new flow has equilibria at $y = \pm 1$ as well, but these are not within the space $(-1,1)$; they correspond to the points $x = \pm\infty$ in the original space. The limiting behavior $\psi_t(y) \underset{t\to-\infty}{\to} \mp 1$, for $y < 0$ and $y > 0$, respectively, reflects the behavior of the original flow, since $\varphi_t(x) \underset{t\to-\infty}{\to} \mp\infty$ for $x < 0$ and $x > 0$, respectively. ■

Although Example 4.29 showed that the flows xe^{-t} and ye^{-3t} are topologically conjugate, we did not show them to be diffeomorphic, since x^3 is not a diffeomorphism. In fact, these two flows cannot be diffeomorphic, as we will see next.

If two flows are diffeomorphic, then the vector fields are related by the derivative of the conjugacy. Suppose that $\dot{x} = f(x)$ generates the flow φ and $\dot{y} = g(y)$ generates ψ. Then

$$\frac{d}{dt}\psi_t(y) = g\left(\psi_t(y)\right) = \frac{d}{dt}h\left(\varphi_t(x)\right) = Dh\left(\varphi_t(x)\right)\frac{d}{dt}\varphi_t(x) = Dh\left(\varphi_t(x)\right)f\left(\varphi_t(x)\right).$$

Setting $t = 0$ in these relations gives a relation between the vector fields:

$$g(y) = g(h(x)) = Dh(x)f(x). \tag{4.34}$$

Equation (4.34), sketched in Figure 4.14, is precisely the result that we would obtain if we simply transform coordinates using the differential equations:

$$y = h(x) \;\Rightarrow\; \frac{dy}{dt} = Dh(x)\frac{dx}{dt} = Dh(x)f(x) = g(y).$$

It is easy to see that the eigenvalues of equilibria are preserved by a diffeomorphism. Suppose that x^* is an equilibrium of φ, and $D_x f(x^*) = A$ is the Jacobian matrix. Since $f, g \in C^1$, (4.34) implies that Dh is as well, and so differentiating this equation by x yields

$$D_y g(y) D_x h(x) = D_x h(x) D_x f(x) + D_x^2 h(x) f(x).$$

Since h is a diffeomorphism, it has a differentiable inverse and so the matrix $H = Dh(x^*)$ is nonsingular. Consequently, since $f(x^*) = 0$ at the equilibrium,

$$B \equiv D_y g(y^*) = HAH^{-1}.$$

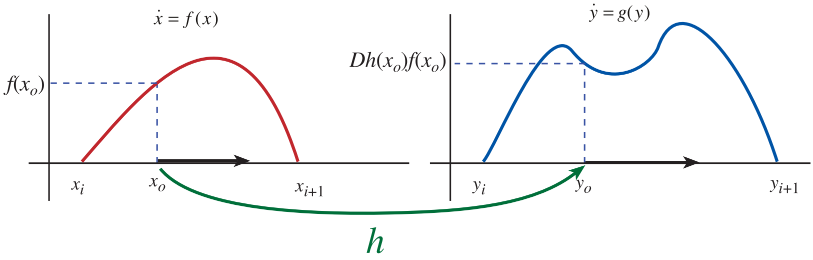

Figure 4.14. *Equivalence between two one-dimensional vector fields,* (4.34).

Therefore, the matrices are related by a similarity transformation and therefore have the same eigenvalues (recall Exercise 2.8).

Note in particular that two linear flows can be diffeomorphic only if the fundamental subspaces E^u, E^s, and E^c have the same dimensions; we will see below that this holds more generally. Conversely, two linear ODEs with distinct eigenvalues cannot be diffeomorphic; see Exercise 6. Indeed, two linear flows are diffeomorphic only if their matrices are similar, as shown below.

Theorem 4.33 (Linear Conjugacy). *The flows φ_t and ψ_t of the linear systems $\dot{x} = Ax$ and $\dot{y} = By$ are diffeomorphic if and only if the matrix A is similar to the matrix B.*

Proof. Assume first that A is similar to B, i.e., there is a nonsingular matrix H such that $HA = BH$. The map $h(x) = Hx$ is clearly a diffeomorphism and

$$h(\varphi_t(x)) = He^{tA}x = e^{tHAH^{-1}}Hx = e^{tB}h(x) = \psi_t(h(x)),$$

which implies that the flows φ and ψ are diffeomorphic. Conversely, suppose there is a diffeomorphism g such that $g(\varphi_t(x)) = \psi_t(g(x))$. Setting $g(0) = c$, then $g(\varphi_t(0)) = c = \psi_t(c)$, so c is an equilibrium of ψ. Let $h(x) = g(x) - c$. Then

$$h(\varphi_t(x)) = \psi_t(g(x)) - c = \psi_t(h(x) + c) - c = \psi_t(h(x)). \tag{4.35}$$

Thus $h(x)$ conjugates the flows and fixes the origin. Define the matrix $H = Dh(0)$, and differentiate (4.35) with respect to x, to obtain, at $x = 0$, $He^{tA} = e^{tB}H$. Taking the time derivative of this relation at $t = 0$ yields $HA = BH$, so the matrices are similar. □

Example 4.34. The matrices

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

are not similar. Indeed, suppose there were an invertible matrix such that $HA = BH$. Then if $(u, v)^T$ is a column of H, we would have $-2u + v = -2u$ and $-2v = -2v$; consequently, $v = 0$ and $u = c$. Since this is true for each column, H would be singular. However, there does exist a topological conjugacy between the flows $\varphi_t(x) = e^{tA}x$ and $\psi_t(y) = e^{tB}y$. To find $y = h(x) = (h_1(x_1, x_2), h_2(x_1, x_2))$, we first find the flows

$$\begin{aligned} \varphi_t(x_1, x_2) &= \left(e^{-2t}x_1, e^{-2t}x_2\right), \\ \psi_t(y_1, y_2) &= \left(e^{-2t}(y_1 + ty_2), e^{-2t}y_2\right). \end{aligned}$$

The second component of the conjugacy $h_2(\varphi_t(x)) = \psi_{2t}(y)$ implies

$$h_2\left(e^{-2t}x_1, e^{-2t}x_2\right) = e^{-2t}y_2 = e^{-2t}h_2(x_1,x_2),$$

which has a particular solution $h_2(x_1,x_2) = x_2$. The first component of the conjugacy requires that $h_1\left(e^{-2t}x_1, e^{-2t}x_2\right) = e^{-2t}\left(h_1(x_1,x_2) + tx_2\right)$. To solve this, set $h_1(x_1,x_2) = x_1 + f(x_2)$, to find

$$f(e^{-2t}x_2) = e^{-2t}\left(f(x_2) + tx_2\right).$$

A solution to this is $f(x) = -{}^1\!/\!_2 x \ln|x|$, and if we define $f(0) = 0$, then f is continuous at $x = 0$. Putting this result together with h_1 gives the homeomorphism

$$(y_1,y_2) = h(x) = \left(x_1 - \tfrac{1}{2}x_2\ln|x_2|, x_2\right);$$

however, h is not a diffeomorphism since its derivative does not exist at the origin. At every other point the vector fields can be transformed using (4.34):

$$\begin{aligned} \dot{y}_1 &= \dot{x}_1 - \tfrac{1}{2}\dot{x}_2\ln|x_2| - \tfrac{1}{2}\dot{x}_2 = -2y_1 + y_2, \\ \dot{y}_2 &= \dot{x}_2 = -2y_2, \end{aligned}$$

showing conjugacy as we expect. ∎

This example can be generalized to show that topological conjugacy of hyperbolic systems depends only on the dimensions of their stable and unstable subspaces: for example a system with complex eigenvalues can be conjugate to one with real eigenvalues; see Exercise 7.

Theorem 4.35. *Suppose A and B are two real, hyperbolic $n \times n$ matrices and $\varphi_t(x) = e^{tA}x$ and $\psi_t(y) = e^{tB}y$ the corresponding flows. Then φ and ψ are topologically conjugate if and only if the dimensions of the stable and unstable spaces of A are equal to the corresponding dimensions for B.*

Proof (Sketch). The necessity of this condition is easy to see. Any homeomorphism $h : \mathbb{R}^n \to \mathbb{R}^n$ must map bounded sets to bounded sets. Moreover, for any $x \in E^s_A$, we have $\lim_{t\to\infty}\varphi_t(x) = 0$; consequently, since h is continuous $\lim_{t\to\infty} h(\varphi_t(x)) = h(0) = \lim_{t\to\infty}\psi_t(h(x))$. Since $h(0)$ is bounded, then $y = h(x)$ must be in E^s_B, and indeed $h(0) = 0$ because every orbit in E^s_B approaches the origin. Consequently, $h : E^s_A \to E^s_B$ is a homeomorphism, which implies that these spaces must have the same dimension. The same can be said for the unstable spaces.

The proof of the converse requires a bit more work: given that the dimensions of the stable and unstable spaces are the same we must construct the conjugacy. Since the stable spaces E^s_A and E^s_B are invariant under the flows, we start by constructing a map $h_s : E^s_A \to E^s_B$. A similar map h_u can be constructed for the unstable spaces. In the end, we write any vector $x = \pi_u(x) + \pi_s(x)$, where π_i are projection operators onto the unstable and stable spaces of A, respectively, and the full conjugacy is then $h(x) = h_s(\pi_s(x)) + h_u(\pi_u(x))$.

The proof is simple for the case when A and B are semisimple and all their eigenvalues are real. Then both A and B are linearly conjugate to real diagonal matrices and so to the systems $\dot{x}_i = \lambda_i x_i$ and $\dot{y}_i = \mu_i y_i$. Order the eigenvalues so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k > 0 > \lambda_{k+1} \geq \cdots \geq \lambda_n$ and similarly for μ_i. By assumption the number, k, of positive eigenvalues must be the same. Now we construct conjugacies for each one-dimensional system, mapping λ_i to μ_i, by choosing

$$h_i(x_i) = \operatorname{sgn}(x_i)|x_i|^{a_i}, \text{ where } a_i = \mu_i/\lambda_i.$$

Then $h_i(e^{\lambda_i t}x_i) = e^{\mu_i t}\operatorname{sgn}(x_i)|x_i|^{a_i} = e^{\mu_i t}h_i(x_i)$. Whenever $\lambda_i \neq \mu_i$, h_i is *not* a diffeomorphism. Note also that we cannot get out of this difficulty by relaxing the conjugacy requirement to one of equivalency, since the ratio of the eigenvalues may be different for different i, and thus we would need a different time scaling for each dimension.

In the general case we construct h_s by first finding norms that are adapted to the matrices A and B. These norms are constructed so that $||e^{tA}\pi_s(x)||_A \leq e^{-\alpha t}||\pi_s(x)||_A$ for $t \geq 0$, i.e., eliminating the constant K in (4.20). The point of these norms is that each trajectory crosses its respective unit sphere $||x|| = 1$ exactly once. The unit spheres in the new norms are then used to define h_s as the "identity" map from the A-sphere to the B-sphere. The homeomorphism is extended from the spheres by flowing, just like we did for the one-dimensional case. The full proof is given, for example, in (Robinson 1999, see §4.7). □

4.8 ▪ Hartman–Grobman Theorem

We showed in §4.7 that linear, hyperbolic systems come in a few equivalence classes, categorized solely by the dimension of their stable and unstable spaces. Now we show that nonlinear systems sometimes "look like" their linearizations near hyperbolic equilibria. The formal statement of this result was proved independently by Hartman in 1960 and Grobman in 1959.

Theorem 4.36 (Hartman–Grobman). *Let x^* be a hyperbolic equilibrium point of a C^1 vector field $f(x)$ with flow $\varphi_t(x)$. Then there is a neighborhood N of x^* such that φ is topologically conjugate to its linearization on N.*

It is interesting to note that while the theorem requires a smooth ODE, it does not say that the flow is diffeomorphic to its linearization. A theorem due to Sternberg does provide a diffeomorphism; however, it requires an additional hypothesis: the eigenvalues must satisfy a "nonresonance" condition (Sternberg 1958).

Note that the Hartman–Grobman theorem requires that the equilibrium be hyperbolic. As we shall see in Chapter 6, the topological classification of nonhyperbolic equilibria will depend upon more than just the linearization of the system.

Proof (Discussion). The construction of the homeomorphism is rather clever and potentially useful, so we sketch it here. As is now usual, we begin with an ODE of the form

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

where A is a hyperbolic matrix and $g \in C^1$ represents the nonlinear terms, so that $g = o(x)$. Define also the flow of the linear equation $\psi_t(x) = e^{tA}x$. Since the theorem is to be proved only locally, we can modify the ODE by defining a new nonlinearity $\tilde{g}$ such that $\tilde{g}(x) = g(x)$ for some neighborhood N of 0, and $\tilde{g}(x) = 0$ for x outside some larger neighborhood $\tilde{N}$. This can be done so that $\tilde{g}$ is still a smooth function. Moreover, $\tilde{g}$ is bounded, since it vanishes outside a compact set. Let φ_t be the flow for the modified ODE. This flow agrees with the linear flow while the orbit stays outside $\tilde{N}$. (See the following examples to understand why this modification is needed.)

Our goal is to find a homeomorphism h satisfying

$$\psi_t(h(x)) = h(\varphi_t(x)), \text{ that is, } h(x) = e^{-tA} \circ h \circ \varphi_t(x). \tag{4.36}$$

Suppose first that H is a homeomorphism that satisfies (4.36) for one value of time, say, $t = 1$, e.g.,

$$H_1(x) = e^{-A} H_1(\varphi_1(x)). \tag{4.37}$$

In addition, suppose we can show that H_1 is the unique such homeomorphism (among the class of continuous functions such that $H_1 - id$ is bounded). Now let

$$H_t(x) = e^{-tA} \circ H_1 \circ \varphi_t(x);$$

a sketch of this relation is shown in Figure 4.15. We then claim that H_t is also a homeomorphism that satisfies (4.37). This follows from the group property of the flow φ:

$$\begin{aligned} e^{-A} \circ H_t \circ \varphi_1(x) &= e^{-A} \circ e^{-tA} \circ H_1 \circ \varphi_t \circ \varphi_1(x) \\ &= e^{-tA} \circ e^{-A} \circ H_1 \circ \varphi_1 \circ \varphi_t(x) \\ &= e^{-tA} \circ H_1 \circ \varphi_t(x) = H_t(x). \end{aligned}$$

Consequently, H_t satisfies (4.37); however, since we asserted that H_1 is the unique such homeomorphism, we must have $H_t = H_1$. Therefore,

$$H_1 = e^{-tA} \circ H_1 \circ \varphi_t(x).$$

So H_1 is also the homeomorphism for any time t! This can be seen as well by considering the following diagram:

$$\begin{tikzcd} x \arrow[r, "\varphi_t"] \arrow[d, "H_1"'] & x(t) \arrow[r, "\varphi_{1-t}"] \arrow[d, "H_1"'] & x(1) \arrow[d, "H_1"'] \\ y \arrow[r, "e^{tA}"] & y(t) \arrow[r, "e^{(1-t)A}"] & y(1) \end{tikzcd}.$$

We have just shown that this diagram commutes; that is, if we go from x to $y(1)$ by any path we obtain the same result.

So we reduce the problem to solving for H_1, the conjugacy at $t = 1$. Basically, we can do this iteratively by starting with the assumption that $H_1^{(0)}(x) = x$, and defining

$$H_1^{(i+1)}(x) = e^{-A} \circ H_1^{(i)} \circ \varphi_1(x), \; i = 0, 1, \ldots. \tag{4.38}$$

The theorem actually proves that there is a neighborhood of the origin for which (a version of) this iteration converges and that H_1 is unique among all homeomorphisms that are near the identity.

The full proof of the theorem is in (Robinson 1999, §5.7). □

Example 4.37. The simple two-dimensional system

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

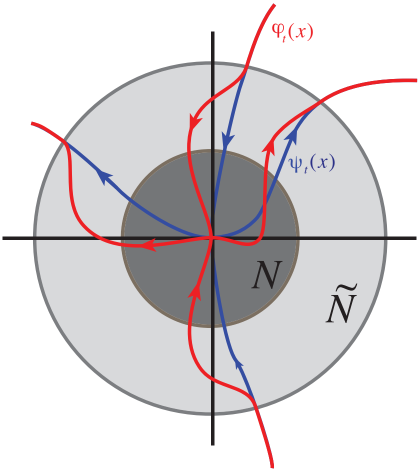

Figure 4.15. *The homeomorphism* (4.36).

has a saddle equilibrium at the origin. The linear matrix for the saddle is $A = \left(\begin{smallmatrix}1 & 0\\ 0 & -1\end{smallmatrix}\right)$, which is conveniently diagonal, so that $e^{tA}(x,y)^T = (e^t x, e^{-t}y)^T$. The nonlinear system can be easily solved analytically to obtain the flow

$$\varphi_t(x,y) = \begin{pmatrix} e^t x \\ e^{-t}y + \frac{1}{3}\left(e^{2t} - e^{-t}\right)x^2 \end{pmatrix}.$$

As a consequence, the homeomorphism $H = H_1$ in (4.37) must satisfy the equation

$$H(x,y) = e^{-A}H(\varphi_1(x,y)) = \begin{pmatrix} e^{-1} & 0 \\ 0 & e \end{pmatrix} H(ex, e^{-1}y + kx^2), \qquad (4.39)$$

upon defining $k = \frac{e^3-1}{3e}$. It is convenient to solve for the two components of H separately; let $H = (K,L)^T$. Then the iterative equation (4.38) for K is

$$K^{(i+1)}(x,y) = \frac{1}{e}K^{(i)}\left(ex, \frac{1}{e}y + kx^2\right).$$

The superscripts on this equation indicate that we will attempt to solve it iteratively. Starting with $K^{(0)}(x,y) = x$, the identity, then $K^{(1)} = \frac{1}{e}(ex) = x$; thus, $K(x,y) = x$ is the solution. The formal iterative equation for L from (4.39) is

$$L^{(i+1)}(x,y) = eL^{(i)}\left(ex, \frac{1}{e}y + kx^2\right),$$

which looks like it should be amenable to iteration in the same way. However, because there is a factor of e in front of the right-hand side, we cannot iterate this equation in the form as written—the result will diverge (try it and see!). Instead we must invert it. To do this, set $\xi = ex$ and $\eta = y/e + kx^2$, so that $x = \xi/e$, and $y = e\left(\eta - k\xi^2/e^2\right)$. Using this to invert the equation above and write it as an iteration yields

$$L^{(i+1)}(x,y) = \frac{1}{e}L^{(i)}\left(\frac{1}{e}x, ey - \frac{k}{e}x^2\right).$$

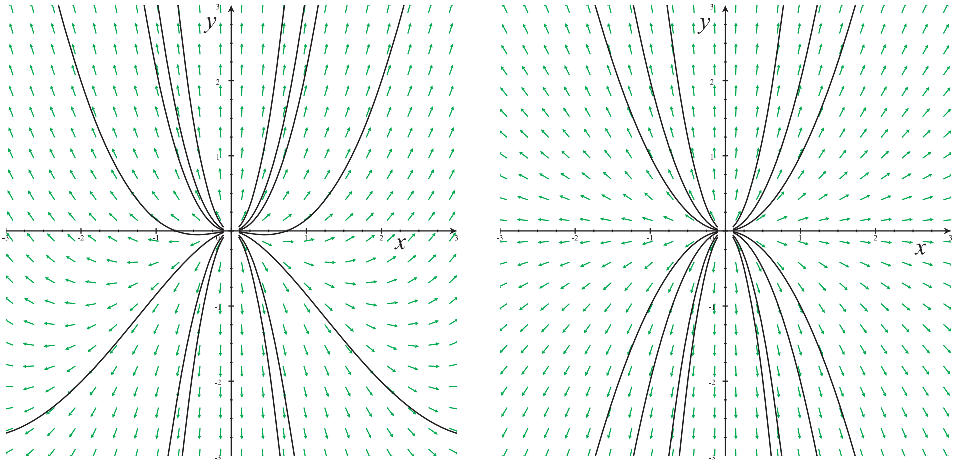

Figure 4.16. *Phase planes for the nonlinear flow (left) and linear flow (right) in* (4.40). *The constructed homeomorphism maps the two families of curves onto each other.*

As before we start the iteration with the identity, $L^{(0)}(x,y)=y$, and now obtain

$$\begin{aligned}
L^{(1)}(x,y)&=\frac{1}{e}L^{(0)}\left(\frac{1}{e}x, ey-\frac{k}{e}x^2\right)=\frac{1}{e}\left(ey-\frac{k}{e}x^2\right)=y-ke^{-2}x^2,\\
L^{(2)}(x,y)&=\frac{1}{e}\left(ey-\frac{k}{e}x^2-ke^{-2}\left(\frac{x}{e}\right)^2\right)=y-ke^{-2}(1+e^{-3})x^2,\\
L^{(3)}(x,y)&=\frac{1}{e}\left(ey-\frac{k}{e}x^2-ke^{-2}(1+e^{-3})\left(\frac{x}{e}\right)^2\right)=y-ke^{-2}(1+e^{-3}+e^{-6})x^2.
\end{aligned}$$

This series limits to

$$L(x,y)=y-ke^{-2}(1+e^{-3}+e^{-6}+e^{-9}+\cdots)x^2=y-\frac{ke^{-2}}{1-e^{-3}}x^2=y-\frac{1}{3}x^2.$$

So the actual homeomorphism is $H(x,y)=\left(x,y-x^2/3\right)$. The reader is encouraged to verify that this actually works by doing the calculation (4.36). ■

Example 4.38. The homeomorphism for the Hartman–Grobman theorem is guaranteed to exist only in a neighborhood of the origin. We can see that this is the case if we consider the ODEs

$$\begin{aligned}
\dot{x}&=2x,\\
\dot{y}&=4y+x^2,
\end{aligned}$$

which have a source at the origin. The flow for this system and its linearization are

$$\varphi_t(x,y)=\begin{pmatrix} e^{2t}x \\ e^{4t}(y+tx^2)\end{pmatrix}, \quad e^{tA}\begin{pmatrix} x \\ y\end{pmatrix}=\begin{pmatrix} e^{2t}x \\ e^{4t}y\end{pmatrix}. \tag{4.40}$$

These flows are shown in Figure 4.16. If we attempt the calculation as we did in the previous example, we will find that $H(x,y)=(x,y+g(x,y))$ but that the iteration for g does not converge. Instead of doing this, we modify the vector field:

$$\begin{aligned}\dot{x} &= 2x,\\ \dot{y} &= 4y + b(x^2),\end{aligned}$$

where the function b is a "bump" function. That is, we want $b(\xi) = \xi$ for small ξ and for it to vanish for large ξ. So we set

$$b(\xi) = \begin{cases} \xi, & |\xi| < \varepsilon, \\ 0, & |\xi| > \delta, \end{cases}$$

for some arbitrarily chosen $0 < \varepsilon < \delta$. We assume that b connects these two values smoothly.[25] The new vector field has a flow identical to the original nonlinear one when $x^2 < \varepsilon$ but is identical to the linear flow when $x^2 > \delta$. The fact that the Hartman–Grobman theorem is only locally valid is made manifest by this modification.

When we integrate the modified equations, we obtain $x(t) = e^{2t}x_o$ and

$$y(t) = e^{4t}\left(y_o + \int_0^t e^{-4s} b(e^{4s}x_o^2)ds\right) \equiv e^{4t}\left(y_o + B(x_o^2, t)\right).$$

The new function $B(x^2, t)$ cannot be obtained explicitly—especially since we have not explicitly specified b! However, we do know that if $x^2(s) < \varepsilon$ for all $0 < s < t$, i.e., if $|x_o| < \sqrt{\varepsilon}e^{-2t}$, then $b(x^2(s)) = x^2(s)$ along the entire integration path and we obtain $B(x_o^2, t) = tx_o^2$. Similarly, if $x^2(s) > \delta$ for all $0 < s < t$, i.e., if $|x_o| > \sqrt{\delta}$, then $b(x^2(s)) = 0$, so that $B(x_o^2, t) = 0$. Setting $t = 1$, and letting $B(x^2) = B(x^2, 1)$, we have

$$B(x^2) = \begin{cases} x^2, & |x| < \sqrt{\varepsilon}e^{-2}, \\ 0, & |x| > \sqrt{\delta}. \end{cases}$$

Putting the new flow into (4.37), we obtain the equation for H:

$$H(x, y) = e^{-A}H(\varphi_1(x, y)) = \begin{pmatrix} e^{-2} & 0 \\ 0 & e^{-4} \end{pmatrix} H\left(e^2x, e^4(y + B(x^2)\right).$$

As before we write $H = (K, L)^T$. The equation for K has the simple solution $K(x, y) = x$. For L we obtain

$$L(x, y) = e^{-4}L\left(e^2x, e^4(y + B(x^2)\right).$$

Iterating this starting with $L^{(0)}(x, y) = y$, we get

$$\begin{aligned}L^{(1)}(x, y) &= y + B(x^2),\\ L^{(2)}(x, y) &= y + B(x^2) + e^{-4}B(e^4x^2),\\ L^{(3)}(x, y) &= y + B(x^2) + e^{-4}B(e^4x^2) + e^{-8}B(e^8x^2).\end{aligned}$$

After N steps this gives the obvious result

$$L^{(N)}(x, y) = y + \sum_{n=0}^{N-1} e^{-4n}B(e^{4n}x^2).$$

[25] It is a standard trick in analysis that such "bump" functions can be made arbitrarily smooth, and even C^∞; see, for example, (Friedman 1982, Problem 3.3.1).

Note that if we set $B(x^2) = x^2$, then this series sums to Nx^2, which does not converge as $N \to \infty$. However, since B vanishes when its argument is large, then the series actually terminates after finitely many terms. Explicitly, choose an N such that $e^{4N}x^2 > \delta$, or $N(x) > \frac{1}{4}\ln(\delta/x^2)$, so that $B(e^{4N}x^2) = 0$ and this term and all the following ones vanish. Using this we can "take the limit" to obtain

$$L(x,y) = y + \sum_{n=0}^{N(x)} e^{-4n} B(e^{4n}x^2).$$

Since the sum is finite, it is convergent. This is the local homeomorphism guaranteed by the theorem. Note that it is not unique because we have considerable freedom in choosing b; however, once we have chosen the function $b(x)$, we get a unique homeomorphism. ∎

The Hartman–Grobman theorem implies Theorem 4.6: if x^* is a hyperbolic equilibrium point with $\mathrm{Re}(\lambda) < 0$, then since the linear system is asymptotically stable, so is the nonlinear system.

The Hartman–Grobman theorem says nothing about the structure of the motion in the neighborhood of a nonhyperbolic equilibrium. This case is considerably more intricate—we will discuss it in Chapter 6 and Chapter 8.

4.9 ▪ Omega-Limit Sets

We now develop some terminology that will help in the classification of orbits. Since—as we saw in §4.3—up to reparameterization of time, ODEs give rise to complete flows, we now consider a general flow, $\varphi_t(x)$. Our goal is to study properties of the orbits,

$$\Gamma_x = \{\varphi_t(x) : t \in \mathbb{R}\}. \tag{4.41}$$

In some cases, as in (4.2), we will consider just the forward orbit of x, the set

$$\Gamma_x^+ = \left\{\varphi_t(x) : t \in \mathbb{R}^+\right\}, \tag{4.42}$$

or the similarly defined backward orbit, Γ_x^-. One of the main goals of theory of dynamical systems is to give a geometrical classification of the types of orbits that occur in a given flow.

One important characterization of orbits is their "ultimate" or asymptotic behavior as $t \to \infty$, if this exists in some sense. Asymptotic behavior is defined in terms of *limit points*; recall §3.1: a point y is a *limit point* of the forward orbit of x if there is a sequence $t_1 < t_2 < \cdots < t_k \ldots$ such that $t_k \to \infty$ and $\varphi_{t_k}(x) \to y$ as $k \to \infty$. The asymptotic behavior of an orbit is its

▷ *omega-limit set*: The collection of all limit points of Γ_x^+ is the *omega-limit set* of x, denoted $\omega(x)$.

It is easy to see from the definition that if $z \in \Gamma_x$, $\omega(z) = \omega(x)$. Thus instead of $\omega(x)$, we can just as well write $\omega(\Gamma_x)$, the ω-limit set of the entire trajectory. Similarly, using a sequence $t_k \to -\infty$, we can define a limit set for $t \to -\infty$:

▷ *alpha-limit set*: $\alpha(x)$ is the collection of all limit points of Γ_x^- .

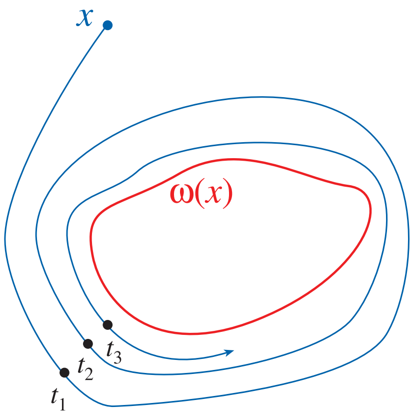

Figure 4.17. *The omega-limit set can be a limit cycle.*

A simple example of an ω-limit set is an asymptotically stable equilibrium, another example is a periodic orbit that attracts a trajectory; see Figure 4.17. Such an orbit is called a

> ▷ *limit cycle*: A periodic orbit γ that is the omega or alpha-limit set of a point $x \notin \gamma$ is a limit cycle.

Thus, a limit cycle is an invariant loop with the property that there is a nearby orbit that spirals either toward it or away from it.[26] As we will see in Chapter 6, limit cycles are common for planar flows and more generally can arise through a "bifurcation" of an equilibrium when it becomes unstable; see §8.8.

Example 4.39. The planar system

$$\begin{aligned}\dot{x} &= x(1-r^2)-y,\\ \dot{y} &= y(1-r^2)+x\end{aligned} \tag{4.43}$$

is most easily analyzed in polar coordinates. The radial equation is

$$\dot{r} = \frac{1}{r}(x\dot{x}+y\dot{y}) = r(1-r^2). \tag{4.44}$$

This one-dimensional system has a source at $r=0$ and a sink at $r=1$ (negative values of r are not allowed). The dynamics of $\theta = \tan^{-1}(y/x)$ are given by

$$\dot{\theta} = \frac{1}{r^2}(x\dot{y}-y\dot{x}) = \frac{1}{r^2}(x^2+y^2) = 1.$$

Thus the dynamics on the circle $\gamma = \{(r,\theta) : r=1\}$ are simply $\theta(t) = \theta_o + t$: it is a periodic orbit. The orbit γ is an asymptotically stable limit cycle because the radial equation shows that $r(t) \to 1$ for any $r(0) \neq 0$. ■

Note that a limit cycle is closed (the loop γ includes all of its limit points) and invariant, $\varphi_t(\gamma) = \gamma$. These properties are generally true for ω-limit sets, as we will

[26] Sometimes limit cycles are defined as isolated periodic orbits. This definition is not equivalent to ours, as a periodic orbit in a planar system could bound a disk of other periodic orbits and still be the limit of a spiraling trajectory from the outside.

see in the following three fundamental lemmas that define the basic structure of ω-limit sets.

Lemma 4.40 (Closure). *$\omega(x) = \bigcap_{T\geq 0} \overline{\Gamma}^+_{\varphi_T(x)}$, where $\overline{\Gamma}^+_x$ is the closure of the forward orbit of x. Hence, $\omega(x)$ is closed.*

Proof. If $z \in \omega(x)$, then $z \in \overline{\Gamma}^+_{\varphi_T(x)} = cl\,\{y : y = \varphi_t(x),\ t \geq T\}$ for any T, since this includes all limit points. Therefore, z is in the intersection of these sets. This proves that $\omega(x) \subset \bigcap_{T\geq 0} \overline{\Gamma}^+_{\varphi_T(x)}$. Now suppose that $z \in \bigcap_{T\geq 0} \overline{\Gamma}^+_{\varphi_T(x)}$, or equivalently for any T, $z \in \overline{\Gamma}^+_{\varphi_T(x)}$. There are now two possibilities: z may be a point in $\Gamma^+_{\varphi_T(x)}$ for each $T \geq 0$ or not. In the first case there must be infinitely many times $t_n \to \infty$ such that $z = \varphi_{t_n}(x)$ implying that z is a limit point and thus in $\omega(x)$. In the latter case there is some time $T \geq 0$ for which $z \notin \Gamma^+_{\varphi_T(x)}$. Since by assumption z is in the closure of $\Gamma^+_{\varphi_T(x)}$, then—by definition of "closure"—it is a limit point of the orbit. Finally, recall that the intersection of a family of closed sets is closed. □

Lemma 4.41 (Invariance). *The ω-limit set is invariant.*

Proof. If $y \in \omega(x)$, then there is a sequence t_k such that $\varphi_{t_k}(x) \to y$. Continuity then implies that for any fixed $s \in \mathbb{R}$, $\varphi_{t_k+s}(x) \to \varphi_s(y)$. Therefore, $\varphi_s(y) \in \omega(x)$. □

Now suppose that there is a metric $\rho(x,y)$ defined on the phase space; recall §3.2. We define the distance between a point, x, and a set, S, by

$$\rho(x,S) = \inf_{y\in S} \rho(x,y). \tag{4.45}$$

We will show next that when an orbit of a flow is bounded, it must approach its ω-limit set, in the sense that $\rho(\varphi_t(x), \omega(x)) \to 0$; in this case we say that $\varphi_t(x) \to \omega(x)$. We will also show that in this case that $\omega(x)$ is

> ▷ *connected*: A set S is connected if it *cannot* be partitioned into two nonempty sets such that each subset has no points in common with the closure of the other.

Thus, $\mathbb{R}^+$ is connected; for example, it can be partitioned into $A = (0,1)$, and $B = [1,\infty)$, but $\overline{A} \cap B = \{1\}$ is not empty.

Lemma 4.42 (Compact and Connected). *If the forward orbit of x is contained in a compact set, then $\omega(x)$ is nonempty, compact, and connected. Furthermore, $\varphi_t(x) \to \omega(x)$ as $t \to \infty$.*

Proof. The sets $\overline{\Gamma}^+_{\varphi_T(x)} = cl\,\{\varphi_t(x) : t \geq T\}$ are nested, since $\overline{\Gamma}^+_{\varphi_{T+s}(x)} \subset \overline{\Gamma}^+_{\varphi_T(x)}$ for any $s > 0$. Since, by assumption, the forward orbit of x is contained in a compact set, C, say, each $\overline{\Gamma}^+_{\varphi_T(x)} \subset C$ is also compact. According to Lemma 4.40, $\omega(x)$ is the intersection of these sets, and the intersection of a collection of nested closed sets is nonempty; then $\omega(x)$ is nonempty. Moreover, since $\omega(x)$ is closed and contained in a compact set, it is compact.

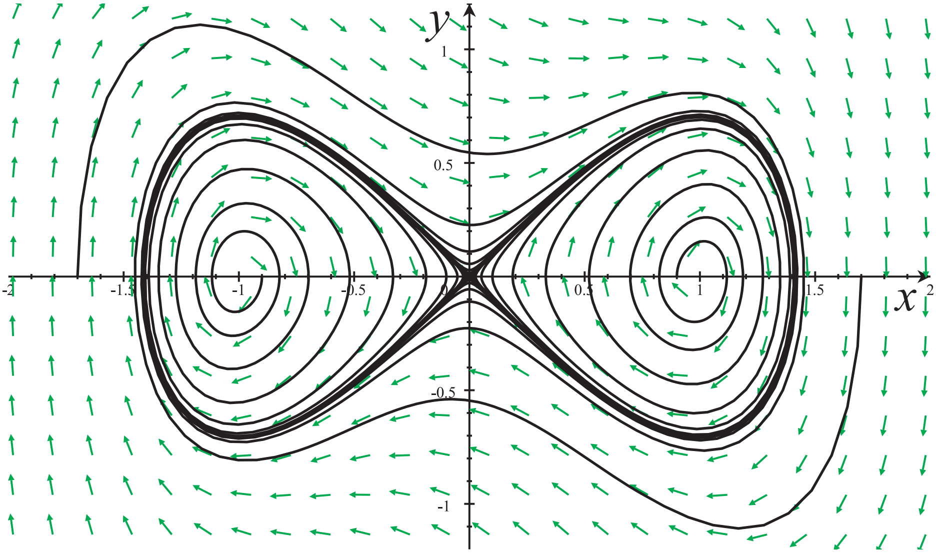

Figure 4.18. *Attracting figure-eight orbit of* (4.46) *for* $\mu = 0.5$.

Now suppose that $\omega(x)$ is not connected, i.e., that there are disjoint open sets $A, B \subset C$ such that $\omega(x) \subset A \cup B$, with $\omega(x) \cap A \neq \emptyset$ and $\omega(x) \cap B \neq \emptyset$. Now consider the set $K = C \setminus (A \cup B)$; it is compact since C is, and the complement of an open set is closed. Since both A and B contain limit points of the trajectory, for any $T \geq 0$ there are times $t_b > t_a > T$, such that $\varphi_{t_a}(x) \in A$ and $\varphi_{t_b}(x) \in B$, Thus there is a time $t_k \in (t_a, t_b)$ for which $\varphi_{t_k}(x) \in K$. Since this is true for any $T > 0$, there is an infinite sequence $\{t_k\}$ for which $\varphi_{t_k}(x) \in K$. Since K is compact, this sequence has limit points. But this contradicts the assumption that $\omega(x) \cap K = \emptyset$. (More generally, any intersection of a nested collection of compact, connected sets is connected.)

Finally, suppose that $\rho(\varphi_t(x), \omega(x))$ does not go to zero. Then there must be a subsequence $\varphi_{t_k}(x)$ of points that stay a distance δ away from $\omega(x)$. However, since this sequence is contained in a compact set, it has a convergent subsequence, which would be a limit point not in $\omega(x)$, but this is a contradiction. In conclusion, $\rho(\varphi_t(x), \omega(x)) \to 0$. □

Example 4.43. Consider the system

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= x - x^3 - \mu y \left(y^2 - x^2 + \tfrac{1}{2}x^4\right). \end{aligned} \tag{4.46}$$

When $\mu = 0$, (4.46) is a Hamiltonian system with $H = {}^1\!/_2(y^2 - x^2 + {}^1\!/_2 x^4)$. The level set $H = 0$ is a figure eight, with $H < 0$ inside its lobes and $H > 0$ outside; see Figure 4.18. The term proportional to μ in the y-equation is specially chosen so that it vanishes on $H = 0$. Thus, any orbit that starts on this curve will stay on it even when $\mu \neq 0$. Note that the rate of change of energy is given by

$$\frac{dH}{dt} = \frac{\partial H}{\partial x} y + \frac{\partial H}{\partial y}\left(x - x^3 - 2\mu y H\right) = -2\mu y^2 H.$$

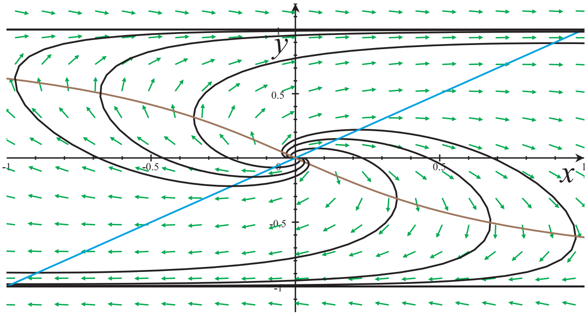

Figure 4.19. *Phase portrait of the system* (4.47), *showing the nullclines (blue and brown).*

Consequently, when $y \neq 0$ and $H < 0$ (inside the lobes of the figure eight), H is increasing and when $H > 0$ (outside the figure eight), H is decreasing. Therefore, trajectories move toward the figure eight contour except possibly when $y = 0$. Only the points $(0,0)$ and $(\pm 1, 0)$ on this set are invariant, so we can conclude, using LaSalle's invariance principle, Theorem 4.25, that $|H(x(t), y(t))|$ monotonically decreases to zero as $t \to \infty$ for every point except the equilibria $(\pm 1, 0)$.

We can, therefore, completely characterize the ω-limit sets for each point in the plane. A point x inside the right lobe of the figure eight (but not at the equilibrium $(-1,0)$) has an ω-limit set given by the entire right lobe—each point on the lobe is a limit point of its trajectory. A similar discussion applies to points inside the left lobe. Any point outside the two lobes (i.e., with $H > 0$) has the entire figure eight as its ω-limit set. The ω-limit set of any point on the figure eight is the origin. Finally, each equilibrium is its own ω-limit set. ∎

If $\omega(x)$ is not compact, then it need not be connected.

Example 4.44. Consider the system

$$\begin{aligned} \dot{x} &= y + x(1-y^2), \\ \dot{y} &= (1-y^2)(y-x). \end{aligned} \tag{4.47}$$

There is a spiral source at the origin, and the lines $y = \pm 1$ are invariant. Let $R = \{(x,y) \neq (0,0) : |y| < 1\}$ be the open region that is bounded by these lines. A numerical phase portrait, see Figure 4.19, shows that trajectories starting in R spiral outward and approach either $y = +1$ or $y = -1$. However, they appear to continually spiral and never settle down on either line. In particular, when the trajectory crosses the nullcline $N_y = \{y = x\}$, then $\dot{y}$ changes sign: if $y > 0$ and is approaching 1, then it will cross this line and begin to diverge from 1. Thus for any point $z \in R$, it appears that $\omega(z) = \{y = 1\} \cup \{y = -1\}$, which is not connected. The conclusion can be made rigorous by consideration of the global phase portrait; see Exercise 6.14. ∎

There are two other characterizations of long-time behavior that are of interest:

▷ *nonwandering*: A point x is nonwandering if for *every* neighborhood W of x and every time $T > 0$ there is time $t > T$ such that $\varphi_t(W) \cap W \neq \emptyset$.

In other words, a nonwandering point has nearby points that continually return. Consequently, any periodic orbit is nonwandering. Moreover, it can be shown that every point in an ω-limit set is nonwandering; see Exercise 14.

▷ *minimal set*: A set S is minimal if it is closed, nonempty, and invariant and does not contain any such set as a proper subset.

For example, a periodic orbit is minimal, but the union of two periodic orbits is not.

Theorem 4.45. *Suppose S is compact; then S is minimal if and only if for each $x \in S$ we have $S = \omega(x)$.*

Proof. First assume that $S = \omega(x)$ but is not minimal. Then there is a closed set $B \subset S$ that is invariant. However, if $x \in B$, then $\omega(x) \subset B$. This is a contradiction, so S must be minimal. Now assume that S is minimal, but there is an $x \in S$ for which $\omega(x) \neq S$. Since S is compact, so is $\omega(x)$, and Lemma 4.41 implies that $\omega(x)$ is invariant, so S has an invariant subset. Again, this is a contradiction. □

4.10 ▪ Attractors and Basins

Informally, an attractor is an invariant set toward which all nearby trajectories move. We saw in §4.5 that any equilibrium that is linearly asymptotically stable satisfies this condition. Our goal is to define the notion of attractor without reference to the kind of orbit or orbits that it contains; indeed, some attractors consist of infinitely many orbits. We start by generalizing the definition of stability that we used for equilibria in §4.5 to arbitrary invariant sets (recall the definition of invariant set in §4.1):

▷ *stability*: An invariant set Λ is stable if, for any neighborhood N of Λ, there is a neighborhood S of Λ such that all points that start in S stay in N for all $t > 0$.

▷ *asymptotic stability*: An invariant set Λ is asymptotically stable if it is stable and has a neighborhood N such that for each $x \in N$, $\rho(\varphi_t(x), \Lambda) \to 0$ as $t \to \infty$.

Since these definitions always refer to a neighborhood of the invariant set, we will define an attractor by constructing a special neighborhood that will envelope it:

▷ *trapping region*. A set N is a trapping region if it is compact and $\varphi_t(N) \subset \text{int}(N)$ for $t > 0$.

Here, "int(N)" denotes the "interior" of the set N. Thus, a trapping region is strictly "forward invariant." Note also that since φ_t is a homeomorphism, for any $s > 0$, $\varphi_{t+s}(N) = \varphi_t(\varphi_s(N)) \subset \varphi_t(\text{int}(N)) = \text{int}(\varphi_t(N))$; thus the sequence of sets $\varphi_{t_i}(N)$ is nested for any increasing sequence t_i.

Trapping regions are computationally and analytically quite easy to find: it is sufficient that the vector field point inward everywhere on the boundary. The maximal invariant set inside a trapping set is called an

▷ *attracting set*: A set Λ is an attracting set if there is a compact trapping region $N \supset \Lambda$ so that

$$\Lambda = \bigcap_{t>0} \varphi_t(N). \tag{4.48}$$

Note that since the collection $\{\varphi_t(N) : t \geq 0\}$ is a set of closed and nested sets, the intersection, Λ, is closed and nonempty. For compact sets there is no difference between the concepts of asymptotic stability and attracting set.

Lemma 4.46. *An attracting set is asymptotically stable. Conversely, if a compact set is asymptotically stable, then it is an attracting set.*

Proof. First, suppose Λ is an attracting set; then by definition every point in any trapping region N stays in N, so Λ is stable. Now suppose there is a point $x \in N$ whose orbit does not approach Λ. Since $\varphi_t(x) \subset N$ for $t \geq 0$ and N is compact, there is a convergent subsequence $\lim_{j\to\infty} \varphi_{t_j}(x) = x^*$ and, by assumption, $x^* \notin \Lambda$. However, since the sets $\varphi_t(N)$ are nested and for any t there is a j such that $t_j \geq t$, then $x^* \in \varphi_{t_j}(N) \subset \varphi_t(N)$, i.e., $x^* \in \Lambda$. This contradiction shows that $\varphi_t(x)$ must approach Λ—so it is asymptotically stable.

Conversely, assume that A is compact and asymptotically stable. To show it is an attracting set we must construct a trapping set. Since A is asymptotically stable, there is a neighborhood U of A for which all points approach A and stay in some larger neighborhood D. Since A is compact, a compact subset of U can be chosen if needed. Now we have to find a subset of U that is forward invariant. Since all points $x \in U$ eventually approach A, there exists a time $T(x)$ for each $x \in U$ such that $\varphi_t(x) \in \text{int}(U)$ for all $t > T(x)$. Moreover, since U is compact, the function $T(x)$ has a maximum:

$$T_{\max} = \max_{x\in U}(T(x)).$$

Given this construction, U is a "weak trapping set", a set for which $\varphi_t(U) \subset \text{int}(U)$ whenever $t \geq T$. In this case, one can show, see Exercise 16, that there exists a strong Lyapunov function $L : U \to \mathbb{R}^+$, i.e., a continuous function such that whenever $t > 0$, then $L(\varphi_t(x)) < L(x)$ and such that for each $x \in A$, $L(x) = 0$. Since U is a neighborhood of A, there is an $\epsilon > 0$ such that the region $N = \{x : L(x) \leq \epsilon\}$ is contained in U. This set defines a trapping region for A, since if $x \in \partial N$, then whenever $t > 0$, $\varphi_t(x) \in \text{int}(N)$ since $L(\varphi_t(x)) < \epsilon$. □

Any attracting set has a maximal trapping region that is called the *stable set* of Λ or the

> ▷ *basin of attraction*, $W^s(\Lambda)$: The basin (or stable set) of an invariant set Λ is the set of all points x for which $\rho(\varphi_t(x), \Lambda) \to 0$ as $t \to \infty$.

Thus if Λ is an attracting set with trapping region N, then

$$W^s(\Lambda) = \bigcup_{t\leq 0} \varphi_t(N).$$

Note that the definition of asymptotic stability is equivalent to the fact that Λ is stable and $\Lambda \subset \text{int}(W^s(\Lambda))$. This concept also provides another way of stating Lemma 4.42: if the forward orbit of x is contained in a compact set, then $x \in W^s(\omega(x))$.

Example 4.47. Consider a diagonalizable linear system with a matrix A whose eigenvalues are all negative. The system can be put in diagonal form by a linear coordinate transformation to obtain $\dot{x}_j = \lambda_j x_j$. The unit square $N = \{x : \left|x_j\right| \leq 1\}$ is mapped to the set $\varphi_t(N) = \left\{x : \left|x_j\right| \le e^{\lambda_j t}\right\} \subset \operatorname{int}(N)$ when $t > 0$, so N is a trapping region. Moreover, the origin is an attracting set and the entire phase space is the basin of the origin: $W^s(\{0\}) = \mathbb{R}^n$. ∎

Following Charles Conley, an attractor is an attracting set with an additional assumption of "irreducibility" (Ruelle 1981). Basically, we would like to decompose attracting sets into their fundamental components. There are several possible requirements that one could add to our definition; for example, an attractor could be minimal (Perko 2000), "chain transitive" (Robinson 1999), or contain a dense orbit (Guckenheimer and Holmes 1983). We follow (Field 1996) to define an

> ▷ *attractor*: A set Λ is an attractor if it is an attracting set and there is some point x such that $\Lambda = \omega(x)$.

Example 4.48. Consider the system

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

There are three equilibria $(0,0)$ (a saddle), and $(\pm 1, 0)$ (sinks). The set $\Lambda = \{-1 \le x \le 1, y = 0\}$ is, by our definition, an attracting set. Its basin is the entire plane. For the trapping set we could take any rectangular disk enclosing Λ. Note that there is no orbit, however, that approaches all the points in Λ; indeed, almost every trajectory approaches one of the two sinks. Thus the only attractors for this example are the equilibria $(\pm 1, 0)$. ∎

The definition of attractor that we give follows the school of Conley (Conley 1978; Easton 1998). A related concept, a *measure attractor*, is due to John Milnor: it is a set that attracts a set of positive measure but does not necessarily have an attracting neighborhood (Milnor 1985a, b). There are interesting examples of sets that attract many but not all points in a neighborhood, and even sets whose basin is *nowhere dense* (Alexander et al. 1996). We will always assume that an attractor has an attracting neighborhood.

Example 4.49. In §4.6 it was shown that the Lorenz system (4.26) has a Lyapunov function about the origin when $\sigma > 0, b > 0$, and $r < 1$. Lorenz studied the system at much different values: $\sigma = 10$, $b = 8/3$, and $r = 28$. Here, it has an attracting set that appears to be a "strange" set: a fractal.[27] We can demonstrate that this system does have an attractor, when $\sigma, b > 0$, by constructing a trapping region. Consider the ball defined by

$$C(x,y,z) \equiv x^2 + y^2 + (z - r - \sigma)^2 \le R^2. \tag{4.49}$$

The vector field on the surface of the ball can be shown to point inward if R is chosen large enough. To see this, compute the derivative of the function $C(x,y,z)$ along the vector field to obtain

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}C &= \sigma xy - \sigma x^2 + rxy - y^2 - xyz + (z - r - \sigma)(xy - bz) \\ &= -\sigma x^2 - y^2 - bz^2 + (r + \sigma)bz. \end{aligned}$$

[27] We will discuss strange sets in §7.3.

Since b and σ are positive, the set on which $dC/dt = 0$ is an ellipsoid defined by the zero contour of the function

$$E(x,y,z) \equiv -\sigma x^2 - y^2 - bz^2 + (r+\sigma)bz.$$

We need to find an R to guarantee that $E(x,y,z) \le 0$ on the sphere $C = R^2$. The smallest such ball just intersects the zero contour of E. One way to find the optimal value is using the method of Lagrange multipliers. The idea is that critical points of the function

$$L(x,y,z,\lambda) = C(x,y,z) + \lambda(E(x,y,z) - 0).$$

correspond to critical points of C subject to the constraint $E = 0$. Differentiation of L gives four equations

$$\begin{aligned}
\frac{\partial L}{\partial x} &= 0 \Rightarrow x(1-\lambda\sigma) = 0,\\
\frac{\partial L}{\partial y} &= 0 \Rightarrow y(1-\lambda) = 0,\\
\frac{\partial L}{\partial z} &= 0 \Rightarrow 2z(1-\lambda b) = (2-\lambda b)(r+\sigma),\\
\frac{\partial L}{\partial \lambda} &= 0 \Rightarrow E = 0.
\end{aligned}$$

There are six solutions, two on the plane $x = 0$, two on $y = 0$, one at the origin, and one on the z-axis. After some algebra these give four possible critical radii

$$R \in \frac{r+\sigma}{2}\left\{0, 2, \frac{b}{\sqrt{b-1}}, \frac{b}{\sqrt{\sigma(b-\sigma)}}\right\}.$$

Only for the largest of these, does the $C = R^2$ sphere enclose ellipsoid $E = 0$, implying that we must choose

$$R > \frac{r+\sigma}{2}\begin{cases} 2 & \alpha \le 2 \\ \frac{\alpha}{\sqrt{\alpha-1}} & \alpha > 2 \end{cases}, \quad \alpha = b\max(1,\sigma^{-1}). \tag{4.50}$$

For the classic Lorenz parameters this requirement is $R > 152/\sqrt{15}$; so for example, B_{40} is a trapping region. The resulting attractor is amazingly complex, as shown in Figure 4.20.

The Lorenz attractor, Λ, is commonly visualized by numerically computing a single trajectory. Thus it appears to be the ω-limit set of an arbitrary point which would imply it is an attractor. It is not obvious, however, from the numerical simulations exactly how complicated the dynamics are on Λ: it is possible that Λ is simply a very long periodic orbit. Indeed showing that there is no attracting periodic orbit for the classic Lorenz system was listed by Stephen Smale as his 14th mathematical problem for the 21st century (Smale 1998). Recently this has been proved using rigorous numerical computation (Tucker 2002). An attractor that is geometrically complicated, such as the Lorenz attractor, is called a *strange attractor*; see §7.3. ∎

Note that not every ω-limit set is an attractor. As an example, the origin in (4.46) is the ω-limit set for any initial condition that starts on the figure eight but it is not an attractor because points in its neighborhood have limit points on the figure eight. The figure eight itself, however, is an attractor according to our definition. Since this attractor is not a minimal set, it does not satisfy Perko's definition of attractor.

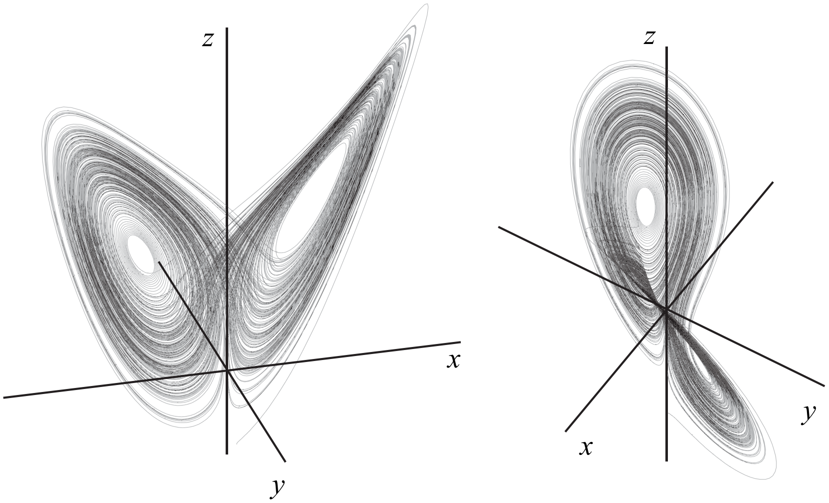

Figure 4.20. *Two views of a numerical approximation of the Lorenz Attractor for* $(\sigma, b, r) = (10, 8/3, 28)$. *The axes shown are centered at* $(0,0,20)$ *and are of length* 50.

4.11 ▪ Stability of Periodic Orbits

A periodic orbit is an invariant set and can be stable (recall example (4.43)) or unstable. It is natural to first study their stability using the same method of linearization that we used for equilibria in §4.4. Indeed, we will show that linearization provides valid results in the same situation as in that case: when the orbit is linearly asymptotically stable.

Suppose that $x(t) = \gamma(t) = \gamma(t+T)$ is a periodic orbit of period T for the differential equation $\dot{x} = f(x)$. If the vector field $f \in C^1$ we can linearize the ODE about γ by setting $x(t) = \gamma(t) + y(t)$ and expanding f in a Taylor series to obtain

$$\frac{d}{dt}(\gamma + y) = f(\gamma(t)) + \frac{d}{dt}y = f(\gamma(t) + y) = f(\gamma(t)) + Df(\gamma(t))y + o(y).$$

If we neglect the $o(y)$ term we obtain the linearization

$$\frac{d}{dt}y = Df(\gamma(t))y = A(t)y, \tag{4.51}$$

where the matrix, $A(t)$, is a periodic function of time. Such systems can be analyzed using Floquet theory, as we did in §2.8.

Recall from (2.46) that the fundamental matrix solution of (4.51) can be written $\Phi(t, t_o)$, and that the matrix $M = \Phi(T, 0)$, is called the monodromy matrix. The eigenvalues of M are the Floquet multipliers, and Floquet's theorem (Theorem 2.36) shows that all of the solutions of (4.51) are bounded whenever the Floquet multipliers have magnitude smaller than one.

For the case (4.51), one of the Floquet multipliers is trivially unity.

Theorem 4.50. *The monodromy matrix M for the linearization of a system $\dot{x} = f(x)$ about a periodic orbit $\gamma(t)$ always has at least one unit eigenvalue.*

Proof. Since $x(t) = \gamma(t)$ is a solution of the original nonlinear equations, so is $x(t) = \gamma(t + \tau)$ for any phase shift τ. Differentiate this solution with respect to τ and set

$\tau = 0$ to give

$$\frac{d}{d\tau}[\dot{\gamma}(t+\tau) = f(\gamma(t+\tau))]\Big|_{\tau=0} \quad \Rightarrow \quad \frac{d}{dt}\dot{\gamma} = Df(\gamma(t))\dot{\gamma}(t).$$

Therefore, $\dot{\gamma}$ is a solution of the linearized equations: $\dot{\gamma}(t) = \Phi(t,0)\dot{\gamma}(0)$. However, since γ is periodic, $\dot{\gamma}(T) = \dot{\gamma}(0)$ and is therefore an eigenvector of the monodromy matrix with eigenvalue (Floquet multiplier) one. □

Note that the vector $\dot{\gamma}(t)$ is tangent to γ at the point $\gamma(t)$. A simple interpretation of Theorem 4.50 is that two nearby points on the same orbit stay close for all time. Since there is always a unit multiplier, a periodic orbit cannot be asymptotically stable in the same sense as an equilibrium. However, the unit multiplier is associated with the "trivial" tangent direction and does not affect the stability of the invariant set γ. Thus we will say a periodic orbit is *linearly stable* if all of its Floquet multipliers have magnitude at most 1, $|\mu_i| \le 1$. Moreover, the orbit is *linearly asymptotically stable* if all of its multipliers apart from the trivial unit multiplier have magnitude strictly less than one, $|\mu_i| < 1$ for $i = 2, \ldots, n$.

Abel's theorem, Theorem 2.34, gave one nontrivial relation between the Floquet multipliers,

$$\det(M) = \exp \int_0^T \operatorname{tr}(Df(\gamma(s)))\,ds. \tag{4.52}$$

Since $\det(M) = \prod_i \mu_i$, this relation determines the product of the multipliers. For the planar case, this is all the information we need: in $\mathbb{R}^2$, the 2×2 monodromy has one unit multiplier, $\mu_1 = 1$. The second nontrivial multiplier thus determines the stability of the periodic orbit, and $\mu_2 = \det(M)$.

Example 4.51. Consider again the planar system (4.43). Consider the limit cycle $\gamma = \{(r,\theta) = (1, \theta_o + t) : t \in \mathbb{R}\}$. Choosing $\theta_o = 0$ and returning to rectangular coordinates so that $\gamma = \{(x,y) = (\cos t, \sin t) : t \in \mathbb{R}\}$ gives the linearized matrix

$$Df(\gamma(t)) = \begin{pmatrix} -2x^2 & -1-2yx \\ 1-2yx & -2y^2 \end{pmatrix} = \begin{pmatrix} -2\cos^2 t & -1-2\sin t\cos t \\ 1-2\sin t\cos t & -2\sin^2 t \end{pmatrix}.$$

As promised, the derivative of the solution, $\dot{\gamma} = (-\sin t, \cos t)^T$, is a solution of the linearized ODE:

$$\frac{d}{dt}\begin{pmatrix} -\sin t \\ \cos t \end{pmatrix} = \begin{pmatrix} -2\cos^2 t & -1-2\sin t\cos t \\ 1-2\sin t\cos t & -2\sin^2 t \end{pmatrix}\begin{pmatrix} -\sin t \\ \cos t \end{pmatrix} = \begin{pmatrix} -\cos t \\ -\sin t \end{pmatrix}.$$

A second solution can be easily obtained by linearizing the r equation (4.44) about its equilibrium $r = 1$, to obtain $\delta\dot{r} = -2\delta r$, showing that a linearized solution should take the form $(\delta x, \delta y) = \delta r_o e^{-2t}(\cos t, \sin t)$. Indeed, substituting this into the linearized ODE yields an identity. We can conclude that the fundamental matrix solution to the linear equation is

$$\Phi(t,0) = \begin{pmatrix} e^{-2t}\cos t & -\sin t \\ e^{-2t}\sin t & \cos t \end{pmatrix},$$

which gives a monodromy matrix

$$M = \Phi(2\pi, 0) = \begin{pmatrix} e^{-4\pi} & 0 \\ 0 & 1 \end{pmatrix}.$$

The Floquet multipliers are simply the elements on the diagonal, $\mu_1 = 1$ and $\mu_2 = e^{-4\pi}$. ∎

If $\operatorname{tr}(Df)$ vanishes identically, then (4.52) implies that $\det(M) = 1$; this means that a planar, "incompressible" flow has both multipliers equal to one (see §9.2).

Example 4.52. Any C^2 Hamiltonian system in the plane, (4.27), has both Floquet multipliers equal to one, since $f = (\partial H/\partial y, -\partial H/\partial x)$, so that $\operatorname{tr}(Df) = \partial^2 H/\partial x \partial y - \partial^2 H/\partial y \partial x = 0$. If one is careful with indices, one can show that $\operatorname{tr}(Df) = 0$ for Hamiltonian systems in any dimension (recall (1.13)), which means that the product of the Floquet multipliers for these systems is always one.

If the Hamiltonian depends explicitly on time, $H(x, y, t)$, the system (4.27) is still called a Hamiltonian system; however, the energy is no longer conserved. Indeed, (4.28) becomes

$$\frac{dH}{dt} = \frac{\partial H}{\partial x}\dot{x} + \frac{\partial H}{\partial y}\dot{y} + \frac{\partial H}{\partial t} = \frac{\partial H}{\partial t} \neq 0.$$

As we discussed in §1.2, a two-dimensional nonautonomous system is equivalent to an autonomous one on a three-dimensional space. If we suppose that H is a periodic function of time, $H(x, y, t) = H(x, y, t + T)$, then the third variable can be taken to be an angle, say, $\theta = t/T$, so the phase space is $\mathbb{R}^2 \times \mathbb{S}^1$, and the ODEs are

$$\dot{x} = \frac{\partial}{\partial y} H(x, y, T\theta), \quad \dot{y} = -\frac{\partial}{\partial x} H(x, y, T\theta), \quad \dot{\theta} = \frac{1}{T}.$$

A periodic orbit of this system is a curve $\gamma(t) = (x(t), y(t), \theta(t))$ whose period must be some multiple of T, since the angle returns to itself "mod 1." Since the third component of the new three-dimensional vector field is constant, the result $\operatorname{tr}(Df) = 0$ still holds. In this case there are three Floquet multipliers. One multiplier will be one, $\mu_1 = 1$, and so $\mu_2 \mu_3 = 1$ as well. ∎

Consequently, periodic orbits of Hamiltonian systems are never asymptotically stable. The only case in which they are linearly stable is if all Floquet multipliers are on the unit circle. This will be discussed in Chapter 9.

The relationship between linear asymptotic stability and true asymptotic stability in the sense of §4.10 is most easily discussed by introducing the concept of Poincaré maps.

4.12 ▪ Poincaré Maps

Maps are dynamical systems in the sense of §4.1 when the set of allowed time values is discrete. While much of the theory of dynamical systems can be developed for maps themselves (Arrowsmith and Place 1992; Devaney 1986; Guckenheimer et al. 1983; Katok and Hasselblatt 1999; Robinson 1999; Strogatz 1994; Wiggins 2003), our primary interest in maps will be to discuss the behavior of a flow in the neighborhood of a periodic orbit. The Poincaré map naturally arises in this context.

A map is defined by a function $P : M \to M$ through the relation $x' = P(x)$, where $x' \in M$ denotes the new point that arises from the initial point $x \in M$.[28] For a map, an

[28] We always use the symbol "D" to represent derivative and reserve the prime symbol $'$ for the iterate of a map.

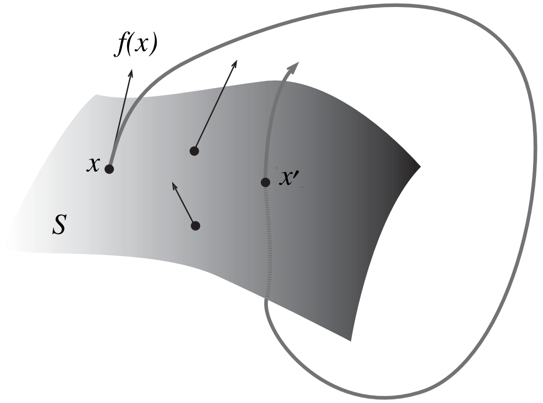

Figure 4.21. *Construction of a Poincaré map from a flow on a section S.*

orbit is no longer a function $x(t)$ of $t \in \mathbb{R}$ but is instead a sequence $\{x_t : t \in \mathbb{Z}\}$. Using this subscript notation, the dynamics is given by the iteration

$$x_t = P(x_{t-1}).$$

Maps arise naturally from flows by taking *sections* of the flow. For a flow in $\mathbb{R}^n$, a section, S, is a smooth surface of dimension $d = n-1$ (i.e., a codimension-one surface) such that the velocity vector is not tangent to S at any point. That is, if $\hat{n}_x$ is the unit normal to S at x, then S is a section if $f(x)\cdot\hat{n}_x \neq 0$ for all $x \in S$.

A *Poincaré map* for a section S is obtained by choosing an $x \in S$, and following the flow $\varphi_t(x)$ to find the first return to S: let $\tau(x)$ be the first positive time for which $\varphi_t(x) \in S$. The map is defined by

$$P(x) = \varphi_{\tau(x)}(x), \tag{4.53}$$

as illustrated in Figure 4.21. Note that $\tau(x)$ might not exist for all $x \in S$, in which case the Poincaré map is not well defined. The best scenario occurs when S is a

> ▷ *global section*: If the orbit of every point $x \in \mathbb{R}^n$ eventually crosses an $n-1$ dimensional surface S and then returns to S at a later time, then S is a global section.

In this case the Poincaré map is defined for all $x \in S$.

Example 4.53. A system with a natural angle variable that is always increasing has a global section. For example, the skew-product[29] system

$$\begin{aligned} \dot{x} &= f(x,\theta), \\ \dot{\theta} &= 1, \end{aligned}$$

where $x \in \mathbb{R}^n$, and $\theta \in \mathbb{S}^1$, has a global section $S = \{(x,\theta) : \theta = \theta_o\} \cong \mathbb{R}^{n-1}$, since all trajectories cross this section with unit speed in the θ direction. This can also be generalized to the case that $\dot{\theta} = g(x,\theta)$, provided that $g > 0$ everywhere. ■

[29] A system $\dot{x} = f(x)$ is a skew product if the variables can be separated as $x = (y,z)$ such that the equations become $\dot{y} = f_1(y,z)$ and $\dot{z} = f_2(z)$.

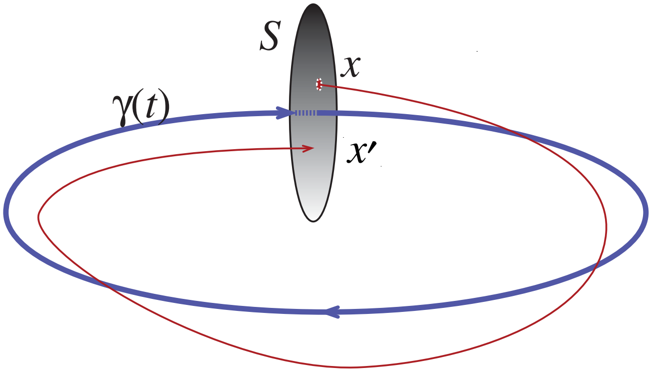

Figure 4.22. *Poincaré section in the neighborhood of a periodic orbit.*

If S and $\tilde{S}$ are two global sections, then the corresponding Poincaré maps are conjugate. This follows since the flow takes every point $x \in S$ to a point on $\tilde{S}$ after some time $\tau(x)$. The homeomorphism $h : S \to \tilde{S}$ is defined by $h(x) = \varphi_{\tau(x)}(x)$. Each global section contains the same information about the flow.

A locally defined Poincaré map always exists in a neighborhood of a periodic orbit γ, as shown in Figure 4.22. The section S is assumed to be a (small) disk containing a point $x_o \in \gamma$ that is oriented perpendicular to the vector field $f(x_o)$. By continuity, there is always some neighborhood of this point for which the vector field will be transverse to the disk. Moreover, continuity with respect to initial conditions, recall §3.4, implies that points "near" γ will stay "near" for any finite time t, and so they must intersect the disk at a time that is near the period, $T = \tau(x_o)$.

For example, suppose that a flow in the plane has a periodic orbit. Then the section is a line segment that is perpendicular to the periodic orbit at a point on the orbit.

Example 4.54. Let (r, θ) be polar coordinates and consider the system

$$\begin{aligned} \dot{r} &= r + \alpha r^3, \\ \dot{\theta} &= \nu. \end{aligned}$$

When $\alpha < 0$ there is a unique periodic orbit at $r^* = (-\alpha)^{-1/2}$. It is not hard to solve explicitly for $r(t)$ by separation of variables:

$$t + c = \int \frac{dr}{r(1+\alpha r^2)} = \frac{1}{2} \ln \left| \frac{r^2}{1+\alpha r^2} \right| \quad \Rightarrow \quad r(t; r_o) = \frac{r_o}{\sqrt{(1+\alpha r_o^2)e^{-2t} - \alpha r_o^2}}.$$

The solution for θ is trivial: $\theta(t) = \theta_o + \nu t$. Let the positive x-axis represent S. The radius r is a good coordinate on S and the Poincaré map $P : S \to S$ is simply the value that r takes after one period of the angle, or at $t = 2\pi/\nu$:

$$r' = P(r) = \frac{r}{\sqrt{(1+\alpha r^2)e^{-4\pi/\nu} - \alpha r^2}}. \tag{4.54}$$

For this one-dimensional case, the Poincaré map and its iteration can be visualized graphically; see Figure 4.23. Consider an initial condition r_o. Move vertically up to $P(r_o)$ to obtain r_1. Put this value onto the r-axis by moving horizontally to the diagonal. To get r_2 move again vertically to the function value $P(r_2)$. The resulting series

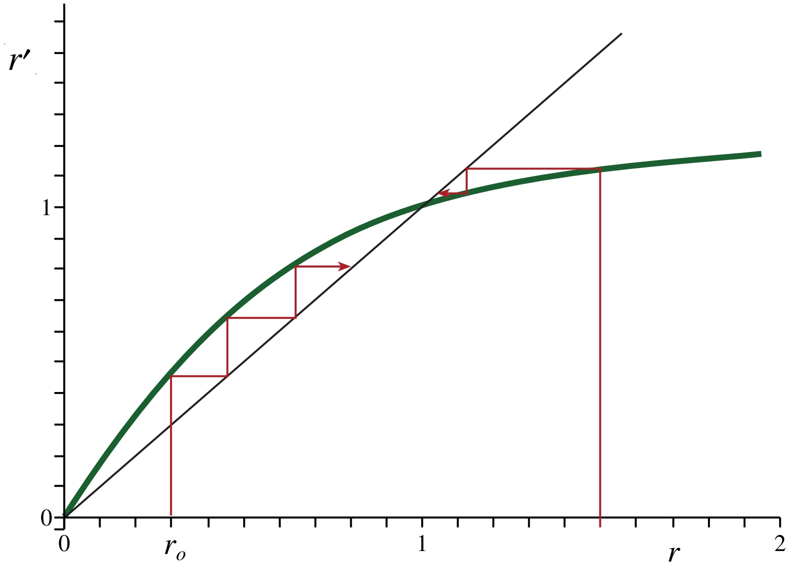

Figure 4.23. *Poincaré map* (4.54) *for* $\alpha = -1$ *and* $\nu = 4\pi$. *The periodic orbit corresponds to the intersection of the graph* $r' = P(r)$ *with the diagonal* $r' = r$. *It is stable because* $DP(1) < 1$. *The stair-stepped curve is the graphical iteration of* $r_o = 0.3$.

of lines, as shown in the figure, resembles a staircase. (For more complicated maps the picture looks like a cobweb and so is typically called the *cobweb diagram*.) The staircase picture implies that if the slope at a fixed point is less than one in magnitude, then the equilibrium is stable, since iterates move monotonically in the direction of the fixed point. ▮

Generally, the computation of the stability of a periodic orbit requires that we consider the linearization of the flow in the neighborhood of the periodic orbit. One must typically resort to numerical methods to solve for the Floquet multipliers, even if the periodic orbit is known analytically. It is often convenient numerically to compute the Poincaré map (4.53) and study stability of an orbit by this method. One advantage is that the Poincaré map acts on the section S that has dimension $n-1$, one less than the flow. Moreover, the removed dimension corresponds to the motion along the periodic orbit and thus to the neutral Floquet multiplier $\mu_1 = 1$. Consequently, stability computed using the Poincaré map is the same as that from the Floquet spectrum:

Theorem 4.55. *Let* γ *be a periodic orbit of a* C^2 *flow* φ, S *be a local section through a point* $x_o \in \gamma$, *and* $x_o \subset S_o \subset S$ *so that* $\varphi_\tau(x)(x) \in S$ *for each* $x \in S_o$. *Then there is a Poincaré return map* $P : S_o \to S$. *If the monodromy matrix of* γ *is* M, *then*

$$\operatorname{spec}(M) = \operatorname{spec}(DP(x_o)) \cup \{1\}.$$

Proof. Suppose $x \in S_o$, and $\tau(x)$ is the time of first return to S. The Poincaré map is given by (4.53), where we restrict x to S_o. For the moment, ignore this restriction, and let $Q(x) = \varphi_{\tau(x)}(x)$ for any x near γ. Differentiating Q with respect to x gives

$$DQ(x) = D_x\varphi_{\tau(x)}(x) + \frac{d}{dt}\varphi_{\tau(x)}(x)(D_x\tau(x))^T.$$

Here the last term is the "outer product" of the flow vector $f(x(\tau(x))$ and the gradient vector $D\tau(x)$. This latter vector represents the change in period with respect to x; it can be called the "twist." When $x = x_o \in \gamma$, $\tau(x_o) = T$, $D\varphi_T(x_o) = M$, and $\varphi_T(x_o) = x_o$ so that

$$DQ(x_o) = M + f(x_o)(D\tau(x_o))^T.$$

We can take the section S to consist of points orthogonal to the flow vector at x_o, i.e., $x = x_o + \xi$, where $f(x_o)^T\xi = 0$. If w_i, $i = 1,2,\ldots,n-1$, are a set of orthonormal vectors perpendicular to $f(x_o)$, and $W = (w_1, w_2, \ldots, w_{n-1})$, then the matrix WW^T is a projection onto the section. The matrix $DP(x_o)$ in the w_i basis has the representation $W^T DQ(x_o)W$. Since $W^T f(x_o) = 0$, we obtain

$$DP(x_o) = W^T M W.$$

By Theorem 4.50, $\hat{f} = f(x_o)/|f(x_o)|$ is an eigenvector of M with eigenvalue 1. Add this vector to W to form the orthogonal matrix $U = (W, \hat{f})$ and define the similar matrix

$$\tilde{M} = U^T M U = \begin{pmatrix} DP(x_o) & 0 \\ \hat{f}^T M W & 1 \end{pmatrix}.$$

Since similar matrices have the same spectrum, $\mathrm{spec}(\tilde{M}) = \mathrm{spec}(M)$. Moreover, since the last column has only one nonzero element, then $\det(\lambda I - \tilde{M}) = (\lambda - 1)\det(\lambda I - DP(x_o))$. □

This theorem shows that, up to the trivial Floquet multiplier, $\mu_1 = 1$, linear stability of a periodic orbit can be computed from the Poincaré map.

Finally we are ready to state the result about linear stability.

Theorem 4.56. *If γ is a periodic orbit of a C^2 flow that is linearly asymptotically stable (the spectrum of its Poincaré map is inside the unit circle), then it is asymptotically stable.*

Proof. The proof of this theorem is similar to the proof of Theorem 4.19. Following that analysis, let $x_o \in \gamma$, and $x_o + y \in N \cap S$, where N is a neighborhood of γ and S is a section. Write the Poincaré map at $x_o + y$ as $P(x_o + y) = x_o + DP(x_o)y + g(y)$. Thus

$$y' = DP(x_o)y + g(y).$$

Since the orbit γ is linearly asymptotically stable, the spectrum of $DP(x_o)$ is contained in the interior of the unit circle. Analogously to (4.20), for any $n \geq 0$ we can bound the orbit of this linear mapping by

$$|DP^n(x_o)y| < K\mu^n\,|y|$$

for some $0 < \mu < 1$ and $K \geq 1$. Since the flow is smooth, $g(y) = o(y)$, that is, for any $\varepsilon > 0$ there is a neighborhood $N_\varepsilon \subset S$ of x_o such that $|g(y)| < \varepsilon\,|y|$ for all $y \in N_\varepsilon$. Using the discrete analogue of the integrating factor and the Grönwall lemma, it is possible to see that there is an ε such that if $y_o \in N_\varepsilon$, then the sequence y_n limits to x_o as $n \to \infty$ and is bounded in distance from x_o. We leave the details to the reader. Since the Poincaré maps through any two local sections to γ are topologically conjugate, this implies that γ is asymptotically stable. □

4.13 ▪ Exercises

1. Show that the following functions are flows on the spaces indicated. Find the vector field for each flow.

   (a) $\varphi_t(x) = \dfrac{x + \tanh t}{1 + x \tanh t}$, $x \in [-1, 1]$,

   (b) $\varphi_t(x, y) = \begin{pmatrix} x\cos(r^2 t) + y\sin(r^2 t) \\ -x\sin(r^2 t) + y\cos(r^2 t) \end{pmatrix}$, $r^2 = x^2 + y^2$, $(x, y) \in \mathbb{R}^2$.

2. Find and analyze the linear behavior near each equilibrium of the following systems on $\mathbb{R}^2$. Classify the equilibria. Are they linearly stable or unstable? Sketch the local behavior you obtained in the phase plane and compare with a numerical phase plane plotter that shows the global solutions.

   (a) $\begin{array}{l} \dot{x} = y \\ \dot{y} = x - x^3 - ay \end{array}$,

   (b) $\begin{array}{l} \dot{x} = x^2 - y^2 - 1 \\ \dot{y} = 2y \end{array}$,

   (c) $\begin{array}{l} \dot{x} = y - x^2 + 2 \\ \dot{y} = 2y^2 - 2xy \end{array}$,

   (d) $\begin{array}{l} \dot{x} = -4x - 2y + 4 \\ \dot{y} = xy \end{array}$.

3. The centrifugal governor (see Figure 4.24) was patented by James Watt in 1789 to control the steam engine. It is described by the set of ODEs (Pontryagin 1962)

$$\begin{aligned} \dot{\varphi} &= \psi, \\ \dot{\psi} &= n^2\omega^2 \sin\varphi\cos\varphi - \Omega^2 \sin\varphi - \tfrac{b}{m}\psi, \\ \dot{\omega} &= \tfrac{1}{I}(\mu\cos\varphi - F), \end{aligned}$$

   similar to those first derived by Vishnegradskii in 1876. Here the dynamical variables are $\varphi \in [0, \pi]$, the angle between the spindle S and the "flyball arms" of length L, ω, the rotational velocity of the flywheel, and ψ, the angular acceleration. Constants in the equation are n the transmission ratio of the gears—the ratio between the angular velocity of the spindle and flywheel, $\Omega = \sqrt{g/L}$ the arm pendulum frequency, b friction of the flywheel, m the flyball mass, I the moment of inertia of the flywheel, F the torque load on the engine, and μ, representing the steam-driven torque caused by closing the valve as the collar rises on the spindle.

   (a) Show that by rescaling time, setting $\tau = \Omega t$, and defining new variables, $(x, y, z) = (\varphi, \psi/\Omega, n\omega/\Omega)$, the equations can be reduced to the system

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= \sin x\left(z^2\cos x - 1\right) - \varepsilon y, \\ \dot{z} &= \alpha(\cos x - \beta) \end{aligned}$$

   for new parameters $(\alpha, \beta, \varepsilon)$, all positive.

   (b) Show that if β is small enough, there is a unique nonnegative equilibrium (x^*, y^*, z^*).

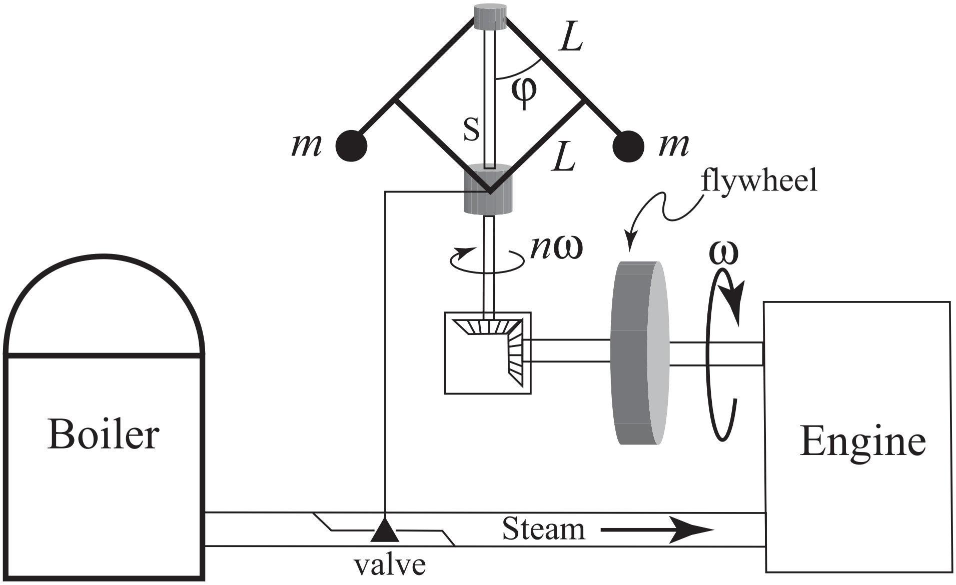

Figure 4.24. *Sketch of Watt's centrifugal governor.*

(c) Linearize about the equilibrium and find the characteristic polynomial.

(d) Show that there is a critical value, $\varepsilon_o(\alpha, \beta)$, such that if $\varepsilon > \varepsilon_o$, then the equilibrium is asymptotically stable, and if $0 < \varepsilon < \varepsilon_o$, then the equilibrium is a saddle.

(e) It can be shown that the system undergoes a Hopf bifurcation (see Chapter 8) at ε_o. Solve the equations numerically and demonstrate that as ε decreases through ε_o the equilibrium becomes unstable and there is an attracting limit cycle.

4. Are the following functions homeomorphisms? Are they diffeomorphisms? If the functions depend upon parameters, then so might your answers. Explain.

(a) $f : [0,1] \to [0,1], \;\; f(x) = ax(1-x)$,

(b) $f : \mathbb{R} \to \mathbb{R}, \;\; f(x) = ax + b\sin(2\pi x)$,

(c) $f : [0,1] \to \mathbb{S}^1, \;\; f(x) = [x + b\sin(2\pi x)] \bmod 1$,

(d) $f : \mathbb{S}^1 \times \mathbb{R} \to \mathbb{S}^1 \times \mathrm{R}, \;\; f(x,y) = ([x + y + b\sin(2\pi x)] \bmod 1, y + b\sin(2\pi x))$,

(e) $f : \mathbb{R}^2 \to \mathbb{R}^2, \;\; f(x,y) = (y + ax(1-x), -bx)$.

5. Use the iterative construction (4.38) of the Hartman–Grobman homeomorphism H to obtain an approximation for the conjugacy for the flow of the system on $\mathbb{R}^3$ given by

$$\begin{aligned} \dot{x} &= -x, \\ \dot{y} &= -y + x^2 z, \\ \dot{z} &= z \end{aligned}$$

to its linearization at $(0,0,0)$. Show that the iteration is not globally convergent. Discuss how to modify the iteration to make it locally convergent, using a "bump function."

6. Which of the ODEs $\dot{x} = Ax$ with the following matrices are topologically conjugate? Which are diffeomorphic? Which are linearly conjugate?

$$\text{(a)} \begin{pmatrix} -2 & -1 \\ 3 & 2 \end{pmatrix}, \quad \text{(b)} \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \quad \text{(c)} \begin{pmatrix} -5 & -2 \\ 5 & 1 \end{pmatrix}, \quad \text{(d)} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix},$$

$$\text{(e)} \begin{pmatrix} 7 & -10 \\ 5 & -8 \end{pmatrix}, \quad \text{(f)} \begin{pmatrix} 3 & 1 \\ -1 & 1 \end{pmatrix}, \quad \text{(g)} \begin{pmatrix} -5 & 1 \\ -6 & 0 \end{pmatrix}, \quad \text{(h)} \begin{pmatrix} 1 & 0 \\ -2 & -1 \end{pmatrix}.$$

7. Construct a topological conjugacy between the linear systems with the matrices

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

(*Hint*: Transform to polar coordinates and assume the homeomorphism has the form $h(r,\theta) = (h_r(r), h_\theta(r,\theta))$. The r-dependence of h_θ will involve $\ln r$.)

8. Construct Lyapunov functions to determine the stability of the equilibrium $(0,0)$ for the following systems on $\mathbb{R}^2$.

$$\text{(a)} \quad \begin{aligned} \dot{x} &= -x + y - y^2 - x^3 \\ \dot{y} &= x - y + xy \end{aligned} \quad ,$$

$$\text{(b)} \quad \begin{aligned} \dot{x} &= y - x^2 + 3y^2 - 2xy \\ \dot{y} &= -x - 3x^2 + y^2 + 2xy \end{aligned} \quad .$$

(*Hints*: Try a power series for L, starting with quadratic terms. Add higher-order terms if necessary. Sometimes it is easier to check for a Hamiltonian than it is to construct L *ab initio*.)

9. An asymptotically stable linear system always has a Lyapunov function of the form $L = x^T S x$.

   (a) Show that when all the eigenvalues of A have negative real parts, then the "Lyapunov equation" (4.24) has the unique, positive definite, symmetric solution
$$S = \int_0^\infty e^{\tau A^T} e^{\tau A} d\tau. \tag{4.55}$$
(*Hint*: Premultiply (4.24) by e^{tA^T} and postmultiply by e^{tA}. Note that the left-hand side of (4.24) then becomes a total derivative. Remember that $e^{A^T + A} \neq e^{A^T} e^A$ in general.)

   (b) Compute S for the matrix $A = \begin{pmatrix} -2 & 1 \\ 0 & -2 \end{pmatrix}$, and demonstrate explicitly that $dL/dt < 0$.

10. The Lyapunov function defined in Exercise 9 also works when nonlinear terms are added to the ODE. Consider the system $\dot{x} = Ax + g(x)$, where $g(x) = o(x)$ and A is a matrix whose eigenvalues have negative real parts. Show that there is a neighborhood U of the origin for which the function $L = x^T S x$, where S is given by (4.55), is a strong Lyapunov function. (*Hint*: You may need to use the Cauchy–Schwarz inequality $|\langle u, v \rangle| \leq \|u\| \, \|v\|$.)

11. In 1965 Goodwin proposed the model

$$\dot{x} = \frac{1}{1 + z^m} - ax, \quad \dot{y} = x - by, \quad \dot{z} = y - cz,$$

for the regulation of enzyme synthesis of a product in a cell. Here a, b, c are positive constants, and m is a positive integer ($m = 1$ for Goodwin's original model) (Murray 1993, §6.2). Here x represents the concentration of messenger RNA, y the enzyme, and z the product. The nonlinear term in these equations represents the negative feedback of the product on the RNA, since as z grows, the growth rate of x decreases.

(a) Show that there is a trapping set of the form $N = \{(x,y,z) : 0 \le x \le X, 0 \le y \le Y, 0 \le z \le Z\}$ for suitably chosen values X, Y, and Z. Take care to think about the dynamics on the coordinate axes.

(b) Find the unique equilibrium in N, and show that it is asymptotically stable when $m = 1$. It also can be shown with more work that this is true for any $m < 8$. (*Hint*: The characteristic polynomial has only stable roots only if it satisfies the Routh-Hurwitz criterion; see Exercise 2.11.) Consequently the attracting set in Λ contains this equilibrium. While this system was initially proposed to model oscillatory behavior, a recent general result implies that no such cycle exists for $m \le 4$ and indeed that the attracting set Λ in N is the equilibrium (Enciso and Sontag 2006).

12. Assume that the flow $\varphi_t : A \to A$ is conjugate to the flow $\psi_t : B \to B$ with conjugacy $h : A \to B$.

(a) Show that if $\omega(x)$ is the omega limit set for $x \in A$ under φ, then $h(\omega(h^{-1}(y)))$ is the omega limit set for $y = h(x) \in B$ under ψ.

(b) Show that if Λ is an invariant set for φ, then $h(\Lambda)$ is an invariant set for ψ.

(c) Show that if $W^s(\Lambda)$ is the basin of Λ, then $h(W^s(\Lambda))$ is the basin of $h(\Lambda)$.

(d) Show that if Λ is an attractor, then so is $h(\Lambda)$.

13. Suppose that φ and ψ are flows on $\mathbb{R}^2$ that each have exactly two equilibria that are both saddles. Suppose for the flow φ that the unstable set of one saddle corresponds to the stable set of the other but that this is not true for ψ. Show that φ and ψ are not topologically equivalent.

14. Show that if $y \in \omega(x)$, then y is nonwandering.

15. An alternative trapping set to (4.49) for the Lorenz system (4.26) is the ellipsoid

$$E_C = \left\{ r x^2 + \sigma y^2 + \sigma(z - 2r)^2 \le C \right\}.$$

Find the minimal value of C such that every trajectory eventually enters E_C. Does this give a better bound than that represented by (4.50)?

16. Suppose that A is an asymptotically stable invariant set, and—as in the proof of Lemma 4.46—A has a "weak" trapping set U, i.e., a compact neighborhood of A for which there is some $T > 0$ such that $\varphi_T(U) \subset \text{int}(U)$. Let $D : U \to \mathbb{R}^+$ be defined as

$$D(x) = \sup_{t \ge 0} \rho(\varphi_t(x), A),$$

recall (4.45). Using arguments similar to those in the proof of Theorem 4.23, show that the function

$$L(x) = \int_0^\infty e^{-s} D(\varphi_s(x)) ds$$

is a strong Lyapunov function on U.

17. Let (r,θ) be a point in the phase space $[0,\infty)\times\mathbb{S}$ that obeys the system

$$\dot{r} = r(1+a\cos\theta - r^2),$$
$$\dot{\theta} = 1,$$

where $|a|<1$.

   (a) Show that the circle $r=0$ is a periodic orbit with period 2π.
   (b) Compute the monodromy matrix $M=\Phi(2\pi,0)$ for the circle $r=0$ and show that its Floquet multipliers are $\mu=1$ and $e^{2\pi}$. (*Hint*: The linear system has solutions $(0,\delta\theta(t))$ and $(\delta r(t),0.)$
   (c) Show that there are two circles $r=r_-$ and $r=r_+$ such that if $0<r<r_-$, then $\dot{r}>0$, and if $r>r_+$, then $\dot{r}<0$. Thus the region $N=\{(r,\theta),r_-<r<r_+\}$ is a trapping region. Our next goal is to show that the attracting set in N is a periodic orbit.
   (d) Let S be the ray $\{(r,0)\}$. Argue that S is a global section. Let $P:\mathbb{R}^+\to\mathbb{R}^+$ be the Poincaré map on S.
   (e) Suppose that the orbit of the point $(r_L,0)$ has the property $0<P(r_L)<r_-$. Argue that $P(r_L)>r_L$. Alternatively, suppose that the orbit of $(r_H,0)$ has the property that $P(r_H)>r_+$. Then argue that $P(r_H)<r_H$.
   (f) Apply the intermediate value theorem to $P(r)$ to show that there is a point $(r^*,0)$, where $r_L<r^*<r_H$, whose orbit is periodic.
   (g) Show that the Floquet multipliers of the new orbit are $\mu=1$ and $e^{-4\pi}$. Consequently, the new periodic orbit is asymptotically stable. (*Hint*: To do the integral $\int_0^{2\pi} r^2(t)dt$ use the differential equation to set $r^2=1+a\cos\theta-\dot{r}/r$.)

18. The Shimizu–Morioka model is a simplified model of the Lorenz system when r is large (Shilnikov 1993). It is given by

$$\dot{x} = y,$$
$$\dot{y} = x-\alpha y - xz,$$
$$\dot{z} = -\beta z + x^2,$$

where $(x,y,z)\in\mathbb{R}^3$, and $\alpha,\beta\in\mathbb{R}$.

   (a) Find all of the equilibria for this system depending on the values of α and β (there can be three).
   (b) Find the eigenvalues of the equilibrium that exists (is a point in $\mathbb{R}^3$) for all parameter values, and classify its stability type as a function of α and β.

19. Consider your adopted system of quadratic differential equations (recall §1.6 and Exercise 1.10). If possible, find a set of values of the reduced parameters for which one of your system's equilibria (x^*,y^*,z^*) is spectrally stable. If there are no such equilibria, then prove so. Otherwise, attempt to construct a Lyapunov function for a neighborhood of your stable equilibrium. It would probably be good to attempt to use a quadratic function

$$L(x,y,z)=\alpha(x-x^*)^2+\beta(y-y^*)^2+\gamma(z-z^*)^2,$$

though you might have to experiment with adding cross terms to the equation, or going to a higher degree. This is a case where you may or may not succeed; indeed, your system may not have a simple Lyapunov function. You will get full credit for making a convincing attempt—for example, by showing that the function above is not a Lyapunov function for any values of α, β, γ.

Chapter 5

Invariant Manifolds

> *Nunquam praescriptos transibunt sidera fines.* (Never will heavenly bodies transgress their prescribed bounds.) (Henri Poincaré 1890)

Hyperbolic fixed points of a linear ordinary differential equation (ODE) have stable, E^s, and unstable spaces, E^u, determined by the eigenvectors of the associated matrix at the fixed point. We showed in §2.6 that these spaces are invariant under the dynamics of the linear system. In this chapter we will show that there are also invariant subspaces W^u and W^s that are generalizations of E^u and E^s for a nonlinear ODE with a hyperbolic fixed point. Some local information about these subspaces can be inferred from Theorem 4.36 (Hartman–Grobman), which implies that when an equilibrium is hyperbolic, the flow in its neighborhood is topologically conjugate to the linearized flow. Here, however, we will obtain much more precise control over the structure of these subspaces, showing that they are "manifolds" that are smoothly tangent to the linear subspaces. We begin by looking at a few simple examples where the manifolds can be found analytically.

5.1 ▪ Stable and Unstable Sets

Stable and unstable sets are collections of orbits that are *forwar*d or *backward asymptotic* to a given orbit. Recall that in §4.10 we defined the stable set, or basin of attraction, of an invariant set Λ as the set of points forward asymptotic to Λ:

$$W^s(\Lambda) = \left\{ x \notin \Lambda : \lim_{t\to\infty} \rho(\varphi_t(x), \Lambda) = 0 \right\}. \tag{5.1}$$

We can also define the backward basin or unstable set of Λ as the set of points that are backward asymptotic to it:

$$W^u(\Lambda) = \left\{ x \notin \Lambda : \lim_{t\to-\infty} \rho(\varphi_t(x), \Lambda) = 0 \right\}. \tag{5.2}$$

Generally the stable and unstable sets are invariant.

Lemma 5.1. *The stable and unstable sets of an invariant set Λ are themselves invariant sets.*

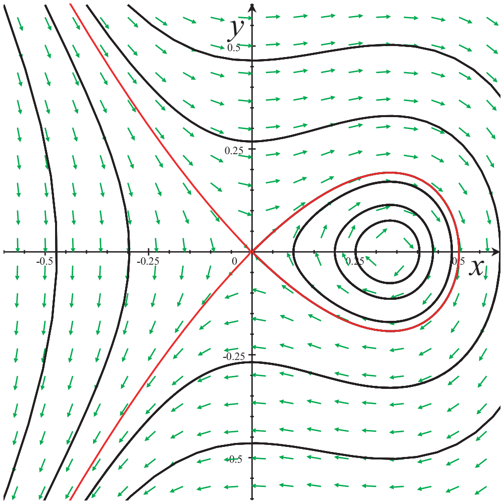

Figure 5.1. *Phase portrait of* (5.3) *with* $a = 1$*. The stable and unstable manifolds of* $(0,0)$ *are the red curves that form the zero energy contour.*

Proof. We must show that whenever $z \in W^s(\Lambda)$ we have $\varphi_s(z) \in W^s(\Lambda)$ for any $s \in \mathbb{R}$. This follows from the group property of the flow: by definition (5.1), $\varphi_s(z)$ is a point such that $\varphi_t(\varphi_s(z)) = \varphi_{s+t}(z) \to \Lambda$ as $t \to \infty$. Since this holds for any s, the stable set is invariant. A similar argument applies to the unstable set. □

In some special cases we can find the stable and unstable sets analytically. For example, consider a Hamiltonian $H(x,y)$ in the plane with a saddle equilibrium at a point (x^*,y^*). The energy contours $H(x,y) = H(x^*,y^*) = E$ that emanate from the saddle correspond to the stable and unstable sets of the saddle—since these are curves they are called the stable and unstable *manifolds*.

Example 5.2. The Hamiltonian for the system (4.29) is

$$H(x,y) = \tfrac{1}{2}(y^2 - x^2) + ax^3, \tag{5.3}$$

where we take $a > 0$. Since the linearization for the equilibrium at the origin has the Jacobian $Df(0) = \left(\begin{smallmatrix}0 & 1\\ 1 & 0\end{smallmatrix}\right)$, it is a saddle. The energy at the saddle point is $H(0,0) = E = 0$; this contour corresponds to the curves $y_\pm = \pm x\sqrt{1-2ax}$, shown in Figure 5.1, that intersect at $x = (2a)^{-1}$. Since orbits lie on contours of constant H, the union of these two curves, like every contour of H, is an invariant set. Noting the direction of the flow (from $\dot{x} = y$), we see that

$$\begin{aligned} W^u(0,0) &= \{(x,y) : H(x,y) = 0,\ x > 0 \text{ or } x,y < 0\}, \\ W^s(0,0) &= \{(x,y) : H(x,y) = 0,\ x > 0 \text{ or } x < 0 \text{ and } y > 0\}. \end{aligned}$$

Here we specifically do not include the equilibrium as part of the stable and unstable sets. Note that the positive-x branches of the two manifolds coincide; moreover, these branches bound the set of orbits that are oscillating about the center equilibrium at $((3a)^{-1},0)$. Orbits outside this closed loop are unbounded. Since this loop separates two topologically distinct types of motion, we call it a "separatrix"; see §5.2. ■

When the ODE is linear and hyperbolic, $\mathbb{R}^n = E^s \oplus E^u$ and the stable and unstable sets of the origin correspond to E^s and E^u. Our task in this chapter is to generalize these subspaces to the nonlinear case. We will see that when the equilibrium is hyperbolic, its linear stable and unstable sets give a "linear approximation" to the stable and unstable manifolds of the equilibrium.

Example 5.3. For the Hamiltonian (5.3), the stable and unstable manifolds of the origin correspond to the curves $y_\pm = \pm x\sqrt{1-2ax}$; recall Figure 5.1. As we will see in §5.4, the stable manifold theorem implies that the local unstable manifold is the unique invariant curve emanating from the origin that is tangent to the unstable eigenvector of $Df(0)$, in this case the vector $v_+ = (1,1)^T$. Since $dy_+\big/dx = 1$ at $x = 0$, this shows that the local unstable manifold of the origin is indeed the set $W^u(0) = \big\{(x, y_+(x)) : x \in (-\infty, 1/2a)\big\}$. Similarly, the local stable manifold is

$$W^s(0) = \big\{(x, y_-(x)) : x \in (-\infty, 1/2a)\big\}$$

and is tangent to the stable eigenvector $v_- = (1,-1)^T$. ■

5.2 ▪ Heteroclinic Orbits

In special situations it is possible that $W^u(\Lambda)$ and $W^s(\Lambda)$ may coincide or perhaps have points of intersection. The realization that there could be such intersections (in particular transverse intersections) is what led Poincaré to understand that the dynamics of the n-body problem (n point masses interacting under their mutual gravitational attraction) could be very complicated. The discovery of this complexity—and indeed the beginnings of what we now call *chaos*—arose from a mistake in a manuscript that Poincaré had submitted in 1888 to King Oscar of Sweden for a mathematics prize to be awarded to the first person to "find a solution" to the n-body problem! Although Poincaré was awarded the prize in 1889, his initial essay had mistakenly asserted that if W^u intersects W^s, then they must coincide.[30] The story of this mistake and its subsequent correction (leading to Poincaré having to pay for the entire print run of the issue of *Acta Mathematica* containing the original essay) is elegantly told in (Diacu and Holmes 1996).

The corrected version of Poincaré's paper (Poincaré 1889) began his extensive study of the complexity induced by two types of orbits; the first type he calls a

> ▷ *heteroclinic orbit*: An orbit Γ is heteroclinic if each $x \in \Gamma$ is backward asymptotic to an invariant set A and forward asymptotic to an invariant set B, i.e., $\Gamma \subset W^u(A) \cap W^s(B)$.

The second class is a special case of the first; Poincaré called the second type a *doubly asymptotic* or

> ▷ *homoclinic orbit*: Γ is homoclinic if each $x \in \Gamma$ is both forward and backward asymptotic to the same invariant set A, i.e., $\Gamma \subset W^u(A) \cap W^s(A)$.

This definition could be generalized to say that an orbit Γ_h is homoclinic to another orbit Γ if every point on Γ_h is both forward and backward asymptotic to Γ.

In a two-dimensional phase space, a saddle equilibrium has both a stable and an unstable set and each is one-dimensional. The uniqueness theorem implies that if a

[30] Some of the consequences of noncoincident intersections are discussed in §8.13 et seq.

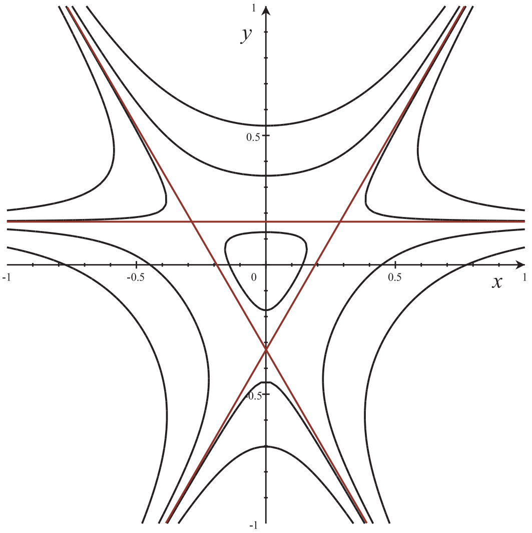

Figure 5.2. *Contours of the Hamiltonian* (5.4).

branch of W^u intersects a branch of W^s, then they must coincide; therefore, in a two-dimensional phase space homoclinic orbits form impenetrable boundaries—we saw such a boundary in Figure 5.1. Orbits such as these are called *separatrices*, as they separate phase space into regions that cannot communicate. Poincaré's mistake in 1888 was the conclusion that this must happen in higher-dimensional systems; we will see how this fails in §8.13.

For the case of Hamiltonian systems in the plane, separatrices are common. Since H is constant along trajectories, recall (4.28), any closed contour of a Hamiltonian H that intersects one or more critical points (note that $\nabla H = 0$ implies also that the point is an equilibrium) gives a separatrix. When a heteroclinic orbit connects two saddle equilibria, it is also called a *saddle connection*.

Example 5.4. Heteroclinic orbits can be constructed by choosing an H that has several saddle points with the same energy. For example, the function $f = {}^1\!/_2 r^2 - r^3 \sin(3\theta)$ in polar coordinates has a triangular contour $f = 1/54$. Translating this back to rectangular coordinates yields the Hamiltonian

$$H = \tfrac{1}{2}\left(x^2 + y^2\right) + y^3 - 3x^2 y. \tag{5.4}$$

As can be seen in Figure 5.2, H has three saddle equilibria $(x,y) = (\pm\sqrt{3}/6, 1/6)$, and $(0, -1/3)$ on the contour $H = 1/54$. There are three heteroclinic orbits connecting these saddles. When such a collection of heteroclinic orbits divides the plane into two regions we call it a *separatrix cycle*. ■

The existence of a saddle connection is unusual for general ODEs in the plane; however, with some care we can construct examples that do have a connection.

Example 5.5. Given a Hamiltonian system with a homoclinic orbit, it is easy to construct a non-Hamiltonian system that has one as well; such an example was given in (4.46). More generally, the contour $H(x,y) = E$ is preserved by the differential equa-

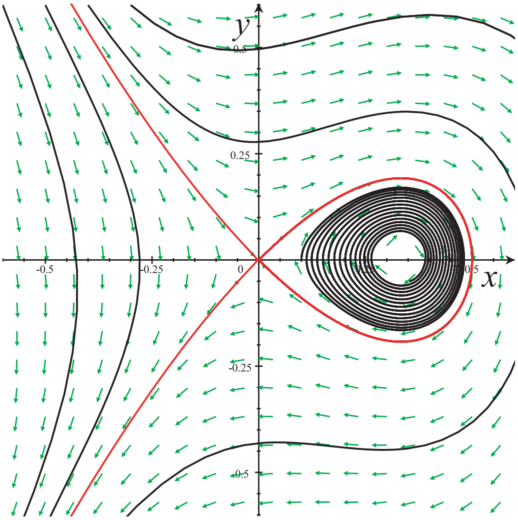

Figure 5.3. *Non-Hamiltonian system* (5.6) *with a homoclinic orbit. Here* $a = 1$.

tions

$$\frac{dx}{dt} = \frac{\partial H}{\partial y} + (H(x,y) - E)g_1(x,y),$$
$$\frac{dy}{dt} = -\frac{\partial H}{\partial x} + (H(x,y) - E)g_2(x,y) \tag{5.5}$$

for *any* functions g_1 and g_2. If this contour contains a homoclinic orbit, then (5.5) will have a homoclinic orbit too. In the example (5.3), the homoclinic orbit was at $E = 0$; therefore, the system

$$\dot{x} = y + H(x,y)x,$$
$$\dot{y} = x - 3ax^2 + H(x,y)y, \tag{5.6}$$

shown in Figure 5.3, still has the same homoclinic loop as the original Hamiltonian flow shown in Figure 5.1. Note that the origin is still a saddle. There are two more equilibria at $y^* = {}^1\!/_2\, ax^{*4}$ where x^* is a real root of the sixth-order polynomial $-4 + 12ax + a^2x^6$. For $a > 0$, the positive root of this polynomial is near the original center; however, this point is now a stable focus and attracts every point inside the homoclinic loop; see Exercise 2. ∎

5.3 ▪ Stable Manifolds

We can sometimes find W^u and W^s analytically even for the non-Hamiltonian case if the system of equations is a *skew product*; for example, if one of the equations of an ODE system in $\mathbb{R}^2$ is independent of the other. This kind of example seems special at first, but will prove to be of great use to us in the next section in the general proof of the stable manifold theorem.

Example 5.6. For example, suppose that $(x,y) \in \mathbb{R}^2$ and

$$\dot{x} = -x,$$
$$\dot{y} = y + g(x). \tag{5.7}$$

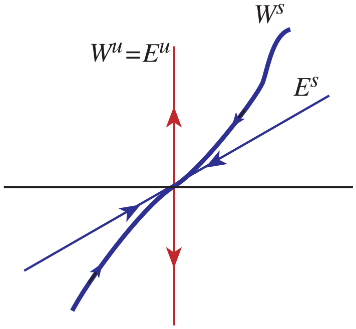

Figure 5.4. *Sketch of stable and unstable manifolds for* (5.7).

Here, we will assume that g is C^1 and that $g(0) = 0$. The latter condition ensures that the origin is an equilibrium. The Jacobian of the origin is

$$Df(0) = \begin{pmatrix} -1 & 0 \\ Dg(0) & 1 \end{pmatrix}.$$

This matrix has eigenvalues $\lambda = \pm 1$ and so is hyperbolic. The unstable eigenvector is $v_u = (0,1)^T$ so that the unstable subspace is the y-axis:

$$E^u = \{(x,y) : x = 0\}.$$

The second eigenvector is $v_s = (2, -Dg(0))^T$, so that the stable subspace is the line

$$E^s = \{(x,y) : Dg(0)x + 2y = 0\}.$$

Our goal is to find the stable and unstable sets of the origin. The ODEs are simple enough that the flow is easily obtained. Solving the x equation gives $x(t) = x_0 e^{-t}$. Substituting this into the y equation yields a nonautonomous linear equation. We can use the integrating factor method (recall Exercise 2.17) to find

$$\frac{d}{dt}(e^{-t}y) = e^{-t}g(x_0 e^{-t}) \quad \Rightarrow e^{-t}y(t) = y_0 + \int_0^t e^{-s} g(x_0 e^{-s})ds.$$

Upon changing integration variables, setting $u = e^{-s}$, and putting the two solutions together, we obtain the expression for the flow:

$$\varphi_t \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} xe^{-t} \\ ye^t + e^t \displaystyle\int_{e^{-t}}^1 g(xu)du \end{pmatrix}.$$

Since this is the general solution, we can find the set of points (x,y) that lie, for example, on the unstable manifold by asking which points have $\varphi_t(x,y) \to (0,0)$ as $t \to -\infty$. This immediately implies that $x = 0$, since otherwise the first component is unbounded. In this case, since $g(0) = 0$, the second component becomes ye^t, which does approach 0. So we have shown that $W^u(0,0)$ is simply the y-axis.

The stable set, $W^s(0,0)$, is the set such that $\varphi_t(x,y) \to (0,0)$ as $t \to \infty$. This means that x can be arbitrary, but y must be chosen specifically since we require

$$0 = \lim_{t\to\infty} y(t) = \lim_{t\to\infty} e^t \left(y + \int_{e^{-t}}^1 g(xu)du \right).$$

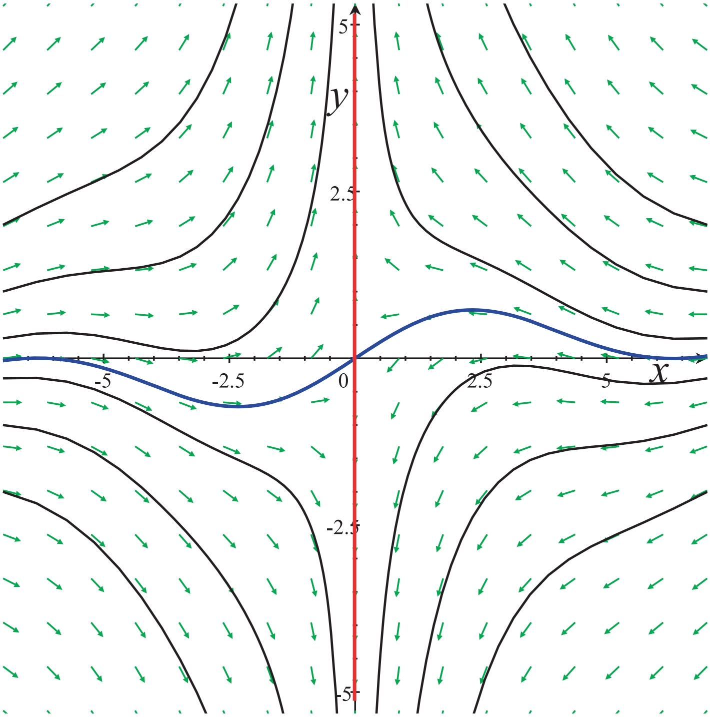

Figure 5.5. *Phase portrait for* (5.7) *with* $g(x)$ *given by* (5.9). *Here the unstable manifold is the* y*-axis (red line) and the stable manifold is the blue curve. Several other trajectories are also shown.*

We claim that for each x there is a solution of the form $y(x) = -\int_0^1 g(xu)du$. To see this, substitute it into the limit to obtain

$$\lim_{t\to\infty} y(t) = \lim_{t\to\infty} e^t \left(\int_{e^{-t}}^1 g(xu)du - \int_0^1 g(xu)du \right) = -\lim_{t\to\infty} e^t \left(\int_0^{e^{-t}} g(xu)du \right).$$

Since $g(0) = 0$ and $g \in C^0$, then for any $\varepsilon > 0$, there is a δ such that $|g(xu)| < \varepsilon$ for all $|xu| < \delta$. If we choose t large enough so that $|x|e^{-t} < \delta$, then the magnitude of the integral is bounded by εe^{-t}. Since this is true for any ε, the limit is zero as required. Thus, we have shown that

$$W^s = \left\{ (x, y(x)) : \; y(x) = -\int_0^1 g(xu)du \right\}, \tag{5.8}$$

as sketched in Figure 5.4. For example, if

$$g(x) = -\sin x, \tag{5.9}$$

we can easily do the integral in (5.8) to obtain the function

$$y(x) = \frac{1}{x}\int_0^x \sin(\xi)d\xi = \frac{1-\cos x}{x}.$$

The phase portrait of this case is shown in Figure 5.5.

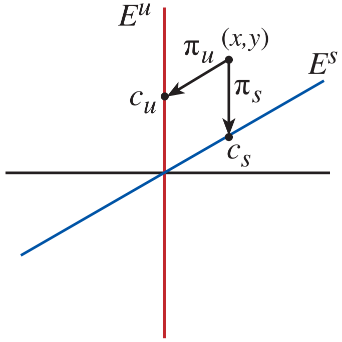

Figure 5.6. *Projections onto E^u and E^s.*

Note that W^s is tangent to E^s at the origin because its slope is

$$\left.\frac{dy}{dx}\right|_{x=0} = -\left.\int_0^1 Dg(xu)u\,du\right|_{x=0} = -\frac{1}{2}Dg(0),$$

which is precisely the slope of E^s. This tangency property will be generalized to the fully nonlinear case below. Since y is expressed as a function of x in (5.8) and each x determines a unique point on E^s, the stable manifold is a graph over E^s. Finally, both W^u and W^s are smooth curves, that is, they are *manifolds*. ■

In the construction of the manifolds in the example above, we noticed that W^s is a graph over E^s. To use this property for a general hyperbolic equilibrium, we define projection operators onto E^s and E^u. A projection is a linear operator $\pi : \mathbb{R}^n \to \mathbb{R}^n$ such that $\pi \circ \pi = \pi$. We will define two projections π_u and π_s such that $\pi_u + \pi_s = id$; see Figure 5.6. These projections formalize the idea of finding components of a vector "along the eigenvectors." Recall from §2.6 that any vector can be written as a linear combination of generalized eigenvectors,

$$x = \sum_{j=1}^{n} c_j v_j.$$

In other words, there is a nonsingular matrix $P = [v_1, v_2, \ldots, v_n]$ such that $x = Pc$ and $c = P^{-1}x$. If the first k of these vectors span E^u, then the projections are given by

$$\pi_u(x) = \sum_{j=1}^{k} c_j v_j, \quad \pi_s(x) = \sum_{j=k+1}^{n} c_j v_j.$$

Example 5.7. For the system (5.7) $P = (v_u, v_s) = \begin{pmatrix} 0 & 2 \\ 1 & -Dg(0) \end{pmatrix}$, so that

$$\begin{pmatrix} c_u \\ c_s \end{pmatrix} = P^{-1}\begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{2}\begin{pmatrix} Dg(0) & 2 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{1}{2}Dg(0)x + y \\ \frac{1}{2}x \end{pmatrix}.$$

Thus, the projection operators onto E^u and E^s are

$$\pi_u\begin{pmatrix} x \\ y \end{pmatrix} = c_u v_u = \begin{pmatrix} 0 \\ y + \frac{1}{2}xDg(0) \end{pmatrix}, \quad \pi_s\begin{pmatrix} x \\ y \end{pmatrix} = c_s v_s = \begin{pmatrix} x \\ -\frac{1}{2}xDg(0) \end{pmatrix}.$$ ■

With these examples under our belt, we proceed to develop a general understanding of the stable and unstable manifolds of a saddle equilibrium. We begin by restricting our study to a neighborhood of the equilibrium to construct the "local" manifolds.

5.4 ▪ Local Stable Manifold Theorem

In this section we will show that the stable and unstable sets of a hyperbolic equilibrium are actually smooth manifolds when the vector field is C^1. Suppose that x^* is a hyperbolic equilibrium with linearization $Df(x^*) = A$. We can always shift coordinates so that the equilibrium is at the origin by replacing $x \to x + x^*$, so that the equations take the form

$$\dot{x} = Ax + g(x), \tag{5.10}$$

where $g(x) = f(x + x^*) - Ax$ represents the nonlinear terms in the equation so that $g(0) = 0$ and $Dg(0) = 0$. Since A is hyperbolic, there is an $\alpha > 0$ such that $|\mathrm{Re}\lambda_i| > \alpha$ for all eigenvalues λ_i of A. The projection operators are $\pi_s : \mathbb{R}^n \to E^s$ and $\pi_u : \mathbb{R}^n \to E^u$. Since A leaves E^s and E^u invariant, by Lemma 2.19, it commutes with the projections:

$$\pi_u A = A\pi_u \text{ and } \pi_s A = A\pi_s .$$

The same is true for the fundamental matrix e^{tA}. Moreover, the estimate (2.44) in §2.7 implies that there is a $K \geq 1$ such that

$$\left.\begin{aligned} \left|e^{tA}\pi_s x\right| &\leq Ke^{-\alpha t}\left|\pi_s x\right|, \\ \left|e^{-tA}\pi_u x\right| &\leq Ke^{-\alpha t}\left|\pi_u x\right|, \end{aligned}\right. \quad t \geq 0. \tag{5.11}$$

Our goal is to prove that the stable set W^s for (5.10) is a smooth manifold, and our main tool is the contraction-mapping theorem (what else!). The first step is to find the appropriate operator, T, and function space. To motivate the construction of T—which generalizes the operator (3.12) used to prove existence and uniqueness—we first study a simpler set of affine ODEs.

Lemma 5.8. *Consider the affine, nonautonomous initial value problem*

$$\dot{x} = Ax + \gamma(t), \quad \pi_s x(0) = \sigma \in E^s. \tag{5.12}$$

Suppose A is hyperbolic and $\gamma(t)$ is bounded and continuous for $t \geq 0$. Then the unique solution, $x(t;\sigma)$, of (5.12) *that is bounded for positive time is*

$$x(t;\sigma) = e^{tA}\sigma + \int_0^t e^{(t-s)A}\pi_s\gamma(s)ds - \int_t^\infty e^{(t-s)A}\pi_u\gamma(s)ds. \tag{5.13}$$

The uniqueness of the solution (5.13) is surprising because only "half" of the initial conditions have been specified, the stable components σ. We will see that the assumption that x is bounded for $t > 0$ is enough to determine its unstable components.

Proof. The general solution of the forced linear equation can be obtained by the integrating factor method or the method of variation of parameters. To implement the latter, guess a solution of the form $x(t) = e^{tA}\xi(t)$. Substitute this into the ODE to obtain $\dot{\xi} = e^{-tA}\gamma(t)$, which can be solved trivially by integrating. If we specify the initial condition $x(\tau)$ at some arbitrary time $t = \tau$, the general solution to (5.12) has the form

$$x(t) = e^{(t-\tau)A}x(\tau) + \int_\tau^t e^{(t-s)A}\gamma(s)ds. \tag{5.14}$$

Our goal is to find a particular case of (5.14) that is bounded in forward time. We write $x(t) = \pi_u x(t) + \pi_s x(t)$ and consider these two projections separately.

First set $\tau = 0$ and take the stable projection of (5.14). Noting that $\pi_s x(0) = \sigma$, we obtain

$$\pi_s x(t) = e^{tA}\sigma + \int_0^t e^{(t-s)A}\pi_s \gamma(s)ds.$$

To show that this expression is bounded as $t \to \infty$, we use the assumption that γ is bounded, i.e., that there is a δ such that $|\gamma(s)| \le \delta$ for all $s \ge 0$. Imposing the bound (5.11) then gives

$$\left|\int_0^t e^{(t-s)A}\pi_s \gamma(s)ds\right| \le K\int_0^t e^{-(t-s)\alpha}\,|\gamma(s)|\,ds \le \frac{K}{\alpha}\delta.$$

Consequently, the stable projection of our solution is indeed bounded.

Projecting (5.14) onto the unstable space yields

$$\pi_u x(t) = e^{tA}\left(e^{-\tau A}\pi_u x(\tau) + \int_\tau^t e^{-sA}\pi_u \gamma(s)ds\right). \tag{5.15}$$

We must choose $\pi_u x(t)$ so that (5.15) remains bounded. Since the exponential $e^{tA}\pi_u$ generally grows without bound, a necessary condition is that the term in parenthesis in (5.15) limits to zero as $t \to \infty$, that is,

$$e^{-\tau A}\pi_u x(\tau) = -\int_\tau^\infty e^{-sA}\pi_u \gamma(s)ds.$$

Since this is true for any τ, we can replace τ by t in this equation to obtain

$$\pi_u x(t) = -\int_t^\infty e^{(t-s)A}\pi_u \gamma(s)ds. \tag{5.16}$$

Substitution of (5.16) back into (5.15) gives an identity; therefore, (5.16) is a solution for the unstable projection. We now show that (5.16) is indeed bounded. The integral in (5.16) can be bounded using (5.11) on $e^{\tau A}\pi_u$ for $\tau = t - s \le 0$:

$$|\pi_u x(t)| = \left|\int_t^\infty e^{(t-s)A}\pi_u \gamma(s)ds\right| \le K\int_t^\infty e^{(t-s)\alpha}\,|\gamma(s)|\,ds \le \frac{K}{\alpha}\delta.$$

Thus, (5.16) is both necessary and sufficient for the unstable projection being bounded.

Adding the stable and unstable projections gives the promised result (5.13). □

We now return to (5.10), where $\gamma(t)$ is replaced by the nonlinear function $g(x)$. If we similarly replace $\gamma(s)$ in the integrand of (5.13) with $g(x(s))$, the resulting integral equation is satisfied by a solution of (5.10). Just as for the integral operator (3.12), which we used to prove existence and uniqueness, the new integral equation can be viewed as an operator on a suitable function space. Indeed we will show that this operator is a contraction map whose fixed point is the stable manifold of (5.10). Since g is nonlinear, we must restrict the analysis to a neighborhood of the equilibrium where g is sufficiently small; thus, we will only prove the existence of a "local" stable manifold, W^s_{loc}: the set of points on W^s that remain in some neighborhood of the equilibrium for all $t \ge 0$. The global stable manifold will be constructed from the local one in §5.5.

Theorem 5.9 (Local Stable Manifold). *Let A be hyperbolic, $g \in C^k(U)$, $k \geq 1$, for some neighborhood U of 0, and $g(x) = o(x)$ as $x \to 0$. Denote the linear stable and unstable subspaces of A by E^s and E^u. Then there is a $\tilde{U} \subset U$ such that local stable manifold of* (5.10),

$$W^s_{loc}(0) \equiv \left\{ x \in W^s(0) : \varphi_t(x) \in \tilde{U},\ t \geq 0 \right\},$$

is a Lipschitz graph over E^s that is tangent to E^s at 0. Moreover, $W^s_{loc}(0)$ is a C^k manifold.

Since this is a rather long proof, we divide it into three parts. In the first part we prove that there is a unique, forward bounded solution for each point $\sigma \in E^s$ close enough to the origin. We then show in the second part that these solutions actually are on the stable manifold, since they are asymptotic to 0. In the final part of the proof, we show that these solutions lie on a smooth, Lipschitz graph.

Proof (Part 1). By analogy with (5.13), define an operator $T : C^0(\mathbb{R}^+, \mathbb{R}^n) \to C^0(\mathbb{R}^+, \mathbb{R}^n)$ for a given point $\sigma \in E^s$ of A by

$$T(x)(t) = e^{tA}\sigma + \int_0^t e^{(t-s)A}\pi_s g(x(s))ds - \int_t^\infty e^{(t-s)A}\pi_u g(x(s))ds. \tag{5.17}$$

It is clear that if $x \in C^0(\mathbb{R}^+, \mathbb{R}^n)$, then so is $T(x)$. It is not hard to show that a sufficiently small, continuous fixed point of T, $x : \mathbb{R}^+ \to U$ is a C^1 solution of the ODE (5.10), call it $x(t;\sigma)$ (see Exercise 5).

We first show that T is a contraction map and therefore that the fixed point of T exists and is unique. To do this, define a closed subset of the function space $C^0(\mathbb{R}^+)$ by

$$V_\delta = \left\{ x \in C^0(\mathbb{R}^+, \mathbb{R}^n) : ||x|| \leq \delta \right\}, \tag{5.18}$$

where $||x||$ is the sup-norm (3.4). As discussed in §3.2, this space with the sup-norm is complete. Since $g(x) = o(x)$ as $x \to 0$ (recall §4.4), then for any $\varepsilon > 0$—no matter how small—there is a δ, such that when $x \in V_\delta$, then $|g(x(t)| \leq \varepsilon\, |x(t)|$. Using the bounds (5.11) in (5.17) we obtain

$$|T(x)(t)| \leq Ke^{-t\alpha}|\sigma| + K\varepsilon \int_0^t e^{-(t-s)\alpha}|x(s)|ds + K\varepsilon \int_t^\infty e^{(t-s)\alpha}|x(s)|ds \leq K|\sigma| + 2\frac{K\varepsilon}{\alpha}\delta$$

for any for $t \geq 0$. The necessary bound $||T(x)|| \leq \delta$ can be satisfied by requiring, e.g.,

$$|\sigma| \leq \delta/2K \text{ and } \varepsilon \leq \alpha/4K. \tag{5.19}$$

These requirements define the neighborhood

$$\tilde{U} = \left\{ x : |g(x)| \leq \frac{\alpha}{4K}|x| \right\} \cap U \tag{5.20}$$

that effectively defines δ, since ε can be made arbitrarily small for a sufficiently small δ

We now show that T is a contraction. Since $g \in C^1$, and $||Dg(x)|| \leq \varepsilon$ for $|x| \leq \delta$, then (3.9) implies that $|g(x) - g(y)| \leq \varepsilon\,|x - y|$ for $x, y \in B_\delta(0)$. Using this and (5.11) gives

$$|T(x) - T(y)| \leq K\varepsilon\, ||x - y|| \left(\int_0^t e^{-(t-s)\alpha}ds + \int_t^\infty e^{(t-s)\alpha}ds \right) \leq 2\frac{K\varepsilon}{\alpha}\, ||x - y||.$$

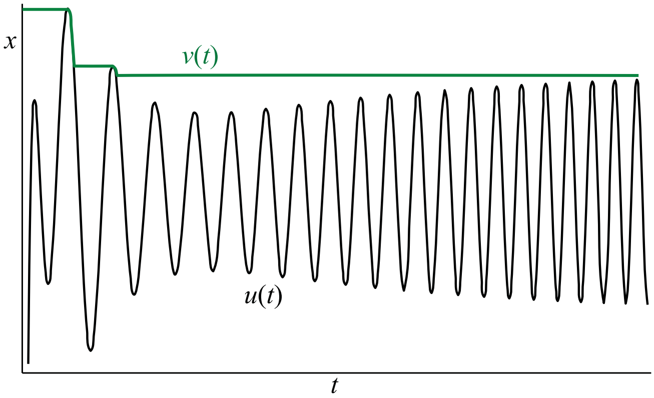

Figure 5.7. *Construction of the function* $v(t)$ *in* (5.23).

Therefore, T is a contraction when $\varepsilon \leq \alpha/4K$, which we already assumed, and the contraction-mapping theorem implies that T has a unique fixed point in V_δ. Since there is a unique fixed point $x(t;\sigma)$ for each $\sigma \in E^s$ providing $|\sigma| < \delta/2K$, the set $x(0;\sigma)$ is a graph over E^s. □

Proof (Part 2). To show that $x(t;\sigma)$ is a point on the stable manifold, we must show it approaches zero as $t \to \infty$. Since is $x(t;\sigma)$ is a fixed point of T, we use (5.11) to bound it by

$$|x(t;\sigma)| \leq Ke^{-\alpha t}|\sigma| + K\varepsilon \int_0^t e^{-\alpha(t-s)}|x(s;\sigma)|\,ds + K\varepsilon \int_t^\infty e^{\alpha(t-s)}|x(s;\sigma)|\,ds. \quad (5.21)$$

We assert that this implies that $x \to 0$ exponentially fast. To show this, we need a generalization of Grönwall's inequality (3.31). □

Lemma 5.10 (Generalized Grönwall). *Suppose* α, M, *and* L *are nonnegative,* $L < \alpha/2$, *and there is a nonnegative, bounded, continuous function* $u : \mathbb{R}^+ \to \mathbb{R}^+$ *satisfying*

$$u(t) \leq e^{-\alpha t}M + L\int_0^t e^{-\alpha(t-s)}u(s)ds + L\int_t^\infty e^{\alpha(t-s)}u(s)ds; \quad (5.22)$$

then $u(t) \leq \frac{M}{\beta}e^{-(\alpha - L/\beta)t}$, *where* $\beta = 1 - 2\frac{L}{\alpha}$.

Putting aside the proof of the lemma for the moment, note that it applies to the inequality (5.21) since we know that the fixed point $x(t;\sigma)$ is continuous. We set $u = |x(t;\sigma)|$, $L = K\varepsilon$, and $M = K|\sigma|$. Since $\varepsilon \leq \alpha/4K$, then $L \leq \alpha/4$ and $\beta = 1 - 2L/\alpha \geq 1/2$, so that $L/\beta \leq \alpha/2$. Thus the hypotheses of Lemma 5.10 apply and give

$$|x(t;\sigma)| \leq 2Ke^{-\alpha t/2}|\sigma|,$$

implying that $x(t;\sigma) \to 0$ exponentially fast.

Proof (of Lemma). By assumption u is bounded; therefore, we can define its supremum. Moreover, the function

$$v(t) = \sup_{s>t} u(s) \quad (5.23)$$

exists and is nonincreasing: $v(t) \le v(s)$ if $s \ge t$; see Figure 5.7. Since u is continuous, for any t and any $\varepsilon > 0$ there is a $T \ge t$ such that $v(t) \le u(T) + \varepsilon$. Thus using (5.22) gives

$$\begin{aligned} v(t) \le u(T) + \varepsilon \le & e^{-T\alpha} M + L \int_0^T e^{-\alpha(T-s)} u(s) ds + L \int_0^\infty e^{-\alpha s} u(T+s) ds + \varepsilon \\ \le & e^{-T\alpha} M + L \int_0^t e^{-\alpha(T-s)} u(s) ds + L \int_t^T e^{-\alpha(T-s)} u(s) ds \\ & + L \int_0^\infty e^{-\alpha s} u(T+s) ds + \varepsilon \\ \le & e^{-T\alpha} M + L \int_0^t e^{-\alpha(T-s)} u(s) ds + 2\frac{L}{\alpha} v(t) + \varepsilon, \end{aligned}$$

where we have used the facts that $u(s) \le v(t)$ and $u(T+s) \le v(t)$ to approximate the last two integrals. Rearranging this gives

$$\left(1 - 2\frac{L}{\alpha}\right) e^{\alpha t} v(t) \le e^{-\alpha(T-t)} M + L \int_0^t e^{-\alpha(T-t)} e^{\alpha s} u(s) ds + \varepsilon e^{\alpha t}.$$

Defining $z(t) = \beta e^{\alpha t} v(t)$, and noting that $e^{-\alpha(T-t)} \le 1$, we have

$$z(t) \le M + \varepsilon e^{\alpha t} + \frac{L}{\beta} \int_0^t z(s) ds.$$

This is of the form of the time-dependent version of Grönwall's lemma in Exercise (3.11), so that $z(t) \le (M + \varepsilon e^{\alpha t}) e^{tL/\beta}$. Since this is true for *any* $\varepsilon > 0$, rewriting it in terms of $u(t) \le v(t)$ gives the promised result. □

Proof (Part 3). It is relatively easily to see that the solutions $x(t;\sigma)$ lie on a Lipschitz graph, i.e., that the unstable components are Lipschitz functions of σ. To show this, consider $\pi_u x$ at two different σ values, subtract the fixed-point equations $x = T(x)$, and take the projections onto E^u. Using the fact that π_u annihilates σ, we obtain

$$|\pi_u(x(t;\sigma_1) - x(t;\sigma_2))| \le K\varepsilon \int_t^\infty e^{(t-s)\alpha} |x(s;\sigma_1) - x(s;\sigma_2)| ds. \tag{5.24}$$

To evaluate this, we must also bound the difference in the integral, which we can do with the same integral equation:

$$\begin{aligned} |x(t;\sigma_1) - x(t;\sigma_2)| \le K e^{-\alpha t} |\sigma_1 - \sigma_2| + & K\varepsilon \int_0^t e^{-(t-s)\alpha} |x(s;\sigma_1) - x(s;\sigma_2)| ds \\ + & K\varepsilon \int_t^\infty e^{(t-s)\alpha} |x(s;\sigma_1) - x(s;\sigma_2)| ds. \end{aligned}$$

This is of the form (5.22), so the generalized Grönwall inequality yields

$$|x(t;\sigma_1) - x(t;\sigma_2)| \le 2K e^{-\alpha t/2} |\sigma_1 - \sigma_2|.$$

Consequently, $x(t;\sigma)$ is a Lipschitz function of σ. We can now use this bound in (5.24) to obtain

$$|\pi_u x(t;\sigma_1) - \pi_u x(t;\sigma_2)| \le \frac{4K^2\varepsilon}{3\alpha} e^{-\alpha t/2} |\sigma_1 - \sigma_2|,$$

giving the promised Lipschitz condition.

Differentiability of the stable set is more difficult to prove. The basic principle we will use is the following generalization of Theorem 3.9, the contraction-mapping theorem: if a contraction map depends smoothly on parameters, its fixed points must as well.

Theorem 5.11 (Uniform Contraction Principle). *Let X and Y be closed subsets of two Banach spaces and let $T \in C^k(X \times Y, X)$, $k \geq 0$, be a uniform contraction map.*[31] *Then there is a unique fixed point, $x(y) = T(x(y), y)$, where $x(y) \in X$ is a C^k function of $y \in Y$.*

Delaying the proof of this theorem for the moment, note that it gives the promised result. It applies to our map T because when g is C^k, the fixed point, $x(t;\sigma)$ is also C^k in both t and σ. It also implies the tangency of W^s to E^s, since the Jacobian matrix obtained from differentiating x with respect to σ at $\sigma = 0$ is

$$D_\sigma x(t;0) = e^{tA}\pi_s + \left(\int_0^t ds e^{(t-s)A}\pi_s - \int_t^\infty ds e^{(t-s)A}\pi_u\right) Dg\left(x(s;0)\right) D_\sigma x(s;0).$$

But since $x(s;0) = 0$ is the unique fixed point when $\sigma = 0$ and $Dg(0) = 0$, this implies $D_\sigma x(t;0) = e^{tA}\pi_s$. Consequently, for any v, $D_\sigma x(t;0)v \in E^s$, so that W^s is tangent to E^s.

***Proof (of Theorem* 5.11).** Let $|| \, ||$ denote the norms on both X, and Y. Since T is a uniform contraction, there is a constant c such that $0 < c < 1$ and $||T(x;y) - T(\xi, y)|| \leq c\,||x - \xi||$ for all $x, \xi \in X$, and $y \in Y$. Moreover, the contraction mapping theorem, Theorem 3.9, implies that for each y there is a unique fixed-point $x(y) = T(x(y);y)$.

Suppose first that T is uniformly C^0. We will show that the fixed point, $x(y)$, is uniformly continuous. The fixed-point equation and triangle inequality imply that for any $h \in Y$

$$\begin{aligned} ||x(y+h) - x(y)|| &= ||T(x(y+h); y+h) - T(x(y);y)|| \\ &\leq ||T(x(y+h); y+h) - T(x(y); y+h)|| \\ &\quad + ||T(x(y); y+h) - T(x(y);y)|| \\ &\leq c\,||x(y+h) - x(y)|| + ||T(x(y); y+h) - T(x(y);y)||. \end{aligned}$$

Since T is uniformly continuous in y, for every $\varepsilon > 0$ there is a $\delta > 0$ such that whenever $||h|| < \delta$ then $||T(x; y+h) - T(x, y)|| < \varepsilon$; using this, the previous inequality gives

$$||x(y+h) - x(y)|| < \frac{\varepsilon}{1-c}$$

for any $\varepsilon > 0$. This shows that x is uniformly continuous, since c and ε are independent of y.

It is much more difficult to prove smoothness; we will consider only the case $k = 1$. Suppose that T is uniformly C^1. If the fixed point $x(y) = T(x(y);y)$ were differentiable, then its derivative would obey the relation

$$D_y x(y) = D_x T(x(y);y) D_y x(y) + D_y T(x(y);y). \tag{5.25}$$

[31] This means that the contraction constant $c < 1$ is independent of y and that $T(x;y)$ is a uniformly C^k function of y.

Replace $D_y x$ by a linear operator $M : X \to X$ and think of this equation as a linear system for an unknown M:

$$(I - D_x T(x(y); y))M = D_y T(x(y); y). \tag{5.26}$$

This system has a unique solution if the left-hand side is nonsingular.[32] This follows since $||D_x T|| \le c < 1$; see Exercise 6. Now we must show that this $M(y)$ is really $D_y x$. Define

$$\xi(h) \equiv x(y+h) - x(y) = T(x(y) + \xi(h); y+h) - T(x(y); y).$$

Combining this with (5.26) gives

$$\begin{aligned}&(I - D_x T(x(y); y))(\xi(h) - M(y)h) = \Delta(\xi, h),\\ &\Delta(\xi, h) \equiv T(x(y) + \xi(h); y+h) - T(x(y); y) - D_x T(x(y); y)\xi(h) - D_y T(x(y); y)h.\end{aligned}$$

If we can show that $||\Delta|| \to 0$ as $||h|| \to 0$, then because $I - D_x T$ is nonsingular, we would have $\xi(h) - Mh \to 0$, which would imply that $x(y)$ is differentiable with derivative M.

Since T is C^1, for any $\varepsilon > 0$ there is a $\delta > 0$ such that when $||h|| < \delta$ and $||\xi(h)|| < \delta$, we have

$$||\Delta(\xi, h)|| < \varepsilon(||\xi(h)|| + ||h||). \tag{5.27}$$

This is not quite good enough since we do not have $\xi = O(h)$ yet. However, this can be obtained using the definition of Δ, which implies

$$\xi(y) = D_x T(x(y); y)\xi + D_y T(x(y); y)h + \Delta.$$

Using the bounds on $D_x T$ and Δ we obtain

$$\begin{aligned}||\xi(h)|| \le c\,||\xi(h)|| + \left\|D_y T(x(y); y)h\right\| + \varepsilon(||\xi(h)|| + ||h||) \quad &\Rightarrow\\ ||\xi(h)|| \le \frac{\left\|D_y T(x(y); y)h\right\| + \varepsilon\,||h||}{1 - c - \varepsilon} &\le C\,||h||,\end{aligned}$$

providing $\varepsilon < 1 - c$. Putting this back into (5.27) gives

$$||\Delta(\xi, h)|| < \varepsilon(C+1)\,||h||.$$

Therefore $||\Delta|| \to 0$ as $||h|| \to 0$.

Showing that x is C^k for $k > 1$ requires an additional inductive step, which we leave to the reader. □

This completes, as well, our rather lengthy proof of Theorem 5.9. □

Example 5.12. The two-dimensional system

$$\begin{aligned}\dot{x} &= 2x + y^2,\\ \dot{y} &= -2y + x^2 + y^2\end{aligned} \tag{5.28}$$

[32] Equation (5.25) can also be thought of as a contraction map on $D_y x$ and so has a unique solution.

has a saddle at the origin with a diagonal Jacobian $Df(0,0) = \text{diag}(2,-2)$. Consequently, the linear spaces are $E^u = \text{span}(1,0)^T$ and $E^s = \text{span}(0,1)^T$ with the corresponding projection matrices

$$\pi_u = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \pi_s = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

These exemplify the general property $\pi_u + \pi_s = I$. Given a point $\sigma = (0, \sigma_y) \in E^s$, the operator (5.17) becomes

$$T(x) = \begin{pmatrix} 0 \\ e^{-2t}\sigma_y \end{pmatrix} + \begin{pmatrix} -e^{2t}\int_t^\infty e^{-2s} y^2(s)\,ds \\ e^{-2t}\int_0^t e^{2s}\left(x^2(s) + y^2(s)\right)ds \end{pmatrix}.$$

According to Theorem 5.9, we can begin with any function in V_δ providing δ is small enough. The crucial estimate is that $|g(x)| < \varepsilon|x|$, for $|x| < \delta$. For the example, $|g(x)| \le \sqrt{2}\delta^2$, so we may set $\delta = \varepsilon/\sqrt{2}$. Since $Df(0,0)$ is diagonal with $|\lambda| = 2$, we may set $K = 1$ and $\alpha = 2$ so the requirements (5.19) become

$$\varepsilon < \frac{1}{2} \text{ and } \delta < \frac{1}{2\sqrt{2}}.$$

Beginning with the initial guess $(x_0(t), y_0(t)) = (0,0)$, clearly in V_δ, the first two iterates of T are

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = T(x_0, y_0) = \begin{pmatrix} 0 \\ e^{-2t}\sigma_y \end{pmatrix},$$

$$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = T(x_1, y_1) = \begin{pmatrix} -\frac{1}{6}e^{-4t}\sigma_y^2 \\ e^{-2t}\sigma_y + \frac{1}{2}e^{-2t}\left(1 - e^{-2t}\right)\sigma_y^2 \end{pmatrix}.$$

Note that the approximate solutions do indeed limit to the origin as $t \to \infty$. To obtain a picture of the stable manifold, it is sufficient to plot the curve as a function of the initial point at any value of t, say, for example, at $t = 0$. In this case we have a parametric representation of the approximate stable manifold:

$$W^s_{loc}(0,0) \approx \left\{ (x_2(y), y); x_2(y) = -\frac{1}{6}y^2,\ |y| < \frac{1}{2}\delta \right\}.$$

The next iterate gives an improved curve

$$x_3(y) = -\frac{1}{6}y^2\left(1 + \frac{1}{4}y + \frac{1}{40}y^2\right). \tag{5.29}$$

A plot of this curve, along with some representative trajectories, is shown in Figure 5.8. Note that the approximate manifold fails to capture the behavior near the spiral focus at $(-0.931, 1.364)$. Continuing this iteration would show that the first two terms in (5.29) are correct, but the coefficient of the third is not. ■

5.5 ▪ Global Stable Manifolds

The stable manifold theorem implies that there is a neighborhood of a hyperbolic equilibrium for which the local stable manifold, W^s_{loc}, is a smooth submanifold of $\mathbb{R}^n$

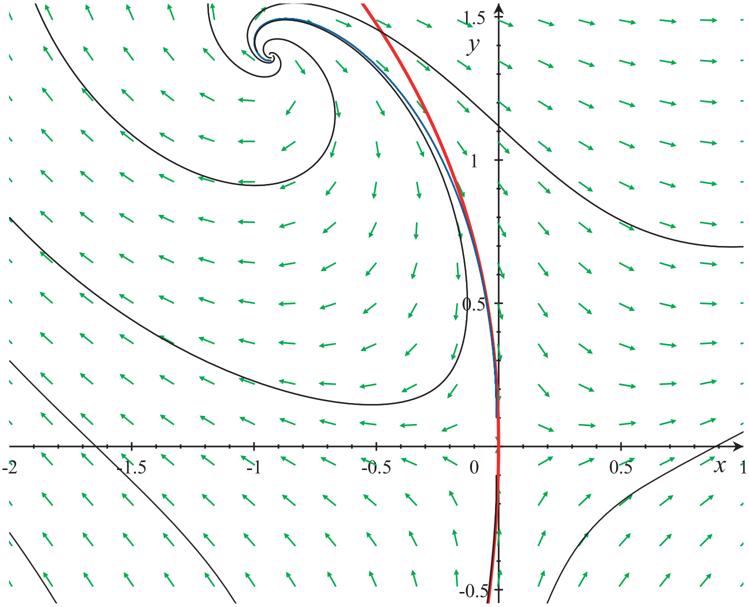

Figure 5.8. *Phase portrait of* (5.28) *and its approximate stable manifold* (5.29).

with the same dimension as the stable subspace, E^s. On the other hand, the global stable set consists of *all* points that eventually limit on the equilibrium in forward time. As Lemma 5.1 implies, W^s is an invariant set: if $z \in W^s(x^*)$, then so are all the points on its orbit, $\varphi_t(z)$. Moreover, since every point on $W^s(x^*)$ must eventually stay in an arbitrarily small neighborhood x^*, the forward orbit of every point in W^s must eventually land in W^s_{loc}. Consequently, if we extend the local stable manifold by allowing each point to flow backward, we obtain the global stable set:

$$W^s = \left\{\varphi_t(x) : \, x \in W^s_{loc}, \; t \in \mathbb{R}\right\}.$$

Since W^s_{loc} is smooth, and the orbits are smooth functions of time, the extension of W^s_{loc} for any *finite* value of t is as smooth as the vector field. However, it is not obvious that the set W^s defined for *all* t is quite so nice. The question that we seek to answer here is, how "nice" is W^s?

To discuss the structure of W^s, we briefly pause to consider several properties of maps from one space to another. Our goal is to define the concept of "embedding," which is, loosely speaking, what we think of when we imagine a smooth surface.

Mathematically, a relation of the form $g : M \to N$ that maps one space into another defines a surface—we say g is a *map*. So that it is *possible* for g to be one-to-one, we will require that $m = \dim(M) \le n = \dim(N)$.

Example 5.13. Consider the map $g : \mathbb{S}^1 \to \mathbb{R}^2$ defined by $g(\theta) = (x(\theta), y(\theta)) = (2\cos\theta, \sin\theta)$. This is a mapping of a circle represented by the points $\theta \in [0, 2\pi)$ into $\mathbb{R}^2$ represented by points (x, y). Geometrically, the map describes an ellipse. Alternatively, the map $g(\theta) = (\sin(2\theta), \sin\theta)$ describes a figure eight; see Figure 5.9. Both are maps of the circle into the plane, but the latter map is not one-to-one. ∎

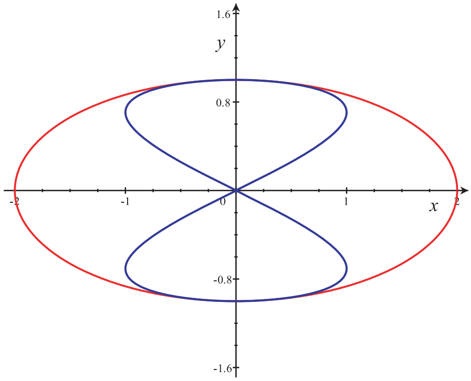

Figure 5.9. *Two maps of the circle into the plane.*

Both maps in the example are locally smooth in the sense that each component of g is a C^1 function. The Jacobian derivative of the map g, at a point $x \in M$, $Dg(x)$, is a matrix of dimension $n \times m$; it takes a vector v of dimension m and gives a new vector $w = Dg(x)v$ attached to the point $g(x) \in N$.[33] Indeed, this vector is tangent to the surface $g(M)$, and the range of $Dg(x)$ corresponds to the tangent plane to the surface. If the rank of $Dg(x)$ is m for all x, then the tangent planes are everywhere m-dimensional. Both maps in the first example have this property: since the derivative is a nonzero vector for all θ, $\text{rank}(Dg(\theta)) = 1$.

Example 5.14. The map $g : \mathbb{R}^2 \to \mathbb{R}^3$ defined by

$$g(s,t) = (\cos t - s, \sin s - t, 2\sin t) \tag{5.30}$$

gives the surface shown in Figure 5.10. Its two tangent vectors are the columns of Dg, $v_1 = (-\sin t, -1, 2\cos t)^T$ and $v_2 = (-1, \cos s, 0)^T$. Since $v_1 \times v_2 \neq 0$, these vectors are never parallel and they define a two-dimensional plane tangent to the surface for each (s,t). ∎

A map with this property is an

▹ *immersion*: A C^1 map $g : M \to N$ is an immersion if $\text{rank}(Dg) = \dim(M)$.

An immersion is locally a smooth surface.

Example 5.15. Consider the map $g : \mathbb{R}^1 \to \mathbb{R}^2$ given by

$$g(t) = (1 + \cos(2t), \cos t). \tag{5.31}$$

The rank of $Dg(t) = -(2\sin(2t), \sin t)$ is 1 except where it vanishes, i.e., when $t = n\pi$. The curve (5.31) has a cusp at these points, as shown in Figure 5.11. Consequently, it fails to be an immersion. ∎

[33] Actually, $Dg(x)$, is a map from the tangent space of M to the tangent space of N. Thus, $Dg(x)v$ is a tangent vector, a point in the tangent space $TN_{g(x)}$. If $N = \mathbb{R}^n$, then we can identify the tangent space with $\mathbb{R}^n$.

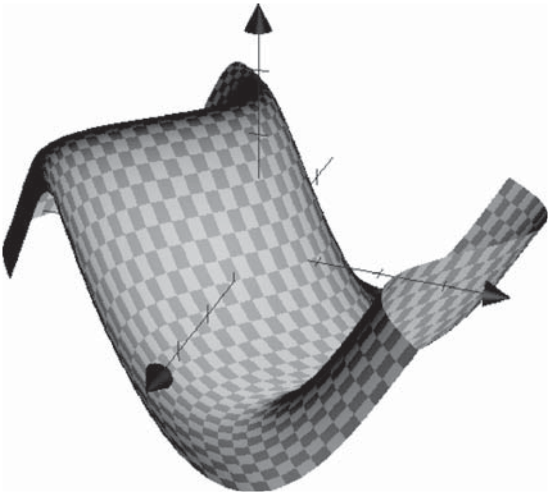

Figure 5.10. *Immersion* (5.30) *into* $\mathbb{R}^3$.

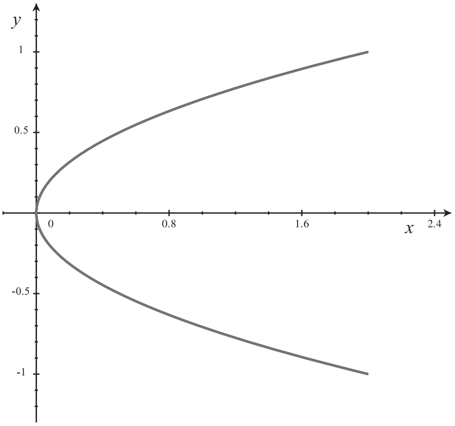

Figure 5.11. *Singular map* (5.31).

The global stable manifold is easily seen to be an immersion:

Lemma 5.16. *Let f be a C^1 vector field on $\mathbb{R}^n$ with hyperbolic equilibrium at the origin having a k-dimensional stable space E^s. Then $W^s(0)$ is a k-dimensional immersion.*

Proof. Let the local stable manifold be defined by the map g: $g : E^s \to \mathbb{R}^n$ where $g(\sigma) = x(0;\sigma)$. The stable manifold theorem implies that W^s is an immersion since it defines a smooth Lipschitz graph over E^s. Hence, the rank of Dg is k. Each neighborhood of the global stable manifold can be obtained by flowing a region on W^s_{loc} backward in time for some fixed time. Thus for any neighborhood of W^s, we can consider the set of points defined by the map $h(\sigma) = \varphi_t(g(\sigma))$. This is smooth since φ is a smooth function of its arguments according to Theorem 3.30. Moreover, the derivative of this map is

$$Dh = D\varphi_t(g(\sigma))Dg(\sigma),$$

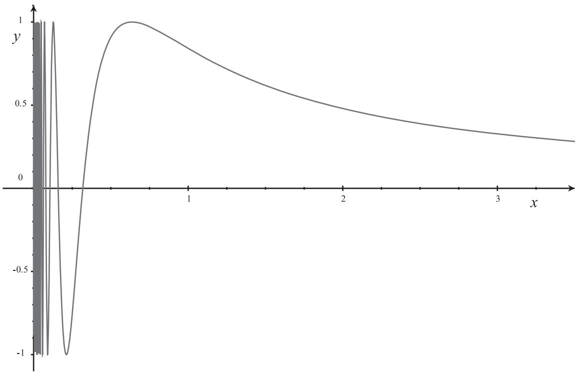

Figure 5.12. *The topologist's sine curve.*

which has rank k since the matrix $\Phi = D\varphi_t$ solves the linearized differential equation (4.51) with initial condition $\Phi(0) = I$ and therefore is a nonsingular matrix. □

Even though an immersion is smooth, it may cross itself. For example, the figure-eight curve in Figure 5.9, though an immersion, is not one-to-one since both $\theta = 0$ and π are mapped to the origin. Even if we eliminate this problem by requiring that an immersion be one-to-one, there can be problems, as follows.

Example 5.17. Consider the immersion $g : \mathbb{R} \to \mathbb{T}^2$ given by $g(t) = (t \bmod 1, \nu t \bmod 1)$, where ν is irrational. This is smooth and one-to-one but gives a dense line on the torus (see §7.1)—not what one would think of as a submanifold. ■

Example 5.18. The topologist's sine curve is the map $g : \mathbb{R}^+ \to \mathbb{R}^2$ defined by $g(t) = (1/t, \sin t)$. This curve is an immersion since $Dg \neq 0$ and is one-to-one. However, as $t \to 0$, the curve has infinitely many oscillations and accumulates upon the interval $[-1, 1]$ on the y-axis, as can be seen in Figure 5.12. ■

We will see later that the global stable manifold can have this accumulation problem: indeed, this is one of the indications of chaos. A map that does not have this pathology is called a

> ▷ *proper map*: A map $g : M \to N$ is proper if the preimage of every compact set in N is compact in M.

Example 5.19. The topologist's sine curve of Figure 5.12 is not proper because any neighborhood of the origin in $\mathbb{R}^2$ has a preimage consisting of infinitely many intervals of t in $(0, \infty)$. ■

We finally arrive at the ultimate definition of a "nice" map:

> ▷ *embedding*: A map $g : M \to N$ is an embedding if it is a one-to-one, proper immersion.

Of our examples above, only the ellipse and the map (5.30) are embeddings. However, any finite piece of W^s is an embedding, as follows from the next theorem.

Theorem 5.20. *If $g : M \to N$ is a C^1, one-to-one immersion, and both M and N are compact then it is automatically proper.*

Proof. Consider a compact subset $U \subset N$. Since U is closed, its complement is open. Since g is continuous, the preimage of any open set is open, and thus the preimage of U is the complement of an open set. Therefore, $g^{-1}(U)$ is closed and must be compact since it is a subset of the compact set M. So g is proper. □

5.6 ▪ Center Manifolds

Linear systems are classified according to their generalized eigenspaces, E^s, E^u, and E^c. The most important distinction was made between hyperbolic systems, where E^c is trivial, and nonhyperbolic systems. We now begin our study of the behavior of a system with a nonhyperbolic fixed point—that is, for cases where $E^c \neq \{0\}$. This study will continue in Chapter 6 for the planar case and also will be a major focus of bifurcation theory in Chapter 8.

In the nonhyperbolic case it is still possible to construct stable and unstable manifolds at the fixed point for the hyperbolic directions. Moreover, the nonhyperbolic part of the dynamics can be reduced to a system of ODEs with the same dimension as the center subspace of the linear system. This is based on the following generalization of the stable manifold theorem.

Theorem 5.21 (Center Manifold). *Suppose that f is a C^k vector field, $k \geq 1$, with a fixed point at the origin. Let the eigenspaces of $Df(0) = A$ be written $E^u \oplus E^c \oplus E^s$. Then there is a neighborhood of the origin in which there exist C^k locally invariant manifolds: the local stable manifold, W^s_{loc}, tangent to E^s, on which $|x(t)| \to 0$ as $t \to \infty$, the local unstable manifold W^u_{loc}, tangent to E^u, on which $|x(t)| \to 0$ as $t \to -\infty$, and a local center manifold W^c, tangent to E^c.*

The proof of this theorem is more complicated than the stable manifold theorem; see (Carr 1981; Chicone 1999; Chow and Hale 1982; Hirsch, Pugh, and Shub 1977).

Note that this theorem does not state that the manifolds are unique, nor does it say that the manifolds are the only sets that have the proper asymptotic behavior. This is to be contrasted with the stable-manifold theorem for hyperbolic equilibria, which asserts that the local stable and unstable manifolds are unique and that they generate the global manifolds.

Example 5.22. Consider the skew-product system

$$\begin{aligned} \dot{x} &= x^2, \\ \dot{y} &= -y. \end{aligned} \tag{5.32}$$

Here, the linearization of the equilibrium at the origin has eigenvalues $\lambda = 0$ and -1, so the stable space E^s is the y-axis and the center space E^c is the x-axis. It is clear that the local stable manifold is the y-axis, since this is tangent to E^s and every point on the y-axis limits to the origin. We are tempted to say that W^c_{loc} is the x-axis, and this is certainly an acceptable center manifold: it is clearly an invariant set and is tangent to E^c. However, if we solve the equation for the phase curves, by dividing the y equation by the x equation, we obtain

$$\frac{dy}{dx} = -\frac{y}{x^2} \quad \Rightarrow y(x) = ce^{x^{-1}}.$$

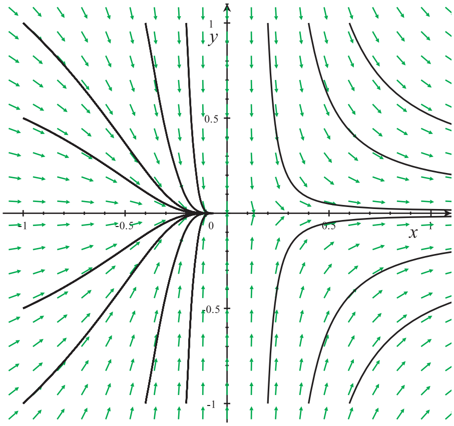

Figure 5.13. *Phase portrait of* (5.32).

When $x < 0$, each of these curves is asymptotic to the origin and is tangent to the x-axis (in fact, the function $y(x)$ has all derivatives zero at $x = 0^-$). So we could define a center manifold by

$$W^c_{loc}(0,0) = \left\{ (x,y) : y = \left[\begin{array}{ll} ce^{x^{-1}} & x < 0 \\ 0 & x \geq 0 \end{array} \right. \right\} \tag{5.33}$$

for *any* value of c. There is a one-parameter family of possible center manifolds; see Figure 5.13! This example shows that the center manifold is *not* unique.

The example also has another pathology: though the local stable manifold is the y-axis, the *global* stable set—namely, the set of points that are asymptotic to the origin—is the left half-plane $W^s(0,0) = \{(x,y) : x \leq 0\}$. Similarly, the *global* unstable set is the positive x-axis $W^u(0,0) = \{(x,0) : x > 0\}$ since this set is asymptotic to the origin as $t \to -\infty$. ∎

In the example, the center manifold was not unique; nevertheless, every choice of c in (5.33) gives a curve with the *same* power series expression, namely, $y(x) = 0 + 0x + 0x^2 + \cdots$. Consequently, as far as the power series is concerned, there is a unique center manifold, the x-axis. Indeed, whenever f is C^∞, there is a unique power series expression for a center manifold.

This series can be easily determined by looking for functions corresponding to a graph that is tangent to E^c and demanding that the resulting surfaces are invariant. It is most convenient to do this by preparing the system so that the linear matrix breaks into blocks corresponding to the stable, unstable, and center subspaces. To do this, write the system as $\dot{\xi} = A\xi + g(\xi)$, where $g = o(\xi)$ represents the nonlinear terms. As we saw in §2.6, the matrix P of generalized eigenvectors transforms A to block diagonal form: $P^{-1}AP = J$, where

$$J = \begin{pmatrix} C & 0 & 0 \\ 0 & S & 0 \\ 0 & 0 & U \end{pmatrix}.$$

Here, C, S, and U are square matrices representing the center, stable, and unstable dynamics; they are diagonal only if A is semisimple. Then we define new coordinates

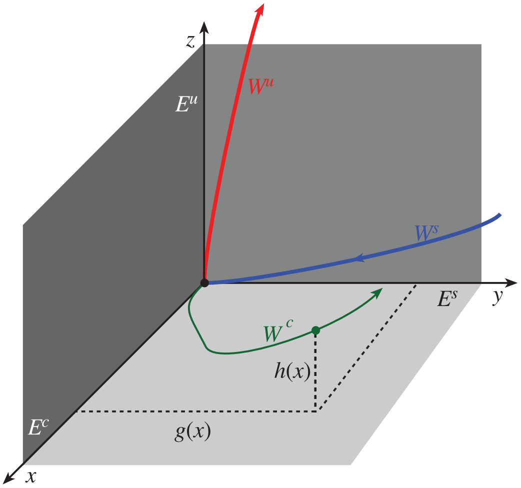

Figure 5.14. *Stable, unstable, and center manifolds.*

$\eta = P^{-1}\xi$, so that

$$\begin{aligned}\dot{\eta} &= P^{-1}A\xi + P^{-1}g(\xi) = P^{-1}AP\eta + P^{-1}g(P\eta)\\ &= J\eta + P^{-1}g(P\eta).\end{aligned}$$

Now set $\eta = (x, y, z)$, where $\dim(x) = \dim(E^c)$, $\dim(y) = \dim(E^s)$, and $\dim(z) = \dim(E^u)$. In terms of the three subsets of variables, the ODEs now take the form

$$\begin{aligned}\dot{x} &= Cx + F(x, y, z),\\ \dot{y} &= Sy + G(x, y, z),\\ \dot{z} &= Uz + H(x, y, z).\end{aligned} \tag{5.34}$$

Since the local center manifold W^c is a graph over E^c, we can define it using two maps $g : E^c \to E^s$, and $h : E^c \to E^u$, so that on W^c we have $y = g(x)$ and $z = h(x)$ (see Figure 5.14), that is,

$$W^c = \{(x, g(x), h(x))\}.$$

The manifold begins at the origin, thus both $h(0) = g(0) = 0$; moreover, the manifold must be tangent to E^c, so $Dh(0) = Dg(0) = 0$. Finally, W^c must be invariant, so that if $(x, y, z) \in W^c$, then so is $\varphi_t(x, y, z)$. This means that the vector field $(\dot{x}, \dot{y}, \dot{z})$ must be in the tangent space of W^c. To compute this we insist that the flow lies on W^c, so that $y(t) = g(x(t))$ and $z(t) = h(x(t))$. Consequently, the derivatives of these functions must also match:

$$\dot{y} = Dg(x)\dot{x}, \quad \dot{z} = Dh(x)\dot{x}.$$

Putting this into (5.34) gives a system of PDEs that can be used to determine g and h:

$$\begin{aligned}Sg(x) + G(x, g(x), h(x)) &= Dg(x)(Cx + F(x, g(x), h(x))),\\ Uh(x) + H(x, g(x), h(x)) &= Dh(x)(Cx + F(x, g(x), h(x))).\end{aligned} \tag{5.35}$$

These PDEs can be solved order by order for the power series of h and g—see the examples below.

The dynamics on the center manifold are given by the equation for x upon restricting y and z to the manifold:

$$\dot{x} = Cx + F(x, g(x), h(x)). \tag{5.36}$$

That this equation describes the local dynamics on W^c follows from a generalization of the Hartman–Grobman theorem of §4.8.

Theorem 5.23 (Nonhyperbolic Hartman–Grobman). *Suppose* (5.34) *is a C^1 vector field with fixed point at the origin, that all the eigenvalues of C have zero real part, that S is a contracting and U is an expanding hyperbolic matrix, and that $F, G, H = o(x, y, z)$. Then there is a neighborhood N of the origin such that $W^c_{loc} = \{(x, g(x), h(x)) : x \in E^c\} \cap N$ and the dynamics in N is topologically conjugate to the system*

$$\begin{aligned} \dot{x} &= Cx + F(x, g(x), h(x)), \\ \dot{y} &= Sy, \\ \dot{z} &= Uz. \end{aligned} \tag{5.37}$$

Thus, the topological type of a nonhyperbolic fixed point is determined by the flow on the center manifold.

We now give several examples of the formal solution of the PDEs (5.35) order by order in the power series for the functions g and h.

Example 5.24. A two-dimensional system with a single zero eigenvalue has the block diagonal form

$$J = \begin{pmatrix} 0 & 0 \\ 0 & \lambda \end{pmatrix}.$$

Here the center matrix is the 1×1 matrix $C = (0)$, and (taking $\lambda > 0$) the unstable matrix is $U = (\lambda)$. Therefore, the linear spaces are $E^c = \text{span}(1, 0)^T$ and $E^u = \text{span}(0, 1)^T$. For example, consider the C^∞ system

$$\begin{aligned} \dot{x} &= x^2 - z^2, \\ \dot{z} &= \lambda z + x^2. \end{aligned} \tag{5.38}$$

Following the general theory, we suppose that the local center manifold is $W^c_{loc}(0,0) = \{(x, h(x)) : x \in \mathbb{R}\}$, where $h(0) = Dh(0) = 0$. Thus, the power series for h has the form $h(x) = \alpha x^2 + \beta x^3 + \cdots$. Putting this into (5.35) gives

$$\lambda\left(\alpha x^2 + \beta x^3 + \cdots\right) + x^2 = \left(2\alpha x + 3\beta x^2 + \cdots\right)\left(x^2 - \left(\alpha x^2 + \beta x^3 + \cdots\right)^2\right).$$

The lowest degree terms in this equation are quadratic and require that $\lambda\alpha + 1 = 0$. This determines α. The cubic terms give the equation $\lambda\beta = 2\alpha$, which determines β. After some algebra we find that

$$h(x) = -\frac{x^2}{\lambda} - 2\frac{x^3}{\lambda^2} - 6\frac{x^4}{\lambda^3} - 22\frac{x^5}{\lambda^4} - 96\frac{x^6}{\lambda^5} + \cdots.$$

The resulting curve $z = h(x)$ is shown in Figure 5.15. This result can be inserted into the differential equation for x, (5.38), to give the center manifold dynamics

$$\dot{x} = x^2 - \frac{x^4}{\lambda^2} - 4\frac{x^5}{\lambda^3} - 16\frac{x^6}{\lambda^4} \cdots. \tag{5.39}$$

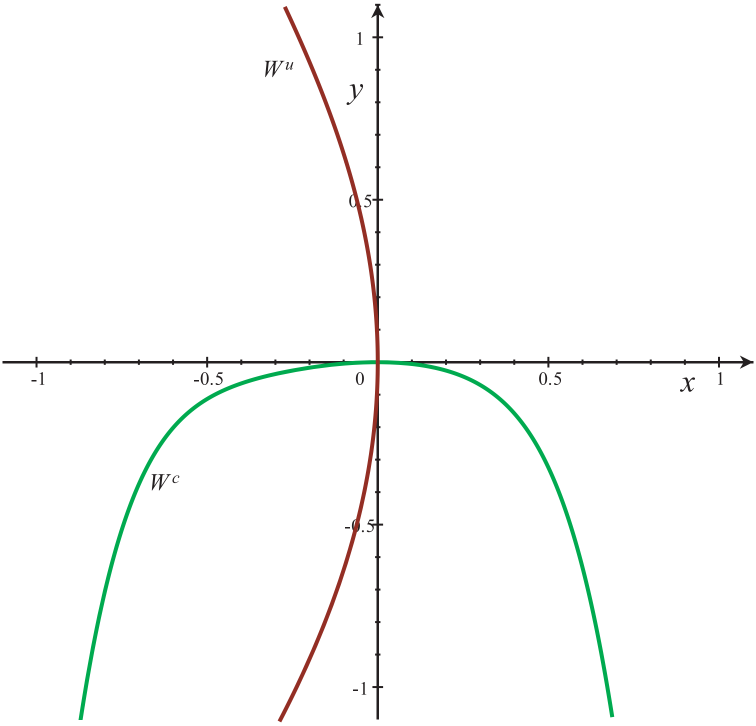

Figure 5.15. *Center and unstable manifolds for* (5.38) *through sixth order for* $\lambda = 2$.

This implies that $\dot{x} > 0$ when x is nonzero and small, which shows that on the center manifold the point $x = 0$ is "semistable"; see Figure 5.16.

The unstable manifold can be similarly found. If we let $x = g(z) = \alpha z^2 + \beta z^3 + \cdots$ and substitute this into the equation $\dot{x} = Dg(z)\dot{z}$, we obtain (after some algebra)

$$g(z) = -\frac{z^2}{2\lambda} + \frac{z^4}{16\lambda^3} + \frac{z^5}{20\lambda^4} - \frac{z^6}{96\lambda^5} + \cdots .$$

The curve $x = g(z)$ is shown in Figure 5.15.

According to Theorem 5.23, we have shown that (5.38) is conjugate to the system

$$\dot{x} = x^2 - \frac{x^4}{\lambda^2} + \cdots ,$$
$$\dot{z} = \lambda z.$$

If we compare the dynamics that we have found with a numerical solution of (5.38), see Figure 5.17, we see that the center and unstable manifolds prominently appear—note that the motion near the origin for decreasing t appears to rapidly compress along the unstable manifold (as e^{-t}) and then move more slowly along the center manifold toward the origin.

The system (5.38) has two additional fixed points, a saddle at $(\lambda, -\lambda)$ and a spiral sink at $(-\lambda, \lambda)$. The phase plane shows that the right branch of the center manifold appears to coincide with the stable manifold of the saddle. The spiral sink traps the bottom branch of $W^u(0)$. ∎

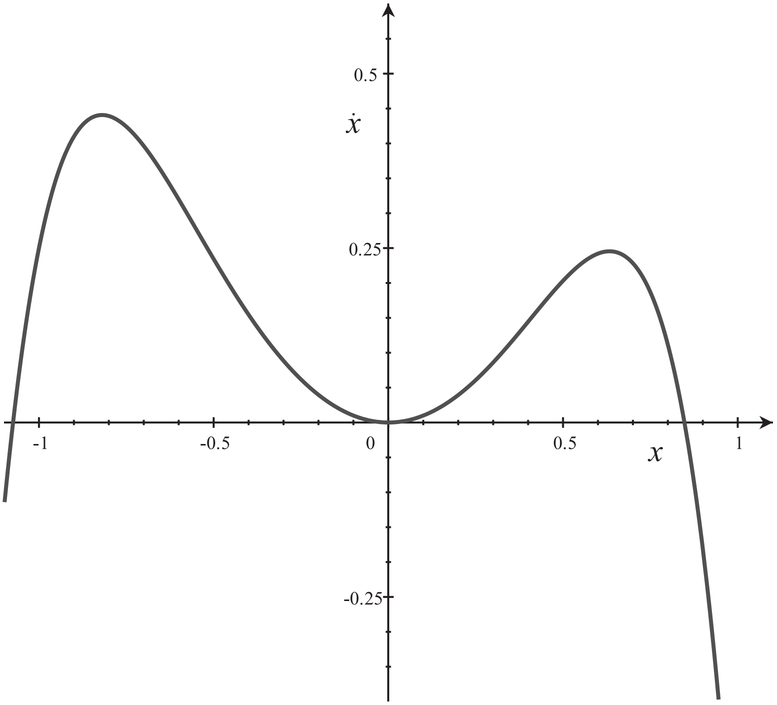

Figure 5.16. *The vector field* (5.39) *as a function of x on the local center manifold for* $\lambda = 2$.

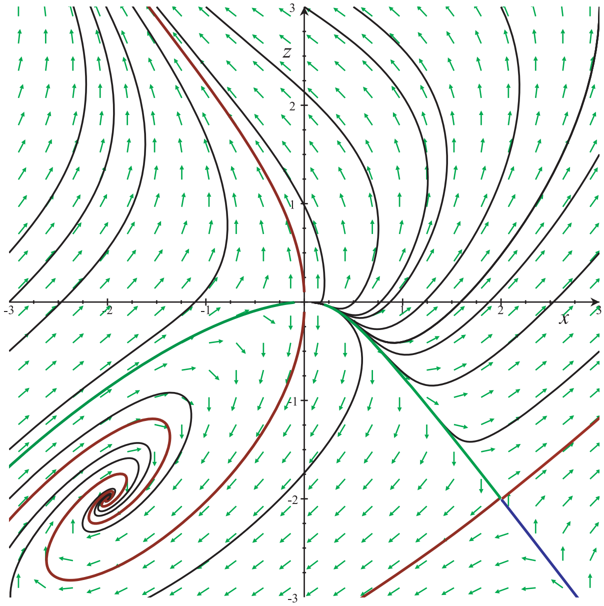

Figure 5.17. *Phase plane of* (5.38) *for* $\lambda = 2$.

Example 5.25. Consider the three-dimensional system

$$\begin{aligned} \dot{x}_1 &= -x_2 + x_1 y, \\ \dot{x}_2 &= x_1 + x_2 y, \\ \dot{y} &= -y - x_1^2 - x_2^2 + y^2, \end{aligned} \qquad Df(0) = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}. \tag{5.40}$$

Here, Df is already in the normal form, and we can immediately see that $E^c = \{(x_1, x_2, 0)\}$ and $E^s = \{(0, 0, y)\}$. Again, look for solutions that are tangent to the center space, so that $W^c = \{(x_1, x_2, g(x_1, x_2))\}$. As before, assume a power series for $g(x) = \alpha x_1^2 + \beta x_1 x_2 + \gamma x_2^2 + \cdots$. Requiring that $y = g(x)$ is an invariant manifold, (5.35), gives

$$\begin{aligned} \dot{y} =& Dg(x)\dot{x} = \frac{\partial g}{\partial x_1}\dot{x}_1 + \frac{\partial g}{\partial x_2}\dot{x}_2, \\ -\alpha x_1^2 - \beta x_1 x_2 - \gamma x_2^2 - x_1^2 - x_2^2 + \cdots =& (2\alpha x_1 + \beta x_2 + \cdots)(-x_2 + \cdots) \\ &+ (\beta x_1 + 2\gamma x_2 + \cdots)(x_1 + \cdots) \end{aligned}$$

to quadratic order. Collecting the terms in x_1^2, x_2^2, and $x_1 x_2$ gives three equations for the three unknowns α, β, and γ. These can be written as a single linear system:

$$\begin{pmatrix} -1 & -1 & 0 \\ 0 & 1 & -1 \\ 2 & -1 & -2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

This matrix is guaranteed to be nonsingular by the center manifold theorem, and indeed we find that is the case. The solution is $\alpha = \gamma = -1$ and $\beta = 0$, so $y = -x_1^2 - x_2^2 + \cdots$. Substituting this back into the original equations for (x_1, x_2) gives the dynamics on the center manifold:

$$\begin{aligned} \dot{x}_1 &= -x_2 - x_1^3 - x_1 x_2^2, \\ \dot{x}_2 &= x_1 - x_2 x_1^2 - x_2^3, \end{aligned} \tag{5.41}$$

up to terms of cubic order. The dynamics of (5.41) is nontrivial, and to study it we must use some additional tricks—we will develop these in the next chapter (see §6.3). We will find that (5.41) has the dynamics of a spiral focus. This implies, according to Theorem 5.23, that the origin of (5.40) is asymptotically stable. ■

5.7 ▪ Exercises

1. Find all the homoclinic and heteroclinic orbits for the Hamiltonian

   $$H(x,y) = \tfrac{1}{2}(y^2 + x^2) - x^4.$$

   What are the stable W^s and unstable W^u sets for each of the three equilibria?

2. Consider the system (5.6) with Hamiltonian (5.3).

   (a) Find the equilibria. You should verify that $p = (x^*, 1/2 a x^{*4})$ is an equilibrium when $x^* = 0$ or is a root of the polynomial $q(x) = -4 + 12ax + a^2 x^6$. Show that when $a \neq 0$, q has exactly two real roots and hence that there are three equilibria.

   (b) Show that the origin is saddle. Find its eigenvalues and eigenvectors.

   (c) Set $a = 1$, and find the new equilibria numerically. Show that that one is a stable focus and the other an unstable focus.

   (d) Investigate, using phase plane software, the dynamics of this system. What are the stable and unstable sets of each equilibrium?

3. Like the Lorenz model (1.33), the Busse–Heikes model describes three spatial modes in a convecting fluid, but in this case the fluid is rotating (Toral, San Miguel, and Gallego 2000). In one limit the model becomes

$$\begin{aligned} \dot{x} &= x(1-x-(1+\delta)y-(1-\delta)z), \\ \dot{y} &= y(1-y-(1+\delta)z-(1-\delta)x), \\ \dot{z} &= z(1-z-(1+\delta)x-(1-\delta)y), \end{aligned} \tag{5.42}$$

where $\delta > 0$, and (x,y,z) represent nonnegative mode amplitudes.

   (a) Find all the equilibria and characterize their stability types as a function of δ. (*Hint*: There are eight equilibria in $\mathbb{R}^3$: the origin, three solutions with one nonzero amplitude, three solutions with two nonzero amplitudes, and one with all three nonzero.)

   (b) Show that the quantity $R = x+y+z$ obeys a simple self-contained equation and that if $R(0) > 0$, then $R(t) \to 1$ as $t \to \infty$.

   (c) Assume that $R = 1$ and reduce (5.42) to a set of two equations for (x,y). Show that these equations are Hamiltonian with $H = \delta xy(1-x-y)$.

   (d) Give a complete discussion of the dynamics of this model in the positive octant.

4. Using the integral (5.13), find the unique bounded solution to the forced system

$$\begin{aligned} \dot{x} &= -x, \\ \dot{y} &= y + \sin(t) \end{aligned}$$

for an initial condition $\sigma = (x_o, 0)^T \in E^s$.

5. Show that any bounded fixed point $x \in C^0(\mathbb{R}^+, U)$ of the operator T defined by (5.17) is a C^1 solution of the differential equation (5.10). (*Hint*: Differentiate $x = T(x)$ with respect to t, remembering to differentiate with respect to all the places that t enters on the right-hand side.)

6. Show that if $L : X \to X$ is a linear operator on a Banach space, and $||L|| \le c < 1$, then the operator $I - L$ is invertible. (*Hint*: Consider the formal series expansion $(I-L)^{-1} = \sum_{k=0}^{\infty} L^k$.)

7. Here, you will show that the stable manifold theorem implies an equivalent unstable manifold theorem.

   (a) First, let $\hat{x}(\tau) = x(-\tau)$ in (5.10) and obtain the ODE for $\hat{x}$. This will give an equation similar to (5.10) but with $A \to -A$. Now, show that stable manifold theorem for the new equation implies the existence of a Lipschitz graph W^u over E^u.

   (b) Transform back to $t = -\tau$, and obtain the explicit operator T equivalent to (5.17) for the unstable manifold. Take care to keep track of all the minus signs!

8. Consider the system on $\mathbb{R}^2$ given by

$$\begin{aligned} \dot{x} &= -x + xy, \\ \dot{y} &= 2y + x^2. \end{aligned}$$

(a) Find E^s and E^u for the fixed point $(0,0)$.

(b) Construct successive approximations $(x_i(t), y_i(t))$, $i = 1,2$, to the stable manifold $W^s(0,0)$ by applying the operator T, (5.17), to the initial guess $(x_o(t), y_o(t)) = (0,0)$.

(c) Compare the approximations in (b) with a power series expansions for the stable and unstable manifolds using the techniques of §5.6.

(d) Using your favorite software, plot the functions you constructed and some numerical solutions of the differential equations. Compare the manifolds that you compute with the solutions.

9. Consider the system

$$\begin{aligned}\dot{x} &= x^3 - 2xy,\\ \dot{y} &= -y + x^2.\end{aligned}$$

(a) Find the first few terms in the power series expansion for the stable and center manifolds of the origin.

(b) Study the reduced dynamics on the center manifold. Show that $x(t) \sim t^{-1/2}$ as $t \to \infty$. Classify the equilibrium.

(c) Compare your analytical expression with numerical orbits generated by your favorite software package.

10. The three-dimensional system

$$\begin{aligned}\dot{x} &= y + 2z + (x+z)^2 + xy - y^2,\\ \dot{y} &= (x+z)^2,\\ \dot{z} &= -2z - (x+z)^2 + y^2\end{aligned} \tag{5.43}$$

has a nonhyperbolic equilibrium at the origin.

(a) Find a linear transformation to write (5.43) in the form (5.34).

(b) Find the quadratic approximation for $W^c(0,0,0)$.

(c) Obtain the reduced dynamics (5.36) on W^c and use your favorite software package to study it. Is the origin stable or unstable?

11. Consider your adopted system of quadratic differential equations (recall §1.7 and Exercise 1.10) for the chaotic values of the reduced parameters. Use the techniques of this chapter to study the stable, unstable, and center manifolds of one of the equilibria.

Chapter 6

The Phase Plane

> *There is plenty in the subject to interest a pure mathematician, although perhaps interesting problems of moderate difficulty are getting scarce... non-linear phenomena are genuinely complicated and no easily applicable general theory can be expected.* (Mary Lucy Cartwright 1952)

The analysis of Chapter 4 allows us to obtain a picture of the dynamical behavior of a flow on a patchwork quilt of local phase portraits near equilibria or periodic orbits. This local, linearized analysis is relevant only for hyperbolic orbits: what can one do for the nonhyperbolic case? In this chapter we will study nonhyperbolic equilibria as well as methods to obtain global phase portraits. The methods will be specific to two dimensions, as many of the tools that we will use require that orbits, being one-dimensional curves, can separate regions in a two-dimensional space.

6.1 ▪ Nonhyperbolic Equilibria in the Plane

A purely topological classification of the nonhyperbolic equilibria for flows in $\mathbb{R}$ was easily obtained in §4.5. Here we attempt to accomplish the same task for nonhyperbolic equilibria in the plane. As introduced in §1.5, a planar system for $z=(x,y)$ has the form

$$\begin{aligned} \dot{x} &= P(x,y), \\ \dot{y} &= Q(x,y). \end{aligned} \tag{6.1}$$

If we choose a particular equilibrium, z^*, to study, the coordinates can always be shifted so that z^* is at the origin. Therefore, whenever there is an equilibrium it can be assumed without loss of generality that

$$P(0,0) = Q(0,0) = 0. \tag{6.2}$$

The first step in the classification of an equilibrium is the study of the linearization of the ordinary differential equation (ODE). As we learned in Chapter 2, the linear case, $\dot{z} = Az$, is classified by the eigenvalues, λ_i, of A. When P and Q are C^1, there are three

▷ *hyperbolic cases* (recall §2.2):

- *node*: λ_1 and λ_2 are real, nonzero, and have the same sign;

- *saddle*: λ_1 and λ_2 are real, nonzero, and have opposite signs;
- *focus*: $\lambda_1 = \bar{\lambda}_2$, and $\mathrm{Re}(\lambda_1) \neq 0$.

The node and focus cases can be either stable or unstable depending upon the sign of $\mathrm{Re}(\lambda)$. A node with equal eigenvalues is called a *proper* node if it has two eigenvectors (geometric multiplicity is two) and an *improper* node if it has only one eigenvector (geometric multiplicity is one). The Hartman–Grobman theorem, Theorem 4.36, implies that the linear results are sufficient to classify the dynamics of (6.1) in a neighborhood of the origin when A is hyperbolic.

In §2.2 we noted there are four additional

▷ *nonhyperbolic cases*:

- *singly degenerate equilibrium*: $\lambda_1 = 0$, but $\lambda_2 \neq 0$. The linear system has a line of equilibria.
- *doubly degenerate equilibrium*: $\lambda_1 = \lambda_2 = 0$, geometric multiplicity one. In this case, A is equivalent to the Jordan form

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

  The linear system has a line of equilibria ($y = 0$).
- *doubly degenerate equilibrium*: $\lambda_1 = \lambda_2 = 0$, geometric multiplicity two. In this case, $A = 0$ and the entire plane consists of equilibria.
- *center*: $\lambda_1 = \bar{\lambda}_2 = i\beta$ are pure imaginary. The linear orbits are ellipses.

A nonlinear system with a singly degenerate equilibrium is amenable to the center manifold analysis of §5.6. In this case the center manifold is one-dimensional, and the restriction of the dynamics to W^c is easily understood by using graphical analysis for a one-dimensional ODE; recall §1.1. When the dimension of E^c is two, as in the last three cases, additional methods must be used to analyze the dynamics.

6.2 ▪ Two Zero Eigenvalues and Nonhyperbolic Nodes

There are two Jordan forms for the linearization about an equilibrium z^* for the doubly degenerate case, when $\lambda_1 = \lambda_2 = 0$; here we will assume $A = Df(z^*) = 0$. Since the linearization is identically zero, this system is particularly hard to treat by our previous methods. The case of a nontrivial Jordan form will be treated in §8.10.

Example 6.1. The system

$$\begin{aligned} \dot{x} &= y^2 - x^2, \\ \dot{y} &= -2xy \end{aligned} \tag{6.3}$$

has only one equilibrium, at the origin. Since $Df(0) = 0$, both eigenvalues are 0, and every point in the plane is an equilibrium of the linearized system: linear stability says nothing about stability of the full system. It is not possible to find a Lyapunov function near the origin; for example, if we assume that L is quadratic, then dL/dt would be a homogeneous cubic polynomial, and thus cannot have one sign. Numerical integration of the flow, shown in Figure 6.1, indicates that the origin is unstable. ∎

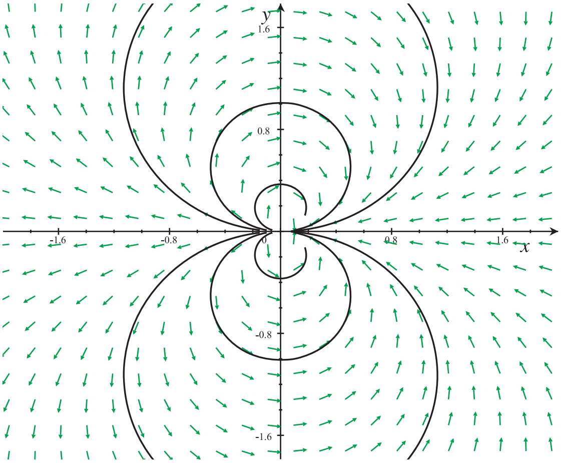

Figure 6.1. *Phase portrait of the example* (6.3).

The main tool for studying the behavior near such an equilibrium is a simple transformation to polar coordinates:

$$\begin{aligned} x &= r\cos\theta, \qquad & r^2 &= x^2+y^2, \\ y &= r\sin\theta, \qquad & \theta &= \arctan(y/x). \end{aligned} \tag{6.4}$$

The time derivatives of (r,θ) are found using the chain rule:

$$\begin{aligned} &\frac{d}{dt}r^2 = 2(x\dot{x}+y\dot{y}) \Rightarrow \frac{d}{dt}r = \frac{1}{r}(x\dot{x}+y\dot{y}), \\ &\frac{d}{dt}\theta = \frac{d}{dt}\arctan\left(\frac{y}{x}\right) = \frac{1}{1+y^2/x^2}\left(\frac{\dot{y}}{x}-\frac{y\dot{x}}{x^2}\right) = \frac{1}{r^2}(x\dot{y}-y\dot{x}). \end{aligned} \tag{6.5}$$

Inserting (6.4) and (6.5) into the system (6.1) and eliminating x and y in favor of r and θ gives

$$\begin{aligned} \frac{d}{dt}r &= P(r\cos\theta, r\sin\theta)\cos\theta + Q(r\cos\theta, r\sin\theta)\sin\theta, \\ \frac{d}{dt}\theta &= \frac{1}{r}[Q(r\cos\theta, r\sin\theta)\cos\theta - P(r\cos\theta, r\sin\theta)\sin\theta]. \end{aligned} \tag{6.6}$$

Dividing the $\dot{r}$ equation by the $\dot{\theta}$ equation gives an equation for the phase curves (recall (1.22):

$$\frac{dr}{d\theta} = r\frac{P(r\cos\theta, r\sin\theta)\cos\theta + Q(r\cos\theta, r\sin\theta)\sin\theta}{Q(r\cos\theta, r\sin\theta)\cos\theta - P(r\cos\theta, r\sin\theta)\sin\theta}. \tag{6.7}$$

As a simple case, suppose that both P and Q are homogeneous nth degree polynomials in their arguments, so that $P(ax,ay) = a^n P(x,y)$ and similarly for Q. In this case $P(r\cos\theta, r\sin\theta) = r^n P(\cos\theta, \sin\theta)$, and separation of variables in (6.7) yields

$$\frac{dr}{r} = \frac{P(\cos\theta,\sin\theta)\cos\theta + Q(\cos\theta,\sin\theta)\sin\theta}{Q(\cos\theta,\sin\theta)\cos\theta - P(\cos\theta,\sin\theta)\sin\theta}d\theta = g(\theta)d\theta,$$

which gives

$$\ln r = \ln r_o + \int_{\theta_o}^{\theta} g(\varphi) d\varphi. \tag{6.8}$$

If the equilibrium were asymptotically stable, then for any $r_o, r(t; r_o) \to 0$ as $t \to \infty$. For this to happen, the integral of the function g must go to minus infinity. Note that $g(\theta + 2\pi) = g(\theta)$, so an important quantity to consider is

$$G = \int_0^{2\pi} g(\theta) d\theta. \tag{6.9}$$

Since this integral may or may not exist, there are three possibilities:

▷ *topological center*: If $G = 0$, then $r \not\to 0$ because the integral (6.8) for any θ is finite. Moreover, in this case (6.8) implies $r(\theta_o + 2\pi) = r(\theta_o)$, so the curve $r(\theta)$ is a closed loop.

▷ *nonhyperbolic focus*: If G exists and is nonzero, then the only way the integral (6.8) can diverge is for $\theta \to \pm\infty$. In this case, (6.8) implies

$$r(2\pi) = r(0)e^G.$$

Therefore, r is multiplied by a factor of e^G each time the angle increases by 2π. If $G > 0$, the curve spirals outward; otherwise it spirals inward. Moreover, the curve must be an infinite spiral as $r \to 0$, since each time θ changes by $-2\pi \operatorname{sgn}(G)$ the radius decreases by the fixed factor.

▷ *nonhyperbolic node*: If G does not exist, then g must have a nonintegrable singularity at some point θ_c where its denominator vanishes:

$$D(\theta_c) = Q(\cos\theta_c, \sin\theta_c)\cos\theta_c - P(\cos\theta_c, \sin\theta_c)\sin\theta_c = 0.$$

In this case the integral in (6.8) is unbounded as $\theta \to \theta_c$. This angle defines an asymptotic direction of approach to the origin as $t \to \pm\infty$.

Example 6.2. A homogeneous cubic example is provided by $P(x,y) = -x^2y - y^3$ and $Q(x,y) = x^3 + xy^2$. In polar coordinates, (6.6), the system becomes

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

so that $g(\theta) \equiv 0$. Consequently, the origin is a topological center; indeed, every orbit apart from the origin lies on a periodic orbit with period $T = 2\pi/r^2$. ■

Example 6.3. When $P(x,y) = -(x^2 + y^2)(x + y)$, and $Q(x,y) = (x^2 + y^2)(x - y)$, we obtain

$$\dot{r} = -r^3,$$
$$\dot{\theta} = r^2.$$

This shows that $g(\theta) = -1$, and the origin is a nonhyperbolic focus. In this case, every trajectory spirals around the origin infinitely many times as $t \to \infty$ and $r \to 0$. ■

For the nonhyperbolic node, note that if θ_c is one root of the denominator D, then $\theta_c + \pi$ is also a root, since $\cos(\theta_c + \pi) = -\cos\theta_c$ and $\sin(\theta_c + \pi) = -\sin\theta_c$, and since

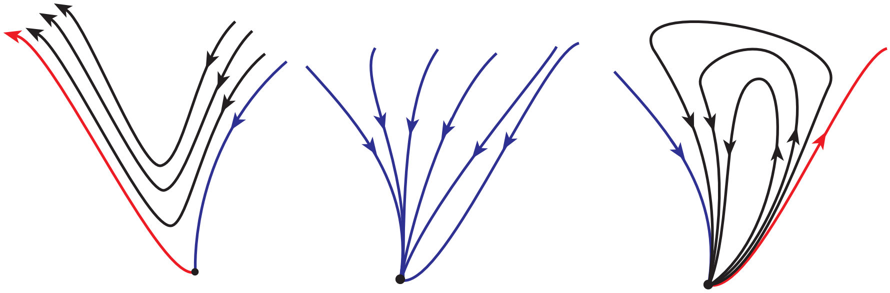

Figure 6.2. *Hyperbolic, parabolic, and elliptic sectors when $\dot{\theta} > 0$. A second parabolic case, not shown, occurs if both directions are diverging.*

we assumed that P and Q are homogeneous, $Q(-\cos\theta_c - \sin\theta_c) = (-1)^n Q(\cos\theta_c, \sin\theta_c)$. So, if there is one asymptotic direction of approach to the origin, then there are two such directions on a line through the origin with slope $\tan\theta_c$. As $r \to 0$, the rate of change of r limits to

$$\frac{dr}{dt} \to r^n \left(P(\cos\theta_c, \sin\theta_c)\cos\theta_c + Q(\cos\theta_c, \sin\theta_c)\sin\theta_c\right) = r^n \frac{P_c}{\cos\theta_c}, \qquad (6.10)$$

where $D(\theta_c) = 0$ has been used to eliminate Q. Consequently, r asymptotically grows or decreases depending upon the sign in (6.10), implying that the ray $\theta = \theta_c$ is either an asymptotically unstable or a stable direction, respectively. Note that $\operatorname{sgn}(dr/dt)$ for $\theta_c + \pi$ is $(-1)^{n+1}$ times that for θ_c. Hence, when P and Q have even degree in r, one sign gives approach and the other divergence, but they have the same behavior when the degree is odd.

Example 6.4. Applying the polar transformation to (6.3) gives

$$\begin{aligned} \dot{r} &= -r^2 \cos\theta, \\ \dot{\theta} &= -r \sin\theta. \end{aligned}$$

This implies that $g(\theta) = \cot\theta$, which is singular at $\theta = 0$ and π. Therefore, every trajectory that approaches the origin must do so along the x-axis and the origin is a nonhyperbolic node. Note that $\dot{r} < 0$ at $\theta = 0$ and $\dot{r} > 0$ at $\theta = \pi$; thus the positive x-axis is a stable direction while the negative x-axis is an unstable direction. Of course, this can also easily be seen by restricting the system to the invariant line $y = 0$, where (6.3) becomes $\dot{x} = -x^2$, showing that the origin is semistable. ∎

When an equilibrium is a node, there are one or more directions corresponding to approaching or diverging orbits. These orbits divide a small disk enclosing z^* into sectors, bounded by the asymptotic curves. The sectors can be one of three types: *elliptic*, *hyperbolic*, or *parabolic*, as sketched in Figure 6.2.[34]

A parabolic sector is bounded by two curves of the same asymptotic type—both are approaching or both are diverging. Hyperbolic and elliptic sectors are bounded by one diverging and one approaching curve. The hyperbolic and elliptic cases are distinguished by the sign of $\dot{\theta}$; for example, in Figure 6.2, $\dot{\theta} > 0$ as $r \to 0$, so that

[34] This use of the word *hyperbolic* is geometrical, as opposed to our characterization of equilibria as hyperbolic by their eigenvalues.

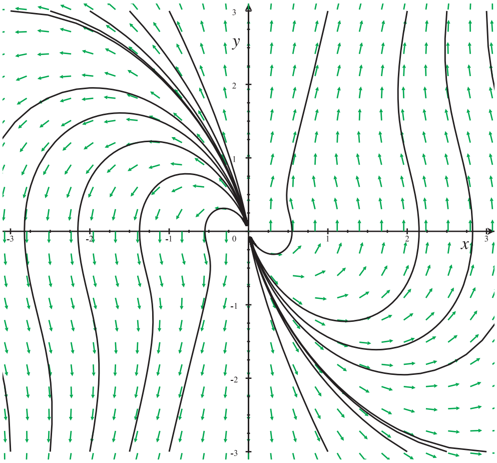

Figure 6.3. *Phase portrait of* (6.11).

if the converging direction is counterclockwise from the diverging one, the sector is elliptic; otherwise it is the hyperbolic. The local dynamics in the hyperbolic case is unbounded: every orbit in the sector that is not on the approaching direction eventually leaves any disk enclosing the equilibrium. For the elliptic case each orbit eventually returns.

A hyperbolic saddle provides the standard example of an equilibrium with hyperbolic sectors—each sector bounded by the eigenvectors is hyperbolic. Nonhyperbolic equilibria can have a combination of sectors depending on the character of (6.6).

Example 6.5. The system

$$\begin{aligned} \dot{x} &= y^2x - x^2y, \\ \dot{y} &= x^3 + y^3 \end{aligned} \tag{6.11}$$

is equivalent to the polar equations

$$\begin{aligned} \dot{r} &= r^3 \sin^2\theta, \\ \dot{\theta} &= r^2 \cos^2\theta. \end{aligned}$$

Therefore, $g(\theta) = \tan^2\theta$, which has singularities at $\theta = \pi/2$ and $3\pi/2$. In both cases, $\dot{r} > 0$ as $r \to 0$. As a consequence the sectors defined by $\theta \in [-\pi/2, \pi/2]$ and $[\pi/2, 3\pi/2]$ are both parabolic. For this case, (6.8) can be solved explicitly for $r(\theta)$ to obtain

$$r(\theta) = ce^{\tan(\theta)-\theta}$$

so that $r(\theta) \to 0$ as $\theta \to \pi/2|_+$ and as $\theta \to 3\pi/2|_+$; here the limits are only one-sided. This shows what is typical in a parabolic sector: all the orbits in the interior of the sector limit on only one of the sector boundaries as $r \to 0$. The phase space is shown in Figure 6.3. ∎

Example 6.6. As $r \to 0$, the Vinograd example (4.17) is equivalent (see Exercise 2) to the system

$$\begin{aligned} \dot{x} &= x^2(y - x), \\ \dot{y} &= y^2(y - 2x). \end{aligned} \tag{6.12}$$

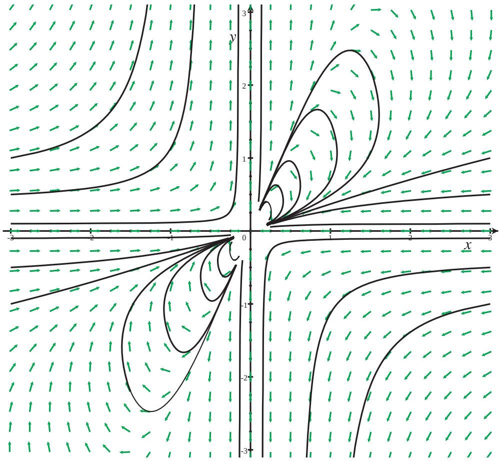

Figure 6.4. *Phase space of* (6.12) *showing two hyperbolic sectors, four parabolic sectors, and two elliptic sectors.*

Upon conversion to polar coordinates, (6.12) becomes

$$\dot{r} = \frac{r^3}{4}\left[(3\sin(2\theta) - 4)\cos(2\theta) - \sin(2\theta)\right],$$
$$\dot{\theta} = \frac{r^2}{4}\sin(2\theta)\left[2 - 3\sin(2\theta)\right].$$

Accordingly, $\dot{\theta} = 0$ for $\theta = n\pi/2$, $\theta^* = {}^1\!/\!_2 \sin^{-1}(2/3) \approx 20.9°$ and $\theta = \pi/2 - \theta^* \approx 69.1°$. Along the x-axis and at θ^*, $\dot{r} < 0$, while along the y-axis and at $\pi/2 - \theta^*$, $\dot{r} > 0$. The sector $[0, \theta^*]$ is parabolic, since both of its asymptotes are converging. The sector $[\theta^*, \pi/2 - \theta^*]$ is elliptic since $\dot{\theta} < 0$ in the sector. The sector $[\pi/2 - \theta^*, \pi/2]$ is parabolic, and finally, $[\pi/2, \pi]$ is hyperbolic since in this sector $\dot{\theta} < 0$. The analysis of the remaining sectors is similar. This gives the configuration shown in Figure 6.4. ∎

The analysis above also applies to a more general case: suppose P and Q are not homogeneous but that they are given by power series that have the lowest degree terms of the same order, say, the nth order:

$$P(r\cos\theta, r\sin\theta) = r^n P_n(\cos\theta, \sin\theta) + O(r^{n+1}),$$
$$Q(r\cos\theta, r\sin\theta) = r^n Q_n(\cos\theta, \sin\theta) + O(r^{n+1}).$$

In this case, as $r \to 0$, these terms dominate the higher-order terms, and the vector field can be approximated by its lowest-order terms. We are most familiar with this when the lowest-order terms are linear. The analysis can also be applied to the case in which the lowest-order terms in P and Q have different degrees (see Exercise 1).

6.3 ▪ Imaginary Eigenvalues: Topological Centers

We now consider a system (6.1) with an equilibrium that has a linearization with imaginary eigenvalues. Without loss of generality, we can change coordinates so that the Jacobian at the equilibrium can be written in the normal form $Df(0) = \left(\begin{smallmatrix} 0 & -\omega \\ \omega & 0 \end{smallmatrix}\right)$, and

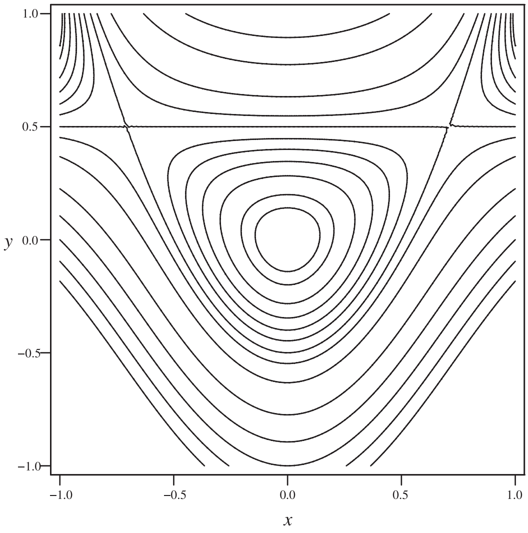

Figure 6.5. *Contours of the Hamiltonian function* (6.14) *for* $V(x,y) = -x^2y$. *Orbits follow the contours since* H *is constant.*

the ODEs become

$$\begin{aligned} \dot{x} &= -\omega y + p(x,y), \\ \dot{y} &= \omega x + q(x,y), \end{aligned} \qquad p,q = o(x,y). \tag{6.13}$$

We will also assume that there is a neighborhood of the origin for which p and q are Lipschitz, so that Theorem 3.19 (existence and uniqueness) applies to (6.13).

Example 6.7. Linear centers often occur for Hamiltonian systems. For example, suppose $V(x,y)$ is a smooth function, and consider the equations

$$\begin{aligned} \dot{x} &= y + V_y, \\ \dot{y} &= -x - V_x. \end{aligned}$$

This system is of the form (4.27) with Hamiltonian

$$H(x,y) = \tfrac{1}{2}\left(x^2 + y^2\right) + V(x,y). \tag{6.14}$$

Moreover, as shown by (4.28), the energy is invariant: $dH/dt = 0$. If V is cubic or higher order in x and y, then the origin has eigenvalues $\lambda = \pm i$: it is a center for the linear system. The Hartman–Grobman theorem says nothing about the behavior of the orbits near the origin when the nonlinear terms are included, since this equilibrium is not hyperbolic. Nevertheless, when V is cubic, the curves of constant H, the energy surfaces, are closed in the neighborhood of the origin and, since H is a weak Lyapunov function (recall §4.6), the origin is a topological center.

An example with a homogeneous cubic potential is shown in Figure 6.5. For this case, in addition to the center at the origin, there are two saddle equilibria at the points $(\pm 1/\sqrt{2}, 1/2)$. ■

Although, as this example shows, a linear center can sometimes be a topological center, it is easy to find examples in which this is not the case.

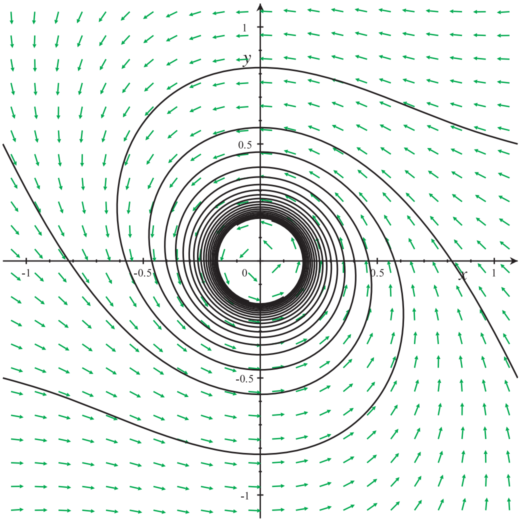

Figure 6.6. *A stable nonhyperbolic focus;* (6.15) *with* $g = -x^2$.

Example 6.8. Suppose that $g : \mathbb{R}^2 \to \mathbb{R}$ is a continuous function, and consider the system

$$\dot{x} = -\omega y + x g(x, y), \quad \dot{y} = \omega x + y g(x, y). \tag{6.15}$$

Using the transformation (6.5) to polar coordinates gives

$$\dot{r} = r g(r\cos\theta, r\sin\theta), \quad \dot{\theta} = \omega. \tag{6.16}$$

So that the origin is a linear center, we must assume that $g = O(r)$. However, this does not ensure that the orbits near the origin are topological circles. There are two simple cases: if $g > 0$ for r sufficiently small, then $\dot{r} > 0$ and the origin is unstable. If, however, $g < 0$ near the origin, then it is asymptotically stable. For example, when $g = -x^2$ the radial equation reduces to

$$\dot{r} = -r^3 \cos^2\theta.$$

For this case, the origin is stable, since $\dot{r} \le 0$, and $\dot{r} = 0$ only momentarily when θ passes through $\pi/2$ or $3\pi/2$. Since $\theta(t)$ is known, the radial part is separable and can even be solved explicitly. Choosing $\theta_o = 0$, we obtain

$$\int \frac{dr}{r^3} = -\int \cos^2 \omega t \, dt \quad \Rightarrow \quad \frac{1}{r^2} = \frac{1}{r_o^2} + t + \frac{1}{2\omega}\sin(2\omega t).$$

Thus $r \to 0$ as $t \to \infty$ for any trajectory. This case is shown in Figure 6.6. ■

As we will show below, (6.13) has precisely three possible scenarios near the origin:

- ▷ *Topological center*: there is a $\delta > 0$ such that every trajectory in $B_\delta(0) \setminus \{0\}$ is a closed loop enclosing the origin.

- ▷ *Nonhyperbolic focus*: there is a $\delta > 0$ such that every trajectory in the ball $B_\delta(0)$ approaches 0 and $|\theta(t)| \to \infty$ as either $t \to +\infty$ or $-\infty$.

▷ *Center-focus*: there is an infinite sequence of nested limit cycles, γ_n, such that $\gamma_n \to 0$ as $n \to \infty$ and every trajectory between two limit cycles spirals toward one limit cycle or the other as $t \to \pm\infty$.

Just as for the double-zero eigenvalue, the tool for studying the possible behaviors of (6.13) is the transformation to polar coordinates:

$$\begin{aligned}\frac{d}{dt} r &= p(r\cos\theta, r\sin\theta)\cos\theta + q(r\cos\theta, r\sin\theta)\sin\theta,\\ \frac{d}{dt}\theta &= \omega + \frac{1}{r}\left[q(r\cos\theta, r\sin\theta)\cos\theta - p(r\cos\theta, r\sin\theta)\sin\theta\right].\end{aligned} \tag{6.17}$$

By assumption, $p, q = o(r)$; consequently, for any $\varepsilon > 0$ there is a δ such that if $r < \delta$, then $|p|, |q| < \varepsilon r$ (recall §4.4) so that

$$\left|\dot\theta - \omega\right| < \varepsilon, \quad r \in B_\delta(0). \tag{6.18}$$

For example, choosing $\varepsilon = \omega/2$, we then have $\dot\theta > \omega/2$. Therefore, if the trajectory remains in $B_\delta(0)$ it must encircle the origin infinitely many times. So if a trajectory approaches $r = 0$ as $t \to \pm\infty$, it must do so on an infinite spiral. In this case the equilibrium is a nonhyperbolic focus.

Lemma 6.9. *If the system* (6.13) *has one trajectory that approaches the origin as* $t \to \infty$ *or as* $t \to -\infty$*, then the origin is a nonhyperbolic focus.*

Proof. We need to show that there is a neighborhood for which all trajectories approach the origin. Assume first that $\varphi_t(r,\theta)$ approaches the origin as $t \to \infty$. Thus there is a δ and a time T such that $\varphi_t(r,\theta) \in B_\delta(0)$ for all $t > T$ and (6.18) holds for some $\varepsilon < \omega$. This implies that $\theta(t)$ is unbounded. Let $t_k > T$ be the sequence of times such that $\theta(t_k) = 2\pi k$, and let $r_k = r(t_k)$; see Figure 6.7. By assumption, $r_k \to 0$ as $k \to \infty$. Uniqueness implies that this sequence is monotone decreasing: the segment of the trajectory between t_{k+1} and t_{k+2} cannot cross the segment between t_k and t_{k+1}. The same argument implies that the orbit of any initial point $(s, 0)$ with $r_1 < s < r_o$ cannot cross the original orbit and has a monotone and unbounded angle. Therefore all these forward orbits must also be infinite spirals that approach the origin. Consequently, the forward orbits of all points in the ball with radius $\min(r(t) : t_o \le t \le t_1)$ also approach the origin. The argument for $t \to -\infty$ is similar. □

Example 6.10. The system (5.41),

$$\begin{aligned}\dot x &= -y - x^3 - xy^2,\\ \dot y &= x - yx^2 - y^3,\end{aligned}$$

describes the dynamics on the center manifold of a three-dimensional ODE. This system is of the form (6.15) and the transformation to polar coordinates, (6.17), yields

$$\dot r = -r^3, \quad \dot\theta = 1.$$

Hence, the origin is a stable, nonhyperbolic focus. Putting this together with the stable dynamics in the third dimension of (5.40) implies that the origin is stable. ■

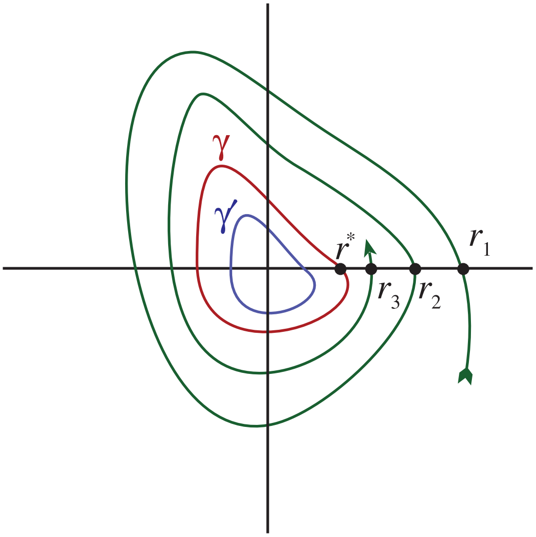

Figure 6.7. *Trajectory approaching a limit cycle.*

Example 6.11. Consider the system

$$\begin{aligned} \dot{x} &= -\omega y + xy^2 + x^2 y + y^3 \\ \dot{y} &= \omega x + y^3 - x^3 - xy^2 \end{aligned} \qquad \Rightarrow \qquad \begin{aligned} \dot{r} &= r^3 \sin^2 \theta \\ \dot{\theta} &= \omega - r^2. \end{aligned} \tag{6.19}$$

When $r < \delta = \sqrt{|\omega|}$, the angle is monotonically growing with time. In this case, $\dot{r} \geq 0$, and it is zero instantaneously only when $\theta = 0$ or π. Since θ is monotonically changing, this implies that $r(t)$ grows without bound in positive time and decreases, limiting to zero as $t \to -\infty$. Thus the origin is an unstable, nonhyperbolic focus. The phase portrait is shown in Figure 6.8. When $\omega > 0$, this system has two other equilibria at $(x, y) = (\pm\sqrt{\omega}, 0)$; see Exercise 7. ■

We now argue that the origin of (6.13) is a topological center, a nonhyperbolic focus, or a center-focus. The examples above have shown that the center and focus cases do occur, and Lemma 6.9 shows that if there is any trajectory that limits on the origin, then it is a focus. Now suppose that there are trajectories that remain bounded but do not approach the origin in either direction of t. We will argue that a bounded trajectory must have limit points, and the orbit of the limit points must be periodic.

Lemma 6.12. *Consider the system* (6.17) *and assume that p and q are continuous. Suppose there is a trajectory whose forward orbit remains in a neighborhood of the origin where* $\operatorname{sgn}(\dot{\theta}) = \operatorname{sgn}(\omega)$ *but does not limit on the origin. Then either the trajectory is periodic or its omega-limit set is a periodic orbit.*

Proof. As before let t_k be the sequence of times such that $\theta(t_k) = 2\pi k$, and let $r_k = r(t_k)$; see Figure 6.7. If $r_{k+1} = r_k$, then uniqueness implies that the trajectory is periodic, i.e., that $r_k = r^*$ for all k. Alternatively, either $r_{k+1} > r_k$ or $r_{k+1} < r_k$. In the first case, uniqueness implies that the segment of the trajectory between t_{k+1} and t_{k+2} cannot cross the segment between t_k and t_{k+1}. Therefore, $r_{k+2} > r_{k+1}$ as well, and the sequence is monotonically growing. Similarly, in the second case r_k is monotonically decreasing. Any monotone bounded sequence has a limit, $r_k \to r^*$. We claim that

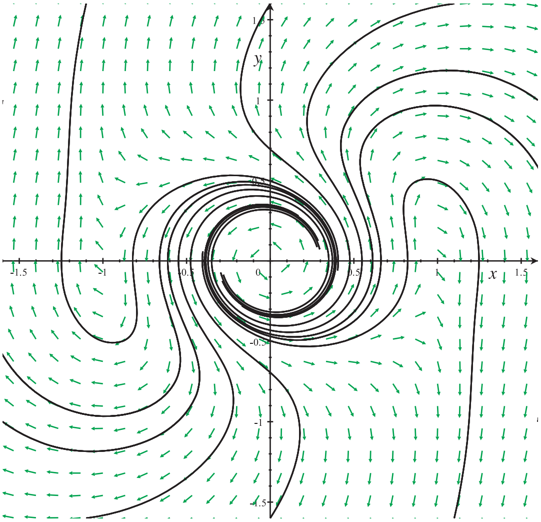

Figure 6.8. *Unstable, nonhyperbolic focus for* (6.19) *when* $\omega = 1$.

$\gamma = \Gamma_{(r^*,0)}$ is a periodic orbit. Note that $\lim_{k\to\infty}(t_{k+1} - t_k) = T$ exists because the trajectories of (6.17) depend continuously on initial conditions; recall Theorem 3.29. Moreover, T is the period of the limit cycle since

$$\varphi_T(r^*,0) = \varphi_T\left(\lim_{k\to\infty} r_k, 0\right) = \lim_{k\to\infty}\left(\varphi_{t_{k+1}-t_k}(r_k,0)\right) = \lim_{k\to\infty}\left((r_{k+1},0)\right) = (r^*,0). \quad \square$$

The style of argument represented by this lemma is similar to that which we will use to prove the Poincaré–Bendixson theorem in §6.6.

The final possibility is the center-focus.

Lemma 6.13. *Suppose the origin for* (6.17) *is neither a topological center nor a nonhyperbolic focus. Then it is a center-focus.*

Proof. One can show that there is a δ and an $\varepsilon < \omega$ such that (6.18) holds and such that there is an initial condition $(r_0, 0) \in B_\delta(0)$ that evolves to a point $(r(T), 2\pi) \in B_\delta(0)$; see Exercise 8. If $r(T) < r_0$, then this implies that the trajectory remains in $B_\delta(0)$ for all $t > 0$ and that $\theta(t) \to \infty$ as $t \to \infty$. A similar conclusion can be made for the backward trajectory if $r(T) > r_0$. If by chance $r(T) = r_0$, the trajectory is a limit cycle (since we have assumed the origin is not a center) and we can choose a smaller initial point r_0 for which $r(T) \neq r_0$. As in Figure 6.7, let r_j be the infinite sequence of radii at which the trajectory crosses $\theta = 0$, choosing a direction of time for which this sequence is strictly decreasing. This monotone sequence has a limit, but, since the origin is assumed to not be a focus, this limit is not 0; thus $r_j \to r^* \neq 0$. The orbit of the point $(r^*, 0)$ must be periodic and thus is a limit cycle, γ. Every trajectory inside γ is bounded and so remains inside the ball of radius $\tilde{\delta}$ that corresponds to the maximum distance of γ from the origin. Since the origin is not a center or focus for the new $\tilde{\delta}$, the same argument yields a new curve $\tilde{\gamma}$ inside γ that is also a limit cycle. This argument can be repeated arbitrarily many times. $\square$

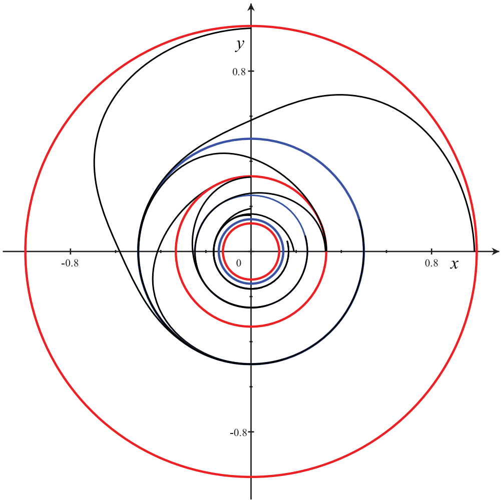

Figure 6.9. *Center-focus phase portrait for* (6.20) *with* (6.21). *The red (blue) circles are unstable (stable) limit cycles.*

Example 6.14. A special case of (6.15) is the system

$$\dot{x} = -y + xh(r),$$
$$\dot{y} = x + yh(r),$$

where $h(r)$ is a function that is continuous and has the limit $h(0) = 0$. In polar coordinates this becomes

$$\dot{r} = rh(r), \quad \dot{\theta} = 1. \tag{6.20}$$

There is a circular trajectory for each zero of h. These trajectories are limit cycles if the zeros of h are isolated. When $h > 0$ the trajectories spiral outward and when $h < 0$ they spiral inward. If h has an infinite sequence of zeros $h(r_j) = 0$ such that $r_j \to 0$, then the origin is a center-focus. One example of this is

$$h(r) = r\sin(\pi r^{-1}), \tag{6.21}$$

which has zeros at $r_j = j^{-1}$ so that the limit cycles are

$$\gamma_j = \left\{(x,y) : \ r = j^{-1}\right\}, \ \ j = 1,2,\ldots.$$

These are alternately stable (even j) and unstable (odd j), as shown in Figure 6.9. ■

Example 6.15. In some cases, the planar system (6.13) has a conserved quantity, i.e., a function $H(x,y)$ that is constant along the trajectories. So that this is the case,

$$0 = \frac{d}{dt}H(x,y) = \frac{\partial H}{\partial x}\left[-\omega y + p(x,y)\right] + \frac{\partial H}{\partial y}\left[\omega x + q(x,y)\right]$$

must have a solution for some function H. This equation is a quasi-linear partial differential equation (PDE). The function H could be found by solving its characteristic equations; however, this is just as hard as solving the original problem! There is

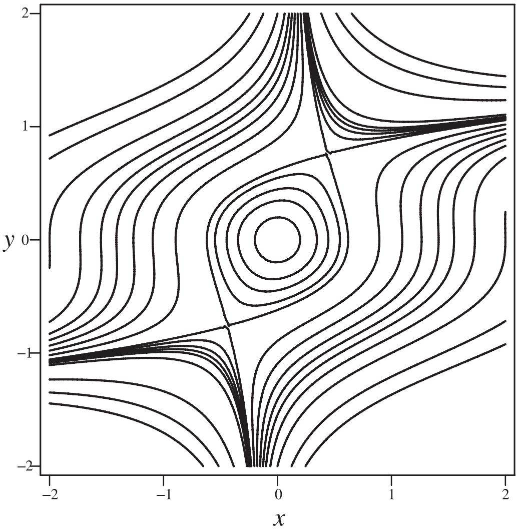

Figure 6.10. *Contours of the Hamiltonian* (6.23).

a special case when this is not true, and that is when the system can be written in Hamiltonian form,

$$\dot{x} = \frac{\partial H}{\partial y}, \quad \dot{y} = -\frac{\partial H}{\partial x};$$

recall (1.13). For (6.13), the Hamiltonian must take the form $H(x,y) = -\frac{\omega}{2}(x^2+y^2)+h(x,y)$ with $p = \partial h/\partial y$ and $q = -\partial h/\partial x$. These two are compatible if and only if

$$\frac{\partial p}{\partial x} = -\frac{\partial q}{\partial y}. \tag{6.22}$$

By assumption, $h = o(r^2)$; therefore, contours of H are closed loops in the neighborhood of the origin. This gives a quick test for a center; however, it is inconclusive if it fails. ■

Example 6.16. It is easy to see that the system

$$\begin{aligned} \dot{x} &= -y + 3xy^2, \\ \dot{y} &= x - y^3 \end{aligned} \tag{6.23}$$

satisfies (6.22) and is therefore Hamiltonian with

$$H = -\tfrac{1}{2}\left(x^2+y^2\right) + xy^3.$$

The contours of H are shown in Figure 6.10. Note that the origin is a topological center and that the stable and unstable manifolds of the two saddle points bound the family of closed loops surrounding the origin. ■

6.4 ▪ Symmetries and Reversors

Although our analysis has shown that topological centers are not common, every linear center of a planar Hamiltonian system is a topological center. Topological centers are also common in systems that have a "reversing symmetry."

A flow is said to have a *symmetry* if there is a diffeomorphism, $S : M \to M$, that conjugates the flow to itself:

$$\varphi_t(S(z)) = S(\varphi_t(z)),\ t \in \mathbb{R}. \tag{6.24}$$

Since we assume that S is smooth, we can take the time derivative of this relation to obtain an equivalent requirement on the vector field associated with φ:

$$f(S(z)) = DS(z)f(z). \tag{6.25}$$

Some symmetries, like a rotation symmetry, depend continuously upon a parameter and are thus called *continuous symmetries.* For example, the system (6.20) is obviously symmetric under the rotation

$$S_\psi(r,\theta) = (r,\theta+\psi) \tag{6.26}$$

for any angle ψ. For this case DS is the identity matrix, so (6.25) becomes $f(r,\theta+\psi) = f(r,\theta)$, which is satisfied for all ψ when f is a function of r only.

The collection of symmetries of a flow forms a group. This follows because the identity map is always a symmetry, and if S_1 and S_2 are symmetries of φ, then so is their composition $S_3 = S_1 \circ S_2$. Similarly, the inverse of a symmetry also satisfies (6.24) and therefore is also a symmetry. For example, the rotation symmetry (6.26) is a representation of the abstract rotation group, $O(2)$.

Discrete symmetries can also occur. For example, the system (6.11) is symmetric under the transformation $S(x,y) = (-x,-y)$, a rotation by π. To see this, note that for this case $DS = -I$, so (6.25) becomes $f(-x,-y) = -f(x,y)$, which is obviously satisfied by (6.11). The symmetry group in this case has two elements, the identity and S, and is called $\mathbb{Z}^2$. Much more about the implications of the existence of a nontrivial symmetry group can be found in (Field and Golubitsky 1995; Golubitsky and Stewart 2002).

Another type of symmetry that commonly occurs is a *time reversal* or *reversing* symmetry—when the motion backward in time is equivalent to that forward in time. Thus, a system is said to have reversing symmetry if there is a diffeomorphism, S (the reversor), that conjugates the flow to its inverse so that $\varphi_{-t}(S(z)) = S(\varphi_t(z))$. Again, this is equivalent to a requirement on the vector field

$$-f(S(z)) = DS(z)f(z). \tag{6.27}$$

This implies that in the new coordinate system, $\zeta = S(z)$, the differential equation $\dot{z} = f(z)$ becomes

$$\dot{\zeta} = DS(z)\dot{z} = DS(z)f(z) = -f(S(z)) = -f(\zeta),$$

which is the same differential equation going backward in time.

In many cases the reversor S is an involution, i.e., $S^2 = S \circ S = id$. For example, for mechanical Hamiltonian systems (recall §1.4) of the form

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

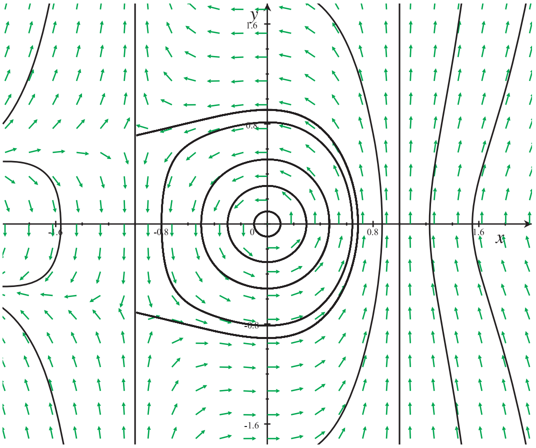

Figure 6.11. *Reversible system* (6.28) *with* $\alpha = 1$ *and* $\beta = 2$. *The origin is a symmetric equilibrium, but the saddles are not.*

the involution $S(x,y) = (x,-y)$ reverses the momentum, y, and is equivalent to reversing time. Note also that in this case S is orientation reversing, $\det(DS) = -1 < 0$.

The fixed set of a reversor S is

$$\mathrm{Fix}(S) = \{z : z = S(z)\}.$$

An orbit that intersects $\mathrm{Fix}(S)$ is a *symmetric orbit*. In particular, a symmetric equilibrium is a point $z^* \in \mathrm{Fix}(S) \cap \{f(z) = 0\}$. Not every orbit is symmetric; however, every orbit has a symmetric partner (see Exercise 5).

It can be shown that the fixed set of any orientation-reversing involution in $\mathbb{R}^2$ is a curve, $C = \mathrm{Fix}(S)$ (MacKay 1993). If this is the case, then whenever z^* is a symmetric, linear center, it must be a true center of the nonlinear system.

Lemma 6.17. *Suppose* $\dot{z} = f(z)$ *is reversible with reversor* S *and* $\mathrm{Fix}(S)$ *is a curve that contains an equilibrium* z^* *that is a linear center. Then* z^* *is a topological center.*

Proof. According to (6.18), the angle θ about the equilibrium must increase monotonically near z^*. The orbit of a point $z(0) \in \mathrm{Fix}(S)$ in this neighborhood must therefore return to $\mathrm{Fix}(S)$ after θ has increased by (roughly) π. Let τ be the time at which this first return happens. Then the reflection $\zeta(t) = S(z(t))$ of this orbit segment also touches $\mathrm{Fix}(S)$ at $z(0)$ and $z(\tau)$. Since $\zeta(t)$ is a solution beginning at $z(0)$ but going backward in time, the curve $\gamma = \{\varphi_t(z(0)) : -\tau \le t < \tau\}$ is a closed loop and by uniqueness must be periodic with period 2τ. Incidentally, each solution must cross the curve $\mathrm{Fix}(S)$ smoothly, so $DS(f(z(0))) = -f(z(0))$; this follows from the conjugacy relation (6.27) when $z \in \mathrm{Fix}(S)$. □

Example 6.18. The system

$$\begin{aligned} \dot{x} &= -y + \alpha x^2 y, \\ \dot{y} &= x + \beta y^2 x^2 \end{aligned} \tag{6.28}$$

has the reversor $S(x,y) = (x,-y)$ since

$$DSf(x,y) = (-y + \alpha x^2 y, -x - \beta y^2 x^2) = -(-(-y) + \alpha x^2(-y), x + \beta(-y)^2 x^2) = -f(S(x,y)).$$

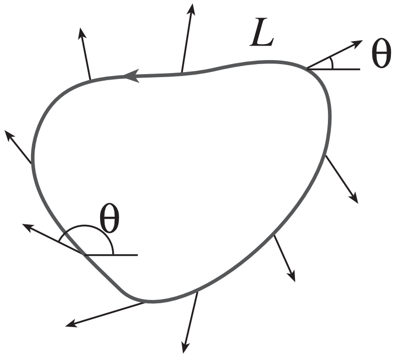

Figure 6.12. *Definition of the Poincaré index. Here* $I_L(f) = 1$.

Note that the fixed curve for S is the x-axis, and since the origin is a symmetric fixed point, Lemma 6.17 implies it is a center. A phase portrait is shown in Figure 6.11. When $\alpha > 0$, this system also has a pair of saddle equilibria.

Note that (6.28) is not Hamiltonian since

$$\partial p / \partial x = 2\alpha x y \neq -\partial q / \partial y = -2\beta y x^2.$$

Thus, the reversible property is independent of being Hamiltonian. ■

6.5 ▪ Index Theory

Another way to classify equilibria is through a topological property called the Poincaré index. An advantage of this concept is that it does not require that the vector field be smooth.

We begin by defining the index of a *simple closed loop L*, a curve defined by a continuous, one-to-one mapping $L : \mathbb{S}^1 \to \mathbb{R}^2$ of the circle into the plane (recall §5.5). Such curves are often called *Jordan curves*. A simple example of such a mapping is $L = \{(\cos t, \sin t) : 0 \le t < 2\pi\}$, the unit circle. The loop L is assigned an orientation by the direction of its traversal; in this example the orientation is counterclockwise.

Taking $f = (P, Q)$ to be a vector field on $\mathbb{R}^2$ as in (6.1), define θ to be the direction of f so that $\tan\theta = Q/P$. The direction is well defined even when the slope is infinite, provided that P and Q do not simultaneously vanish—that is, everywhere except for the equilibria; see Figure 6.12. Using the direction field, Poincaré defined an index of L relative to the vector field.

> ▷ *Poincaré index*: Suppose $f \in C^0(\mathbb{R}^2, \mathbb{R}^2)$, L is an oriented, Jordan curve, and there are no equilibria of f on L. The index, $I_L(f)$, is the integer number of rotations of the vector $f(x)$ as x traverses the loop in the positive direction,
>
> $$I_L(f) \equiv \frac{\Delta\theta}{2\pi}, \tag{6.29}$$
>
> where $\Delta\theta$ is the net change in direction of f upon traversal of the loop.

When the vector field is C^1, $\Delta\theta$ can be obtained by integrating along the curve:

$I_L(f) = \frac{1}{2\pi}\oint_L d\theta$. Given that $\tan\theta = Q/P$, we differentiate to obtain

$$\sec^2\theta d\theta = \frac{PdQ - QdP}{P^2}.$$

Since $\sec^2\theta = 1 + (Q/P)^2$, the index is then defined by the line integral

$$I_L(f) = \frac{1}{2\pi}\oint_L d\theta = \frac{1}{2\pi}\oint_L \frac{PdQ - QdP}{P^2 + Q^2}, \tag{6.30}$$

which can be evaluated explicitly if the loop L is given in parametric form.

Example 6.19. If $f(x,y) = (x,y)$ and L is the circle of radius r with counterclockwise orientation, then $(x(s), y(s)) = (r\cos s, r\sin s)$, and the index is

$$I_L(f) = \frac{1}{2\pi}\oint_C \frac{xdy - ydx}{x^2 + y^2} = \frac{1}{2\pi}\int_0^{2\pi} \frac{r^2\cos^2 s + r^2\sin^2 s}{r^2} ds = 1.$$

Similarly, if $f = (y,x)$, we obtain $I_L(f) = -1$. Note that neither of these results depends upon the radius of the loop. ▮

It is often easier to simply sketch the vector field and visually construct the index by "watching" the vector field rotate as its base point traverses L, instead of computing the integral (6.30).

Example 6.20. Suppose $f(x) = Ax$ and A is a hyperbolic matrix. Let L be a circle enclosing the equilibrium, $(0,0)$, with counterclockwise orientation. The computation of the index for four different hyperbolic equilibria of a linear equation is sketched in Figure 6.13. The upper left panel shows the vector field for a sink. Here, the direction is primarily inward along the loop L,and thus θ increases upon counterclockwise traversal of L, undergoing one complete rotation. Thus the index in this case is $+1$. Similarly, the source and spiral source also have index $+1$. The spiral sink, not shown, also has index $+1$. The last panel shows the saddle; here θ rotates clockwise so that the saddle has index -1. This visual analysis suggests that the index is independent of the details of the vector field and distinguishes the saddle from the other cases. We will prove this fact below. ▮

The Poincaré index can be used to obtain restrictions on the type of equilibria that are contained in a region bounded by a closed loop. We first show that the index of a curve typically does not change as it moves.

Lemma 6.21. *If a curve L is deformed and does not cross an equilibrium, then $I_L(f)$ does not change. Similarly, if the curve is held fixed and the vector field is varied, then the index does not change, so long as no equilibria fall on L throughout the deformation.*

Proof. The index is a continuous function of the curve L, as can be seen from (6.30), providing that $P^2 + Q^2|_L \neq 0$. This is just the condition that there be no equilibria on L. Since $I_L(f)$ is an integer and is continuous, it cannot change. The same considerations imply that the index is a continuous function of f. □

One application of this lemma is to prove the observations of the previous example.

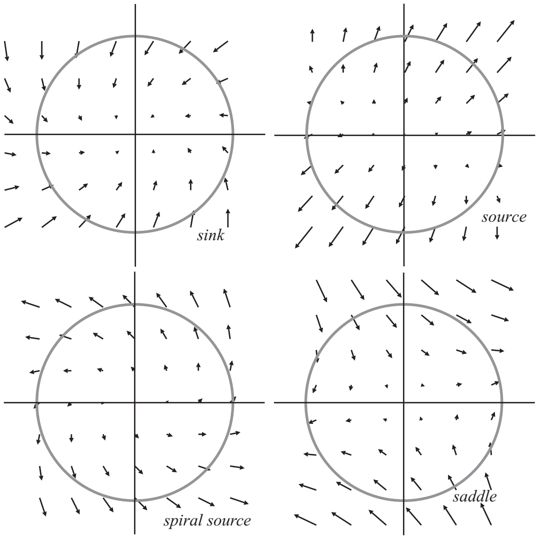

Figure 6.13. *Index of four types of hyperbolic matrices.*

Lemma 6.22. *If $f(x) = Ax$, where A is a nonsingular, 2×2 matrix and L is any counterclockwise loop enclosing the origin, then $I_L(f) = \operatorname{sgn}(\det(A))$.*

Proof. Let $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Since $\det(A) \neq 0$, either $ad \neq 0$ or $bc \neq 0$. We can thus deform A without changing $I_L(f)$ into one of two forms, $\begin{pmatrix} \operatorname{sgn}(a) & 0 \\ 0 & \operatorname{sgn}(d) \end{pmatrix}$ or $\begin{pmatrix} 0 & \operatorname{sgn}(b) \\ \operatorname{sgn}(c) & 0 \end{pmatrix}$. Now, deform the loop L to the unit circle. The index in either case is easy to compute from (6.30), verifying the assertion. □

Consequently, the index of a loop distinguishes between linear systems with saddle and nonsaddle equilibria. We can also use it to detect the very existence of equilibria.

Theorem 6.23. *If L is a Jordan curve and does not enclose an equilibrium of f, then $I_L(f) = 0$.*

Proof. According to Lemma 6.21, L can be shrunk to an infinitesimal loop without changing the index. The vector field limits to a nonzero constant on the loop since the loop does not enclose an equilibrium. Thus the index of this infinitesimal loop is zero. □

Unfortunately, the converse is not true: if $I_L(f) = 0$, we cannot conclude that there are no equilibria inside L, since it is equally possible that L contains an even number of equilibria, half with positive and half with negative index. One way to determine if this is the case is to refine the loop by dividing it into subloops.

Lemma 6.24. *The index of a sum of curves is the sum of the indices of the curves.*

Proof. This follows from the definition again: divide the loop L into two loops with a common connecting piece: write $L = L_1 + L_2$, as shown in Figure 6.14. The direction of traversal of the new loops is inherited from that of L. This implies that the net

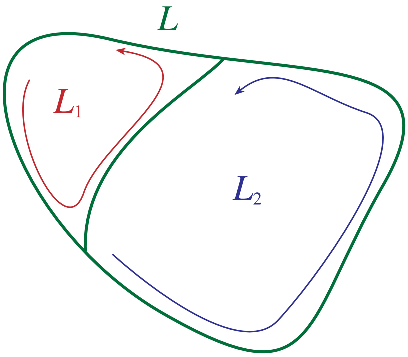

Figure 6.14. *Index of a sum of two curves.*

contribution from the connecting piece vanishes because the two subloops traverse it in opposite directions. □

An equilibrium has an index:

▷ *index of an isolated equilibrium*: $I_{x^*}(f)$ is the index of any curve that encircles the equilibrium x^* and no others.

According to Lemma 6.21 the index $I_{x^*}(f)$ is independent of the encircling loop, since the loop can be deformed to any other enclosing loop. Any loop that encloses a set of isolated equilibria can be partitioned into loops that enclose each individual equilibrium. Then Lemma 6.24 implies that the index of the original loop is the sum of the indices of the enclosed equilibria.

We have already computed the index of each type of hyperbolic equilibrium. It is also possible to find the index of an isolated nonhyperbolic node (recall §6.2) using

▷ *Bendixson's formula*: The index of an isolated node of a continuous vector field is

$$I_{x_o}(f) = 1 + \tfrac{1}{2}(e - h), \tag{6.31}$$

where e is the number of elliptic sectors and h the number of hyperbolic sectors. Parabolic sectors do not contribute to the index.

Finally, the index can be used to detect the existence of periodic orbits.

Theorem 6.25. *If γ is a periodic orbit of f, then $I_\gamma(f) = 1$.*

Proof. First, assume that the flow direction on γ is in the positive direction (e.g., the loop is traversed counterclockwise). Since f is tangent to γ it clearly makes a positive circuit as γ is traversed. Similarly, if the flow is in the opposite direction to the positive traversal of γ, then f still rotates once in a positive direction. □

Corollary 6.26. *Any periodic orbit of f must enclose at least one equilibrium.*

Higher Dimensions: The Degree

The concept of index can be generalized to higher dimensions, where it is more properly viewed as a special case of the concept of the *degree* of a mapping. A vector field

f can be thought of as a map from the phase space M to the vector space $\mathbb{R}^n$. As we discussed in §5.5, a map is simply a function

$$f : M \to N.$$

For the case of vector fields, both spaces have the same dimension. Even in this case a map is not necessarily one-to-one; this is quantified by its

> ▷ *degree*: Suppose M and N are compact, oriented manifolds of the same dimension and $y \in N$ is a regular value of $f \in C^1(M,N)$. The degree of f, $\deg(f)$, is the number of preimages of y counted with orientation:
>
> $$\deg(f) = \sum_{x \in f^{-1}(y)} \operatorname{sgn}\bigl(\det(Df(x))\bigr). \tag{6.32}$$

By definition, a point y in the range of f is a regular value if the rank of the Jacobian $Df(x)$ at every point $x = f^{-1}(y)$ is $\dim(N)$. When the manifolds M and N are compact, then the number of preimages of any regular point is finite because f^{-1} is locally a diffeomorphism near each preimage, and compactness implies that there can be only finitely many regions where f is one-to-one. It can be shown that (6.32) is independent of the choice of the regular value y.

Example 6.27. Consider the map $f(\theta) = 2\theta \bmod 2\pi$ from $\mathbb{S}^1$ to itself. Each point on the circle has two preimages, θ and $\theta + \pi$. The map is increasing at both points. Therefore, $\deg(f) = 2$. ∎

The orientation of a map corresponds to whether it maintains or reverses the orientation of a local coordinate system. For example, when both the domain and the range of f are $\mathbb{R}^3$, we can place a local set of unit vectors, e_1, e_2, e_3, at a point x whose orientation is defined by the right-hand-rule, $e_3 = e_1 \times e_2$. If the images of these vectors still have the same orientation after mapping by f, then f has positive orientation.

A similar concept of orientation applies to manifolds, though we must think of the axes as being a local coordinate system attached to the tangent space of a point. The orientation of a set of independent vectors is $\operatorname{sgn}(\det(P))$, where $P = (e_1, e_2, \dots, e_n)$ is the matrix formed from the vectors. Since the Jacobian $Df(x)$ determines the image of an infinitesimal set of vectors, the orientation of the image is given by

$$\det(P') = \det(Df(x)P) = \det(Df(x))\det(P).$$

Thus the orientation reverses if $\det(Df(x)) < 0$. Consequently, if f is a smooth vector field on an n-dimensional manifold and $Df(x)$ is nonsingular, the degree of f at x is defined to be

$$\deg_x(f) \equiv \operatorname{sgn}\bigl(\det(Df(x))\bigr). \tag{6.33}$$

Thus the right-hand side of (6.32) counts the number of times the point y is covered with a sign determined by the orientation.

Example 6.28. Since $\det(A) = \Pi\lambda_i$, the degree of a nonsingular, linear map $f(x) = Ax$ on $\mathbb{R}^n$ is $(-1)^m$, where m is the number of negative, real eigenvalues. Note that if there are any complex eigenvalues, they come in conjugate pairs and therefore do not contribute to the sign. ∎

The degree of f actually depends only on its *direction field*, namely, the normalized vectors $g(x) = f(x)/|f(x)|$, since an isolated equilibrium x^* has a neighborhood N, where $f \neq 0$ except at x^*. When $x \in \partial B_\delta(x^*) \cong \mathbb{S}^{n-1}$ the direction is well defined and can be thought of as a map $g : \mathbb{S}^{n-1} \to \mathbb{S}^{n-1}$.

Example 6.29. Suppose $f(x,y) = (y,x)$, and the point (x,y) is on the unit circle. The direction of f is obtained by normalization:

$$g = \frac{f}{|f|} = \frac{1}{\sqrt{x^2+y^2}}(y,x) = (\sin\theta, \cos\theta) = \left(\cos\left(\frac{\pi}{2}-\theta\right), \sin\left(\frac{\pi}{2}-\theta\right)\right).$$

Consequently, g maps a point $\theta \in \mathbb{S}^1$ to the new angle $\psi = \pi/2 - \theta$. The direction of increasing θ is transformed to decreasing ψ. Since each point has one preimage, but the orientation is reversed, $\deg(g) = -1$. ■

Using the direction field we can define the

▷ *index of an isolated equilibrium*: Suppose x^* is an isolated equilibrium of a C^1 vector field f and N is a neighborhood of x^* for which $f(x) \neq 0$ whenever $x \in N \setminus \{x^*\}$. For each such x, let $\xi = (x-x^*)/|x-x^*| \in \mathbb{S}^{n-1}$ and $g(\xi) \equiv f(x)/|f(x)|$ so that $g : \mathbb{S}^{n-1} \to \mathbb{S}^{n-1}$. The index of x^* is

$$I_{x^*}(f) \equiv \deg(g). \tag{6.34}$$

Example 6.30. For $x \in \mathbb{R}^n$, the vector field $f = -x$ has a single equilibrium, $x = 0$. The direction field is $g(\xi) = -\xi$ for a unit vector ξ. Each unit vector has a unique preimage under g and $\det(Dg) = (-1)^n$. Thus, $I_0(-x) = (-1)^n$. ■

These concepts lead to a profound result that relates a topological property of manifolds to vector fields.

Theorem 6.31 (Poincaré index). *If M is a compact manifold, then the sum of the indices of all equilibria of any smooth vector field that has at most a finite number of equilibria on M is independent of the choice of the vector field and is determined by M alone. This sum is the Euler characteristic of M.*

This is proved in most topology texts; see, e.g., (Hirsch 1976; Hocking and Young 1961).

6.6 ▪ Poincaré–Bendixson Theorem

The classification of the possible behaviors of a dynamical system requires the classification of its possible ω-limit sets. In general, this is extremely difficult—in fact, *chaotic* dynamical systems are complicated precisely because their ω-limit sets are complicated (see Chapter 7). However, the remarkable Poincaré–Bendixson theorem for flows in two dimensions essentially says

There is no chaos in two dimensions.

More specifically, this theorem implies that there are only three possibilities for ω-limit sets in the plane: equilibria, periodic orbits, and separatrix cycles (recall §5.2).

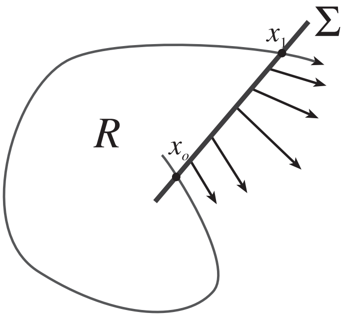

Figure 6.15. *A flow leaving R through a section* Σ.

Many of our previous examples have shown that an equilibrium can be an ω-limit set (for example, any asymptotically stable equilibrium). By contrast, if the ω-limit set contains no equilibria, then it turns out that the only other (compact) possibility is a periodic orbit.

Theorem 6.32 (Poincaré–Bendixson). *Let D be a connected subset of $\mathbb{R}^2$ and φ be a flow on D. Suppose that the forward orbit of some $p \in D$ is contained in a compact set and that $\omega(p)$ contains no equilibria. Then $\omega(p)$ is a periodic orbit.*

Proof. By Lemma 4.42, since the orbit of p is contained in a compact set, its ω-limit set is nonempty, compact, and connected. Choose a point $z \in \omega(p)$. Now $\omega(z) \subset \omega(p)$, since by Lemmas 4.40–4.41, the ω-limit set is closed and invariant, and $\Gamma_z \subset \omega(p)$, as are all its limit points. Since $\omega(p)$ contains no equilibria, $f(z) \neq 0$, and there exists a cross section of the flow near z (recall §4.12). Let Σ be a finite curve segment through z such that $f(x)$ is transverse to Σ for all $x \in \Sigma$. We will show that Γ_z intersects Σ only once and must be periodic. This is proved using four lemmas.

Lemma 6.33. *For any $x_o \in \Sigma$, a set of intersections of $\varphi_t(x_o)$ with Σ is monotone and ordered with t; that is, if $t_1 < t_2 < t_3$ are three intersection times, then the point $\varphi_{t_2}(x_o)$ is between $\varphi_{t_1}(x_o)$ and $\varphi_{t_3}(x_o)$ on Σ.*

Proof. Suppose the forward orbit of x_o intersects Σ more than once. Let $x_1 = \varphi_{t_1}(x_o)$ be the first intersection for $t > 0$. Note that $t_1 > 0$ because Σ is transverse to the flow. Then consider the curve $C = \{\varphi_t(x_o) : 0 \leq t \leq t_1\} \cup \{\Sigma$ between x_o and $x_1\}$—this curve is a closed non-self-intersecting loop and so divides the plane into two regions; see Figure 6.15. This result is a consequence of the following nontrivial theorem.

Theorem 6.34 (Jordan curve). *A simple closed curve (a set that is homeomorphic to $\mathbb{S}^1$) in $\mathbb{R}^2$ separates the plane into two connected components: one bounded, called the interior, and one unbounded, called the exterior.*

This theorem, though stated by Camille Jordan, was first correctly proved by Oswald Veblen in 1905; it is proved in most topology textbooks (Hocking et al. 1961, Theorem 2.28).

Continuing with the proof of the lemma, let R be the region interior to C. Since there are no equilibria on Σ and the flow cannot cross Γ_{x_o}, the flow must either leave or enter R through Σ, as shown in Figure 6.15. In the former case, the flow leaves R and cannot enter again; thus the next intersection of $\varphi_t(x_o)$ with Σ, i.e., x_2 cannot be in between x_o and x_1, because then the trajectory would have to be in R for some $t_1 < t < t_2$. Similarly, if the flow enters R on Σ, then it can never leave again, and therefore x_2 cannot be between x_o and x_1. In conclusion, the set of intersections x_k, $k \in \mathbb{Z}$, is ordered monotonically along Σ. □

The second lemma uses this monotonicity to show that the ω-limit set intersects Σ in a simple way.

Lemma 6.35. *$\omega(p)$ intersects a transversal Σ to a point in $z \in \omega(p)$ exactly once.*

Proof. As above let $z \in \omega(p)$ and Σ be a transversal to f at z. By definition of the ω-limit set, there is an infinite set of times t_n, $n = 1, 2, \ldots$, $t_n \to \infty$, such that $x_n = \varphi_{t_n}(p) \to z$. Since there are no equilibria in $\omega(p)$, and f is continuous, then $f(x) \neq 0$ for x near z; therefore, the orbit of every point x in some neighborhood of z must cross Σ. Hence, the times t_n can be chosen so that $x_n \in \Sigma$. By Lemma 6.33, these intersection points are ordered and therefore monotonically approach z. Since any monotone sequence on $\mathbb{R}$ has at most one limit point,[35] and since there is one already, $\Sigma \cap \omega(p) = z$. □

The next lemma implies that because of this single intersection, there must be a periodic orbit on the ω-limit set. Let z be a point on $\omega(p)$. By invariance $\Gamma_z \subset \omega(p)$, and since $\omega(p)$ is closed, $\omega(z) \subset \omega(p)$. If these two subsets intersect, then a periodic orbit ensues.

Lemma 6.36. *If Γ_z and $\omega(z)$ have a point in common, then Γ_z must be a periodic orbit.*

Proof. Let $x \in \Gamma_z \cap \omega(z)$ be the assumed common point. Then, since x is not an equilibrium by assumption, there is a transversal Σ at x. Now, $x \in \omega(z)$ so there is a sequence of times t_n such that $\varphi_{t_n}(z) \in \Sigma$ limits on x. Since $x \in \Gamma_z$, there is an s such that $\varphi_s(z) = x$. Letting $s_n = t_n - s$, then $\varphi_{t_n}(z) = \varphi_{t_n}(\varphi_{-s}(x)) = \varphi_{s_n}(x) \to x$. Suppose that $\varphi_{s_1}(x) \neq x$; then Lemma 6.33 implies that the next point, $\varphi_{s_2}(x)$ must be monotonically ordered on Σ and therefore be further from x. Consequently, the sequence $\varphi_{s_n}(x)$ must be ordered on Σ and move monotonically away from x, but this contradicts the fact that they limit on x. Thus, $\varphi_{s_n}(x) = x$. If s_1 is the first such time, it must be nonzero since $f(x) \neq 0$; moreover, by uniqueness, the next time is $s_2 = 2s_1$. In conclusion, $\Gamma_x = \Gamma_z$ is a periodic orbit with period s_1. □

Lemma 6.37. *If $\omega(p)$ contains a periodic orbit γ, then $\omega(p) = \gamma$.*

Proof. Let $y \in \gamma$, and construct a transversal Σ through y. Since $\omega(p)$ is closed and connected, if there is a point in $\omega(p)$ that is not in γ, then by connectedness there must be a sequence of points $y_n \in \omega(p) \backslash \gamma$ that limit on y. A point y_n close enough to y must have an orbit that intersects Σ; however, this contradicts Lemma 6.35, which says $\omega(p)$ intersects Σ precisely once. □

[35] Let x_n be a monotone sequence on $\mathbb{R}$ so that $x_n \leq x_{n+1}$ for all n. Suppose that x^* is a limit point; then $x_n \leq x^*$ for all n, since otherwise there would have to be a finite n for which $x_n \leq x^* < x_{n+1}$, and then it could not be a limit point. Suppose $y^* \neq x^*$ is another limit point. Without loss of generality we can assume $x^* < y^*$; however, in this case there are no points limiting on y^*, since all $x_n \leq x^*$. Thus there is no other limit point.

With these four lemmas in hand, we are finally ready to prove the Poincaré–Bendixson theorem.

Proof (of Theorem 6.32 continued). For any point $z \in \omega(p)$, let $x \in \omega(z) \subset \omega(p)$. Since x is not an equilibrium, there is a transversal Σ and, according to Lemma 6.35, $\omega(p)$ must intersect Σ precisely once. However, x is a limit point of the orbit of z, so there must be an infinite sequence of times t_n for which $\Sigma \ni \varphi_{t_n}(z) \to x$. Since $\omega(p)$ is invariant, $\Gamma_z \subset \omega(p)$, and since it intersects Σ precisely once, $\varphi_{t_n}(z) = x$. As a result $x = \Gamma_z \cap \omega(z)$, and so by Lemma 6.36 Γ_z is a periodic orbit. Accordingly, $\omega(p)$ contains a periodic orbit, and finally by Lemma 6.37, $\omega(p) = \Gamma_z$. □

A simple corollary of Theorem 6.32 allows us to show that limit cycles must exist in certain situations.

Corollary 6.38. *If R is a bounded, positively invariant subset of D that contains no equilibria, then it contains a limit cycle. The same holds for a negatively invariant subset.*

Proof. The orbit of every point p in R satisfies the Poincaré–Bendixson theorem, and in consequence $\omega(R)$ is a periodic orbit. □

Example 6.39. Let

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -x + y(1 - x^2 - 2y^2). \end{aligned} \tag{6.35}$$

The only equilibrium is the origin. The rate of change of the polar radius for (6.35) is

$$\dot{r} = \frac{y^2}{r}(1 - x^2 - 2y^2).$$

When $r^2 = 1$, then $\dot{r} = -y^4/r \leq 0$, and when $r^2 = 1/2$, then $\dot{r} = y^2x^2/r \geq 0$. This implies that the annulus $R = \{(x,y) : 2^{-1/2} \leq r \leq 1\}$ is positively invariant—note that even though $\dot{r} = 0$ at some points on the boundary of R, orbits cannot leave the annulus. We conclude there is at least one limit cycle in R. A numerical solution confirms our analysis; see Figure 6.16. ■

Example 6.40 (Hilbert's sixteenth problem). Part of the sixteenth of Hilbert's 22 problems for the twentieth century is to find an upper bound for the number of limit cycles of a polynomial vector field on $\mathbb{R}^2$ (Hilbert 1900). Perhaps the simplest case corresponds to quadratic differential equations in $\mathbb{R}^2$; it is known that these can have as many as four limit cycles, but it has never been proved that this is the maximum number, even though the proof has been announced numerous times (Ilyashenko and Yakovenko 1995)! An example of four limit cycles was given in (Shi 1980):

$$\begin{aligned} \dot{x} &= \lambda x - y + ax^2 + bxy + y^2, \\ \dot{y} &= x + x^2 + exy, \end{aligned} \tag{6.36}$$

where $a = -10$, $b = 5 - 10^{-13}$, $e = -25 + 9 \cdot 10^{-13} - 8 \cdot 10^{-52}$, and $\lambda = -10^{-200}$. Obviously, it is virtually impossible to study this system numerically! This system has two equilibria (the origin and $(0,1)$) that are foci (stable and unstable, respectively). The first is surrounded by a single local limit cycle, and the second has three—it is

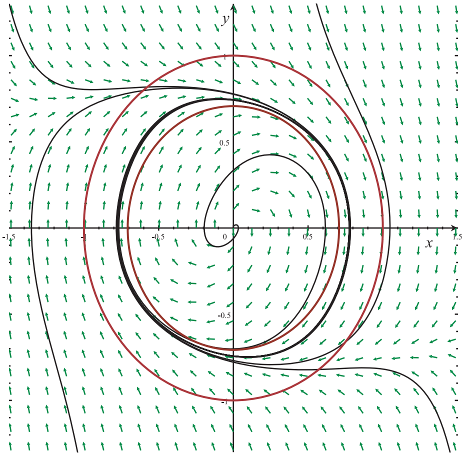

Figure 6.16. *Phase portrait of* (6.35) *showing the limit cycle and the boundaries of R.*

known that this is the maximum possible number of "local" limit cycles for a quadratic system. However, since not every limit cycle need enclose a single equilibrium, a more general system might also have more limit cycles. ■

We have seen that ω-limit sets in $\mathbb{R}^2$ can be equilibria, periodic orbits, or heteroclinic orbits. As we show next, these are the only possibilities.

Theorem 6.41. *Let D be a connected, open subset of $\mathbb{R}^2$, and suppose φ is a flow on D that has only* finitely *many equilibria. Suppose that the forward orbit of some $p \in D$ is contained in a compact set. Then $\omega(p)$ is either* (1) *an equilibrium,* (2) *a periodic orbit, or* (3) *the union of heteroclinic orbits (a separatrix cycle).*

Proof. If $\omega(p)$ contains only equilibria, then it must be a single equilibrium, since it is connected and the equilibria are isolated. If $\omega(p)$ contains no equilibria, then by Theorem 6.32 it is a periodic orbit. The only remaining case is when $\omega(p)$ contains both equilibria and "regular points," that is, points for which $f(x) \neq 0$. In this case $\omega(p)$ cannot contain any periodic orbits, since by Lemma 6.37 it would then be periodic and not contain any equilibria. Since $\omega(p)$ is connected, if it contains an equilibrium, it must contain an orbit that limits on the equilibrium. Moreover, if $z \in \omega(p)$ is a regular point, then $\omega(z)$ contains an equilibrium since otherwise $\omega(z)$ would be periodic, but this violates our assumption. Moreover, $\omega(z)$ cannot contain any regular points, since if it did, there would be infinitely many points on $\Gamma_z \subset \omega(p)$ that intersect a section at the regular point, violating Lemma 6.35. Therefore, every regular orbit in $\omega(p)$ must limit on an equilibrium. Similarly, since $\omega(p)$ is connected and invariant, the α-limit set of each regular orbit must be an equilibrium in $\omega(p)$. □

The Poincaré–Bendixson theorem can be proved more generally when the set D is any two-dimensional manifold—for C^2 flows there is only one change: if the manifold is the torus, then $\omega(x)$ could be the entire torus. This happens, for example, for the flow (take θ_i mod 2π),

$$\dot{\theta}_1 = 1, \quad \dot{\theta}_2 = \nu,$$

on the two-dimensional torus, when the rotation number, ν, is irrational; see §7.2.

6.7 ▪ Liénard Systems

A nonlinear oscillator of the form

$$\ddot{x} + f(x)\dot{x} = -g(x)$$

corresponds to a system with a nonlinear restoring force, $-g$, and generalized damping/forcing f. It is a generalization of the van der Pol oscillator that was introduced in §1.4.

Using the nonstandard change of variables $y = \dot{x} + F(x)$, where F is the antiderivative of f, this second-order equation can be written as the *Liénard* system,

$$\begin{aligned} \dot{x} &= y - F(x), \\ \dot{y} &= -g(x). \end{aligned} \tag{6.37}$$

It is easy to see that if $F(x) \equiv 0$, the system is Hamiltonian (recall (1.13)), with the energy

$$H(x,y) = \tfrac{1}{2}y^2 + G(x), \quad G(x) \equiv \int g(x)dx.$$

When $F \neq 0$, the energy changes at the rate

$$\frac{dH}{dt} = -g(x)F(x). \tag{6.38}$$

Therefore, energy drains from the system when $gF > 0$; otherwise the energy grows. For the van der Pol case, $gF = x^2\left(x^2/3 - 2\mu\right)$, so that the system is driven for small x and dissipative for large x.

The French engineer A. Liénard studied the special case $g(x) = x$ in 1928. His result goes beyond the Poincaré–Bendixson theorem because it gives conditions under which his system has a *unique* limit cycle.

Theorem 6.42 (Liénard). *Suppose that F and g are C^1 and*

(i) *F and g are odd, so that $F(0) = g(0) = 0$;*
(ii) *$xg(x) > 0$ for $x \neq 0$;*
(iii) *a is the unique positive zero of F and $F(x) < 0$ for $0 < x < a$;*
(iv) *$F(x)$ increases monotonically for $a < x$ and $F(x) \to \infty$ as $x \to \infty$.*

Then the system (6.37) *has a unique, stable limit cycle.*

Liénard's hypotheses imply that $g(x) = 0$ only at $x = 0$. The sign assumption means that physically g represents a *restoring force*, so that when $x > 0$, then $\dot{y} < 0$. This implies that the only equilibrium is the origin. To prove the theorem, we first show that every trajectory encircles this equilibrium.

Lemma 6.43. *Divide the plane into four regions bounded by the nullclines: $R_1 = \{(x,y) : x > 0, y > F(x)\}$, $R_2 = \{(x,y) : x > 0, y < F(x)\}$, $R_3 = \{(x,y) : x < 0, y < F(x)\}$, and $R_4 = \{(x,y) : x < 0, y > F(x)\}$. Then every trajectory beginning in R_1 moves successively to R_2, R_3, R_4, and finally back to R_1.*

Proof. The regions are sketched in Figure 6.17. The flow is to the right in R_1 and R_4 since $y > F(x)$ and is to the left in R_2 and R_3. It is down in R_1 and R_2 since $x > 0$

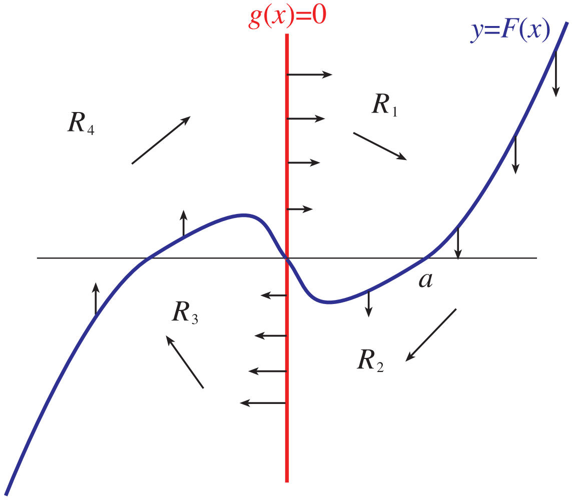

Figure 6.17. *Nullclines of Liénard's system* (6.37).

and up in R_3 and R_4. Consider a trajectory Γ that begins in R_1. It is moving to the right and therefore must eventually hit the $y = F(x)$ nullcline and subsequently enter R_2. Note that since $x > 0$, y is monotonically decreasing; at the nullcline, the flow is vertically down. In fact, Γ must leave a neighborhood of this nullcline in a finite time and cannot return so long as $x > 0$, because to do so it would have to be coming from above. Thus, x decreases, and after crossing the nullcline it must decrease at a rate that is bounded from above by some negative constant, so that $\dot{x} \le -c$. Therefore Γ reaches $x = 0$ in a finite time. Note that x is monotonically decreasing and so bounded by the value where it first enters R_2. Thus, the trajectory intersects the negative y-axis at a finite value and enters R_3. The equations are symmetric under the symmetry $S(x,y) = (-x,-y)$, and so the same arguments imply that Γ must continue through R_3 to R_4 and finally back to R_1. □

Lemma 6.44. *A trajectory* Γ *beginning at* $(0,y_o)$ *is periodic if and only if it intersects the negative* y*-axis at* $(0,-y_o)$.

Proof. This follows from symmetry of the equations. Let $y' = P(y_o)$ be the point at which Γ first intersects the negative y-axis. By uniqueness, the trajectories cannot cross and so y' must vary monotonically with y_o—in fact, P must be monotone decreasing, since once there is a trajectory Γ that goes from y_o to y', then trajectories for larger y_o must be outside this; hence a larger y_o leads to a more negative y'. By symmetry, a trajectory starting at $(0,y')$ will hit the positive y-axis as though it started at the point $-y'$ flowing forward with the map P and then flipped signs again, i.e., at the point $y'' = -P(-y')$. So that the orbit is periodic, $y'' = y_o = -P(-P(y_o))$. One solution occurs when $P(y_o) = -y_o$, which is the desired value. On the other hand, when $-P(y_o) < y_o$, then since $-P$ is monotone increasing,[36] $-P(-P(y_o)) < -P(-y_o) < y_o$, so the orbit is not periodic. Similarly, $-P(y_o) > y_o$ also means the orbit is not periodic. In conclusion, the orbit is periodic only when $y' = -y_o$. □

[36] A function is monotone increasing (resp., decreasing) when $x > y$ implies that $f(x) > f(y)$ ($f(x) < f(y)$).

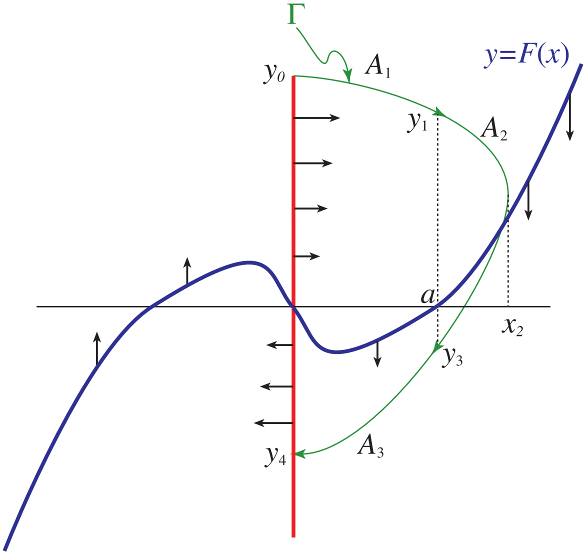

Figure 6.18. *Construction of the limit cycle for* (6.37).

Proof (of Theorem 6.42). Consider a trajectory Γ beginning at $(0, y_o)$ that crosses the nullcline at $(x_2, F(x_2))$ and then the negative y-axis at $(0, y_4)$. Our goal is to show that there is a unique choice $y_o = y^*$ for which $y_4 = -y^*$, for then by Lemma 6.44, Γ is periodic.

Consider the time rate of change of the energy H along the trajectory, (6.38). The change in energy along the trajectory up to a time t_4 when $y = y_4$ is

$$\Delta H(y_o) = H(0, y_4) - H(0, y_o) = \int_0^{t_4} \frac{dH}{dt} dt = -\int_0^{t_4} g(x(t))F(x(t))dt.$$

So that the trajectory is closed, we must have $y_o = -y_4 = y^*$, but then $H(0, y_4) = H(0, y_o)$, and so $\Delta H(y^*) = 0$. Since $g(x) > 0$, the only way this can happen is if F is positive on some part of the trajectory and negative elsewhere. In particular $x_2 > a$ since otherwise F is always negative on Γ, which would give a positive value to ΔH.

We want to argue that ΔH is a monotonically decreasing function of y_o when $x_2 > a$. Break the trajectory between y_0 and y_4 into three pieces, A_1 from $(0, y_o)$ to (a, y_1), A_2 from (a, y_1) to (a, y_3), and finally A_3 from (a, y_3) to $(0, y_4)$; see Figure 6.18. The integral for ΔH then has three terms:

$$\Delta H = \Delta H_1 + \Delta H_2 + \Delta H_3.$$

The pieces A_1 and A_3 must be graphs over x since $x(t)$ is either monotone increasing or decreasing. Consequently, the change of integration variables

$$dt = \frac{dx}{y - F(x)}$$

is well defined. For ΔH_1 this gives

$$\Delta H_1 = \int_0^a \frac{g(x)(-F(x))}{y(x) - F(x)} dx.$$

Note that if y_o increases, then $y(x)$ is larger on the entire segment A_1, and the remainder of the integrand is unchanged; thus the (positive) integrand decreases, and ΔH_1 is a monotone decreasing function of y_o. Similarly, as y_o increases, y_4 must become more negative. Since $y - F(x) < 0$ on A_3, the term

$$\Delta H_3 = \int_a^0 \frac{g(x)(-F(x))}{y(x)-F(x)}dx = \int_0^a \frac{g(x)|F(x)|}{|y(x)-F(x)|}dx$$

is again a decreasing function of y_o, since the denominator increases in magnitude. Finally, consider the middle term. Here, y can be used as the integration variable, setting

$$dt = -\frac{dy}{g(x)},$$

so that

$$\Delta H_2 = -\int_{y_3}^{y_1} F(x(y))dy. \qquad (6.39)$$

Uniqueness again implies that for each y, $x(y)$ must monotonically increase with y_o, since otherwise the trajectories would cross. Since $F(x)$ grows monotonically with x for $x > a$, the integrand is negative and becomes more negative as y_o increases. Since each term in ΔH decreases as y_o increases, ΔH is a monotone decreasing function.

To show that ΔH has a zero, we argue that $\Delta H_2 \to -\infty$ as $y_o \to \infty$. The reason is that y_1 must approach infinity with y_o. This follows because the time t_1 to reach a is finite; indeed, since $\dot{x} > y$, $t_1 < a/y_1$. Letting $g_{\max} = \max_{0\le x\le a} g(x)$, then

$$y_1 = y_o - \int_0^{t_1} g(x(t))dt > y_o - \frac{a}{y_1} g_{\max} \to \infty$$

as $y_o \to \infty$. A similar argument shows that $y_3 \to -\infty$. Because F is positive on A_2 and the limits of integration grow without bound, (6.39) is the integral of a negative function over an arbitrarily large interval as $y_o \to \infty$. Indeed, the integrand can be bounded away from zero apart from a small interval at the endpoints. Consider the trajectory for $x \in [a, b]$ for some $a < b < x_2$. Since F is increasing and this segment is above the nullcline, its slope has the bound

$$-\frac{dy}{dx} = \frac{g(x)}{y - F(x)} \le \frac{g(x)}{y - F(b)}.$$

Denoting $y(b) = y_1 - \delta$, separating variables, and integrating along the trajectory then gives

$$-\int_{y_1}^{y_1-\delta} (y - F(b))dy \le \int_a^b g(x)dx \quad \Rightarrow \quad \delta\left(y_1 - \frac{1}{2}\delta - F(b)\right) \le g_{\max}(b-a),$$

where $g_{\max}$ is the maximum value of g on $[a, b]$. Since $y(b) > 0$, then $\delta < y_1$ and finally

$$\delta \le \frac{2g_{\max}(b-a)}{y_1 - 2F(b)}.$$

Consequently as $y_1 \to \infty$, $\delta \to 0$, and the energy change becomes

$$|\Delta H_2| = \int_{y_3}^{y_1} F(x(y))dy > \int_{y_3+\delta}^{y_1-\delta} F(x(y))dy > F(b)|y_1 - y_3 - 2\delta|,$$

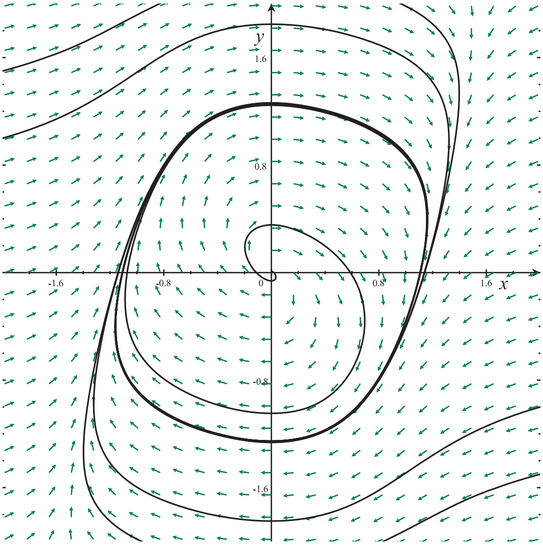

Figure 6.19. *Limit cycle for* (6.40).

which is unbounded.

Since the function ΔH is positive for small y_o and monotonically approaches $-\infty$ as y_o increases, it has a unique zero, $y = y^*$. This corresponds to the promised limit cycle.

When $y_o < y^*$, $\Delta H > 0$ and $y_4 < -y_o$. By uniqueness, the trajectory next intersects the positive y-axis at a point between y_o and y^*. This implies that trajectories inside the limit cycle spiral outward. Similarly, trajectories outside spiral inward. We conclude that the limit cycle is stable. □

Example 6.45. Consider the system

$$\begin{aligned} \dot{x} &= y - x(x^2 - 1), \\ \dot{y} &= -x. \end{aligned} \tag{6.40}$$

These functions satisfy the conditions of Theorem 6.42 and so there is a unique limit cycle. The phase portrait is shown in Figure 6.19. ■

Example 6.46. Liénard's result does not apply to this system:

$$\begin{aligned} \dot{x} &= y + x\cos x, \\ \dot{y} &= -\sin x. \end{aligned} \tag{6.41}$$

Here there are equilibria at $(n\pi, (-1)^{n+1} n\pi)$; these are saddle points for odd n and spiral sources for even n. A numerical plot of the phase portrait, Figure 6.20, shows that there is a stable limit cycle encircling the origin. Note that one branch of the unstable manifold of the saddles at $\pm(\pi, \pi)$ has the limit cycle as an ω-limit set. The reader is encouraged to examine the phase space for larger values of (x, y) to see that there is a succession of nested, ever-larger limit cycles. ■

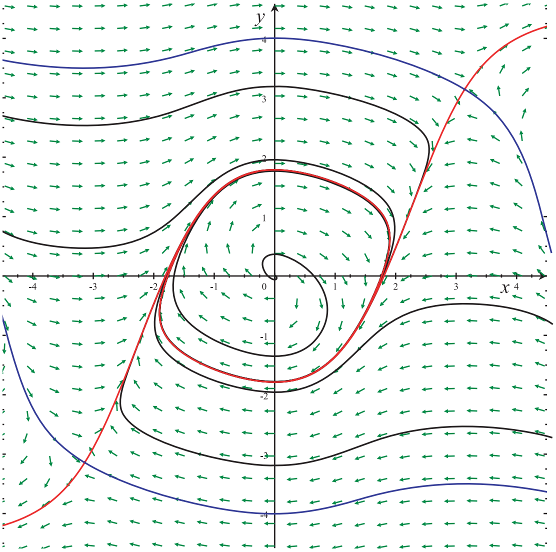

Figure 6.20. *Phase portrait of* (6.41).

6.8 ▪ Behavior at Infinity: The Poincaré Sphere

In previous sections and chapters a local picture of the dynamics in the plane was obtained by techniques such as linearization, center manifold reduction, polar coordinate transformations, and Poincaré maps. In addition, the Poincaré–Bendixson theorem provides a complete classification of the asymptotic behavior of the bounded orbits. A remaining task is the study of unbounded orbits. Although unboundedness often indicates a breakdown in the model of a physical system, studying the character of "equilibria" at infinity can augment our understanding of dynamics in the finite plane.

The behavior near $\pm\infty$ for an ODE on $\mathbb{R}$ is extremely simple. If all the equilibria of f are contained in a bounded interval, then when x is large enough, the sign of f is fixed, and orbits move monotonically toward or away from infinity depending upon $\operatorname{sgn}(f)$. Nevertheless, as a warm-up for the planar case, it is useful to study this behavior analytically.

One simple idea for analyzing the motion near infinity is to transform coordinates so that infinity is mapped to a finite point. For example, the transformation $x \to \xi = 1/x$ maps ∞ to the origin and the dynamics to $\dot{\xi} = -\xi^2 f(\xi^{-1})$. However, this transformation has the misleading property that both $+\infty$ and $-\infty$ map to the same point, $\xi = 0$, and there is no reason why a function should have similar behaviors at both places.[37] Poincaré developed a transformation without this drawback; it maps the extended line, including the two "points" at infinity, to a compact interval. The construction begins by embedding the phase line $\mathbb{R}$ into the plane with coordinates (X, Z) with the map $x \to \{(X, Z) : x = X, Z = 1\}$, as shown in Figure 6.21. This line is then projected onto the half-circle $\mathbb{S}^+ = \{(X, Z) : X^2 + Z^2 = 1, Z \geq 0\}$, using a line from the origin as shown in the figure. This has the effect of mapping $x = +\infty$ to $\theta = 0$ and $x = -\infty$ to $\theta = \pi$. The projection is a homeomorphism

$$\Pi : \mathbb{R} \cup \{\infty, -\infty\} \to \mathbb{S}^+$$

whose domain includes the two points at infinity. Similar triangles imply that if $\Pi(x) =$

[37] The stereographic projection has the same problem: infinity is mapped to the North Pole.

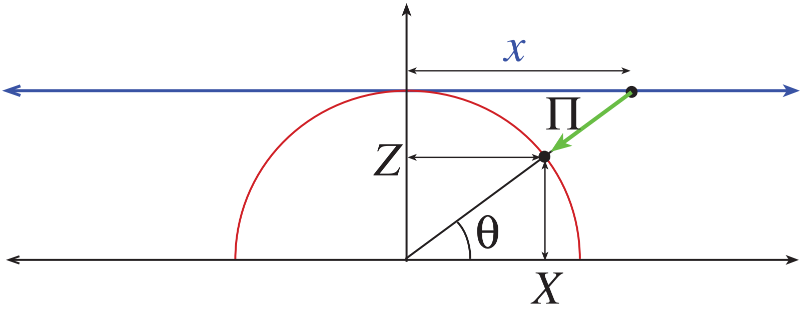

Figure 6.21. *Coordinates for the Poincaré circle.*

(X, Z), then

$$\frac{Z}{1} = \frac{1}{\sqrt{1+x^2}} = \cos\theta, \tag{6.42}$$

where $\theta \in [0, \pi]$ is the polar angle. Equivalently, $x = \cot\theta$.

For a one-dimensional ODE, $\dot{x} = f(x)$, the transformation $x = \cot\theta$ gives the system

$$\frac{d\theta}{dt} = -\frac{1}{\csc^2\theta}\frac{dx}{dt} = -\sin^2\theta f(\cot\theta) = g(\theta).$$

If f has a power law behavior near infinity, $f \sim ax^m + O(x^{m-1})$, then

$$g(\theta) \sim -a\sin^2\theta\cos^m\theta,$$

which for $\theta \to 0^+$ gives

$$\dot{\theta} = -a\theta^{2-m} + O(\theta^{1-m}). \tag{6.43}$$

If $m > 2$, (6.43) has a singular point at $\theta = 0$, and when $a > 0$, this implies that θ reaches zero in a finite time—therefore, the resulting solution is not a complete flow; recall §4.2. Completeness can be restored, however, by rescaling time to find a topologically equivalent system for which the vector field is bounded as $\theta \to 0$; recall §4.3. Defining the new time variable τ so that $d\tau = \theta^{1-m}dt > 0$ for $\theta > 0$ transforms (6.43) into

$$\frac{d\theta}{d\tau} = \frac{dt}{d\tau}\frac{d\theta}{dt} = -a\theta.$$

The new system is no longer singular at $\theta = 0$; instead, this point is stable when $a > 0$ and unstable when $a < 0$, as qualitatively seen from the direction field. Hence, the original ODE effectively has an equilibrium "at infinity" with this same property. This extension of the dynamics to infinity by defining a topologically equivalent dynamics is called *blowing up* the singularity.

To do the same for systems in $\mathbb{R}^2$, the projection Π must be generalized to one more dimension; this is accomplished by a projection from $\mathbb{R}^2$ to the northern hemisphere of a sphere, called the *Poincaré sphere*,

$$\mathbb{S}^{2+} = \{(X, Y, Z) : X^2 + Y^2 + Z^2 = 1, Z \geq 0\}.$$

Geometrically, this corresponds to embedding the (x, y) plane in $\mathbb{R}^3$ as the plane $Z = 1$ that is tangent to the North Pole of $\mathbb{S}^{2+}$. For each (x, y), a unique point in $\mathbb{S}^{2+}$ is

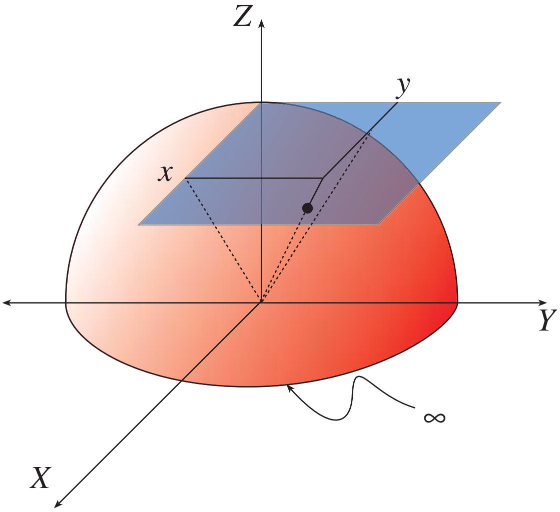

Figure 6.22. *Coordinates for the Poincaré sphere.*

obtained by projecting through the center of the sphere, as shown in Figure 6.22. In this case the projection is

$$x = \frac{X}{Z}, \quad y = \frac{Y}{Z}. \tag{6.44}$$

Note that "infinity" now corresponds to the equatorial circle, $Z = 0$. As in (6.42), similar triangles imply that $Z = \left(1 + x^2 + y^2\right)^{-1/2}$. Combining this with (6.44) yields

$$X = \frac{x}{\sqrt{1 + x^2 + y^2}}, \quad Y = \frac{y}{\sqrt{1 + x^2 + y^2}}, \quad Z = \frac{1}{\sqrt{1 + x^2 + y^2}}.$$

The planar system (6.1) is transformed into a set of equations in (X, Y, Z) that represent motion along the surface of the Poincaré sphere:

$$\begin{aligned} \dot{X} &= \frac{\dot{x}}{\sqrt{1 + x^2 + y^2}} - \frac{x(x\dot{x} + y\dot{y})}{(1 + x^2 + y^2)^{3/2}} = Z\left((1 - X^2)P - XYQ\right), \\ \dot{Y} &= Z\left(-XYP + (1 - Y^2)Q\right), \\ \dot{Z} &= -Z^2(XP + YQ), \end{aligned} \tag{6.45}$$

where P and Q are evaluated at $(X/Z, Y/Z)$. These equations have an invariant, because the motion takes place on the sphere:

$$\frac{d}{dt}\left(X^2 + Y^2 + Z^2\right) = 0.$$

Consequently, the system (6.45) contains one superfluous equation.

The topological properties of the flow near ∞ correspond to those of the system (6.45) near $Z = 0$. If, for example, the vector field (P, Q) has power law behavior near ∞, say, with maximum degree m, then $P(X/Z, Y/Z) \sim Z^{-m} + O(Z^{-m+1})$ near $Z = 0$. This implies that (6.45) has a singularity near $Z = 0$ of the form Z^{1-m}. As before, a topologically equivalent system can be obtained by rescaling time; define the regularization

$$\tau = \int Z^{1-m} dt, \quad \Rightarrow \quad \frac{d}{dt} = Z^{1-m}\frac{d}{d\tau}$$

for $Z > 0$. Note that $\tau(t)$ is monotone increasing, so this transformation is an appropriate one for topological equivalence. When $Z = 0$, the transformation is no longer an equivalence; however, it does give a system that exhibits the limiting behavior of the original one as $Z \to 0$ in a natural way. Defining the functions

$$P^*(X,Y,Z) = Z^m P\left(\frac{X}{Z},\frac{Y}{Z}\right), \quad Q^*(X,Y,Z) = Z^m Q\left(\frac{X}{Z},\frac{Y}{Z}\right),$$

(6.45) becomes

$$\begin{aligned}
\frac{dX}{d\tau} &= (1-X^2)P^* - XYQ^* = (Y^2+Z^2)P^* - XYQ^*,\\
\frac{dY}{d\tau} &= (1-Y^2)Q^* - XYP^* = -XYP^* + (X^2+Z^2)Q^*,\\
\frac{dZ}{d\tau} &= -Z(XP^* + YQ^*).
\end{aligned} \tag{6.46}$$

The equator is no longer a singularity for (6.46)—it is an invariant circle instead. Moreover, for $Z = 0$ all the terms in the equations for P^* and Q^* are zero except the highest order:

$$P^*(X,Y,0) = P_m(X,Y), \; Q^*(X,Y,0) = Q_m(X,Y),$$

where P_m and Q_m are the degree m terms in the original functions P and Q. The (X,Y) motion on this circle is given by

$$\begin{aligned}
\frac{dX}{d\tau} &= -Y(XQ_m - YP_m),\\
\frac{dY}{d\tau} &= X(XQ_m - YP_m).
\end{aligned}$$

"Infinity" has become an invariant manifold, a circle, with nontrivial dynamics. There are *equilibria at infinity* only when

$$XQ_m - YP_m = 0 \tag{6.47}$$

(since $X^2 + Y^2 = 1$ on the equator, X and Y cannot both be zero), and the motion is counterclockwise when $XQ_m - YP_m > 0$. Note that if (X,Y) is an equilibrium, then so is $(-X,-Y)$, i.e., the diametrically opposite point, since (6.47) is homogeneous of degree $m+1$. Moreover, the sign of $XQ_m - YP_m$ flips upon reflection through the origin if m is even but has the same sign when m is odd. For this reason, when m is odd the diametrically opposed points have the same topological types on the circle at infinity, but they have opposite types when m is even.

One way to treat the motion near an equilibrium at infinity is to shift coordinates so that the origin of the new coordinate system is at the equilibrium. It is easier to simply do another Poincaré projection onto a plane tangent to the Poincaré sphere at

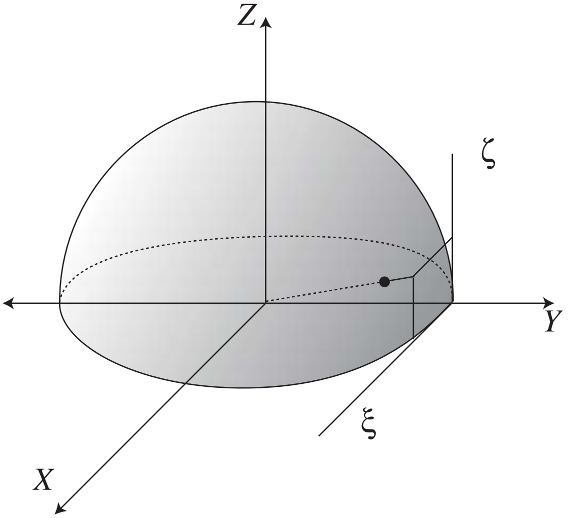

Figure 6.23. *Coordinates on the equator.*

either the X-axis or the Y-axis. For example, if the equilibrium occurs for some $Y > 0$, the transformation

$$\xi = \frac{X}{Y}, \;\; \zeta = \frac{Z}{Y}$$

projects (X, Y, Z) onto the plane tangent to the sphere at $Y = +1$; see Figure 6.23. The differential equations in the new coordinate system become:

$$\dot{\xi} = \frac{\dot{X}}{Y} - \frac{X}{Y^2}\dot{Y} = \frac{1}{Y}\big((Y^2+Z^2)P^* - XYQ^*\big) - \frac{X}{Y^2}\big(-XYP^* + (X^2+Z^2)Q^*\big)$$

$$= \frac{1}{Y}\left(P^* - \frac{X}{Y}Q^*\right) = \frac{1}{Y}(P^* - \xi Q^*),$$

$$\dot{\zeta} = \frac{\dot{Z}}{Y} - \frac{Z}{Y^2}\dot{Y} = \frac{1}{Y}(-ZXP^* - ZYQ^*) - \frac{Z}{Y^2}\big(-XYP^* + (X^2+Z^2)Q_*\big)^*$$

$$= -\frac{Z}{Y^2}Q^* = -\frac{1}{Y}\zeta Q^*.$$

Recalling that $P^* = Z^m P = Y^m \zeta^m P$ and similarly for Q^*, each of these ODEs has a leading factor Y^{m-1} which can be eliminated by rescaling time once more. Noting that $X/Z = \xi/\zeta$, $Y/Z = 1/\zeta$, obtain

$$\dot{\xi} = \zeta^m P\left(\frac{\xi}{\zeta}, \frac{1}{\zeta}\right) - \xi\zeta^m Q\left(\frac{\xi}{\zeta}, \frac{1}{\zeta}\right),$$

$$\dot{\zeta} = -\zeta^{m+1} Q\left(\frac{\xi}{\zeta}, \frac{1}{\zeta}\right).$$

An equilibrium with $Y < 0$ can be treated with the same definitions for ξ and ζ. However, this means that positive ξ and ζ correspond to *negative* X and Z—the projection is through the origin, so that the diametrically opposite points are equivalent. Finally,

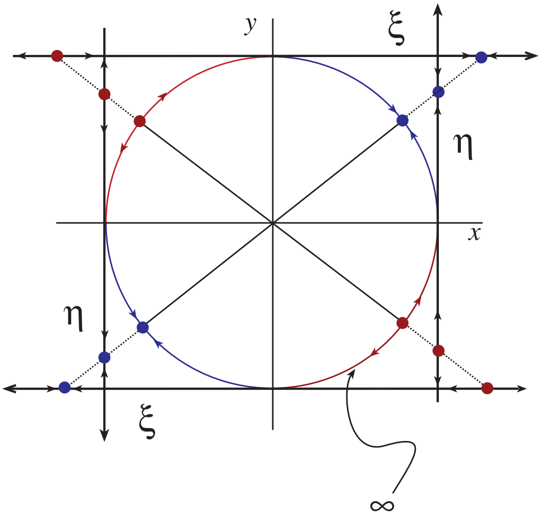

Figure 6.24. *Global phase portrait, looking down from the North Pole. When m is odd, the direction of time is not reversed for diametrically opposed equilibria.*

since time has been rescaled to eliminate the factor Y^{m-1}, when m is even this factor is negative and the direction of time is reversed.

If there is an equilibrium at $Y = 0$, we could similarly define $\eta = Y/X$ and $\zeta = Z/X$ to obtain, finally,

$$\dot{\eta} = Q^* - \eta P^* = \zeta^m Q\left(\frac{1}{\zeta}, \frac{\eta}{\zeta}\right) - \eta\zeta^m P\left(\frac{1}{\zeta}, \frac{\eta}{\zeta}\right),$$

$$\dot{\zeta} = -\zeta P^* = -\zeta^{m+1} P\left(\frac{1}{\zeta}, \frac{\eta}{\zeta}\right).$$

The dynamics can be summarized with a sketch obtained by looking down on the Poincaré sphere from the North Pole to view the (X, Y) plane—Figure 6.24. This gives a picture of the entire plane, together with the circle at infinity.

Example 6.47. For a linear system

$$\begin{aligned} \dot{x} &= ax + by, \\ \dot{y} &= cx + dy, \end{aligned} \tag{6.48}$$

the equilibria at ∞ are determined by

$$0 = YP^* - XQ^* = -cX^2 + (a-d)XY + bY^2,$$

where $P^* = ZP = aX + bY$ and $Q^* = cX + dY$. The intersections of this quadratic curve with the circle $X^2 + Y^2 = 1$ are a bit messy in the general case. Consider a concrete case $\begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix}$. Equilibria occur when $-2X^2 + Y^2 = 1 - 3X^2 = 0$, giving $(\pm 1/\sqrt{3}, \pm 2/\sqrt{3}, 0)$. Since these equilibria are not at $Y = 0$, the second transformation gives the dynamics on the (η, ζ) plane, $\zeta P = a + b\eta$, $\zeta Q = c + d\eta$, so

$$\dot{\eta} = \zeta Q - \eta\zeta P = 2 - \eta^2,$$

$$\dot{\zeta} = -\zeta^2 P = -\zeta - \eta\zeta.$$

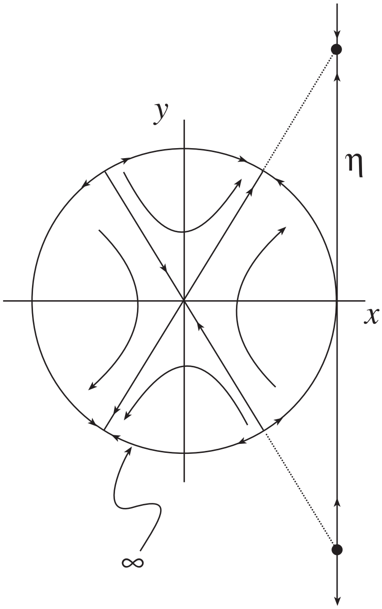

Figure 6.25. *Global phase portrait for the linear system* (6.48) *with $a = b = d$ and $c = 2$.*

Note that the equilibria are at $(\eta, \zeta) = (\pm 2, 0)$, which is the same obtained by noting that $\eta = Y/X$. The linearization about the equilibrium is

$$Df = \begin{pmatrix} -2\eta^* & 0 \\ 0 & -1-\eta^* \end{pmatrix}.$$

Thus the equilibrium $(2, 0)$ is a stable node with $\lambda = -4, -3$ and $(-2, 0)$ is an unstable node with eigenvalues $\lambda = 4, 1$. Looking at the Poincaré sphere from the top gives the global phase portrait in Figure 6.25. ∎

Example 6.48.

$$\begin{aligned} \dot{x} &= -4y + 2xy - 8, \\ \dot{y} &= 4y^2 - x^2. \end{aligned} \tag{6.49}$$

The phase portrait for this system in the finite plane is shown in Figure 6.26. There are two finite equilibria, $(4, 2)$ and $(-2, -1)$, and the Jacobian is

$$Df = \begin{pmatrix} 2y^* & -4 + 2x^* \\ -2x^* & 8y^* \end{pmatrix},$$

so that the linear matrices are

$$Df|_{(4,2)} = \begin{pmatrix} 4 & 4 \\ -8 & 16 \end{pmatrix}, \ \lambda = 8, \ v = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \ \ \lambda = 12, \ v = \begin{pmatrix} 1 \\ 2 \end{pmatrix},$$

$$Df|_{(-2,-1)} = \begin{pmatrix} -2 & -8 \\ 4 & -8 \end{pmatrix}, \ \lambda = -5 \pm i\sqrt{23},$$

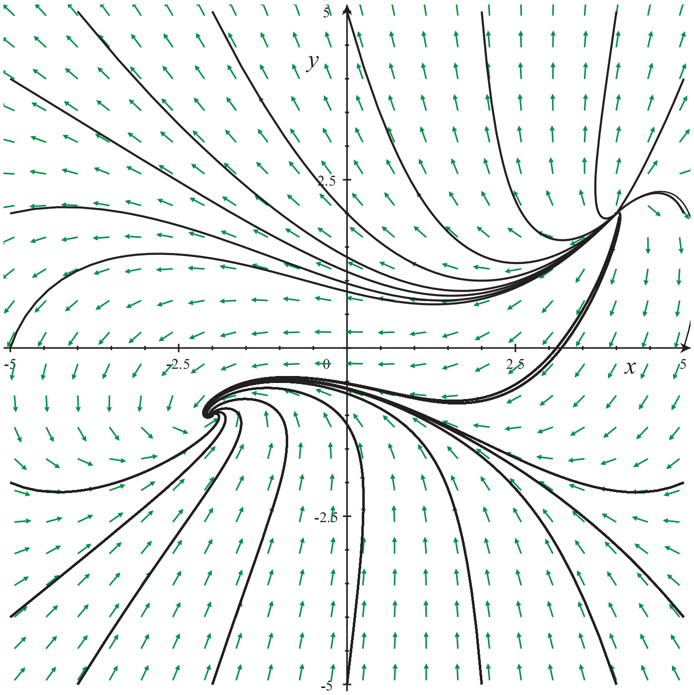

Figure 6.26. *Phase portrait for* (6.49). *The basin of attraction of the stable focus appears to be all points below some curve emanating from the unstable node.*

so that there is an unstable node and a stable focus. The equilibria at infinity are determined by

$$0 = YP_m - XQ_m = Y(2XY) - X(4Y^2 - X^2) = X\left(-2Y^2 + X^2\right).$$

There are six equilibria at infinity given by

$$(X,Y) = \left\{ \left(s_1\sqrt{\frac{2}{3}}, s_2\frac{1}{\sqrt{3}} \right), (0, s_3) :\ s_i \in \{-1, 1\} \right\}.$$

Converting to equations on the (ξ,ζ) plane gives (for the equilibria with $Y > 0$)

$$\zeta^2 P\left(\frac{\xi}{\zeta}, \frac{1}{\zeta}\right) = 2\xi - 4\zeta - 8\zeta^2,$$
$$\zeta^2 Q\left(\frac{\xi}{\zeta}, \frac{1}{\zeta}\right) = 4 - \xi^2.$$

So the differential equations are

$$\dot{\xi} = \zeta^2 P - \xi\zeta^2 Q = 2\xi - 4\zeta - 8\zeta^2 - 4\xi + \xi^3 = -2\xi - 4\zeta - 8\zeta^2 + \xi^3,$$
$$\dot{\zeta} = -\zeta^3 Q = -\zeta(4 - \xi^2).$$

Note that there are equilibria at $(\xi,\zeta) = (0,0)$ and $(\pm\sqrt{2}, 0)$, as expected. Linearizing

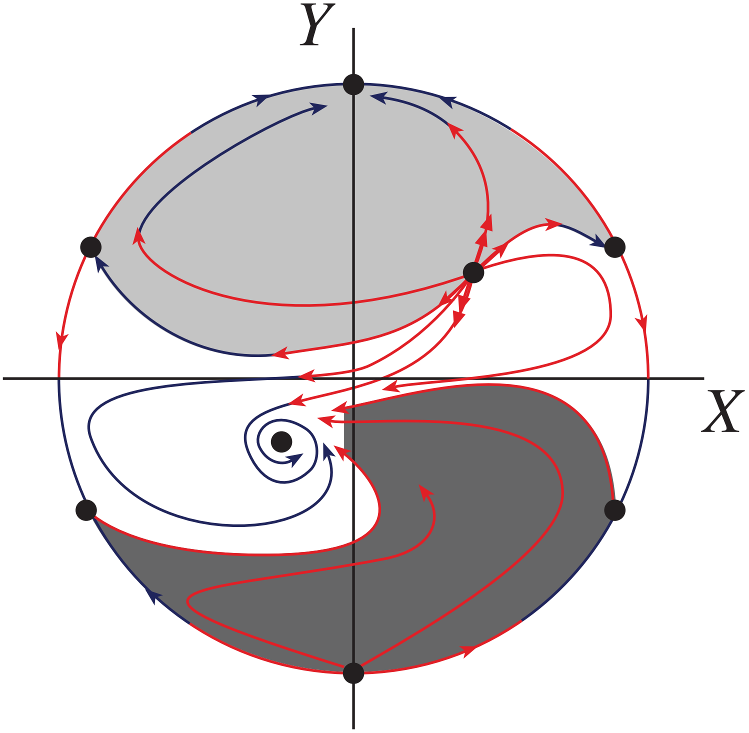

Figure 6.27. *Global phase portrait for* (6.49). *There are six fixed points at infinity; four are saddles,* $(0,-\infty)$ *is a source, and* $(0,\infty)$ *is a sink. The basin of the sink is shown in light gray, and the unstable set of the source is shown in dark gray. The boundaries of these sets are formed from the stable and unstable manifolds of the saddles at infinity. The spiral sink at* $(-2,-1)$ *has a basin that includes both the dark gray and white regions. Therefore, the points* $(-2,-1)$ *and* $(0,\infty)$ *are the ω-limit sets of every orbit, apart from those on the separatrices that form the basin boundaries.*

gives

$$Df|_{(0,0)} = \begin{pmatrix} -2 & -4 \\ 0 & -4 \end{pmatrix} \ (node),$$

$$Df|_{(\pm\sqrt{2},0)} = \begin{pmatrix} 4 & -4 \\ 0 & -4 \end{pmatrix} \ (saddle).$$

Since m is even, the stability for the diametrically opposed points with $Y < 0$ are the opposite of the corresponding $Y > 0$ points, since the direction of the flow is reversed by our transformations. The global phase portrait can be constructed by noting that the stable and unstable manifolds of the four saddle points at infinity define separatrices that divide the plane into sectors; see Figure 6.27. ∎

6.9 ▪ Exercises

1. Suppose that P is homogeneous of degree n and Q is homogeneous of degree m in system (6.6). Use the separation of variables technique as in (6.8)–(6.9) to classify the structure of the flow as $r \to 0$ depending upon n and m.

2. Show that as $r \to 0$ the leading terms in the Vinograd example (4.17) are topologically equivalent to (6.12) on the punctured plane $\mathbb{R}^2 \setminus \{0\}$.

3. Determine the nature of the equilibria of the following systems on $\mathbb{R}^2$. Be as specific as you can. Compare your analysis with numerical phase plane plots.

   (a) $\begin{array}{l}\dot{x}=2y-xy-4\\ \dot{y}=-4y^2+x^2\end{array}$, (b) $\begin{array}{l}\dot{x}=2x^2+y^2-1\\ \dot{y}=-x\end{array}$,

   (c) $\begin{array}{l}\dot{x}=x^2+y^2\\ \dot{y}=y+x^2\end{array}$, (d) $\begin{array}{l}\dot{x}=y^2+x^3\\ \dot{y}=y+x^2\end{array}$.

4. As discussed in §1.2, a Lotka–Volterra model for predator–prey interactions is given by

$$\begin{aligned}\dot{x}&=x(\alpha-\beta y),\\ \dot{y}&=y(-\gamma+\delta x),\end{aligned}\tag{6.50}$$

   where $\alpha,\beta,\gamma,\delta>0$. Here, x represents the prey population with a positive net birth rate α and y represents the predator population, which dies off if its food source is absent.

   (a) Show that this model has two equilibria, a saddle and center, and find the global stable and unstable manifolds of the saddle.

   (b) Show that the linear center is actually a topological center by using polar coordinates centered on the equilibrium and computing G (6.9).

   (c) Show that (6.50) is not a Hamiltonian system.

   (d) Show that there exists an invariant for (6.50) using the one-form (1.26) and setting $\alpha=F(x,y)dH$ for suitable choice of F. Plot the contours of H in the positive quadrant.

   (e) From (d) you can conclude that every orbit is periodic. Find the average predator population over the period, T, of an orbit by using

$$\int_0^T dt\frac{\dot{x}(t)}{x(t)}=\int_0^T dt\,(\alpha-\beta y(t))^{\cdot}$$

   Similarly, find the average prey population.

5. Suppose a flow φ has a reversor S and an orbit $\Gamma=\{\varphi_t(x):t\in\mathbb{R}\}$.

   (a) Show that $\tilde{\Gamma}=\{S\circ\varphi_{-t}(x):t\in\mathbb{R}\}$ is also an orbit of φ.

   (b) Show that the saddle equilibria of (6.28) are a symmetry-related pair.

   (c) Suppose the orbit Γ is symmetric: $\Gamma\cap\mathrm{Fix}(S)\neq\emptyset$. Show that Γ and $\tilde{\Gamma}$ coincide.

   (d) Suppose γ is a symmetric periodic orbit of φ. Show that γ has at least two points on $\mathrm{Fix}(S)$.

   (e) Suppose that x^* is a symmetric equilibrium. Show that $S(W^s(x^*))=W^u(x^*)$.

6. (a) Show that if x^* is a symmetric equilibrium of a reversible system, then whenever λ is an eigenvalue of the linearization at x^*, so is $-\lambda$.

   (b) Suppose $x\in\mathbb{R}^3$. Using the result of (a), find the most general form of the characteristic polynomial of any symmetric equilibrium.

(c) Show that the three-dimensional system

$$\begin{aligned}\dot{x} &= y + bz + ax(y - bz),\\ \dot{y} &= cx + x^2 + 2yz,\\ \dot{z} &= b^{-1}\left(cx - x^2 - 2yz\right)\end{aligned}$$

is reversible with the reversor $S(x,y,z) = (-x, bz, b^{-1}y)$.

(d) Find the fixed sets of S. Are there symmetric equilibria? Verify the eigenvalue property from (a) for each symmetric equilibrium.

7. Study the behavior of the system (6.19) near the equilibria at $(x,y) = (\pm\sqrt{\omega}, 0)$. Note that the system is symmetric under the reflection $S(x,y) = (-x,-y)$, so, using the results of Exercise 5, it is necessary to analyze only one of these equilibria.

8. Prove that when $p, q = o(r)$, there is a ball $B_\delta(0)$ such that (6.13) has a trajectory $\varphi_t(r,0) \in B_\delta(0)$ for $0 \le t \le T$ where $\varphi_T(r,0) = (r(T), 2\pi)$. This fact is used in the proof of Lemma 6.13 to assert that there is a trajectory that remains in $B_\delta(0)$ either as $t \to \infty$ or as $t \to -\infty$.

9. The tokamak is a toroidally shaped magnetic confinement device for plasma and will probably be the first device to produce net energy from controlled nuclear fusion reactions. One of the problems with confining plasma using magnetic fields is the plethora of instabilities that occur. One of these is called a "sawtooth oscillation." This oscillation is caused by helical disturbance in the plasma current and magnetic fields that results in a redistribution of the plasma temperature. A simple model that accounts for this physics is (Bora and Sarmah 2008)

$$\begin{aligned}\frac{3}{2} n\dot{T}_e &= \sigma E_{\|}^2 T_e^{3/2} - \nu n T_e^{1/2} A - \beta n T_e^{5/2},\\ \dot{A} &= \gamma\left(\frac{T_e}{T_s} - 1\right) A.\end{aligned} \tag{6.51}$$

The dynamical variables are T_e the electron temperature and A the amplitude of the instability mode. The remaining parameters are assumed to be constant: n the plasma density, $\sigma E_{\|}^2$ the ohmic heating due to a toroidal electric field, ν a rate of temperature redistribution, β a rate of energy diffusion, γ the growth rate of the instability, and T_s its temperature threshold.

(a) Show that by defining appropriate scaled variables, $\tau = at$, $x = bT_e$, and $y = cA$, the system (6.51) can be reduced to

$$\begin{aligned}\dot{x} &= (1 - \mu x)x^{3/2} - x^{1/2} y,\\ \dot{y} &= \rho(x - 1)y.\end{aligned}$$

Physically, x and y are nonnegative, so the phase space for this system is the positive quadrant. We will assume that $0 < \mu < 1$ and $\rho \gg 1$.

(b) Show that this system has three equilibria. Linearize about the equilibrium in the interior of the phase space and study its stability properties. Show that it is an unstable focus when ρ is large enough, provided that $\mu < 1/2$.

(c) Find a positively invariant region that encloses the unstable focus. This can be essentially done with a triangle formed from lines $y = 0$, $y = c(1-\mu x)$, and $y = dx - e$ for suitable choices of c, d, e, except for a neighborhood of the origin. Exclude the origin from your region by a curve that connects the first and second lines.

(d) Argue that the Poincaré–Bendixson theorem implies that this system has a limit cycle inside your region. This limit cycle is the "sawtooth oscillation."

(e) Confirm your conclusions with a computer study of the dynamics. A graph of $x(t)$ will show a sawtooth shape if ρ is large enough.

10. Consider the system

$$\dot{x} = \lambda x - y - x r^2 + \lambda \frac{x^3}{r},$$
$$\dot{y} = x + \lambda y - y r^2 + \lambda \frac{x^2 y}{r},$$

where r is the polar radius. Prove that this has a stable limit cycle when $\lambda > 0$. (*Hint*: Transformation to polar coordinates will be useful; you should be able to find an annulus that is guaranteed to contain a limit cycle.) Plot some orbits numerically as λ varies to verify your conclusions.

11. Suppose x^* is an isolated, nonhyperbolic node. Prove the validity of Bendixson's formula (6.31).

12. Consider the system

$$\dot{x} = x - y - x^2(x+2y) - xy^2,$$
$$\dot{y} = x + y + x^2(x-y) - y^2(x+y).$$

(a) Show that the equilibrium at the origin is an unstable focus.

(b) Using polar coordinates, find an annulus that is guaranteed, by Corollary 6.38, to contain a limit cycle.

(c) Investigate the dynamics numerically to confirm your conclusions.

13. Investigate Shi's problem (6.36) with $(a, b, e) = (-10, 5, -25)$ as λ increases from zero. Explore the limit cycles using your favorite phase plane software.

14. Construct the global phase portrait for the system (4.47):

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

You should verify the claim in §4.9 that the ω-limit set for points in the strip $R = \{(x,y) : |y| < 1,\ (x,y) \neq (0,0)\}$ is disconnected.

Chapter 7
Chaotic Dynamics

> *It may happen that slight differences in the initial conditions produce very great differences in the final phenomena; a slight error in the former would make an enormous error in the latter. Prediction becomes impossible and we have the fortuitous phenomena.* (Henri Poincaré 1914)
>
> *When our results concerning the instability of nonperiodic flow are applied to the atmosphere, which is ostensibly nonperiodic, they indicate that prediction of the sufficiently distant future is impossible by any method, unless the present conditions are known exactly. In view of the inevitable inaccuracy and incompleteness of weather observations, precise very-long-range forecasting would seem to be non-existent.* (Edward Lorenz 1963)

The Poincaré–Bendixson theorem in §6.6 implies that the ω-limit sets of the bounded motion of a flow in the plane are quite simple: equilibria, periodic orbits, or separatrix cycles. There is no such simple categorization of the possible limiting behavior of dynamics in $\mathbb{R}^3$. Indeed, as we discussed in §4.10, the Lorenz system has an attractor that appears in numerical studies to be aperiodic and have an extremely complicated geometric structure. The Lorenz attractor is a prototype for a *chaotic* and *strange* set.

Informally, the term *chaos* means effectively unpredictable long-time behavior in a deterministic dynamical system because of sensitivity to initial conditions. To formulate this mathematically, we have to give precise meanings to "unpredictable" and "sensitive dependence." Each of these terms has several possible mathematical definitions that more or less capture the concept and are more or less easy to verify and to compute.

7.1 ▪ Chaos

A dynamical system is "chaotic" on a given invariant set X for a flow φ when it satisfies certain properties. Thus to apply this concept, we must first identify an invariant set. Of course, X could be a very small set in the phase space (even one of zero measure), and then the assertion of chaos on X would not necessarily be of much practical importance.

The least restrictive definition of sensitive dependence is that nearby trajectories eventually separate:

> ▷ *Sensitive dependence on initial conditions*: A flow φ exhibits sensitive dependence on an invariant set X if there is a fixed r such that for each $x \in X$ and any $\varepsilon > 0$, there is a nearby $y \in B_\varepsilon(x) \cap X$ such that $|\varphi_t(x) - \varphi_t(y)| > r$ for some $t \geq 0$.

The dynamics of a system with sensitive dependence is difficult to predict: no matter how precisely an initial condition is specified, any small error may lead to a large one (at least of size r) after enough time. Sensitive dependence does not guarantee that the error will grow, just that there exists a nearby point with this property. A system with sensitive dependence is difficult to simulate on a computer, since a small error, such as that arising from representing a real number in floating point, may eventually give rise to a "big" error—the practical questions are, of course, how long does this take and how often does it occur?

However, a system with sensitive dependence alone does not necessarily behave in a complicated way.

Example 7.1. A linear system $\dot{x} = Ax$ exhibits sensitive dependence on the invariant set $X = \mathbb{R}^n$ if any of the eigenvalues of A have a positive real part. Indeed, since the system is linear, the distance between any two points obeys the same equation. If $y = x + \varepsilon v$, where v is an eigenvector that corresponds to an unstable eigenvalue, then $|\varphi_t(y) - \varphi_t(x)| = \varepsilon\,|v|\,e^{\mathrm{Re}(\lambda)t}$. This sensitive dependence is connected to the fact that the motion is unbounded. ■

Example 7.2. Let $(\theta, y) \in \mathbb{S}^1 \times \mathbb{R}^1$ be a point on the cylinder and consider the ordinary differential equation (ODE)

$$\begin{aligned} \dot{\theta} &= y, \\ \dot{y} &= 0. \end{aligned} \tag{7.1}$$

For this system the flow is $\varphi_t(\theta, y) = (\theta + ty \bmod 2\pi, y)$. Let X be the invariant annulus $X = \{(\theta, y) : a < y < b\}$. Consider any ball of radius $\varepsilon > 0$ about a point in X. Since various points in the ball have different y values, they will move at different speeds, and $|\varphi_t(\theta, y + \varepsilon) - \varphi_t(\theta, y)| = \varepsilon t$ providing $\varepsilon t < \pi$. Therefore for any $r < \pi$, these two trajectories will spread apart by a distance r at $t = r/\varepsilon$. Notwithstanding this sensitivity, the system (7.1) does not have complicated motion and certainly would not merit the designation "chaotic." ■

We can also define more stringent notions of sensitivity. One option is to insist that sensitive dependence be replaced by "positive Lyapunov exponents" (see §7.2). This requirement rules out the second example above.

In addition to sensitive dependence, the definition of chaos must include some notion of aperiodicity, or "wanders everywhere." The most general version of this, illustrated in Figure 7.1, is called

> ▷ *transitive*: A flow φ is topologically transitive on an invariant set X if for every pair of nonempty, open sets $U, V \subset X$ there is a $t > 0$ such that $\varphi_t(U) \cap V \neq \emptyset$.

It is interesting that this definition implies that there is a point whose orbit is dense in X—this is called the Birkhoff transitivity theorem.

Theorem 7.3 (Birkhoff Transitivity). *A flow φ is transitive on a compact invariant set X if and only if φ has an orbit that is dense on X.*

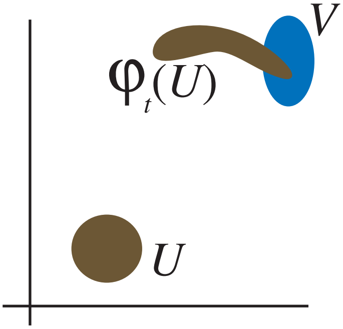

Figure 7.1. *Transitivity.*

Proof. First suppose φ has a dense orbit, $\Gamma(x)$, and let $U, V \subset X$ be any two open sets. For each $u \in U$ there is an $\varepsilon > 0$ such that $B_\varepsilon(u) \cap X \subset U$, and correspondingly for each $v \in V$ there is an $\varepsilon' > 0$ so that $B_{\varepsilon'}(v) \cap X \subset V$. Since $\Gamma(x)$ is dense there are times t, t' for which $\varphi_t(x) \in B_\varepsilon(u)$ and $\varphi_{t'}(x) \in B_{\varepsilon'}(v)$. Thus $\varphi_{t'-t}(U) \cap V \neq \emptyset$.

Conversely suppose that φ is transitive on X. Since X is compact, for any $\varepsilon > 0$ there is a cover of X by finitely many balls $V_i = B_\varepsilon(x_i)$, $i = 1, 2, \ldots N$. By assumption, for any open set U, there is a t_1 such that $\varphi_{t_1}(U) \cap \mathrm{int}(V_1) \neq \emptyset$. Let $U_1 = \varphi_{-t_1}(\mathrm{int}(V_1)) \cap U$. Again by transitivity there is a t_2 such that $\varphi_{t_2}(U_1) \cap \mathrm{int}(V_2) \neq \emptyset$, so that we can let $U_2 = \varphi_{-t_2}(\mathrm{int}(V_2)) \cap U_1 \subset U_1 \subset U$. Repeating this gives a nested sequence $U_N \subset U_{N-1} \subset \ldots \subset U_1 \subset U$ of nonempty sets, for which $\varphi_{t_i}(U_N) \cap V_i \neq \emptyset$, $i = 1, 2, \ldots, N$. Since $\{V_i\}$ cover X and ε was arbitrary, this implies there is a dense orbit. □

A related concept, *ergodicity*, applies to systems that have an invariant measure, such as Hamiltonian or volume preserving systems (see Chapter 9). An invariant set is ergodic if every invariant subset has either full or zero measure.

Example 7.4. Perhaps the simplest example of a system of ODEs with a transitive flow is

$$\begin{aligned} \dot{\theta}_1 &= 1, \\ \dot{\theta}_2 &= \nu, \end{aligned} \tag{7.2}$$

where $(\theta_1, \theta_2) \in \mathbb{T}^2$ and ν is irrational. We will show that every orbit of this system is dense in $\mathbb{T}^2$; that is, the ω-limit set of this arbitrary initial point is $\omega(\theta) = \mathbb{T}^2$. The flow for this system,

$$\varphi_t(\theta_1, \theta_2) = (\theta_1 + t \mod 2\pi, \theta_2 + \nu t \mod 2\pi), \tag{7.3}$$

is complicated only because of the mod 2π operations. The orbit of a point θ is dense if for any point $\alpha = (\alpha_1, \alpha_2)$ and any $\varepsilon > 0$, there is a time t such that

$$|\varphi_t(\theta) - \alpha| < \varepsilon.$$

To prove this, note that (7.3) implies there is an infinite sequence of times,

$$t_n = \alpha_1 - \theta_1 + 2\pi n, n \in \mathbb{Z},$$

at which $\theta_1(t_n) = \alpha_1$. At these times, the vertical component is at

$$\theta_2(t_n) = \theta_2 + \nu(\alpha_1 - \theta_1) + 2\pi\nu n \mod 2\pi.$$

To complete the proof, it is sufficient to show that for any $\varepsilon > 0$ there is an n such that $|\theta_2(t_n) - \alpha_2| < \varepsilon$, or equivalently that there is an integer n such that $\delta_n = |\nu n \bmod 1 - x| < \varepsilon$ with $x = \frac{1}{2\pi}(\alpha_2 - \theta_2 - \nu(\alpha_1 - \theta_1)) \bmod 1$. Since α_2 is arbitrary, then so is $x \in [0,1)$.

We start with a simple fact of elementary number theory (Hardy and Wright 1979, §11.3), sometimes called the pigeonhole principle after the English phrase that denotes staff mailboxes in an office.

Lemma 7.5 (pigeonhole principle). *If ν is irrational, then for any $\varepsilon > 0$ there is an integer q such that $|\nu q \bmod 1| < \varepsilon$.*

Proof. Choose $Q \in \mathbb{N}$ such that $Q > 1/\varepsilon$, and consider the set of $Q+1$ numbers $\left\{a_j \equiv \nu j \bmod 1 : j = 0, 1, \ldots, Q\right\} \subset [0,1)$. Note that the values a_j are distinct because ν is irrational. Indeed, if it were the case that $a_{j_1} = a_{j_2}$, then $\nu(j_1 - j_2) = p \in \mathbb{Z}$ so that ν would be rational. The interval $[0,1)$ is covered by the Q disjoint subintervals $[k/Q, (k+1)/Q)$ for $k = 0, 1, \ldots, Q-1$. Since there are $Q+1$ distinct a_j, there must be at least one subinterval that contains more than one of them. Hence, there are integers j_1, j_2 such that $0 < a_{j_1} - a_{j_2} < \varepsilon$ or equivalently that $0 < \nu(j_1 - j_2) \bmod 1 < \varepsilon$. Set $q = j_1 - j_2$. $\square$

To show that there is an n such that $\delta_n < \varepsilon$, select q as in Lemma 7.5 so that $a_q = \nu q \bmod 1 < \varepsilon$. The neighboring points of the sequence $m a_q : m \in \mathbb{N}$ differ by an amount smaller than ε; therefore, there is an m such that $\left|m a_q - x\right| < \varepsilon$. Set $n = mq$.

We conclude that there is a time for which the orbit of (θ_1, θ_2) is arbitrarily close to any other point in $\mathbb{T}^2$; therefore, the orbit is dense. As a consequence, the flow of (7.2) is transitive. ■

The main ingredients of chaos, therefore, are sensitive dependence and transitivity. We will insist that the invariant set X be bounded so that the sensitivity is not simply due to escape to infinity. Finally, it is also necessary to require that X be closed to ensure that chaos is a topological invariant. Putting these together provides the following definition:

> ▷ *Chaos*: A flow φ is chaotic on a compact invariant set X if φ is transitive and exhibits sensitive dependence on X.

Thus, a chaotic flow mixes things up and is hard to predict. Although this definition is reasonably useful, it is also important to note that the term "chaos" in the literature is used with many definitions; some researchers simply use the loose sense we first discussed, and some require stronger conditions than sensitive dependence. While it is clear from the quotes at the beginning of this chapter that Poincaré and Lorenz had clear notions of sensitive dependence, Li and Yorke first gave a mathematical definition of chaos in 1975; the definition that we use is due to Auslander and Yorke (1980). However, all definitions include elements comparable to transitivity and sensitive dependence; see (Blanchard et al. 2002; Devaney 1986; Robinson 1999; Wiggins 2003).

Our definition of chaos is topological and so it is preserved by conjugacy.

Theorem 7.6. *Suppose the flow $\varphi_t : X \to X$ is chaotic on the compact set X. Then if the*

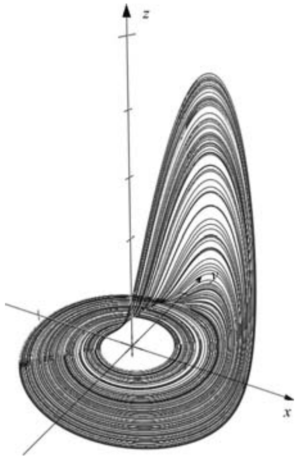

Figure 7.2. *Attractor for the Rössler system* (7.4) *with* $a = b = 0.2$ *and* $c = 5.7$.

flow $\psi_t : Y \to Y$ *is is conjugate to* φ*, it is chaotic on* Y.

Proof. Recall from §4.7 that two flows are conjugate if there exists a homeomorphism $h : X \to Y$ such that $\psi_t \circ h = h \circ \varphi_t$. We first show that transitivity is preserved by the conjugacy. Suppose $I, J \subset X$, and define $U = h(I) \subset Y$ and $V = h(J) \subset Y$. Since φ is topologically transitive, there is a t such that $\varphi_t(I) \cap J \neq \emptyset$. Consequently, $h(\varphi_t(I)) \cap h(J) \neq \emptyset$. Conjugacy then implies that $h(\varphi_t(I)) = \psi_t \circ h(I) = \psi_t(U)$ so that $\psi_t(U) \cap V \neq \emptyset$.

To show sensitive dependence of the flow ψ, the homeomorphism h must be *uniformly* continuous (recall §3.1). This is where the assumption of compactness is used: any continuous function on a compact space is uniformly continuous (recall Exercise 3.2). Thus for any $x, x' \in X$ and any given $\varepsilon > 0$ there is an ε' such that $|x - x'| < \varepsilon' \Rightarrow |h(x) - h(x')| < \varepsilon$, independently of the choice of points. The inverse h^{-1} is uniformly continuous as well; thus given any r' there is an r such that $|h(x) - h(x')| < r \Rightarrow |x - x'| < r'$ or conversely $|x - x'| \geq r' \Rightarrow |h(x) - h(x')| \geq r$.

Now consider points $x, x' \in X$ and their images $y = h(x)$, $y' = h(x') \in Y$. For any $\varepsilon > 0$, uniform continuity implies there is an ε' such that $|x - x'| < \varepsilon' \Rightarrow |y - y'| < \varepsilon$. Sensitive dependence implies that there exists an r' and a t such that $|\varphi_t(x) - \varphi_t(x')| \geq r'$. Finally, uniform continuity of h^{-1} implies there is an r such that $|h \circ \varphi_t(x) - h \circ \varphi_t(x')| = |\psi_t(y) - \psi_t(y')| \geq r$. Thus for any $\varepsilon > 0$, we have found the r and t for sensitive dependence of ψ. □

It is surprisingly difficult to give a simple example of a flow that can be proved to be chaotic. However, many systems give the appearance of chaos when solved numerically. As we mentioned in §4.10, Warwick Tucker used rigorous computations to verify that the Lorenz system (4.26) with $r = 28$ has a chaotic attractor (Tucker 2002).

Example 7.7. Another simple system that is a prototype for chaotic motion is the Rössler system (Rössler 1976):

$$\begin{aligned} \dot{x} &= -y - z, \\ \dot{y} &= x + ay, \\ \dot{z} &= b + z(x - c). \end{aligned} \tag{7.4}$$

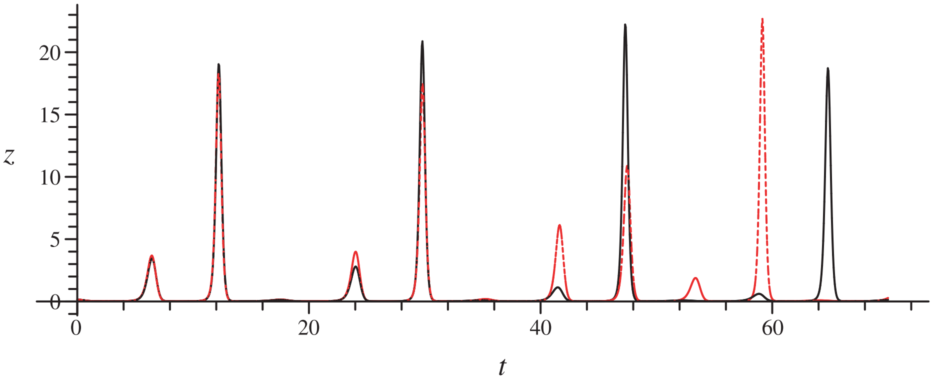

Figure 7.3. *Plot of $z(t)$ for the Rössler system (7.4) for two initial conditions with y values differing by* 0.1*. At $t = 24$, the z values differ by* 1*, and near $t = 60$ they differ by more than* 20.

This system undergoes a complex sequence of "bifurcations" (see Chapter 8) as the parameters are changed and often has an apparently fractal, chaotic attractor (Alligood, Sauer, and Yorke 1997, §9.3). Rössler primarily studied the case $a = b = 0.2$, varying the parameter c. The attractor for $c = 5.7$ is shown in Figure 7.2. This system exhibits sensitive dependence; nearby initial conditions trace out a qualitatively similar set, but their z components undergo large-amplitude excursions at different times; see Figure 7.3. ■

7.2 ▪ Lyapunov Exponents

The concept of "sensitive dependence" requires that nearby orbits eventually separate; however, the rate of separation is not specified. As we saw in the previous section, there are systems whose trajectories separate exponentially in time (a linear hyperbolic flow) and those that separate at a polynomial rate (like the flow of (7.1) on the cylinder). Indeed, there appears to be a dichotomy between systems for which nearby orbits separate linearly and "truly chaotic" systems whose orbits separate exponentially. We say that two orbits separate exponentially when

$$|\varphi_t(y) - \varphi_t(x)| \sim c e^{\lambda t}$$

for some $\lambda > 0$. However, since chaotic dynamics must take place on a compact invariant set, this exponential growth cannot continue forever. To get around this difficulty, we will require only that infinitesimally close orbits separate exponentially. This concept is familiar from our study of the linearization of equilibria: when the Jacobian matrix of the vector field at an equilibrium has a positive eigenvalue, then the linearized system has trajectories that grow exponentially. Although the linearization applies only when trajectories are formally infinitesimally close to the equilibrium, it indicates that nearby trajectories do indeed separate more or less exponentially while they remain close to the equilibrium. To apply this criterion more generally we will need a notion of linear stability for an arbitrary orbit that will allow us to compute the analogue of the eigenvalues of an equilibrium.

However, there is a problem that we must address before we can simply linearize an ODE about an arbitrary trajectory. The linearization method in §4.4 requires that

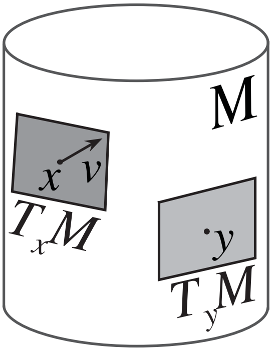

Figure 7.4. *Tangent spaces to a cylinder at two points x and y.*

we compute the stability of an orbit by finding the eigenvalue of some matrix. We obtained such a matrix for an equilibrium by computing the Jacobian, Df, of the vector field at the equilibrium or the monodromy matrix obtained from Floquet theory for a periodic orbit; recall §2.8. For an aperiodic orbit, neither of these constructions will work. However, it is still useful to study the motion of nearby orbits using the linearized vector field, Df, but now this gives rise to a matrix $A(t)=Df(\varphi_t(x))$ that depends on time in an intricate way.

More formally, focus on a particular orbit, $\varphi_t(x_o)$, of a flow φ on an n-dimensional phase space M. We will call this the *fiducial* trajectory.[38] To study orbits near to $\varphi_t(x_o)$ we will linearize the vector field about this orbit. For each point $x \in M$, let T_xM denote the collection of "infinitesimal" vectors attached to x, that is, the *tangent vectors* at x; see Figure 7.4. The tangent space at x is an n-dimensional vector space, isomorphic to $\mathbb{R}^n$. Even though the vectors are formally infinitesimal, as far as the linearization is concerned their lengths can be thought of as arbitrary. The vector space character of T_xM holds even if the phase space has a more complicated topology—that of a cylinder or torus, for example. If $v \in T_xM$ is such a vector, then the collection of all such vectors at all points $x \in M$ is called the *tangent bundle* of M and is denoted

$$TM = \{(x,v) : x \in M, v \in T_xM\}.$$

Note that TM has dimension $2n$, twice that of M.

Consider a trajectory $\varphi_t(x_o+\varepsilon v_o)$ that starts near the fiducial point x_o. Expanding the C^1 flow in ε gives $\varphi_t(x_o+\varepsilon v_o)=\varphi_t(x_o)+\varepsilon D_x\varphi_t(x_o)v_o+o(\varepsilon)$, which implies that the initial deviation vector v_o evolves to

$$v(t)=D_x\varphi_t(x_o)v_o. \tag{7.5}$$

Substituting the expansion into the ODE for φ and assuming that the vector field, f, is also C^1 gives

$$\frac{d}{dt}(\varphi_t(x_o)+\varepsilon v(t))=f(\varphi_t(x_o))+\varepsilon Df(\varphi_t(x_o))v(t)+o(\varepsilon).$$

The first terms on the left- and right-hand sides cancel, so that, to first order in ε,

$$\dot{v}=Df(\varphi_t(x_o))v \equiv A(t)v. \tag{7.6}$$

[38] *Fiducial* is from the Latin word *fiducialis*, "trust"; in this case it means the "standard for comparison."

The Jacobian matrix $Df(\varphi_t(x_o))$ can be thought of as a linear operator that acts on a vector $v(t) \in T_{\varphi_t(x_o)}M$ to give the velocity of $y(t) = \varphi_t(x_o) + \varepsilon v(t)$, relative to that of $x(t)$, when ε is infinitesimally small. The time dependence of $A(t)$ in (7.6) is fixed by the fiducial trajectory.

Since (7.5) and (7.6) hold for any initial vector v_o, the fundamental matrix solution of (7.6) is

$$\Phi(t; x_o) = D_x \varphi_t(x_o). \tag{7.7}$$

It obeys the system[39]

$$\dot{\Phi} = A(t)\Phi, \qquad \Phi(0; x_o) = I. \tag{7.8}$$

Moreover, given *any* initial vector v, the solution to (7.6) is $\Phi(t; x_o)v$. The fundamental matrix Φ is a linear operator that takes a tangent vector at the initial point and gives a tangent vector at the point $\varphi_t(x_o)$:

$$\Phi(t; x_o) : T_{x_o}M \to T_{\varphi_t(x_o)}M. \tag{7.9}$$

If $\varphi_t(x_o)$ is periodic with period T, then the $\varphi_T(x_o) = x_o$, so the (monodromy) matrix $\Phi(T; x_o)$ is a map from $T_{x_o}M$ to itself. It now makes sense to compute the eigenvalues of $\Phi(T; x_o)$; this is what we did in §2.8 to find the Floquet exponents. By contrast, when the fiducial trajectory is aperiodic, the domain and range of Φ are distinct spaces so an equation of the form $\lambda v = \Phi v$ does not make sense.

Definition

In the same spirit as the Floquet exponents, Aleksandr Lyapunov defined his exponents as the asymptotic growth rate of the length of tangent vectors $v(t)$:

$$|\Phi(t; x)v| \sim e^{\mu t} |v|.$$

How do we know that this exponential estimate is appropriate? When an orbit is contained in a compact set and $f \in C^1$, the Jacobian Df is uniformly bounded. Since chaos is defined only for compact invariant sets, this bound is quite natural. Moreover, if the Jacobian is uniformly bounded on the orbit, it is easy to see that the growth of any vector is at most exponential.

Lemma 7.8. *Suppose $\Phi(t; x)$ is the fundamental matrix solution of (7.8) and $||A(t)|| \le K$ for all $t \ge 0$. Then for any v there are positive constants c and c' such that*

$$c' e^{-Kt} \le |\Phi(t; x)v| \le c e^{Kt}$$

for all $t \ge 0$.

Proof. Using (7.8) and the fact that $w^T A w \le ||A|| \, |w|^2$,

$$\frac{d}{dt} |\Phi(t; x)v|^2 = \frac{d}{dt} \left(v^T \Phi^T \Phi v \right) = v^T \Phi^T \left(A^T + A \right) \Phi v \le 2 ||A|| \, |\Phi v|^2 .$$

Thus

$$\frac{d}{dt} e^{-Kt} |\Phi(t; x)v| \le e^{-Kt} (||A(t)|| - K) |\Phi v| \le 0.$$

[39] Although we should allow for the initial condition at an arbitrary time t_o (recall (2.46)), we always impose the initial condition at $t = 0$ in this section.

Consequently, $|\Phi(t;x)v| \le ce^{Kt}$ for some constant c and all $t \ge 0$. Similarly, defining $\tau = -t$, then $d\Phi/d\tau = -A(-\tau)\Phi$, then since $||-A|| = ||A||$, the same inequality,

$$\frac{d}{d\tau} e^{-K\tau} |\Phi(-\tau;x)v| \le 0,$$

holds for any $\tau \le 0$. Replacing τ by $-t$ then implies that $e^{Kt}|\Phi v|$ is a nondecreasing function of t, giving the second half of the promised inequality. □

According to Lemma 7.8, the function $\ln|\Phi v|/t$ is bounded both above and below for $t > 0$. Since any bounded sequence has limit points we may define the *Lyapunov spectrum*[40] as the set of limit points of this function:

$$Sp(x,v) = \left\{ \lambda = \lim_{j\to\infty} \frac{1}{t_j} \ln\left|\Phi(t_j;x)v\right| \text{ for some sequence } t_j \underset{j\to\infty}{\to} \infty \right\}. \tag{7.10}$$

As indicated in (7.10), this spectrum may depend upon both the choice of fiducial trajectory and upon the initial deviation. There are two privileged limits for a bounded sequence, the "liminf" and the "limsup." The latter is defined to be the least upper bound of the tail of the sequence,

$$\limsup_{t\to\infty} s(t) \equiv \lim_{T\to\infty} \left(\sup_{t>T} s(t) \right).$$

When $s(t)$ is bounded, the quantity $\sup_{t>T} s(t)$ exists, is nonincreasing, and bounded below, so the limsup always exists. The liminf is similarly the limit of the infimum and also exists for bounded sequences. Note that all other limit points of a bounded sequence must be between these, and since the functions that we consider are continuous, every possible value between these two limits must occur. Thus the Lyapunov spectrum $Sp(x,v)$ is a closed interval and degenerates to a point when the two limits are equal.

Example 7.9. The simple linear one-dimensional ODE

$$\dot{v} = (\cos(\ln|t|) + \sin(\ln|t|))\, v$$

has the general solution $v(t) = \exp(t\sin(\ln|t|))\, v_0$ (ignoring the fact that the vector field is not defined at $t = 0$). For this system the fundamental matrix is simply the scalar $\exp(t\sin(\ln|t|))$, and the Lyapunov spectrum is

$$\left\{ \lim_{j\to\infty} \frac{1}{t_j} t_j \sin(\ln|t_j|) \right\} = [-1,1]. \quad \blacksquare$$

The largest growth rate is most often of interest; consequently it is useful to define the *Lyapunov exponent* as the supremum limit:

$$\mu(x,v) \equiv \limsup_{t\to\infty} \frac{1}{t} \ln|\Phi(t;x)v|. \tag{7.11}$$

[40] Note that this use of the word *spectrum* is different from the spectrum of a matrix, which is the collection of its eigenvalues.

Since this limit will occur often below, it is nice to give it a more compressed notation. For any function $f(t)$, define the *characteristic exponent* of f by

$$\chi(f) \equiv \limsup_{t\to\infty} \frac{1}{t} \ln|f(t)|. \tag{7.12}$$

Using this notation, $\mu(x,v) = \chi(\Phi(t;x)v)$.

If $Sp(x,v)$ is a point, then the limsup in (7.11) can be replaced by a simple limit. In this case the Lyapunov spectrum is termed *regular*.[41]

Properties of Lyapunov Exponents

Lyapunov showed that the characteristic exponent (7.12) of any function obeys several simple properties (Adrianova 1995). For any scalar functions $f(t)$ and $g(t)$ and any constant $c \neq 0$ (see Exercise 7),

$$\chi(cf) = \chi(f), \tag{7.13}$$

$$\chi(f+g) \le \max(\chi(f), \chi(g)), \tag{7.14}$$

$$\chi(fg) \le \chi(f) + \chi(g). \tag{7.15}$$

In particular (7.13) implies that, though the Lyapunov exponent may depend on the direction of the initial vector, it does not depend on its length. More generally, we have the following lemma.

Lemma 7.10. *The Lyapunov exponent* (7.11) *is independent of the choice of norm on* $\mathbb{R}^n$.

Proof. Let $|v|_1$ and $|v|_2$ represent two different vector norms[42] and μ_1, μ_2 be the corresponding Lyapunov exponents. As is well known, all norms on $\mathbb{R}^n$ are compatible, meaning that there exist constants s and $S > 0$ such that for every vector v,

$$s\,|v|_1 \le |v|_2 \le S\,|v|_1 .$$

Consequently,

$$\chi\left(s\,|\Phi(t;x)v|_1\right) \le \chi\left(|\Phi(t;x)v|_2\right),$$

which implies, using (7.13), that $\mu_1 \le \mu_2$. Using the same result for the upper bound gives $\mu_2 \le \mu_1$ so that we can conclude that $\mu_1 = \mu_2$. □

Since the sup-norm is one possible choice, Lemma 7.10 implies that the characteristic exponent of a vector $v(t)$ is given by $\chi(v) = \max_{1\le i\le n} \chi(v^{(i)})$, where $v^{(i)}$ are the components of v (see Exercise 7).

The Lyapunov exponent is invariant under the flow, since

$$\mu(x,v) = \mu\left(\varphi_t(x), \Phi(t;x)v\right)$$

for any t (see Exercise 4), so we can associate the exponent with an orbit, rather than just an initial condition.

An orbit has at most n distinct Lyapunov exponents.

[41] When there is an invariant measure, Oseledec's *multiplicative ergodic theorem* implies that the spectrum is regular for almost all initial points; see (Robinson 1999).

[42] For example, the Euclidean and sup-norms.

Lemma 7.11. *If $\varphi_t(x)$ is a bounded trajectory of a C^2 flow φ on an n-dimensional manifold, then it has at most n distinct Lyapunov exponents* (7.11).

Proof. Since $\varphi_t(x)$ is bounded and C^2, the Jacobian $Df(\varphi_t)$ is bounded, and thus the limit (7.11) exists for each v. Suppose, for example, there are two different exponents $\mu_1 > \mu_2$ for linearly independent vectors v_1 and v_2. Then since (7.6) is linear, the length of any linear combination $v = \alpha v_1 + \beta v_2$ grows asymptotically at the rate μ_1, provided only that $\alpha \neq 0$; see Exercise 7. Since there are n linearly independent vectors in T_xM, there are at most n distinct values, μ_i. □

Just as constant matrices may have degenerate eigenvalues, a time-dependent matrix may not have n distinct Lyapunov exponents. It is conventional to order the exponents so that

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n. \tag{7.16}$$

Any set of independent vectors $\{v_1, v_2, \ldots, v_n\}$ so that

$$\sum_{i=1}^{n} \mu_i(x, v_i)$$

is as small as possible is called a *Lyapunov basis*. If a Lyapunov exponent has degeneracy k, then its corresponding basis vectors span a k-dimensional subspace. Most bases are not Lyapunov bases, since each vector generally will contain some component along the most rapidly growing direction (see Exercise 7); however, a Lyapunov basis can always be constructed.

Lemma 7.12 (Lyapunov basis). *If $\Phi = [v_1, v_2, \ldots, v_n]$ is any fundamental matrix solution of* (7.6) *obeying* (7.16), *then there is a special upper triangular matrix U ($u_{ii} = 1$) such that ΦU is a Lyapunov basis.*

Proof. The columns w_i of ΦU are $w_i = v_i + \sum_{j=1}^{i-1} u_{ji} v_j$. Consequently $\chi(w_1) = \chi(v_1)$, and using (7.14) $\chi(w_2) = \chi(v_2 + u_{12} v_1) \leq \max(\chi(v_1), \chi(v_2))$. We choose u_{12} to minimize $\chi(w_2)$. There are three cases. First, if $\chi(v_2) < \chi(v_1)$, then we set $u_{12} = 0$, so that $\chi(w_2) = \chi(v_2)$ If however $\chi(v_2) > \chi(v_1)$ then $\chi(w_2) = \chi(v_2)$ for any choice of u_{12}. Finally if $\chi(v_2) = \chi(v_1)$ then there may be a choice of u_{12} for which $\chi(w_2) < \chi(v_2)$ or else all choices lead to equality. For each i, the u_{ji} are chosen to minimize $\chi(w_i)$. We claim that the sum $\sum_{i=1}^{n} \chi(w_i)$ is minimal. Indeed, the exponent of any linear combination $\sum_{i=1}^{k} a_i w_i$with $a_k \neq 0$ is the exponent of the combination $a_k v_k + \sum_{j=1}^{k-1} b_j v_j$ for some coefficients b_j depending upon a and u_{ji}. However, the minimal combination of these first k vectors was already selected. □

An orbit almost always has one "trivial" Lyapunov exponent.

Lemma 7.13. *If $\varphi_t(x_o)$ is a bounded orbit of the flow φ that is not forward asymptotic to an equilibrium, then it has a zero Lyapunov exponent.*

Proof. Consider the vector $v(t) = f(\varphi_t(x_o))$, where f is the vector field for φ. Differentiation gives

$$\frac{d}{dt} v(t) = \frac{d}{dt} f(\varphi_t(x_o)) = Df(\varphi_t(x)) \frac{d}{dt} \varphi_t(x_o) = Df(\varphi_t(x_o)) v(t).$$

Thus $v(t)$ is a solution of (7.6) with initial condition $v_o = f(x_o)$. Since φ is bounded, then v is also bounded. Finally since φ_t is not asymptotic to an equilibrium, $\limsup |v(t)| > 0$. Therefore $\mu(x_o, v_o) = 0$. □

Example 7.14. Consider the one-degree-of-freedom Hamiltonian system (5.3). Each orbit inside the separatrix loop is bounded and periodic. Therefore, the tangent vector $v(t) = f(\varphi_t(x))$ is also periodic and thus has zero Lyapunov exponent.

By contrast, any orbit on the separatrix $y^2 = x^2(1-2ax)$ with $x > 0$ is asymptotic to the origin, and $v(t) \to f(0) = 0$ as $t \to \infty$. Moreover, as the orbit approaches the origin it aligns with the linear stable set $E^s = \operatorname{span}(1,-1)^T$ and since $Df(0)|_{E^s} = -1$, the tangent vector approaches $v(t) \to ce^{-t}(1,-1)^T$ for some constant c that depends upon the initial point; consequently its characteristic exponent is $\chi(v) = -1$. ■

A constraining relation among the Lyapunov exponents can be obtained from Abel's theorem (2.50),

$$\det(\Phi(t;x)) = \exp \int_0^t \operatorname{tr} Df(\varphi_s(x))\,ds. \tag{7.17}$$

Theorem 7.15 (Lyapunov). *Suppose $\varphi_t(x)$ is a bounded orbit of a flow φ and $[v_1, v_2, \ldots, v_n]$ is an independent set of vectors with Lyapunov exponents $\mu_i = \mu(x, v_i)$. If the limit*

$$\delta = \limsup_{t\to\infty} \frac{1}{t} \int_0^t \operatorname{tr} Df(\varphi_s(x))\,ds \tag{7.18}$$

exists, then

$$\delta \le \sum_{i=1}^n \mu_i. \tag{7.19}$$

Proof. Let $P(t) = [v_1, v_2, \ldots, v_n] = \Phi(t;x)P(0)$; then according to (7.17) and (7.18),

$$\delta = \chi(\det\Phi(t;x)) = \chi(\det P(t)).$$

The determinant of an $n \times n$ matrix is the sum of $n!$ terms, each of which is the product of n different elements of the matrix, one from each column. Using (7.14) and (7.15) and Lemma 7.10 gives $\chi(\det P) \le \sum_{j=1}^n \max_{1\le i\le n} \chi(P_{ij}) = \sum_{i=1}^n \chi(v_i)$. □

We showed in §4.7 that if two flows are diffeomorphic, then the spectra at corresponding equilibrium points must be the same. This is also true for the Lyapunov spectrum.

Lemma 7.16. *Suppose that the flows φ and ψ are conjugate under a diffeomorphism h, such that Dh and Dh^{-1} are uniformly bounded. Then the Lyapunov exponents for ψ at $h(x)$ are the same as those for φ at x.*

Proof. Let $v(t)$ be given by (7.5) for an initial vector v_o based at x. We will show that the Lyapunov exponent $\mu_\varphi(x, v_o)$ is the same as the exponent $\mu_\psi(y, w_o)$ for the vector $w_o = Dh(x)v_o$ based at the point $y = h(x)$ under the flow ψ. Differentiating the conjugacy relation $h \circ \varphi_t = \psi_t \circ h$ with respect to x gives $Dh(\varphi_t)D\varphi_t(x) =$

$D\psi_t(h(x))Dh(x)$. Consequently,

$$Dh(\varphi_t)v(t) = D\psi_t(y)Dh(x)v_o = D\psi_t(y)w_o = w(t).$$

By the uniformly bounded assumptions on the derivatives, there are positive constants s and S such that

$$s\,|v| \le |Dh(x)v| \le S\,|v| \Rightarrow \ln s \le \ln|Dh(x)v| - \ln|v| \le \ln S.$$

Therefore

$$\mu_\psi(h(x), w_o) = \chi(w) = \chi\left(Dh(\varphi_t(x))v(t)\right) = \chi\left(v\right) = \mu_\varphi(x, v_o). \quad \square$$

It would be nice if the existence of positive Lyapunov exponents for an invariant set implied that it has sensitive dependence as defined in §7.1. However, this is not the case.

Example 7.17. The separatrix loop, L, of the Hamiltonian system (5.3) has one negative Lyapunov exponent—for the tangent to L—as we noted in the previous example. However, by the same argument, the Lyapunov exponent for any vector transverse to L will be the positive eigenvalue of $Df(0)$, namely $\mu = +1$. However, if we consider L as the invariant set, then L does not have sensitive dependence. Consider any two points $x, y \in L$ for which $|x - y| < \varepsilon$. Since $\varphi_t(x)$ and $\varphi_t(y)$ are both asymptotic to 0, there is a fixed time T such that both $\varphi_t(x)$ and $\varphi_t(y)$ are in $B_r(0)$ for all $t > T$. Thus, if the orbits were to diverge, they must do so for $t \in [0, T)$. However, recall from Grönwall's lemma that nearby initial conditions have bounded divergence, (3.33),

$$|\varphi_t(x) - \varphi_t(y)| \le |x - y|\, e^{Kt},$$

where K is the Lipschitz constant for the vector field f. Hence if $\varepsilon < re^{-KT}$, then these trajectories will never diverge by a distance r.

Thus, as an invariant set L does not have sensitive dependence, even though it has a (transverse) positive Lyapunov exponent. ■

This example is exotic (homoclinic orbits are not generic, as we will see in Chapter 8), and in any practical sense, the existence of positive Lyapunov exponents for an invariant set is a reliable indicator of sensitive dependence. The main barrier to their use is the difficulty in devising accurate computational algorithms.

Computing Exponents

To compute the maximal Lyapunov exponent of a system of ODEs we must integrate both the original system and its linearization (7.6). Essentially any initial vector v_o can be used because almost all vectors will have some component along the direction of the maximal Lyapunov direction. We cannot compute the limit in (7.11) but instead simply integrate for some "long" time T and estimate

$$\mu_{\max}(T) \approx \frac{1}{T} \ln \frac{|v(T)|}{|v_o|}. \tag{7.20}$$

With luck, this quantity will rapidly converge to the maximal exponent; to estimate the error in the computation, it is useful to plot $\mu_{\max}$ as a function of T.

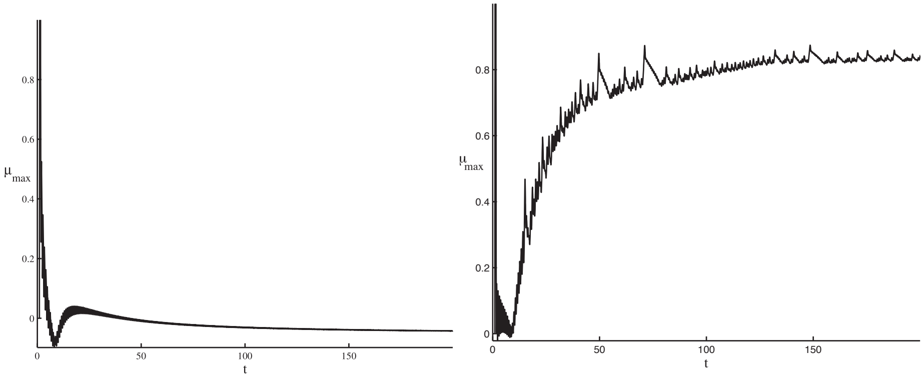

Figure 7.5. *Maximal Lyapunov exponent for the Lorenz system with* $\sigma = 10$ *and* $b = 8/3$*. Left,* $r = 23$ *and* $\mu_{\max} \approx -0.05$*; right,* $r = 28$*, where* $\mu_{\max} \approx 0.9$.

Example 7.18. Consider the Lorenz system (4.26). The linearized equations for a vector $v \in T_x\mathbb{R}^3$ are

$$\dot{v} = \begin{pmatrix} -\sigma & \sigma & 0 \\ r - z & -1 & -x \\ y & x & -b \end{pmatrix} v. \tag{7.21}$$

To integrate these equations, we must simultaneously integrate the Lorenz system itself; a simple algorithm to do this and to compute (7.20) is given in the appendix. A plot of the short time behavior of $\mu_{\max}(T)$ is shown in Figure 7.5 for two values of r. For the standard parameter values, $\mu_{\max}(T)$ appears to (slowly) converge to a positive value; integrating to $t = 1000$ gives $\mu_{\max} \approx 0.88$, and integrating for the longer time $t = 10^4$ gives $\mu_{\max} \approx 0.90$. It is difficult to compute the value accurately because the convergence is slow, though it appears that this value is correct to two places. ∎

Even though only the largest Lyapunov exponent for the Lorenz system was computed in the example, (7.19) can be used to estimate the other two exponents. For the Lorenz case, the trace of the Jacobian matrix (7.21) is constant, so that $\delta = \operatorname{tr}(Df) = -1 - \sigma - b$. Since one exponent vanishes, $\mu_2 = 0$, for the standard parameters,

$$\mu_3 \geq \delta - \mu_1 = -13.66 - \mu_1. \tag{7.22}$$

If the Lorenz system were known to be *regular*, then the supremum limits could be replaced by ordinary limits and the inequality in (7.19) would become an equality. Under this assumption $\mu_3 \approx -14.6$.

To compute all of the Lyapunov exponents, it is necessary to find a Lyapunov basis. Consider the linear system

$$\dot{v} = A(t)v, \tag{7.23}$$

where $A(t)$ is continuous and uniformly bounded. Generalizing the basis change (2.15) to account for time dependence, $v = P(t)w$, the system (7.23) becomes

$$\dot{w} = \left(P^{-1}AP - P^{-1}\dot{P}\right)w = B(t)w. \tag{7.24}$$

If the transformation P is well behaved, the characteristic exponents of the new system are the same as those of the old.

Lemma 7.19 (Lyapunov transformation). *If $P \in C^1$ and P, P^{-1}, and $\dot{P}$, are bounded for all $t > 0$, then the Lyapunov exponents of the transformed system* (7.24) *are the same as those of the original system* (7.23).

Proof. If $A(t)$ is bounded and the hypotheses hold, then $B(t)$ is bounded so its Lyapunov exponents exist. Using $||P(t)|| \le M$ and the definition $v = Pw$ gives

$$\chi(v) \le \chi\left(||P(t)|| \, |w(t)|\right) = \chi(w).$$

Applying the same analysis to $w = P^{-1}v$ and using $\left\|P^{-1}\right\| \le M$ implies that $\chi(w) \le \chi(v)$. Thus these two characteristic exponents must be equal. Since P is nonsingular, all the exponents of B must be the same as those of A. □

Just as for the case of constant matrices, where we can transform to a generalized eigenvector basis, it is always possible to find a new basis, w, such that the system (7.24) has a simple form.

Theorem 7.20 (Perron triangulation). *There is an orthogonal transformation of* (7.23) *to a basis for which B in* (7.24) *is upper triangular. Moreover, if $A(t)$ is bounded, then the characteristic exponents for B are the same as those of A.*

Proof. The fundamental matrix solution $\Phi(t;x)$ of (7.23) is nonsingular for each t, so there exists a QR factorization $\Phi = Q(t)R(t)$ into the product of an orthogonal matrix Q and an upper triangular matrix R. Let $v(t) = Q(t)w$ define a new basis for (7.23). Then since $v(t) = Q(t)R(t)v_o$, $w(t) = R(t)v_o$. Moreover, using $\dot{\Phi} = \dot{Q}R + Q\dot{R} = AQR$ in (7.24) gives

$$B = Q^T AQ - Q^T \dot{Q} = Q^T\left(\dot{Q} + Q\dot{R}R^{-1}\right) - Q^T\dot{Q} = \dot{R}R^{-1}, \tag{7.25}$$

thus, $\dot{R} = BR$. Since R is upper triangular by definition, then so is B.

Define the matrix

$$S(Q) \equiv Q^T\dot{Q}. \tag{7.26}$$

It is easy to see that S is skew-symmetric: since $Q^TQ = I$,

$$0 = \frac{d}{dt}\left(Q^TQ\right) = Q^T\dot{Q} + \dot{Q}^TQ = S + S^T.$$

Since B is upper triangular, (7.25) implies that

$$S_{ij} = \begin{cases} \left(Q^TAQ\right)_{ij}, & i > j, \\ 0, & i = j, \\ -\left(Q^TAQ\right)_{ij}, & i < j. \end{cases} \tag{7.27}$$

To show that the transformation has the same Lyapunov exponents we need only show that Q, Q^{-1}, and $\dot{Q}$ are bounded. The first two matrices are automatically bounded since Q is orthogonal: $||Q|| = 1$. By assumption $A(t)$ is bounded so that Q^TAQ is as well. Then (7.27) implies that the skew-symmetric matrix S is bounded. Consequently $\dot{Q}$ is bounded. □

When an upper triangular system is "regular," its Lyapunov exponents are easily obtained.

Theorem 7.21. *If $B(t)$ is a uniformly bounded, upper triangular matrix, and the limits*

$$\mu_i = \lim_{t\to\infty} \frac{1}{t}\int_0^t b_{ii}(s)ds \tag{7.28}$$

exist, then $\dot{x} = B(t)x$ has a regular Lyapunov spectrum with exponents μ_i.[43]

Proof. Our goal is to construct a Lyapunov basis and show that its exponents are given by (7.28). Define

$$\beta_i(t) \equiv \exp\left(\int_0^t b_{ii}(s)ds\right)$$

so that $\mu_i = \chi(\beta_i)$. The upper triangular system (7.24) can be solved by "back substitution" to give an upper triangular fundamental matrix solution. A first solution has the form $v_1 = [v_{11}, 0, \ldots, 0]^T$, where $v_{11}(t) = \beta_1(t)v_{11}(0)$. The second, $v_2 = [v_{12}, v_{22}, \ldots, 0]^T$, has $v_{22}(t) = \beta_2(t)v_{22}(0)$, and

$$\dot{v}_{12} = b_{11}v_{12} + b_{12}v_{22},$$

which has the solution

$$v_{12}(t) = \beta_1(t)\left(v_{12}(0) + \int_0^t \beta_1^{-1}(s)b_{12}(s)v_{22}(s)ds\right). \tag{7.29}$$

Continuing in this way, we obtain a fundamental matrix $P(t)$ with elements

$$v_{ij} = \begin{cases} \beta_i(t)\left(v_{ij}(0) + \int_0^t \beta_i^{-1}(s)\sum_{k=i+1}^{j} b_{ik}(s)v_{kj}(s)ds\right), & i<j, \\ \beta_i(t)v_{ii}(0), & i=j, \\ 0, & i>j. \end{cases}$$

Note that $P(0)$ is nonsingular whenever the $v_{ii}(0) \neq 0$. To construct a Lyapunov basis, we choose the initial conditions $v_{ii}(0) = 1$ and set $v_{ij}(0)$ for $i<j$ to

$$v_{ij}(0) = \begin{cases} 0, & \mu_i \le \mu_j, \\ -\int_0^\infty \beta_i^{-1}(s)\sum_{k=i+1}^{j} b_{ik}(s)v_{kj}(s)ds, & \mu_i > \mu_j. \end{cases} \tag{7.30}$$

We show this is a Lyapunov basis by induction. First, it is clear that $\chi(v_1) = \mu_1$. The characteristic exponent $\chi(v_2) = \max(\mu_2, \chi(v_{12}))$. When $\mu_1 \le \mu_2$, we use $v_{12}(0) = 0$, and then (7.29), and the results of Exercise 7, give

$$\chi(v_{12}) \le \max(\mu_1, \chi(\beta_1) + \chi(\beta_1^{-1}) + \chi(b_{12}) + \chi(v_{22})).$$

Since the limit (7.28) exists, $\chi(\beta_1) + \chi(\beta_1^{-1}) = 0$. Since $B(t)$ is bounded, $\chi(b_{ij}) = 0$. Thus $\chi(v_{12}) \le \mu_2$. When $\mu_1 > \mu_2$, we use the integral in (7.30) for $v_{12}(0)$. Substituting this into (7.29) gives

$$v_{12}(t) = -\beta_1(t)\int_t^\infty \beta_1^{-1}(s)b_{12}(s)v_{22}(s)ds.$$

[43] This theorem also applies to a more general case of "integral separation" of the diagonal elements; see Dieci and van Vleck (2002). Thus the QR method can be used to compute exponents for some irregular systems whose exponents are distinct.

The results of Exercise 7 imply that this integral converges, and that $\chi(v_{12}) \leq \mu_2$ as before. Consequently $\chi(v_2) = \mu_2$.

Proceeding inductively, suppose that $\chi(v_{ij}) \leq \mu_j$ for $i = k+1, \ldots, j$. Then

$$\chi(v_{kj}) \leq \chi(\beta_k) + \chi(\beta_k^{-1}) + \max_{k<i}\left(\chi(b_{ij})\right) + \max_{k<i}\left(\chi(v_{ij})\right) = \mu_j.$$

Consequently, $\chi(v_j) = \mu_j$.

Finally, by the inequality (7.19), the characteristic exponent $\delta = \chi(\mathrm{tr}(B))$ is a lower bound to the sum of the characteristic exponents of any fundamental set of solutions. When the limits (7.28) exist and B is upper triangular, $\delta = \sum_{i=1}^{n} \chi(b_{ii})$. Since the fundamental matrix we have constructed has $\delta = \sum_{i=1}^{n} \mu_i$, this is the minimal value. Thus $P(t)$ is a Lyapunov basis, and the exponents are μ_i. □

The QR procedure can be turned into an effective computational strategy. While it is possible to compute the QR factorization of Φ for each t, it is better to obtain a smooth factorization by solving for Q using the definition (7.26),

$$\dot{Q} = QS(Q), \tag{7.31}$$

where $S(Q)$ is given by (7.27). Since $\Phi(0;x) = I$, it is appropriate to start with the initial condition $Q(0) = I$. Knowledge of Q then allows us to easily find b_{ii} using (7.25):

$$b_{ii} = \left(Q^T AQ - S(Q)\right)_{ii} = \left(Q^T AQ\right)_{ii}.$$

The subspace $Q^T Q = I$ is an invariant subspace of (7.31) since

$$\frac{d}{dt}(Q^T Q) = Q^T QS + S^T Q^T Q = S + S^T = 0$$

when S is skew-symmetric: thus the solution of (7.31) is guaranteed to be orthogonal if $Q(0)$ is orthogonal. However, numerical errors can cause Q to drift away from orthogonality, and some care must be exercised to prevent this (Dieci et al. 2002).

7.3 ▪ Strange Attractors

A chaotic attractor can have the geometry of an "ordinary" Euclidean set, such as a plane, or of its curved analogue, a smooth submanifold of the phase space; recall §5.5. However, attractors can also be geometrically "strange," that is, fractal sets. Attractors can be strange and not chaotic, as well as chaotic but not strange. Thus attractors that have both of these properties, such as the Lorenz and Rössler attractors, are called "strange, chaotic attractors."

It is difficult to give a precise definition of a fractal. The simplest fractals are self-similar objects that have a recursive construction—for example, the Koch snowflake; see Figure 7.6.

> ▷ *Self-similar*: A set $A \subset \mathbb{R}^n$ is self-similar if it is *similar* to a part of itself: that is, there exists a strict subset $B \subset A$ and a similarity transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ such that $T(B) = A$. A transformation is a similarity if it multiplies distance by a fixed factor: $|T(x) - T(y)| = r\,|x - y|$.

A similarity is the composition of elementary scaling, rotation, reflection, and translation transformations. Not every self-similar object is a fractal; for example, a line

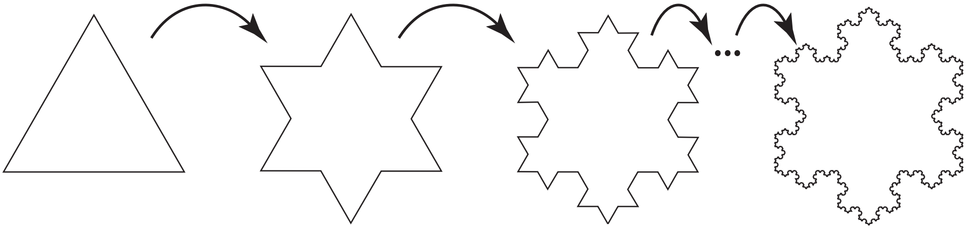

Figure 7.6. *Several levels in the construction of the Koch snowflake beginning with an equilateral triangle. At each level each straight side is replaced by four lines one-third the original size. The resulting limiting curve has infinite length, and fractal dimension* $\ln 4/\ln 3$.

segment is self-similar. However, some fractals, like the Koch snowflake, have a nontrivial scaling property. More generally, a fractal is a set that, when viewed with a microscope of arbitrarily large power, never limits to a Euclidean object—it has fine structure on every scale. This rules out submanifolds, because these sets approach hyperplanes on a small enough scale. The famous "Mandelbrot set" is a fractal by this definition: it does seem to have approximate self-similarity, but it is not strictly self-similar (Falconer 1990).

Hausdorff Dimension

Often fractals have an effective dimension (a "fractal dimension") larger than their topological dimension. One simple way to think about the dimension of a bounded set $S \subset \mathbb{R}^n$ is to ask what is the growth rate in the number of n-dimensional boxes of size ε that are needed to cover the set. Suppose that $B(\varepsilon)$ is a box with side ε, and for a given size ε *at most* $N(\varepsilon)$ boxes, $B_1(\varepsilon), B_2(\varepsilon), \ldots, B_{N(\varepsilon)}(\varepsilon)$, are needed to cover S:

$$\bigcup_{k=1}^{N(\varepsilon)} B_k(\varepsilon) \supset S.$$

If S were a submanifold of (topological) dimension d, then as $\varepsilon \to 0$,

$$N(\varepsilon) \sim \varepsilon^{-d} = e^{-d \ln \varepsilon}. \tag{7.32}$$

Note that $N \to \infty$ exponentially with $-\ln \varepsilon$; the "box counting" dimension of S is defined as the rate of divergence,

$$d_{box} = -\lim_{\varepsilon \to 0} \frac{\ln N(\varepsilon)}{\ln \varepsilon}, \tag{7.33}$$

if this limit exists. When the limit does not exist, one can define (as we did for the Lyapunov exponent) "upper" and "lower" box counting dimensions using the limsup and liminf, respectively (Falconer 1990). It turns out that when the box counting dimension exists, the limit (7.33) is independent of the choice of the centers of the boxes. For example, if one uses a grid of size ε, then the number of grid cells needed to cover S will be somewhat larger than the optimal number of boxes; nevertheless, the resulting computed value of d_{box} will still be the same as that computed from the optimal N. Moreover, it does not matter if the boxes are cubes or spheres or if the boxes are rotated arbitrarily—the dimension calculated is the same.

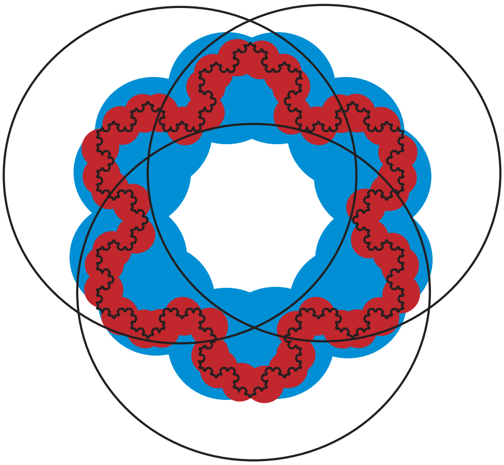

Figure 7.7. *Similarity transformations and coverings of the Koch snowflake.*

Example 7.22. Consider the Koch snowflake, K, of Figure 7.6, and suppose the sides of the original triangle have length L. The set K is contained in a ball of radius $r = 3L/4$ since this is the sum $\frac{L}{2}\sum_{n=0}^{\infty} 3^{-n}$. Each "side" of the Koch snowflake can be covered by a ball B_r whose center is at the center of the side of the original, level-zero triangle; see Figure 7.7. The Koch snowflake is self-similar under a transformation with scaling factor $1/3$, and the scaled side is congruent to each of the four parts of the original side. Thus each side can also be covered by four balls of radius $r/3$, and K is covered by $N(r/3) = 3\cdot 4$ balls. Proceeding to the next level requires four times as many balls of radius $r/9$. After j such transformations, $N(3^{-j}r) = 3\cdot 4^j$. Thus the box counting dimension is

$$d_{box} = -\lim_{j\to\infty}\frac{\ln(3\cdot 4^j)}{\ln(3^{-j}r)} = \frac{\ln 4}{\ln 3}. \quad \blacksquare$$

A more formal definition of dimension was introduced by Felix Hausdorff in 1918 and refined by Abram Besicovitch. Given a metric space with metric ρ, the diameter of a set U is defined as

$$\operatorname{diam}(U) = \sup_{x,y\in U}\rho(x,y). \tag{7.34}$$

We say a set S has an ε-cover if it is covered by a countable collection of open sets $U_j(\varepsilon)$, $j\in\mathbb{N}$, such that $\operatorname{diam}(U)\le\varepsilon$. The s-dimensional *Hausdorff measure* of S is the quantity

$$H^s(S) = \lim_{\varepsilon\to 0}\inf\sum_{i=1}^{\infty}(\operatorname{diam}U_i(\varepsilon))^s,$$

where the infimum is taken over all countable ε-covers of S. It is not hard to see, using the simple fact that if $s<t$, then $\varepsilon^s > \varepsilon^t$ when $\varepsilon < 1$, that H^s is a nonincreasing function of s.

One simple cover consists of the $N(\varepsilon)$ boxes used in the construction of d_{box}. These boxes all have the same diameter, namely, ε, and so make up an ε-cover. Boxes of the same size do not generally provide the optimal cover, but they do give the upper bound, $\sum_i(\operatorname{diam}(U_i(\varepsilon))^s \le N(\varepsilon)\varepsilon^s$. Estimating $N(\varepsilon)$ by (7.32), we obtain $H^s \le C\lim_{\varepsilon\to 0}\varepsilon^{s-d_{box}}$. Note that the right-hand side of this expression is ∞ when $s < d_{box}$

and is 0 if $s > d_{box}$. This property of a transition of H^s from ∞ to 0 at some critical s is a general property of the Hausdorff measure (Falconer 1990). This results in the definition of the

▷ *Hausdorff dimension*: $d_H(S) = \inf\{s : H^s(S) = 0\}$.

Thus d_H is the value of s for which the Hausdorff measure changes from ∞ to 0.

The previous discussion implies that $H^s = 0$ if $s > d_{box}$, and thus $d_H \le d_{box}$. d_H might be smaller because the number of elements of the cover can be optimized by varying their size.

Numerical computations of the dimension of the Lorenz attractor at the standard parameter values give $d_H \approx 2.062$ (Viswanath 2004). It is difficult to compute a value with this implied accuracy using (7.33). Instead, this value is obtained by using a hypothesized relation between the stability multipliers of periodic orbits embedded in the attractor and fractal dimension (Cvitanovic 1995). It is generally agreed that the numerically computed dimension is larger than 2; thus the Lorenz attractor appears to be a fractal. Given that the calculations in §7.2 showed that it has a positive Lyapunov exponent, we can say it is a strange, chaotic attractor.

Strange, Nonchaotic Attractors

Strange attractors can also be nonchaotic in some sense, for example, have no positive Lyapunov exponents. Such objects occur commonly when a nonlinear system is forced quasiperiodically. A function $g(t)$ is *quasiperiodic* when it has a Fourier series-like expansion

$$g(t) = \sum_{m \in \mathbb{Z}^d} a_m e^{im\cdot\omega t}$$

with a frequency vector, $\omega \in \mathbb{R}^d$, that is *incommensurate*:

$$\omega \cdot m \neq 0 \;\; \forall m \in \mathbb{Z}^d \setminus \{0\}. \tag{7.35}$$

Thus a quasiperiodic function has d independent frequencies (under integer combinations). A quasiperiodic function of a single variable, t, can always be thought of as a periodic function of d angle variables, $\theta \in \mathbb{T}^d$, by defining

$$g(t) = G(\omega t), \quad G(\theta) = \sum_{m \in \mathbb{Z}^d} a_m e^{im\cdot\theta}$$

so that G is periodic in each angle. Consequently, any quasiperiodically forced system, $\dot{x} = f(x,t)$, for $x \in M$ can always be rewritten as an autonomous system on $M \times \mathbb{T}^d$ by introducing angle variables $\theta \in \mathbb{T}^d$ and setting

$$\begin{aligned} \dot{x} &= F(x,\theta), \\ \dot{\theta} &= \omega, \end{aligned}$$

where $f(x,t) = F(x,\omega t)$ and F is a periodic function of θ.

Example 7.23. A model of a quasiperiodically forced pendulum is

$$\begin{aligned} \ddot{\theta} + \nu\dot{\theta} - a\cos\theta &= g(\psi), \\ \dot{\psi} &= \omega, \end{aligned}$$

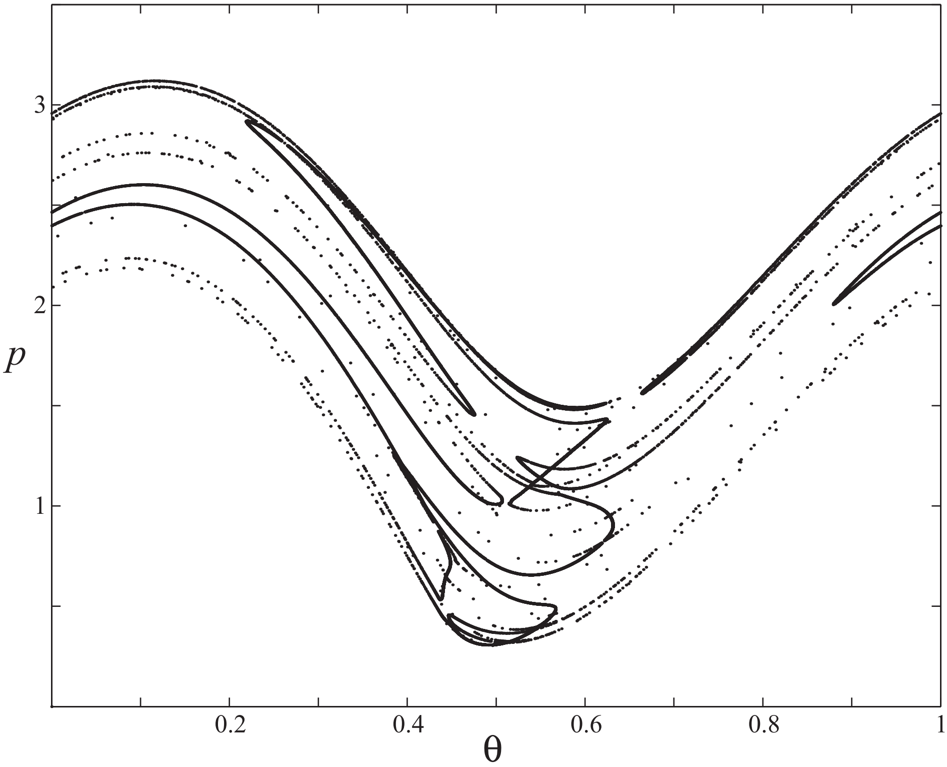

Figure 7.8. *Section through a strange, nonchaotic attractor of the quasiperiodic pendulum* (7.36) *with g given by* (7.38)*. Parameter values are* $\nu = a = 6\pi$*,* $b = 25.07$*, and* $c = 10.37$*. Plotted are* 10^5 *points on the section* $\psi_2 = 0$*, projected onto the* (θ, p) *plane.*

where ν is the damping coefficient, g is the forcing function, and $\psi \in \mathbb{T}^2$, so that $d = 2$. This model applies to a Josephson junction driven by two independent AC current sources. Converting this system to first order in the usual way gives a four-dimensional phase space $\mathbb{R} \times \mathbb{T}^3$ and the ODEs

$$\begin{aligned} \dot{p} &= -\nu p + a\cos\theta + g(\psi_1, \psi_2), \\ \dot{\theta} &= p, \quad \dot{\psi}_1 = \omega_1, \quad \dot{\psi}_2 = \omega_2. \end{aligned} \tag{7.36}$$

By scaling time, one of the frequencies can be set to unity, e.g., $\omega_2 = 1$. The frequency vector is then incommensurate whenever ω_1 is irrational, for example,

$$\omega_1 = \tfrac{1}{2}(-1 + \sqrt{5}), \tag{7.37}$$

the inverse of the golden mean.

This system (7.36) always has a global Poincaré section (recall §4.12): since the ψ dynamics is monotone, the flow returns to the section $\psi_2 = c$ for any c, and every trajectory must cross each such section. We can choose, for example, to visualize the dynamics by plotting the trajectories only when $\psi_2 = 0$. This still leaves a three-dimensional picture that can be difficult to visualize. As an aid in visualization it is also possible to plot only two coordinates, say, (θ, p), and project out the angle, ψ_1.

The linearization of (7.36) maps all vectors into the two-dimensional subspace $v = (v_1, v_2, 0, 0)^T$; thus (7.36) has two zero Lyapunov exponents. The remaining two exponents in the four-dimensional phase space are related by (7.19). Finally, since the trace of Df is constant,

$$\mu_1 + \mu_2 \geq \mathrm{tr}(Df) = -\nu.$$

Thus if the spectrum is regular, there is at most one positive Lyapunov exponent.

An example with

$$g(\psi) = b + c\,(\cos(2\pi\psi_1) + \cos(2\pi\psi_2)) \tag{7.38}$$

was studied by Romeiras and Ott (1987). For some parameter values this system exhibits attractors that appear to be two- or three-dimensional tori on which the motion is quasiperiodic. For others, the attractor is geometrically more complex; see Figure 7.8. This attractor has a complex geometric structure though its Lyapunov exponents are negative (the largest is $\mu_1 \approx -1.35$). It was conjectured by Romeiras and Ott that the set shown in Figure 7.8 has $d_{box} > 1$, a property that can be proved for other simple models that have strange nonchaotic attractors (Kim et al. 2003).

As the damping coefficient, ν, in (7.36) is decreased, one of its Lyapunov exponents becomes positive and the attractor becomes chaotic. ∎

In some cases, one can show that even though these strange attractors are "nonchaotic" in that all their Lyapunov exponents are negative, they still exhibit sensitive dependence (Glendinning, Jäger, and Keller 2006). Consequently, they would actually be called "chaotic" in the weak sense of our definition in §7.1. Perhaps it is best to think of these attractors as on the threshold of chaos.

7.4 ▪ Exercises

1. Prove that the orbits of the system

$$\dot{\theta} = \nu$$

for $\theta \in \mathbb{T}^n$ are transitive if and only if ν is incommensurate, i.e., for every nonzero integer vector $m \in \mathbb{Z}^n$, $m \cdot \nu \neq 0$. (*Hint*: An inductive argument using the pigeonhole principle, Lemma 7.5 might be useful.)

2. Prove that if a flow φ is chaotic on X and is topologically *equivalent* to the flow ψ on Y (recall §4.7), then ψ is chaotic on Y.

3. Prove that if γ is a periodic orbit of a flow φ with period T and λ is a Floquet exponent of the linearized flow about γ, then $\mu = \frac{1}{T}\mathrm{Re}(\lambda)$ is a Lyapunov exponent of γ. (*Hint*: Use Floquet's theorem, Theorem 2.36.)

4. Prove that the Lyapunov exponents are invariant under the flow, i.e., $\mu(x, v) = \mu(\varphi_t(x), \Phi(t; x)v)$.

5. Compute the Lyapunov spectrum for the system

$$\begin{aligned}\dot{x} &= (\sin\ln|t| + \cos\ln|t|)\,y,\\ \dot{y} &= (\sin\ln|t| + \cos\ln|t|)\,x.\end{aligned}$$

Show that the inequality $\delta < \mu_1 + \mu_2$ for (7.18) is strict for this case.

6. Suppose that φ is the flow of an autonomous Hamiltonian system. Show that every bounded orbit that is not an equilibrium nor is asymptotic to an equilibrium has a double zero Lyapunov exponent; that is, there are two linearly independent vectors for which $\mu(x,v)=0$. (*Hint*: Consider the vector ∇H.)

7. Using the definition (7.12) of characteristic exponent, prove the following:

   (a) Prove the results (7.14) and (7.15).

   (b) Show that if $\chi(f) > \chi(g)$, then $\chi(f+g) = \chi(f)$.

   (c) If $F(t) = \int_0^t f(s)ds$, show that $\chi(F) \le \max(0, \chi(f))$. (*Hint*: Show that if $\chi(f) = \alpha$, then for every $\varepsilon > 0$, $\lim_{t\to\infty} f(t)e^{-(\alpha+\varepsilon)t} = 0$.)

   (d) Suppose $v(t) = \left(v^{(1)}, v^{(2)}, \dots, v^{(n)}\right)^T$ is a vector function. Show that $\chi(v) = \max_{1\le i\le n} \chi(v^{(i)})$.

   (e) Consider the three vectors $v_1 = (e^t, 0, 0)^T$, $v_2 = (0, e^{2t}, 0)^T$, and $v_3 = (0, 0, e^{3t})^T$. What is $\chi(|\sum_{i=1}^3 c_i v_i|)$?

   (f) Suppose that A is a constant matrix with eigenvalues $\lambda = 1, 2$, and 3. Show that if $\Phi = [v_1, v_2, v_3]$ is any fundamental matrix of solutions, then

   $$\sum_{i=1}^{3} \chi(v_i) \le 9.$$

   (g) Again, for a constant matrix A, show that an eigenvector basis is one that minimizes the sum $\sum_{i=1}^{n} \chi(v_i)$.

8. Is the Lyapunov spectrum of the ω-limit set of a bounded orbit the same as that of the orbit?

9. Prove that if $\varphi_t(x)$ is the flow of a C^2 vector field, and the orbit Γ of x_o is bounded and has all negative Lyapunov exponents, then Γ is asymptotically stable. (Hints: Mimic the proof of Theorem 4.19, setting $y(t) = \varphi_t(x) - \varphi_t(x_o)$ and showing that y obeys a nonautonomous version of (4.18). Note that, by assumption, $\chi(||\Phi(t; x_o)||) < 0$ using (7.7).)

10. Compute the box counting dimension of the following self-similar sets:

    (a) The middle-α Cantor set, C, is constructed beginning with the closed unit interval, I. Remove the open set of length α from the middle of I. This leaves two closed intervals, each of length $L = (1-\alpha)/2$. Remove the middle interval of length αL from each of these, and continue....

    (b) The Sierpinski gasket, S, is constructed from a solid equilateral triangle T with sides of length one. Remove the equilateral triangle whose vertices are the midpoints of each of the sides of T. The remaining set is the union of three equilateral triangles with sides of length $1/2$. Now remove the middle triangle from each of these, and continue....

    (c) For the Menger sponge, M, begin with the unit cube B in $\mathbb{R}^3$. This can be thought of as the union of 27 cubes whose sides have $1/3$. Now remove seven of these cubes: the six that have a face in the center of each face of B and the seventh embedded in the center of B. This leaves a set that is the union of 20 smaller cubes with sides of length $1/3$. Continue removing seven cubes of size $1/9$ from each of these....

11. Write a program to compute the box counting dimension. As a trial, use it to compute the dimension of the sets in Exercise 10. Now compute the dimension of the Lorenz attractor.

12. Explore the dynamics of your adopted quadratic system (recall Exercise 1.10) for the chaotic values of its reduced parameters.

    (a) Compute the maximum Lyapunov exponent.

    (b) Plot the chaotic attractor.

    (c) Compute its box counting dimension.

    (d) Vary the values of the reduced parameter(s) and discuss how the chaotic attractor is destroyed.

Chapter 8

Bifurcation Theory

In this chapter we will study systems of differential equations $\dot{x} = f(x; \mu)$ that depend on a set of parameters μ. For example, the vector field for the pendulum nominally depends upon two parameters: its length and the strength of gravity. Our goal is to investigate what happens to the flow of the system when parameters vary slightly. Do the properties of the orbits just change slightly, or can orbits be destroyed, created, or otherwise changed dramatically? A bifurcation occurs when there is a dramatic change in the dynamics:

> ▷ *Bifurcation*: a qualitative change in dynamics occurring upon a small change in a parameter.

One of the simplest bifurcations corresponds to the creation or destruction of an equilibrium. A typical case is called the *saddle-node* bifurcation; we will study it first. Another bifurcation corresponds to the change in stability of an orbit—this is often accompanied by the creation or destruction of other nearby orbits. Such bifurcations are called "local" because they can be studied by expanding the vector field in a Taylor series about a reference orbit in the phase space. There are also "global bifurcations" such as the *homoclinic* bifurcation, which corresponds to the creation or destruction of a homoclinic orbit (see §8.11). These bifurcations are much harder to study because they are intrinsically nonlocal. Our treatment starts with the local case and with those bifurcations that typically happen when one varies a single parameter; such bifurcations are called *codimension-one*. One of the triumphs of bifurcation theory is the classification of bifurcations with low codimension. We will find that there are only two local, codimension-one bifurcations for flows: the saddle-node and the Hopf bifurcations.

Before discussing the theory in general, we consider the one-dimensional case.

8.1 ▪ Bifurcations of Equilibria

The logistic model (1.7) is perhaps the simplest, nonlinear population dynamics model. The nonlinearity models competition for a fixed resource. Suppose that x represents the population of fish in a fishery and that, in addition to the competition, the fish are harvested at a constant rate h. The logistic model then becomes

$$\dot{x} = rx(1-x) - h. \tag{8.1}$$

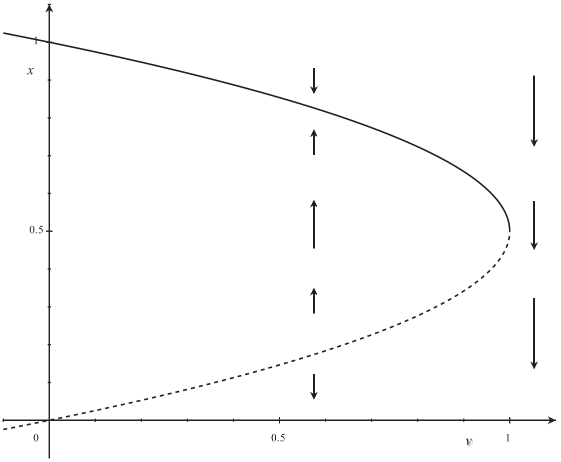

Figure 8.1. *Saddle-node bifurcation for* (8.1).

The vector field f of this model depends not only on the dynamical variable $x \in \mathbb{R}^+$ but also on two parameters $\mu = (r, h) \in \mathbb{R}^{+^2}$; thus we can write it more generally as $\dot{x} = f(x;\mu)$, where the semicolon separates the dynamical variables from the parameters. The simplest bifurcations correspond to qualitative changes in equilibria, namely, in their number and stability type. For (8.1) the equilibria are

$$x^*_\pm = {}^1\!/\!_2\left(1 \pm \sqrt{1-4h/r}\right).$$

Note that there are two equilibria when $4h < r$, one when $4h = r$, and none when $4h > r$. Thus there is a bifurcation—a change in the number of equilibria—on the line $4h = r$ in the parameter space; this is the *bifurcation set*. The existence of the equilibria depends only on one combination of the two parameters, $\nu = 4h/r$; consequently this bifurcation is governed by a single effective parameter. We can conveniently collect the information about the equilibria in a *bifurcation diagram* that shows the two functions $x^*_\pm(\nu)$ as a function of the single parameter ν; see Figure 8.1.

The bifurcation diagram represents the qualitative behavior of our system. Traditionally, the abscissa of the graph corresponds to the parameters and the ordinate to the phase space. Thus, each vertical slice is a picture of the vector field for fixed parameters, and the vector fields with varying parameters are stacked together to obtain the full diagram. A dashed curve is traditionally used to represent an unstable orbit, while a solid curve represents a stable one. When $4h < r$ in (8.1) x^*_+ is stable, and x^*_- is unstable since the slope of f changes sign at $x = {}^1\!/\!_2$.

The dynamics in the bifurcation diagram occurs along vertical lines at fixed values of the reduced parameter ν; we sketch two representative vector fields in Figure 8.1. Note that when the harvesting is too strong, i.e., when $4h > r$, we have $\dot{x} < 0$ for all x and the population crashes, reaching extinction in a finite time. The model then ceases to be valid: the assumption that the harvesting occurs at a constant rate must fail well before this point.

The bifurcation occurs at the point $(x,\nu) = ({}^1\!/\!_2, 1)$ where the two equilibria collide. We can focus on this point by centering the picture at this value. To do this, define a new variable $y = {}^1\!/\!_2 - x$, a new parameter $\mu = h/r - 1/4$, and (to eliminate r) a new time $\tau = rt$. In the new variables the ordinary differential equation (ODE) is

$$-\frac{dy}{d\tau} = \frac{1}{r}\frac{dx}{dt} = \left(\frac{1}{2} - y\right)\left(\frac{1}{2} + y\right) - \frac{1}{4} - \mu \;\Rightarrow\; \dot{y} = \mu + y^2. \tag{8.2}$$

We call (8.2) a *normal form* for the bifurcation. It has the bifurcation point $(0,0)$ where a stable and an unstable equilibrium collide and are destroyed. The resulting bifurcation is called a "saddle-node" bifurcation.[44] As we will see, the normal form describes the local behavior near any saddle-node bifurcation.

Example 8.1. Consider the system

$$\dot{x} = \mu + x - \ln(1+x). \tag{8.3}$$

The equation for the equilibria is transcendental and thus cannot be solved analytically for $x(\mu)$.[45] However, insight into its solutions can be obtained by graphing the two functions $g(x) = \ln(1+x)$ and $h(x) = \mu + x$ for varying values of μ; intersections of the two graphs correspond to equilibria. As μ is varied, the graph of h translates vertically and the intersections move. When $\mu > 0$ there are no equilibria, while for $\mu < 0$ there are two; call them $x^*_\pm$ as before. Even if the equilibria cannot be obtained explicitly, the bifurcation point can often be found. To do this, note that at a point where the equilibria are created or destroyed, the two curves g and h must be tangent, so that $Dh(x^*) = Dg(x^*)$:

$$\frac{d}{dx}(\mu + x) = \frac{d}{dx}(\ln(1+x)) \;\Rightarrow\; 1 = \frac{1}{1+x^*} \;\Rightarrow\; x^* = 0.$$

Combining this with the equilibrium equation $\mu^* + x^* = \ln(1+x^*)$ provides two equations for the two unknowns, (x^*, μ^*). Since $x^* = 0$, the equilibrium equation implies that $\mu^* = 0$, too. Thus, the bifurcation occurs at $(0,0)$. To get a qualitative picture of what happens for other values of μ, note that the graph and the equation $f(x;\mu) = 0$ imply that as $\mu \to -\infty$, $x^*_- \to -1$ and $x^*_+ \to -\mu$, since $\ln x \ll x$. Of course it is also easy to plot the solution numerically (see the appendix), as shown in Figure 8.2.

Upon expanding the ODE about the bifurcation point, $(0,0)$, we obtain a description of the dynamics near the bifurcation:

$$\dot{x} = \mu + x - \left(x - \tfrac{1}{2}x^2\right) + O(x^3) = \mu + \tfrac{1}{2}x^2 + O(x^3).$$

Note that this can be transformed into the "normal form" (8.2) by a scaling. ■

We will show in §8.4 that there is a conjugacy between the normal form (8.2) and the original vector field in a neighborhood (in both x and μ) of the saddle-node bifurcation point, provided that some "nondegeneracy" and "transversality" conditions are satisfied. The nondegeneracy condition is that the coefficient of the quadratic term, x^2, is nonzero in the Taylor expansion about the bifurcation point.

Transversality conditions guarantee that the parameters in f occur in a sufficiently general way so as to be able to cause the bifurcation. Loosely speaking, each parameter

[44] The terminology is not really appropriate for the one-dimensional case, but the reason for using this name becomes clear when we consider higher dimensions.

[45] However, it is easy to obtain $\mu(x)$, which is just as good. We will ignore this for the example as it is not always possible.

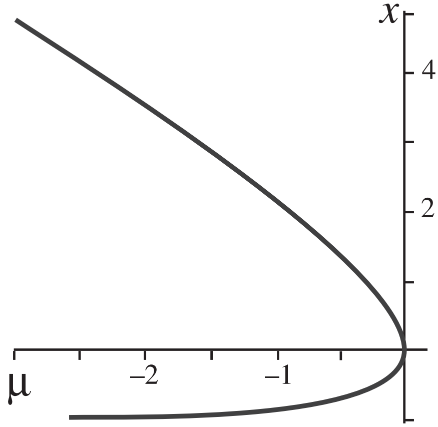

Figure 8.2. *The set $f(x;\mu)=0$ for* (8.3).

is a knob that gives some control over the dynamics. For the saddle node, if the knobs are transverse, then equilibria can be created or destroyed at will, for example, see (8.2). This means that we need to move the minimum of the function $f(x;0)$ up and down or equivalently that $D_\mu f(0;0)\neq 0$. If this is not satisfied, then the bifurcation can be somewhat different in character.

Example 8.2. For example, consider the ODE

$$\dot{x}=\mu x+x^2. \tag{8.4}$$

Here the two equilibria are $x_1^*=0$ and $x_2^*=-\mu$, and the corresponding bifurcation diagram is shown in Figure 8.3. Note that the equilibria coalesce at $\mu=0$ but are not destroyed. However, something does happen at the collision point: since $D_x f(x_1^*;\mu)=\mu$ and $D_x f(x_2^*;\mu)=-\mu$, the two fixed points have opposite stability types, and they switch type at $\mu=0$. This is a "qualitative" change in the dynamics and so qualifies as a bifurcation. It is called an *exchange of stabilities* or *transcritical* bifurcation. ∎

Our goal is to show that when a vector field satisfies the appropriate nondegeneracy and transversality conditions, a saddle-node bifurcation is certain to occur. Additionally we will classify the various "conjugacy classes" of systems near bifurcation points by identifying these conditions.

8.2 ▪ Preservation of Equilibria

To understand when bifurcations happen, it is important to first understand when they do not happen. As we will soon see, nothing dramatic can happen to nondegenerate equilibria when a parameter is slightly changed. Recall from §2.2 that an equilibrium is called *degenerate* if at least one of the eigenvalues of its linearization is zero. Thus we will see that an equilibrium whose eigenvalues are all nonzero is "structurally stable"—it cannot be removed by small changes in the equations.

Generally, a flow φ is structurally stable if every flow in a neighborhood of φ is topologically equivalent. Here the neighborhood corresponds to a set of vector fields in some function space, for example, C^r for some r, near to the vector field of φ.

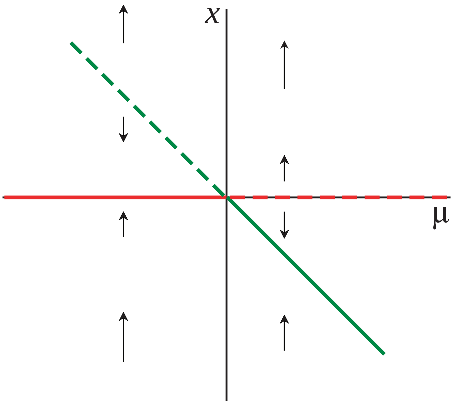

Figure 8.3. *Transcritical bifurcation of* (8.4).

Practically one also usually must consider a neighborhood in phase space about some particular orbit. Here we consider the simplest orbit, an equilibrium.

An essential tool to demonstrate this—as well as many other results in bifurcation theory—is the implicit function theorem. As its title indicates, this theorem deals with "implicitly" defined functions. For example, we might expect that the equation $f(x;\mu)=0$ "typically" can be solved for x to define a "function," $x(\mu)$. However, as we saw in §8.1, there is not necessarily a unique such function (there we obtained two, $x_\pm(\mu)$), and it is also easy to construct examples where there is no such function, e.g., $f(x;\mu)=\operatorname{sech} x+\mu^2$. The implicit function theorem gives sufficient conditions on f such that the implicitly defined function does exist and is unique.

Theorem 8.3 (Implicit Function). *Let U be an open set in $\mathbb{R}^n\times\mathbb{R}^k$ and $F\in C^r(U,\mathbb{R}^n)$ with $r\geq 1$. Suppose there is a point $(x_o,\mu_o)\in U$ such that $F(x_o;\mu_o)=c$ and $D_xF(x_o;\mu_o)$ is a nonsingular $n\times n$ matrix. Then there are open sets $V\subset\mathbb{R}^n$ and $W\subset\mathbb{R}^k$ and a unique C^r function $\xi(\mu):W\to V$ for which $x_o=\xi(\mu_o)$ and $F(\xi(\mu);\mu)=c$.*

This theorem, and its generalization to functions on Banach spaces can be derived from (you guessed it!) the contraction-mapping theorem. It is proved in any respectable course on advanced calculus or analysis (Markley 2004; Taylor and Mann 1983).

Theorem 8.3 states that if we know a solution for some special parameter value μ_o, then there is a unique surface of solutions that goes through the special solution, provided that the Jacobian is nonsingular. It is easy to obtain a rough understanding as to why the condition on the Jacobian D_xF is necessary. We expand $F=c$ about (x_o,μ_o) and neglect terms of higher order than the first derivatives:

$$c=F(x_o+\delta x;\mu_o+\delta\mu)=c+D_xF(x_o;\mu_o)\delta x+D_\mu F(x_o;\mu_o)\delta\mu+O(2).$$

If it were okay to ignore the higher-order terms, we could solve for δx to obtain

$$\delta x\approx-(D_xF)^{-1}D_\mu F\,\delta\mu;$$

this can be done for arbitrary $\delta\mu$ only if D_xF is nonsingular. This calculation gives the lowest-order approximation to the function $\xi(\mu)=x_o+\delta x(\mu)$. The theorem

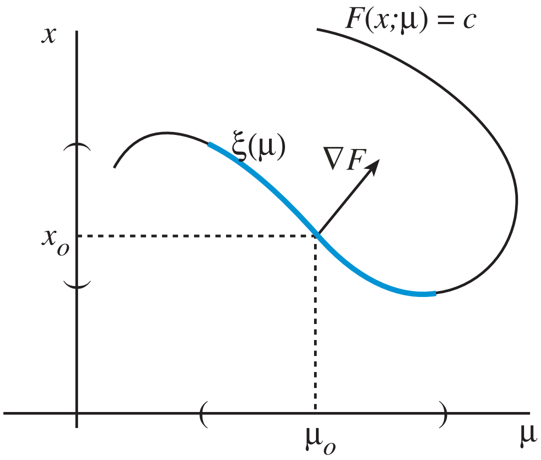

Figure 8.4. *Illustration of the implicit function theorem for the case $n = k = 1$.*

asserts that this approximation can be extended to a smooth function that is an exact solution to $F = c$ in some neighborhood of (x_o, μ_o).

A geometrical understanding of this result is easily obtained in two dimensions; see Figure 8.4. If $(x, \mu) \in \mathbb{R}^1 \times \mathbb{R}^1$, then the contour $F(x;\mu) = c$ is generically a curve. The gradient vector $\nabla F = (D_x F, D_\mu F)$ is perpendicular to the contour. At any point where ∇F is not in the μ-direction, the contour is locally a graph over μ and uniquely defines the function $x = \xi(\mu)$. When $D_x F = 0$, no local graph $\xi(\mu)$ exists. Note that in this case the implicit function theorem could be applied for the "variable" μ as a function of the "parameter" x to obtain $\mu(x)$ provided that $D_\mu F \neq 0$.

The implicit function theorem immediately implies that nondegenerate equilibria are structurally stable.

Corollary 8.4 (preservation of a nondegenerate equilibrium). *Suppose the vector field $f(x;\mu)$ is C^1 in both x and μ and that x_o is a nondegenerate equilibrium point for parameter μ_o (i.e., all the eigenvalues of this equilibrium are nonzero). Then there exists a unique C^1 curve of equilibria $x^*(\mu)$ passing through x_o at μ_o.*

Proof. Recall that the matrix $A = D_x f(x_o;\mu_o)$ governs the stability of the equilibrium, and since A has all its eigenvalues nonzero, then A is nonsingular. Theorem 8.3 then implies that there is a neighborhood of μ_o for which there is a curve of equilibria $x^*(\mu)$. □

This result applies for an arbitrary number of parameters μ—no matter how many knobs you have to turn, you cannot destroy a nondegenerate equilibrium by small turns! For example, a linear center will be preserved under perturbation (though its stability may change). The only time an equilibrium may immediately disappear is when $D_x f$ has a zero eigenvalue.

8.3 ▪ Unfolding Vector Fields

Bifurcation theory begins with a particular vector field, say, $f_0(x)$. To study the dependence of the dynamics on parameters, this vector field is then *unfolded*:

> ▷ *Unfolding*: A family of vector fields $f(x;\mu)$ is an *unfolding* of $f_o(x)$ if $f(x;0)=f_o(x)$.

In the spirit of the implicit function theorem, Theorem 8.3, we focus on a neighborhood of a special parameter value that, without loss of generality, is chosen to be $\mu=0$. Typically, the vector field $f_o(x)$ will be assumed to have a degenerate orbit at this special parameter value; this is called a *singularity* condition. For the next few sections, we will restrict our consideration to bifurcations that are local in phase space, that is, to some neighborhood of the special orbit.

The issue of what space of functions are allowed in an unfolding is an important one, as is a careful definition of the particular neighborhood of f_o that is of interest. For the moment, we will ignore these issues; they will be clarified in our treatment of specific bifurcations.

Just as we used the concepts of conjugacy and equivalence in §4.7 to discover whether two systems were effectively the same, we can extend these concepts to families of vector fields. In particular, two families $f(x;\mu)$ and $g(x;\mu)$ are *conjugate* if there is a family of conjugacies $h(x;\mu)$ between their flows (recall (4.32))—the only difference is that the homeomorphism is now allowed to depend upon the parameters μ. Similarly, two families of vector fields are *equivalent* if, for each value of μ, their orbits are topologically conjugate, preserving the direction of time; recall (4.33).

While two equivalent dynamical systems ostensibly depend upon the same parameters, it is possible that some of the parameters enter one of the systems in a trivial way. For example, the vector fields $f(x;\mu_1,\mu_2)=\mu_1+x^2$ and $g(y;\mu_1,\mu_2)=\mu_1/\mu_2+\mu_2 y^2$ are conjugate under the transformation $y=h(x;\mu_1,\mu_2)=x/\mu_2$ whenever $\mu_2\neq 0$. Thus, even though f formally depends upon both parameters, in reality it depends only upon the first. This is one mechanism that is used below to reduce a system of ODEs to a normal form containing a minimal number of parameters.

It is also useful to have notions of conjugacy that allow reparameterization of the vector fields. This notion is called

> ▷ *induced*: A family $g(x;\nu)$ is *induced* by a family $f(x;\mu)$ if there is a continuous map $\mu=p(\nu)$ such that $g(x;\nu)=f(x;p(\nu))$.

The range of dynamics of the induced vector field g can be as rich as those of f, but may also be simpler.

Example 8.5. The vector field $g(x;\nu)=\nu_1+\nu_2^2-x^2$ with two parameters on $\mathbb{R}^1$ is induced by $f(x;\mu_1)=\mu_1-x^2$ using the map $\mu_1=p(\nu)=\nu_1+\nu_2^2$. Although g depends upon two parameters, only one is essential. Alternatively, the vector field $k(x;\lambda)=\lambda_1+2\lambda_2 x-x^2$ is not induced by f; rather we have the converse—f is induced by k through the map $p(\mu_1)=(\mu_1,0)$. In this sense f is a simpler version of k.

Nevertheless, the vector field k is conjugate to g using the shift $y=h(x;\lambda)=x-\lambda_2$ since $g(y;\lambda)=k(y+\lambda_2;\lambda)=\lambda_1+\lambda_2^2-y^2$. Since g *is* induced by f, we can assert that the flow of k is conjugate to a flow induced by f. Consequently f describes the dynamics of both of the two-parameter families g and k. ∎

An unfolding that describes every possible nearby behavior is called a

> ▷ *versal unfolding*: An unfolding $f(x;\mu)$ is *versal*[46] if *every* other unfolding in some neighborhood of f_o is equivalent to a family induced by $f(x;\mu)$.

[46] The *Oxford English Dictionary* says that "versal" is an illiterate or colloquial abbreviation of universal. It means universal or whole. The latter meaning seems appropriate here. Shakespeare used it in *Romeo and Juliet*, though not in the mathematical sense.

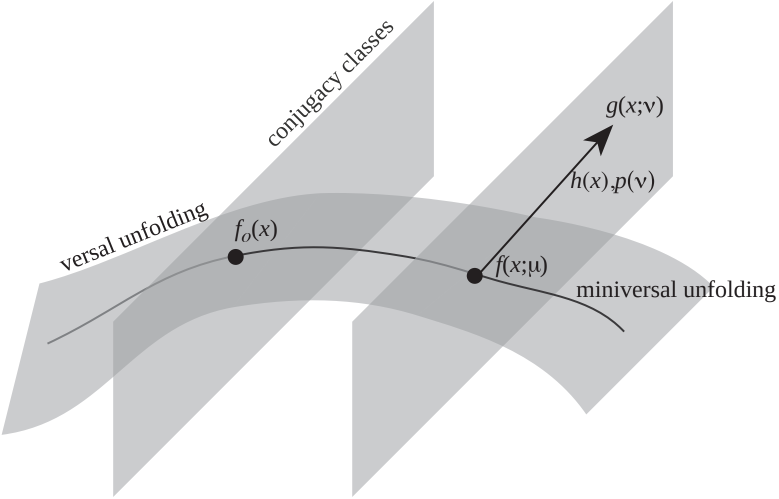

Figure 8.5. *Unfolding a vector field $f_o(x)$.*

If we assume that the equivalence is actually a diffeomorphic conjugacy, then this statement can be made on the level of the vector fields. In this case, if $f(x;\mu)$ is a versal unfolding of f_o, then for every other unfolding $g(x;\nu)$ there must exist a diffeomorphism, h, and a map, p, such that

$$g(h(x;p(\nu));\nu) = Dh(x;p(\nu))f(x;p(\nu))$$

for a neighborhood of $(0,0)$, recall (4.34). One goal is to obtain a complete description of the neighborhood of a special vector field that uses the smallest possible number of parameters. If we achieve this we say that we have a

> ▹ *miniversal unfolding*: An unfolding is *miniversal* if it is a versal unfolding with the minimum number of parameters.

These ideas are presented geometrically in Figure 8.5; here the infinite-dimensional spaces of all functions conjugate to $f(x;\mu)$ are drawn as "planes" and a miniversal unfolding of $f_o(x)$ as a "curve."

Example 8.6. Suppose $x \in \mathbb{R}^1$ and that $f_o = 0$. Consider the behavior of the special degenerate equilibrium at $x = 0$ (even though every point is such an equilibrium!). The family $f(x;\mu) = -\mu_1 x + \mu_2 x^2$ is an unfolding of f_o. However, it is not versal because, for example, the vector field $g(x;\nu) = \nu$ is an unfolding of f_o that has no equilibria when ν is nonzero and hence cannot be conjugate to f, which always has an equilibrium at $x = 0$. Even though it is not versal, the unfolding f in some sense has too many parameters. Indeed, the conjugacy $y = h(x) = \mu_2 x$ transforms f into the vector field $k(y;\mu) = -\mu_1 y + y^2$, so a single parameter family suffices to describe the dynamics of f when $\mu_2 \neq 0$. ■

Unfolding Two-Dimensional Linear Flows

The simplest case of unfolding is in the context of linear systems. Here we consider a linear vector field on $\mathbb{R}^2$, setting $z=(x,y)^T$,

$$\dot{z}=Az=\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}.$$

The set of all 2×2 matrices is a manifold isomorphic to $\mathbb{R}^4$ with coordinates (a,b,c,d). Under a linear change of coordinates $z=P\zeta=P(\xi,\eta)^T$, the matrix A transforms into the similar matrix $B=P^{-1}AP$, and the dynamics of the new system $\dot{\zeta}=B\zeta$ is linearly conjugate to the original dynamics. There are only two combinations of the parameters (a,b,c,d) of A that are invariant under this linear conjugacy: the trace and determinant

$$\tau=\operatorname{tr}(A)=a+d, \quad \delta=\det(A)=ad-bc; \tag{8.5}$$

recall §2.2. When A is semisimple, it can be diagonalized by this coordinate transformation; consequently, every matrix in the two-dimensional subspaces of $\mathbb{R}^4$ with the same trace and determinant has the same dynamics. As we will see, under topological conjugacy the number of essential parameters can be reduced even more.

There are three "singularities" that are of interest in bifurcation theory; they correspond to the three types of nonhyperbolic equilibrium:

(a) single zero eigenvalue, $\det(A_o)=0$, $\operatorname{tr}(A_o)\neq 0$,

(b) pair of imaginary eigenvalues: $\operatorname{tr}(A_o)=0$, $\det(A_o)\neq 0$, and

(c) double zero eigenvalue, $\operatorname{tr}(A_o)=\det(A_o)=0$.

To study these cases we first change coordinates so that the matrix is in its simplest form under linear conjugacy. When there is a single zero eigenvalue, the matrix is always semisimple and can thus be diagonalized, $A_o=PJP^{-1}$, where

$$J=\begin{pmatrix} 0 & 0 \\ 0 & \lambda \end{pmatrix}. \tag{8.6}$$

As λ varies, J defines a one-dimensional curve in the four-dimensional space of 2×2 matrices. This means that all flows in case (a) are linearly conjugate to the flow of (8.6) for some value of λ. However, J is not the simplest form for the class (a); further simplification can be obtained using a topological conjugacy, (4.32). As an extension of Theorem 4.33, the flow of (8.6) is topologically conjugate to a simpler flow with λ replaced by $\operatorname{sgn}(\lambda)=\pm 1$. To see this, denote the two flows by $\varphi_t(x,y)=(x,ye^{\lambda t})$ and $\psi_t(\xi,\eta)=(\xi,\eta e^{\operatorname{sgn}(\lambda)t})$. The homeomorphism

$$(\xi,\eta)=h(x,y)=(x,\operatorname{sgn}(y)\,|y|^{\alpha}) \tag{8.7}$$

with $\alpha=1/|\lambda|$ provides a conjugacy between φ and ψ:

$$h\circ\varphi_t(x,y)=(x,\operatorname{sgn}(y)\left|ye^{\lambda t}\right|^{\alpha})=(x,\operatorname{sgn}(y)\,|y|^{\alpha}\,e^{\operatorname{sgn}(\lambda)t})=\psi_t\circ h(x,y).$$

The new flow has a vector field defined by the matrix

$$\hat{J}_{\pm}=\begin{pmatrix} 0 & 0 \\ 0 & \pm 1 \end{pmatrix}. \tag{8.8}$$

Note that ψ is *not* diffeomorphic to the original flow, so we cannot transform the vector fields directly.

Thus the flows generated by $\hat{J}_\pm$ are conjugate to the flows of any matrix that satisfy condition (a). The "+" matrix represents those with an unstable direction, and the "−" matrix represents those with a stable direction. These are distinct conjugacy classes since the origin has different stability properties. The matrices (8.8) are the normal forms for the linear flows with a single zero eigenvalue. As we will see below, it is typical that normal forms depend upon some parameter that takes a discrete set of possible values; these parameters are called *moduli*.

The matrices that satisfy condition (a),

$$F(a;b,c,d) = \det(A_o) = ad - bc = 0, \tag{8.9}$$

make up a three-dimensional surface in $\mathbb{R}^4$. Theorem 8.3 (the implicit function theorem) implies that in a neighborhood of $\hat{J}_+$ or of $\hat{J}_-$ the set $F = 0$ is a smooth three-dimensional surface in $\mathbb{R}^4$ (a submanifold). Indeed, the condition (8.9) can be thought of as implicitly determining a as a function of b, c, and d. Since $D_a F|_{\hat{J}_\pm} = d = \pm 1 \neq 0$, the implicit function theorem states that there is a unique, smooth function $a(b,c,d)$ on which $\det(A_o) = 0$ in the neighborhood of $A_o = \hat{J}_\pm$. This representation of the surface as a graph over (b,c,d) fails only when $d = 0$—indeed the surface $F = 0$ when $d = 0$ corresponds to the union of two planes: $\{(a,b,0,0)\} \cup \{(a,0,c,0)\}$.

To obtain a versal unfolding of $\hat{J}_\pm$ in the space of linear vector fields on $\mathbb{R}^2$, we need only add one parameter to represent the change in value of the zero eigenvalue. Any matrix A near the surface $F = 0$ has eigenvalues (μ, λ) with μ small and λ close to ±1; consequently for an appropriately chosen neighborhood, $\mu \neq \lambda$. By the same argument as before, the flow of A is conjugate to that of the matrix

$$A_\mu = \begin{pmatrix} \mu & 0 \\ 0 & \operatorname{sgn}(\lambda) \end{pmatrix}. \tag{8.10}$$

Thus A_μ gives a versal unfolding of $\hat{J}_\pm$. Note that it would not be useful to use a conjugacy to scale μ to ±1, since we are interested in varying μ through 0. It is clear that at least one parameter must be added to unfold a three-dimensional surface in $\mathbb{R}^4$; thus the unfolding (8.10) is miniversal.

Recall from §2.5 that any matrix in case (b) is linearly conjugate to the matrix $\left(\begin{smallmatrix} 0 & -\omega \\ \omega & 0 \end{smallmatrix}\right)$. By rescaling time and, if $\omega < 0$, flipping the sign of x, an equivalent system with the matrix

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \tag{8.11}$$

is obtained. Since these matrices are defined again by a single condition, $\operatorname{tr}(A_o) = 0$, the set of matrices equivalent to (8.11) again forms a three-dimensional surface in $\mathbb{R}^4$. Any matrix near (8.11) will have eigenvalues $\lambda = \mu \pm i\nu$, with μ small and ν near 1. Upon rescaling time, these matrices have flows that are equivalent to

$$A_\mu = \begin{pmatrix} \mu & -1 \\ 1 & \mu \end{pmatrix}. \tag{8.12}$$

Note that matrices with negative determinant cannot be obtained in this unfolding; however, we are really interested only in a neighborhood of the nonhyperbolic point.

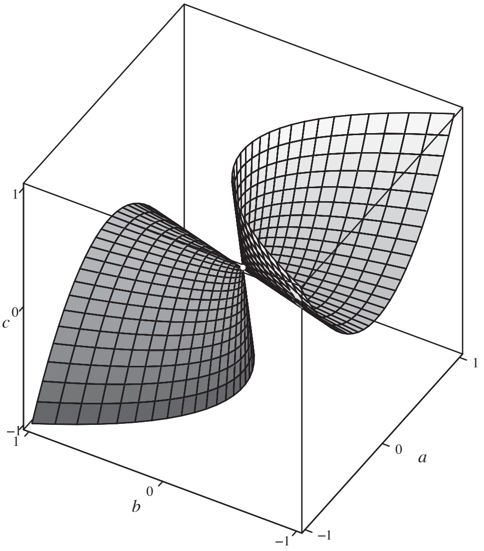

Figure 8.6. *The projection of the two-dimensional surface of matrices conjugate to the normal form* (8.13) *onto* (a,b,c) *using the fact that* $a+d=0$. *Note that the origin is not included in this set.*

Finally consider case (c), the special point of a double zero eigenvalue. The matrix is typically not semisimple, and for this nondiagonalizable case (called the Takens–Bogdanov point), the normal form is the Jordan form

$$J=\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}. \tag{8.13}$$

This normal form corresponds to a point in $\mathbb{R}^4$, but is conjugate to a two-dimensional subspace of matrices: those that satisfy the two relations (c) and have a single eigenvector; see Figure 8.6. Two parameters are required to unfold this matrix—enough to represent the change in both $\det(A)$ and $\operatorname{tr}(A)$. It is not hard to see that the unfolding can be given by

$$A_\mu=\begin{pmatrix} \mu_1 & 1 \\ \mu_2 & 0 \end{pmatrix}, \tag{8.14}$$

for example (see Exercise 4). There are many other valid choices for this unfolding, and some will be more convenient (when we consider the nonlinear terms) than others.

If the degenerate matrix in case (c) has two eigenvectors, then it must be the zero matrix—every matrix that is similar to 0 is itself 0. This case is a single point in the space of 2×2 matrices (it is represented by the origin in Figure 8.6) and is hence much more unusual than the two-dimensional surface conjugate to (8.13). Four parameters are required to unfold the zero matrix, and these might as well be the four entries (a,b,c,d).

8.4 ▪ Saddle-Node Bifurcation in One Dimension

The saddle-node bifurcation corresponds to the creation or destruction of a pair of equilibria; several examples were studied in §8.1. Here we present a theorem that gives conditions under which this bifurcation necessarily occurs. Bifurcation theorems typically involve three types of assumptions. The first is a *singularity* assumption—in this case, that there is a vector field f_o with a nonhyperbolic equilibrium. The second is a *nondegeneracy* or *genericity* assumption, in this case that f_o has quadratic terms near the equilibrium. The final assumption is one of *transversality*—that the parameters are sufficiently general to unfold the vector field and cause the bifurcation.

For an ODE $\dot{x} = f_o(x)$ on $\mathbb{R}^1$, the singular case that gives rise to the saddle node is defined by

$$f_o(0) = 0, \quad Df_o(0) = 0 \qquad \textit{(singularity)}$$

so that there is a nonhyperbolic equilibrium at the origin. The nondegeneracy assumption is that the quadratic term in the power series is nonzero:

$$D_{xx} f_o(0) \neq 0. \qquad \textit{(nondegeneracy)}$$

This assumption limits the complexity of the resulting behavior. The final assumption of transversality serves to guarantee that the parameter μ in the unfolding $f(x;\mu)$ of f_o moves the vector field "transversely" to the singular state. In this case the necessary condition will turn out to be

$$D_\mu f(0;0) \neq 0. \qquad \textit{(transversality)}$$

We begin first, however, with a theorem that assumes only nondegeneracy.

Theorem 8.7. *Suppose that* $f(x;\mu) \in C^2(\mathbb{R} \times \mathbb{R}^k, \mathbb{R})$ *with a nonhyperbolic equilibrium at the origin,* $f(0;0) = 0$, $D_x f(0;0) = 0$, *and that* f *satisfies the nondegeneracy condition*

$$c \equiv D_{xx} f(0;0) \neq 0. \tag{8.15}$$

Then, there is a $\delta > 0$ *such that when* $|\mu| < \delta$, *there is an open interval,* $I(\mu)$, *containing* 0 *such that there is a unique extremal value*

$$m(\mu) \equiv \operatorname*{Ext}_{x \in I}(f(x;\mu)). \tag{8.16}$$

There are two equilibria in I *when* $m(\mu)c < 0$, *one when* $m(\mu)c = 0$ *and zero when* $m(\mu)c > 0$.

Proof. The singularity and nondegeneracy conditions imply that $f_o(x) = {}^1\!/\!_2 cx^2 + g(x)$. Since f_o is C^2, the nonlinear term, g, is small: $g = o(x^2)$ (recall §4.4). Thus

$$g(0;0) = D_x g(0;0) = D_{xx} g(0;0) = 0.$$

A general unfolding of f_o will have the form

$$f(x;\mu) = a(\mu) + b(\mu)x + \tfrac{1}{2}c(\mu)x^2 + g(x;\mu), \tag{8.17}$$

where $a(0) = f(0;0) = 0$, $b(0) = D_x f(0;0) = 0$, $c(0) = c \neq 0$, and $g(x;\mu) = o(x^2)$. To solve for the equilibria we would ordinarily try to solve for $x(\mu)$; however, the implicit function theorem fails because $D_x f(0;0) = 0$.[47] However, it is possible to solve, for the critical points of f, the zeros of the function

$$F(x;\mu) \equiv D_x f(x;\mu) = b(\mu) + c(\mu)x + D_x g.$$

[47] One way to get around this is to solve for $\mu(x)$, which is possible by the implicit function theorem. See (Robinson 1999) for this approach.

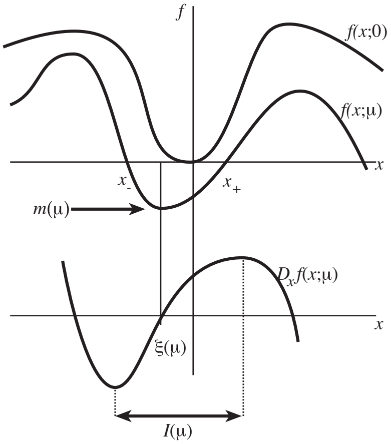

Figure 8.7. *Illustration of a saddle-node bifurcation in* $\mathbb{R}^1$.

The conditions of the implicit function theorem are satisfied for $F(x;\mu)$ since $F(0;0) = b(0) = 0$ and $D_x F(0;0) = c(0) \neq 0$. Thus there are neighborhoods V and W of the origin such that when $\mu \in W$ there is a unique $x = \xi(\mu) \in V$ such that $F(\xi(\mu);\mu) = 0$ and $\xi(0) = 0$. Since $D_x F(0;0) \neq 0$, there is a possibly smaller neighborhood of $\mu = 0$ and an interval $I(\mu)$, containing $\xi(\mu)$, for which $F(x;\mu)$ is a monotone function of x, i.e., for which

$$\operatorname{sgn}(D_{xx} f(x;\mu)) = \operatorname{sgn}(c(\mu)) = \operatorname{sgn}(c).$$

Therefore $m(\mu) = f(\xi(\mu);\mu)$ in (8.16) is the unique extremal value of f for $x \in I$, and $m(0) = f(0;0) = 0$. Note that $\operatorname{sgn}(c)$ determines whether the critical point ξ is a minimum or a maximum. Moreover, since $\operatorname{sgn}(f_o(x)) = \operatorname{sgn}(c)$ when x is on the boundary of $I(0)$, this remains true by continuity, for small enough μ: $\operatorname{sgn}(f(x;\mu)) = \operatorname{sgn}(c)$ for $x \in \partial I(\mu)$. If $c > 0$, for example, then f has a minimum at ξ and is positive on the boundaries so that when $m(\mu) > 0$ there are no zeros of f, and if $m(\mu) < 0$ there are two zeros. Similar considerations apply when $c < 0$. Finally if $m(\mu) = 0$, then since $f(\xi(\mu);\mu) = 0$ and is otherwise nonzero when $x \in I(\mu)$, there is one equilibrium, $x^* = \xi(\mu)$. □

The saddle-node bifurcation "creates" a pair of equilibria as mc crosses from positive to negative values; see Figure 8.7. Indeed, near the critical point f takes the form $f(x;\mu) \approx m(\mu) + {}^1\!/_2 c(\mu)(x - \xi(\mu))^2$ so the positions of the equilibria are approximately

$$x^*_\pm(\mu) \approx \xi(\mu) \pm \sqrt{-m(\mu)/c(\mu)}.$$

Note that, by (8.15) $c(0) \neq 0$, and by (8.16) $m(0) = 0$, thus the two equilibria approach each other as $\mu \to 0$, limiting to $\xi(0)$. The stability of the two new equilibria can be computed by noting that for $c > 0$, f has a minimum at $\xi(\mu)$, and so it has negative slope at x^*_- and positive slope at x^*_+. This implies that x^*_- is a stable equilibrium and x^*_+ is an unstable equilibrium. The stabilities are reversed if $c < 0$.

The most amazing fact that we have discovered is that this bifurcation depends on a *single function* $m(\mu)$, *for any number of parameters* μ. Such a bifurcation is called *codimension-one*. This means that the condition $m(\mu) = 0$ defining the bifurcation set yields a codimension-one surface in the space of parameters.

▷ *Codimension*: A bifurcation is *codimension-k* if the bifurcation set is determined by k independent conditions on the parameters.

The bifurcation occurs when $m(\mu)$ changes sign. Since m has a somewhat obscure derivation, a more convenient criterion is needed.

Corollary 8.8. *If f satisfies the hypotheses of Theorem* 8.7*, and there is a single parameter μ_1 such that the transversality condition $D_{\mu_1} f(0;0) \neq 0$ holds, then a saddle-node bifurcation occurs as μ_1 crosses zero.*

Proof. According to Theorem 8.7, if $\partial m/\partial \mu_1 \neq 0$, then the bifurcation occurs, since we can then choose μ_1 to change the sign of m. Using the definition (8.16), so $m(\mu) = f(\xi(\mu);\mu)$, this derivative is

$$\frac{\partial m}{\partial \mu_1} = \frac{\partial}{\partial \mu_1} f(\xi;\mu) + D_x f(\xi;\mu)\frac{\partial \xi}{\partial \mu_1} = \frac{\partial}{\partial \mu_1} f(\xi;\mu),$$

since by definition of ξ, $D_x f(\xi;\mu) = 0$. □

Example 8.9. Let $f(x;\mu) = \mu_1 + \mu_2 x + x^2$. Our goal is to obtain the saddle-node bifurcation set in (μ_1, μ_2) space. First compute $\xi(\mu)$ by solving $D_x f = \mu_2 + 2x = 0$, which gives $\xi = -\mu_2/2$. Thus $m(\mu) = f(\xi;\mu) = \mu_1 - \mu_2^2/4$. So the saddle-node set is the codimension-one set (a curve), $\mu_1 = \mu_2^2/4$. Since $c = D_{xx} f(0;0) = 2 > 0$, there are two equilibria when $m < 0$, and none when $m > 0$. Of course for this case the equilibria are easily found explicitly,

$$x_{\pm}^{*} = -\frac{\mu_2}{2} \pm \sqrt{\frac{\mu_2^2}{4} - \mu_1} = -\frac{\mu_2}{2} \pm \sqrt{-m(\mu)},$$

which necessarily gives the same result. ∎

Finally we can obtain the miniversal unfolding of the saddle-node bifurcation, as predicted in (8.2).

Theorem 8.10. *The saddle-node bifurcation has a miniversal unfolding*

$$k(y;\nu) = \nu + y^2. \tag{8.18}$$

Proof. Note first that since the saddle node is a codimension-one bifurcation, a miniversal unfolding necessarily will have one parameter, equivalent to mc. Using the variable $x - \xi(\mu)$ instead of x, and noting that $f(\xi;\mu) = m$, $D_x f(\xi;\mu) = 0$, we can write

$$f(x;\mu) = m(\mu) + \tfrac{1}{2}c(\mu)(x - \xi(\mu))^2 + o(x - \xi)^2.$$

This can be simplified using the map $\nu = p(\mu) = m(\mu)c(\mu)$, and the conjugacy $y = h(x) = 1/2\, c(\mu)(x - \xi(\mu)) + o(x - \xi)$. Then f is induced by the simpler vector field $Dhf(h^{-1}(y);\mu) = \nu + y^2 + o(y^2)$. According to the one-dimensional equivalence theorem, Theorem 4.31, there is a neighborhood of the origin for which dynamics of this system is topologically conjugate to those of (8.18) because both systems have two equilibria of the same type, arranged in the same order on the line, with the same stability types. Note that ν is precisely the single parameter identified in Theorem 8.7. □

Before proceeding to the n-dimensional generalization of the saddle-node bifurcation theorem, we pause in the next section to consider the choice of the singular vector field f_o. An appropriate normal form was easily obtained from the singularity assumption in the one-dimensional case; however, the analysis in higher dimensions is not quite as simple. The n-dimensional normal form will be selected from all possible vector fields that satisfy a given singularity condition by a careful choice of coordinates. We will return to the saddle-node bifurcation in §8.6.

8.5 ▪ Normal Forms

To proceed systematically to the study of bifurcations in multidimensional systems, it is important to first simplify a dynamical system as much as possible so that its possible behaviors can be easily classified. The problem here is to find the "simplest" representative of a family of flows that are equivalent up to a coordinate transformation—we call such a system a *normal form*. For example, in §4.8 we showed that the Hartman–Grobman theorem implies that in a neighborhood of a hyperbolic equilibrium, any flow is conjugate to its linearization. Thus, an appropriate normal form in this case is the normal form of the linearization. Bifurcation theory, however, is predominantly concerned with nonhyperbolic orbits since hyperbolic orbits persist under small parameter variation.

Ideally we would like to construct a homeomorphism that linearizes the vector field just as the Hartman–Grobman theorem does. There are two problems. The first is that linearization typically fails to give a complete description of the dynamics near a nonhyperbolic orbit; consequently nonlinear terms will appear in the normal forms. The second problem is one of practicality: the group of homeomorphisms is too big to permit a systematic simplification. To obviate this, we will limit ourselves to diffeomorphisms so that power series methods can be used. Unfortunately, the construction of a diffeomorphism fails more often than would be implied by the Hartman–Grobman theorem; nevertheless, it succeeds often enough to be useful.

For the moment, we will work formally with power series representations and will not worry about their convergence. In later sections we will show that the formal, normal forms give valid, local representations of the dynamics for specific bifurcations.

Homological Operator

Suppose $x \in \mathbb{R}^n$ and, without loss of generality, assume that $\dot{x} = f(x)$ has an equilibrium at $x = 0$. Expanding f in a power series gives

$$f(x) = \sum_{k=1}^{N} f_k(x) + O(N+1). \tag{8.19}$$

Here f_k is a vector of homogeneous polynomials of degree k in x, that is,

$$f_k(\alpha x) = f_k(\alpha x_1, \alpha x_2, \ldots, \alpha x_n) = \alpha^k f(x), \tag{8.20}$$

for any $\alpha \in \mathbb{R}$. The term $O(N+1)$ represents polynomials of degree $N+1$ or larger. The space of homogeneous polynomials is denoted

$$\mathbb{H}_k = \{\text{homogeneous polynomials of degree } k \text{ in } x \in \mathbb{R}^n\}. \tag{8.21}$$

It is easy to see that $\mathbb{H}_k$ is a vector space, since a linear combination of any two homogeneous polynomials is still such a polynomial (see Exercise 5). A basis for $\mathbb{H}_k$ is the set of monomials

$$x^m \equiv x_1^{m_1} x_2^{m_2} \ldots x_n^{m_n}. \tag{8.22}$$

Here m is a vector of natural numbers, $m \in \mathbb{N}^n$, and $|m| \equiv \sum_{i=1}^n m_i = k$ is the degree. Thus for example, $\mathbb{H}_2 = \operatorname{span}\{x^2, xy, y^2\}$ is three-dimensional. We also let $\mathbb{H}_k^n = \mathbb{H}_k \times \mathbb{H}_k \times \cdots \times \mathbb{H}_k$ be the space of vectors of homogeneous polynomials on $\mathbb{R}^n$. For example, $\mathbb{H}_2^2$ has dimension 6, and the basis

$$p_1 = \begin{pmatrix} x^2 \\ 0 \end{pmatrix}, p_2 = \begin{pmatrix} xy \\ 0 \end{pmatrix}, p_3 = \begin{pmatrix} y^2 \\ 0 \end{pmatrix}, p_4 = \begin{pmatrix} 0 \\ x^2 \end{pmatrix}, p_5 = \begin{pmatrix} 0 \\ xy \end{pmatrix}, p_6 = \begin{pmatrix} 0 \\ y^2 \end{pmatrix}. \tag{8.23}$$

Thus $\mathbb{H}_2^2 = \operatorname{span}\{p_1, p_2, \ldots, p_6\}$. Of course, we could have written $\mathbb{H}_2^2$ in terms of a different basis, since the basis for any vector space is not unique. Denoting the standard unit basis of $\mathbb{R}^n$ by e_i, $i = 1, 2, \ldots, n$, the vector monomials,

$$p_{m,i} \equiv x^m e_i, \ |m| = k, \tag{8.24}$$

provide a basis for $\mathbb{H}_k^n$. The dimension of this space is the number of such vector monomials; see Exercise 5. Using this notation, the degree k terms in the power series (8.19) can be written

$$f_k = \sum_{i=1}^{n} \sum_{|m|=k} f_{m,i} p_{m,i}. \tag{8.25}$$

For example, when $k = |m| = 1$, then all the $m_i = 0$ except for one, say, $m_j = 1$, and the double sum (8.25) reduce to a sum over j and i. Therefore $f_1 = \sum_{i=1}^n \sum_{j=1}^n e_i A_{ij} x_j = Ax \in \mathbb{H}_1^n$ is the linearization, $Df(0)x$. For $k = 2$, either two of the $m_j = 1$ or one of them is 2, and the remainder are zero. The sum can be written

$$f_2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} e_i B_{ijk} x_j x_k.$$

Here the coefficients B_{ijk} are the n^3 components of the tensor $D^2 f(0)$ in the monomial basis.

Our quest is to construct the "simplest" vector field g that is conjugate to f by a near identity transformation. Let ξ represent the new variables so that $\dot{\xi} = g(\xi)$ and

$$\xi = h(x) = x + h_2(x) + O(3). \tag{8.26}$$

Recalling (4.34), we see that

$$\dot{\xi} = Dh(x)\dot{x} \quad \Rightarrow \quad g(h(x)) = Dh(x) f(x). \tag{8.27}$$

It would be simplest to choose g to be the linearization, $g(\xi) = A\xi$. Indeed, the Hartman–Grobman theorem implies that this can be achieved when A is hyperbolic (though not with a diffeomorphism). Assume for the moment that this can done in (8.27) for power series. The transformation can be constructed order by order, choosing h_k to eliminate all the nonlinear terms f_k. As we will see, certain terms in f cannot be eliminated; these must remain in g and define the nonlinear normal form.

First consider only the quadratic terms, setting $h(x) = x + h_2(x)$. We attempt to eliminate f_2 so that $g(\xi) = A\xi + O(3)$. Putting the expansions into (8.27) gives

$$\begin{aligned} Ax + Ah_2(x) + O(3) &= Dh(x) f(x) \\ &= f(x) + Dh_2(x) f(x) \\ &= Ax + f_2(x) + Dh_2(x)Ax + O(3). \end{aligned}$$

Note that the linear terms are satisfied identically. Collecting the quadratic terms gives

$$L_A(h_2) \equiv Dh_2(x)Ax - Ah_2(x) = -f_2(x), \tag{8.28}$$

which is an equation for the unknown function h_2. Arnold calls L_A the *homological operator* (Arnold 1983, Chapter 5).

The homological operator is a linear operator (recall §2.3) on the space of degree k vector fields: $L_A : \mathbb{H}_k^n \to \mathbb{H}_k^n$ (see Exercise 6). More generally, given a pair of vector fields X, and Y, the Lie bracket is defined as $L_X(Y) = [X,Y] = DY(X) - DX(Y)$. When $X = Ax$ is linear, the Lie bracket reduces to the homological operator. This operator is also sometimes called the adjoint operator and denoted ad_X.

In principle, h_2 could be obtained by inverting the homological operator to obtain $h_2 = -L_A^{-1}(f_2)$. However, the kernel of L_A is typically not trivial so that it does not have an inverse. Just as in the case of a matrix, (2.11), the kernel of a linear operator $L : \mathbb{H} \to \mathbb{H}$ is its null space:

$$\ker(L) \equiv \{p \in \mathbb{H} : L(p) = 0\}.$$

When L has a nontrivial kernel, (8.28) is solvable only when $f_2 \in \operatorname{rng}(L)$. This solvability condition has another formulation if we are given an inner product, $\langle r,p\rangle$, on $\mathbb{H}$. The adjoint, $L^\dagger$, of L is then defined by $\langle p, Lr\rangle \equiv \left\langle L^\dagger p, r\right\rangle$ for any $r,p \in \mathbb{H}$, and its cokernel is the null space of this adjoint:

$$\operatorname{coker}(L) \equiv \left\{r \in \mathbb{H} : L^\dagger(r) = 0\right\}. \tag{8.29}$$

One possible inner product is discussed in Exercise 7.

A system $Lp = f$ is solvable if and only if f is orthogonal to the cokernel of L, i.e.,

$$\langle r, f\rangle = 0 \text{ for all } r \in \operatorname{coker}(L); \tag{8.30}$$

this is called the Fredholm condition (Olver and Shakiban 2006). Indeed, if there is a solution, $Lp = f$, then for any $r \in \operatorname{coker}(L), \langle f, r\rangle = \langle Lp, r\rangle = \left\langle p, L^\dagger r\right\rangle = 0$ so f satisfies (8.30). Moreover, if $f \in \operatorname{rng}(L)$, then by definition there exists a p such that $Lp = f$, so f satisfies (8.30). Consequently,

$$\mathbb{H} = \operatorname{rng}(L) \oplus \mathbb{G}, \tag{8.31}$$

where $\mathbb{G}$ is a complement to $\operatorname{rng}(L)$.[48] Whenever f does not satisfy (8.30), then the equation $Lp = f$ is inconsistent. To emphasize this, split f into two parts,

$$f = \tilde{f} + f^R, \quad \tilde{f} \in \operatorname{rng}(L), \quad f^R \in \mathbb{G}.$$

The function f^R is the *resonant* part of f.

Using this splitting, we then begin anew with (8.27) but ask only that the normal form eliminate all nonresonant terms, so that

$$g(\xi) = A\xi + f^R(\xi). \tag{8.32}$$

[48] The fundamental theorem of linear algebra (2.13) implies that it is possible to choose $\mathbb{G} = \operatorname{coker}(L) = \operatorname{rng}(L)^\perp$; however, this may not be most convenient, so at this stage we leave the choice open. Consequently, the resonant terms are not uniquely defined. This gives rise to the possibility of a number of different normal forms for a given bifurcation, as we will see when we treat the Takens–Bogdanov case below.

Using this form in (8.28) results in the equation

$$L_A(h_2) = -\tilde{f}_2, \tag{8.33}$$

which is guaranteed to have a solution for h_2.

The problem of constructing a normal form can be reduced to the following set of tasks:

- For a given linearization, A, find a representation of the homological operator L_A on $\mathbb{H}_k^n$.
- Resolve f into components in $\text{rng}(L_A)$ and a complementary space $\mathbb{G}$.
- Solve for the transformation h, eliminating all nonresonant terms, leaving the normal form $g(x) = Ax + f^R$.

Matrix Representation

For the homological operator acting on $\mathbb{H}_k^n$ the calculation of the resonant terms can be reduced to matrix algebra; indeed, any linear operator on a finite-dimensional space has a matrix representation. Specifically, suppose $L : \mathbb{H} \to \mathbb{H}$, where $\dim(\mathbb{H}) = d$, and let p_j, $i = 1, 2, \ldots, d$, represent a basis of $\mathbb{H}$. Since L is a linear operator, then $L(p_i) \in \mathbb{H}$ and is necessarily given by a linear combination of the basis vectors:

$$L(p_j) = \sum_{i=1}^{d} p_i L_{ij}. \tag{8.34}$$

This defines the $d \times d$ matrix, L, as the *representation* of the action of the operator L on $\mathbb{H}$; recall §2.3. Writing a general vector in this basis as $h = \sum_{i=1}^{d} h_i p_i$, the equation $L(h) = f$ becomes

$$L(h) = L\left(\sum_{j=1}^{d} h_j p_j\right) = \sum_{i,j}^{d} p_i L_{ij} h_j = \sum_{i=1}^{d} f_i p_i \;\Rightarrow\; \sum_{i=1}^{d}\left(\sum_{j=1}^{d} L_{ij} h_j - f_i\right) p_i = 0.$$

Since the basis vectors are linearly independent, this is equivalent to the matrix equation $Lh = f$ for the d-dimensional coefficient vectors h and f—that this looks almost exactly the same as the original operator equation is intentional.

Accordingly, the kernel of the operator L in the p-basis is simply the kernel of the matrix L. If the vectors are real, then we can define the inner product as $\langle h, f\rangle = \sum_{i=1}^{d} h_i f_i$ so that the transposed matrix represents the adjoint of L and the cokernel of L is simply the kernel of the transpose, L^T.

The simplest example corresponds to the case that A is real and diagonal, i.e., $A = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$. We compute the action of L_A on the monomial basis of $\mathbb{H}_k^n$ using the basis vectors (8.24). Note that $Ap_{m,i} = \lambda_i p_{m,i}$ and

$$Dp_{m,i}(x)Ax = \sum_{j=1}^{n} e_i \frac{\partial}{\partial x_j}(x^m)\lambda_j x_j = e_i \sum_{j=1}^{n} (m_j \lambda_j) x^m = m \cdot \lambda p_{m,i}.$$

Therefore, using (8.28),

$$L_A(p_{m,i}) = (m \cdot \lambda - \lambda_i) p_{m,i} = \mu_{m,i} p_{m,i}. \tag{8.35}$$

This shows that the vector monomials are eigenfunctions of L_A on $\mathbb{H}_k^n$, with eigenvalues

$$\mu_{m,i} = m \cdot \lambda - \lambda_i . \tag{8.36}$$

Since the vector monomials $p_{m,i}$ provide a basis for $\mathbb{H}_k^n$, if all the $\mu_{m,i}$ are nonzero then $\ker(L_A) = \operatorname{coker}(L_A) = \{0\}$. In this case we can invert L_A to obtain h_2 from (8.28):

$$h_2 = \sum_{|m|=2,i} \frac{f_{2,m,i}}{\lambda_i - m \cdot \lambda} x^m e_i .$$

Example 8.11. Consider the one-dimensional case with a hyperbolic equilibrium: $A = (\lambda)$. Then

$$\dot{x} = \lambda x + a x^2 + b x^3 + \cdots ,$$

we set $\xi = h(x) = x + \alpha x^2 + \beta x^3 + \cdots$, and the homological operator is

$$L_A(h) = Dh(x)\lambda x - \lambda h(x),$$

so $L_A(x^m) = \mu_m x^m$ with $\mu_m = (m-1)\lambda$. Since we consider $m \geq 2$, there are no resonant terms when $\lambda \neq 0$. At quadratic order we must solve

$$L_A(\alpha x^2) = \lambda \alpha x^2 = -f_2 = -a x^2 ,$$

so $\alpha = -a/\lambda$. Thus to second order we choose $\xi = x - a x^2/\lambda$. To show that the ODE is indeed transformed, it is necessary to invert the $\xi = h(x)$; this can locally be done by recursion:

$$x = \xi + \frac{a}{\lambda} x^2 = \xi + \frac{a}{\lambda}\left(\xi + \frac{a}{\lambda} x^2\right)^2 = \xi + \frac{a}{\lambda}\xi^2 + 2\frac{a^2}{\lambda^2}\xi^3 + O(\xi^4).$$

The new dynamical equation is

$$\begin{aligned}
\dot{\xi} &= \frac{d}{dt}\left(x - \frac{a}{\lambda}x^2\right) = \lambda x + a x^2 + b x^3 + \cdots - 2\frac{a}{\lambda} x \left(\lambda x + a x^2 + \cdots\right), \\
&= \lambda x - a x^2 + \left(b - 2a^2/\lambda\right) x^3 + \cdots , \\
&= \lambda\left(\xi + a\xi^2/\lambda + 2a^2\xi^3/\lambda^2 + \cdots\right) - a\left(\xi + a\xi^2/\lambda + \cdots\right)^2 + \left(b - 2a^2/\lambda\right)\xi^3 + \cdots , \\
&= \lambda\xi + \left(b - 2a^2/\lambda\right)\xi^3 + O(\xi^4).
\end{aligned}$$

Thus, the quadratic term has been successfully eliminated. We could now proceed to eliminate the cubic terms with an h_3.

The only problem with this analysis happens when $\lambda = 0$. Now none of the nonlinear terms can be eliminated since $L_A \equiv 0$! Indeed, when $\lambda = 0$ *every* monomial x^m is resonant, and the nonhyperbolic equation $\dot{x} = a x^2 + O(3)$ cannot be simplified by this technique. Luckily, we have already dealt with this situation in §8.4. ∎

More generally, it can happen that one or more of the $\mu_{m,i}$ in (8.36) are zero. This occurs if, for some m, i we have $\lambda_i = m \cdot \lambda$. For example, the eigenvalues for $n = k = 2$ are shown in Table 8.1. Whenever $\lambda_i = 0$ or $\lambda_i = 2\lambda_j$ there are resonances, and $\ker(L_A)$ is nontrivial. The first case should be expected to cause a problem since then the fixed

Table 8.1. *Eigenvectors of* (8.35).

m,i	$(2,0),1$	$(1,1),1$	$(0,2),1$	$(2,0),2$	$(1,1),2$	$(0,2),2$
Basis vector	$\begin{pmatrix} x^2 \\ 0 \end{pmatrix}$	$\begin{pmatrix} xy \\ 0 \end{pmatrix}$	$\begin{pmatrix} y^2 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0 \\ x^2 \end{pmatrix}$	$\begin{pmatrix} 0 \\ xy \end{pmatrix}$	$\begin{pmatrix} 0 \\ y^2 \end{pmatrix}$
$\mu_{m,i}$	λ_1	λ_2	$2\lambda_2-\lambda_1$	$2\lambda_1-\lambda_2$	λ_1	λ_2

point is not hyperbolic. The second case is not so obvious; it arises from the use of power series for the conjugacy.

When there are resonances, $\mathbb{H}_k^n$ can be decomposed as (8.31) into the range of L_A and a complementary subspace $\mathbb{G}$. The simplest choice for $\mathbb{G}$, the cokernel of L_A, is fine here. Indeed, since L_A has a diagonal representation, its kernel and cokernel are identical. Consequently $\mathbb{G}$ is spanned by the zero eigenvectors of L_A.

Example 8.12. Consider a two-dimensional system whose linearization has a single zero eigenvalue. As we argued in §8.3, the linearization can be put in the form (8.8). Thus the power series for f_o is

$$\begin{aligned} \dot{x} &= ax^2 + bxy + cy^2 + \cdots, \\ \dot{y} &= \lambda y + dx^2 + exy + fy^2 + \cdots, \end{aligned} \tag{8.37}$$

where $\lambda = \pm 1$. Denoting the components of $h(x,y) = (h_x, h_y)$, we see that the homological operator (8.28) becomes

$$L_A(h) = \begin{pmatrix} \partial_x h_x & \partial_y h_x \\ \partial_x h_y & \partial_y h_y \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & \lambda \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} 0 & 0 \\ 0 & \lambda \end{pmatrix} \begin{pmatrix} h_x \\ h_y \end{pmatrix} = \lambda \begin{pmatrix} y\partial_y h_x \\ y\partial_y h_y - h_y \end{pmatrix}.$$

This is a special case of the diagonal A that was treated earlier, and each of the monomials $p_{m_1,m_2,i} = x^{m_1} y^{m_2} e_i$ is an eigenvector of L_A with eigenvalues (8.36). For $k = 2$, Table 8.1 shows that L_A has precisely two zero eigenvectors, p_1 and p_5:

$$L_A \begin{pmatrix} x^2 \\ 0 \end{pmatrix} = 0, \quad L_A \begin{pmatrix} 0 \\ xy \end{pmatrix} = 0.$$

These are a basis for coker(L_A), and are the two resonant terms that cannot be eliminated. This implies that f^R is given by a linear combination of these vectors and that the normal form is

$$\begin{aligned} \dot{\xi} &= a\xi^2 + O(3), \\ \dot{\eta} &= \lambda\eta + e\xi\eta + O(3). \end{aligned} \tag{8.38}$$

To this order, the normal form is a skew product, and whenever $a \neq 0$, the ξ motion is "semistable" (recall §4.5), while the η motion is hyperbolic. The system (8.38) is called a *saddle node*; we will study its unfolding in §8.6. ■

Higher-Order Normal Forms

It is not necessary to stop with the elimination of the quadratic terms. To see this we use induction. First, suppose the normal form is known to some order, i.e., that all terms in the range of L_A have been eliminated below order k. To this order, the normal form is

$$\dot{x} = Ax + f_{k-1}^R(x) + f_k(x) + \cdots,$$

where f_{k-1}^R contains the resonant terms through order $k-1$. Now let

$$\xi = x + h_k(x)$$

and demand that the dynamics for ξ has only resonant terms through order k, so that

$$\dot{\xi} = g(\xi) = A\xi + g_k^R(\xi) + O(k+1).$$

Using (8.27) through $O(k)$ we obtain

$$\begin{aligned} Ah(x) + g_k^R(h(x)) + O(k+1) &= Dh(x)\left(Ax + f_{k-1}^R(x) + f_k(x)\right) + O(k+1), \\ Ax + Ah_k(x) + g_k^R(x) + O(k+1) &= Ax + f_{k-1}^R(x) + f_k(x) + Dh_k(x)Ax + O(k+1). \end{aligned}$$

To solve this, set $g_k^R = f_{k-1}^R + f_k^R$, and determine h_k from the equation

$$L_A(h_k) = f_k^R - f_k, \tag{8.39}$$

where L_A is again the homological operator (8.28). Moreover, (8.39) is solvable since its right-hand side is constructed to be in the range of L_A.

Example 8.13. We already worked out the normal form of the flow of (8.37) to quadratic order. Every higher-order monomial $x^{m_1}y^{m_2}e_i \in \mathbb{H}_{|m|}^n$ can be eliminated, provided it does not satisfy one of the resonance conditions $\mu_{m,i} = m \cdot \lambda - \lambda_i = 0$ with $\lambda_1 = 0$ and $\lambda_2 = \pm 1$. For $i = 1$ the resonant terms correspond to $\mu_{m,1} = \pm m_2 = 0$, so $m = (k, 0)$, and for $i = 2$, they correspond to $\mu_{m,2} = \pm(m_2 - 1) = 0$, so $m = (k-1, 1)$ for $k = 2, 3, \ldots$. Thus the normal form to arbitrary order N is

$$\begin{aligned} \dot{\xi} &= \sum_{k=2}^{N} c_k \xi^k, \\ \dot{\eta} &= \lambda\eta + \eta \sum_{k=2}^{N} d_k \xi^{k-1}. \end{aligned} \tag{8.40}$$

Just like the quadratic normal form (8.38), this is a skew-product system. ■

Example 8.14. A linear center on $\mathbb{R}^2$ has the real normal form (8.11) so that f_o becomes

$$\begin{aligned} \dot{x} &= -y + ax^2 + bxy + cy^2 + \cdots, \\ \dot{y} &= x + dx^2 + exy + fy^2 + \cdots. \end{aligned}$$

For the matrix (8.11) the homological operator is

$$L_A(h) = \begin{pmatrix} \partial_x h_x & \partial_y h_x \\ \partial_x h_y & \partial_y h_y \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} h_x \\ h_y \end{pmatrix} = \begin{pmatrix} x\partial_y h_x - y\partial_x h_x + h_y \\ x\partial_y h_y - y\partial_x h_y - h_x \end{pmatrix}. \tag{8.41}$$

In this case the monomial basis vectors are not eigenvectors: the matrix representation for L is not diagonal.[49] The elements of this matrix in the standard basis (8.23) for $\mathbb{H}_2^2$

[49] We could achieve a diagonal representation for L if we used a complex basis. We will do this in §8.8.

are obtained from

$$L_A(p_1) = -\begin{pmatrix} 2xy \\ x^2 \end{pmatrix} = -2p_2 - p_4, \qquad L_A(p_2) = \begin{pmatrix} x^2 - y^2 \\ -xy \end{pmatrix} = p_1 - p_3 - p_5,$$

$$L_A(p_3) = \begin{pmatrix} 2xy \\ -y^2 \end{pmatrix} = 2p_2 - p_6, \qquad L_A(p_4) = \begin{pmatrix} x^2 \\ -2xy \end{pmatrix} = p_1 - 2p_5,$$

$$L_A(p_5) = \begin{pmatrix} xy \\ x^2 - y^2 \end{pmatrix} = p_2 + p_4 - p_6, \quad L_A(p_6) = \begin{pmatrix} y^2 \\ 2xy \end{pmatrix} = p_3 + 2p_5.$$

The matrix representation (8.34) becomes

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

This matrix is diagonalizable and has eigenvalues, $\mu = \pm i$ (double) and $\pm 3i$—none of which are zero. Thus at this order there are no resonances and all the quadratic terms can be eliminated.

At cubic order, there are eight basis vectors, and so the operator L_A has an 8×8 representation. Considerable algebra leads to the conclusion that there are only two eigenvectors with zero eigenvalue; thus two vectors span $\mathbb{G}$ (see Exercise 8). The two polynomials

$$v_1 = \begin{pmatrix} (3x^2 + y^2)x \\ (x^2 + 3y^2)y \end{pmatrix}, \quad v_2 = \begin{pmatrix} (x^2 + 3y^2)y \\ -(3x^2 + y^2)x \end{pmatrix} \tag{8.42}$$

give a basis for coker(L_A) since $L_A^\dagger v_i = 0$. It turns out that this is not the most convenient choice for $\mathbb{G}$, however. A better choice corresponds to the two null right eigenvectors:

$$L_A \begin{pmatrix} (x^2+y^2)x \\ (x^2+y^2)y \end{pmatrix} = \begin{pmatrix} (x^2+y^2)y - (3x^2+y^2)y + 2yx^2 \\ -(x^2+y^2)x - 2xy^2 + (x^2+3y^2)x \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

$$L_A \begin{pmatrix} -(x^2+y^2)y \\ (x^2+y^2)x \end{pmatrix} = \begin{pmatrix} (x^2+y^2)x + 2xy^2 - (x^2+3y^2)x \\ (x^2+y^2)y - (3x^2+y^2)y + 2yx^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

In Exercise 8, you will show that these two vectors, together with the column space of L_A, span $\mathbb{H}_3^2$. Thus to cubic order, we can choose these vectors to span $\mathbb{G}$, and the resulting normal form is

$$\begin{aligned} \dot{x} &= -y + (x^2+y^2)(\alpha x - \beta y) + O(4), \\ \dot{y} &= x + (x^2+y^2)(\alpha y + \beta x) + O(4). \end{aligned} \tag{8.43}$$

It is easier to see what this equation means in polar coordinates. Applying the polar transformation, (6.4), to (8.43) yields

$$\begin{aligned} \dot{r} &= \alpha r^3, \\ \dot{\theta} &= 1 + \beta r^2. \end{aligned}$$

Thus when $\alpha > 0$ the origin is a spiral source, and when $\alpha < 0$ it is a spiral sink. Note that the coefficient α depends on the coefficients (a, b, c, d, e, f) of the original vector field—to actually compute α, the quadratic transformation has to be carried out explicitly since this cubic term will be modified by this calculation. We will do this in §8.8. ■

8.6 ▪ Saddle-Node Bifurcation in $\mathbb{R}^n$

A saddle-node bifurcation typically occurs when a *single* eigenvalue of a linearization crosses zero. This higher-dimensional case is essentially the same as that for one dimension discussed in §8.4, though in this case the name "saddle node" makes more sense. For example, suppose $n = 2$, and the linear part is

$$Df_o(0) = \begin{pmatrix} 0 & 0 \\ 0 & \lambda \end{pmatrix}$$

with $\lambda < 0$, so that the vector field has the power series expansion (8.37). The bifurcation corresponds to the creation of two equilibria. Both will have a stable direction corresponding to λ; one will have a second stable direction and thus be a node, while the second will have an unstable direction and thus be a saddle.

We have already constructed the normal form at the bifurcation point to all orders in (8.40). Note that this is a skew product since the x component is independent of y; thus as far as the normal form is concerned, the dynamics reduces to the one-dimensional case. However, to understand the bifurcation, we now have to unfold f_o. But, as we shall see, nothing untoward happens.

Theorem 8.15 (saddle node). *Let $f \in C^2(\mathbb{R}^n \times \mathbb{R}^k, \mathbb{R}^n)$, and suppose that $f(z;\mu)$ satisfies*

$$f(0;0) = 0,\ \mathrm{spec}(D_z f(0;0)) = \{0, \lambda_2, \lambda_3, \ldots, \lambda_n : \lambda_k \neq 0, k \neq 1\}. \qquad (singularity)$$

Choose coordinates so that $D_z f(0;0)$ is diagonal in the zero eigenvalue and set $z = (x,y)$ where $x \in \mathbb{R}^1$ corresponds to the zero eigenvalue and $y \in \mathbb{R}^{n-1}$ are the remaining coordinates. Then

$$\begin{aligned} \dot{x} &= g_1(x,y;\mu), \\ \dot{y} &= My + g_2(x,y;\mu), \end{aligned} \tag{8.44}$$

where $g(0,0;0) = 0$ and $D_z g(0,0;0) = 0$. Suppose that

$$D_{xx} g_1(0,0;0) = c \neq 0. \qquad (nondegeneracy)$$

Then there exists an interval $I(\mu)$ containing 0, functions $y = \eta(x;\mu)$ and $m(\mu) =$ $\mathrm{Ext}_{x \in I(\mu)}[g_1(x;\eta(x;\mu);\mu)]$, and a neighborhood of $\mu = 0$ such that if $m(\mu)c > 0$ there are no equilibria and if $m(\mu)c < 0$ there are two. Suppose that M has a u-dimensional unstable space and an $(n-u-1)$-dimensional stable space. Then, when there are two equilibria, one has a u-dimensional unstable manifold and an $(n-u)$-dimensional stable manifold and the other has a $(u+1)$-dimensional unstable manifold and an $(n-u-1)$-dimensional stable manifold.

Proof. The equilibria are solutions of

$$\begin{aligned} F_1(x,y;\mu) &= g_1(x,y;\mu) = 0, \\ F_2(x,y;\mu) &= My + g_2(x,y;\mu) = 0. \end{aligned}$$

By assumption, $D_y F_2(0,0;0) = M$ is nonsingular; thus Theorem 8.3 ensures that there is a neighborhood of $(x,\mu) = (0,0)$ where there exists a unique function $y = \eta(x;\mu)$ such that

$$F_2(x;\eta(x;\mu);\mu) = 0 \tag{8.45}$$

and $\eta(0;0) = 0$. Substitute this into $F_1 = 0$ to obtain

$$F(x;\mu) = g_1(x,\eta(x;\mu);\mu) = 0.$$

Consequently, the problem has been reduced to the one-dimensional case; we need only check that F satisfies the same criteria as Theorem 8.7, the one-dimensional case. It is easy to see that $F(0;0) = 0$. Since f is C^2, so is η, and differentiation of (8.45) with respect to x gives

$$M\frac{d\eta}{dx} + D_x g_2 + D_y g_2 \frac{d\eta}{dx} = 0.$$

Since $D_x g(0,0;0) = D_y g(0,0;0) = 0$, this implies that $\frac{d\eta}{dx}(0;0) = 0$. This relation helps compute the required derivatives of F:

$$\begin{aligned} D_x F(0;0) &= D_x g_1 + D_y g_1 \frac{d\eta}{dx} = 0, \\ D_{xx} F(0;0) &= D_{xx} g_1 + 2D_{xy} g_1 \frac{d\eta}{dx} + D_{yy} g_1 \left(\frac{d\eta}{dx}\right)^2 + D_y g_1 \frac{d^2\eta}{dx^2}, \\ &= D_{xx} g_1(0,0;0) = c \neq 0. \end{aligned}$$

Thus the needed hypotheses for Theorem 8.7 are satisfied, and there exists an extremal value $m(\mu)$ such that when m crosses zero the number of equilibria changes from zero to two. The stability of the new equilibria follows easily along the lines of the proof of Theorem 8.7. □

Transversality

As we discussed in §8.4, the saddle-node bifurcation has codimension-one because there is a single condition on the parameters, $m = 0$, that determines the bifurcation set. However, Theorem 8.15 does not guarantee that a saddle-node bifurcation occurs. Indeed, even though the extremal value $m(\mu)$ vanishes when the parameters are zero, m need not change sign as any of parameters cross zero (see Exercise 10). It is not hard, however, to obtain a simple criterion that guarantees that the bifurcation takes place.

Corollary 8.16 (transversality). *If μ_1 is any single parameter such that*

$$D_{\mu_1} g_1(0,0;0) \neq 0, \qquad \textit{(transversality)}$$

then a saddle-node bifurcation takes place when μ_1 crosses zero.

Proof. We must show that $\partial m/\partial \mu_1 \neq 0$. Using $\xi(\mu)$ to denote the critical point of $F(x;\mu)$ in x, then we have

$$\begin{aligned} \left.\frac{\partial m}{\partial \mu_1}\right|_{\mu=0} &= \frac{\partial}{\partial \mu_1} F_1(\xi(\mu),\eta(\xi;\mu);\mu)|_{\mu=0}, \\ &= D_x g_1(0,0;0)\frac{\partial \xi}{\partial \mu_1} + D_y g_1(0,0;0)\left(D_\xi \eta \frac{\partial \xi}{\partial \mu_1} + \frac{\partial \eta}{\partial \mu_1}\right) + \frac{\partial}{\partial \mu_1} g_1(0,0;0). \end{aligned}$$

The first derivatives $D_x g_1$ and $D_y g_1$ both vanish by assumption, and the transversality assumption gives $D_{\mu_1} F = D_{\mu_1} g_1 \neq 0$. □

Example 8.17. Consider the system

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -y + x^2 - \mu. \end{aligned}$$

This is almost too simple for our full analysis, but let us proceed anyway. When $\mu = 0$, there is a nonhyperbolic equilibrium point at the origin with eigenvalues $\lambda_1 = 0$ and $\lambda_2 = -1$. The corresponding eigenvectors are $v_1 = (1,0)^T$ and $v_2 = (-1,1)^T$. To proceed, we put the system into the canonical form (8.44) with the transformation $x = P(\xi, \eta)^T$, where $P = (v_1, v_2)$. This gives $x = \xi - \eta$ and $y = \eta$. The transformed equations are

$$\begin{pmatrix} \dot{\xi} \\ \dot{\eta} \end{pmatrix} = \begin{pmatrix} (\xi-\eta)^2 - \mu \\ -\eta + (\xi-\eta)^2 - \mu \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix} + \begin{pmatrix} (\xi-\eta)^2 - \mu \\ (\xi-\eta)^2 - \mu \end{pmatrix}.$$

This system satisfies the nondegeneracy condition since $c = D_{\xi\xi} g_1(0;0) = 2 \neq 0$. Furthermore $D_\mu g_1 = -1$, so the transversality condition is also satisfied. Since $c > 0$ and $D_\mu g_1 < 0$, F has a minimum and the minimum decreases through zero as μ increases. Consequently, there are no equilibria when $\mu < 0$ and two when $\mu > 0$.

Going back to the original system, we can easily solve for the equilibria to get $y = 0$ and $x = \pm\sqrt{\mu}$, confirming our result. ■

Example 8.18. Consider the equations

$$\begin{aligned} \dot{x} &= \mu - x^2 + xy - xy^2, \\ \dot{y} &= \lambda - y - x^2 + yx^2. \end{aligned} \tag{8.46}$$

This system has two parameters, but if there is a saddle-node bifurcation, only one will be relevant.

There is a nonhyperbolic equilibrium when $\lambda = \mu = 0$ at the origin that is already in the canonical form (8.44); see Figure 8.8. We compute $c = D_{xx} g_1(0;0) = -2$ and $D_\mu g_1 = 1$. Thus there is a saddle-node bifurcation when μ goes from negative to positive values. Note that variation of μ alone can create the bifurcation, but it is not immediately clear whether variation of λ can do this. To determine this, compute the bifurcation function by solving for y from the second equation (of course, Theorem 8.3 guarantees there is a solution):

$$y = \eta(x;\lambda) = \frac{\lambda - x^2}{1 - x^2} = \lambda + (\lambda - 1)x^2 + O(x^4).$$

Here we have expanded the expression, since we are interested only in small x. Substitution into g_1 gives

$$F(x;\mu,\lambda) = g_1(x, \eta(x;\lambda); \mu, \lambda) = \mu + \lambda(1-\lambda)x - x^2 + O(x^3).$$

To quadratic order, the critical point and critical value of F are, respectively,

$$\xi(\mu,\lambda) \approx \frac{\lambda(1-\lambda)}{2}, \quad m(\mu,\lambda) \approx \mu + \frac{\lambda^2(1-\lambda)^2}{4}.$$

Thus there is a single equilibrium near the origin along the curve $m(\mu,\lambda) = 0$, or equivalently when $\mu = -\lambda^2(1-\lambda)^2/4$. Since $c < 0$, there are no equilibria when $\mu < -\lambda^2(1-\lambda)^2/4$, and two when μ is greater. Two phase portraits of this system are shown in Figure 8.9. ■

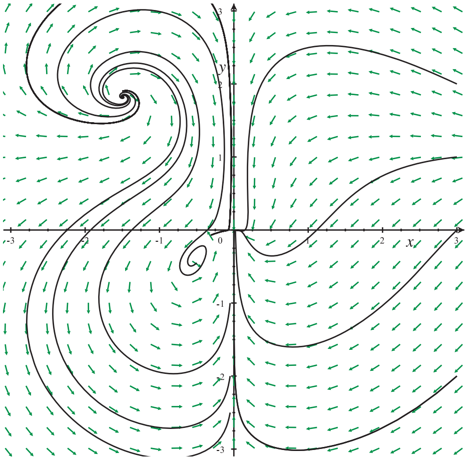

Figure 8.8. *Phase space for* (8.46) *with* $\mu=\lambda=0$*. The origin is a nonhyperbolic equilibrium. Two other equilibria (foci) are also shown.*

Center Manifold Methods

An alternative way to find the saddle-node bifurcation is to introduce additional, trivial, differential equations for the parameters, so that the system (8.44) becomes

$$\begin{aligned}\dot{x} &= g_1(x,y,\mu),\\ \dot{\mu} &= 0,\\ \dot{y} &= My + g_2(x,y,\mu).\end{aligned} \tag{8.47}$$

If there are k parameters, $\mu \in \mathbb{R}^k$, then this system has a $(k+1)$-dimensional center space at the equilibrium $(x,\mu,y)=(0,0,0)$. The center manifold can be computed using the methods of §5.6. Indeed, the system (8.47) is already in a form similar to (5.34). The manifold W^c is a graph over the center coordinates, $y=h(x,\mu)$, and must be invariant, so

$$\dot{y} = D_x h \frac{dx}{dt} + D_\mu h \frac{d\mu}{dt} \;\Rightarrow\; Mh + g_2(x,h,\mu) = D_x h\, g_1(x,h,\mu). \tag{8.48}$$

The reduced dynamics on the center manifold are $\dot{x}=g_1(x,h(x,\mu),\mu)$, and of course $\dot{\mu}=0$. Thus the graph $h(x,\mu)$ replaces the function $\eta(x;\mu)$ of Theorem 8.15.

Example 8.19. For example, consider the model

$$\begin{aligned}\dot{x} &= \mu - x^2 + xy,\\ \dot{\mu} &= 0,\\ \dot{y} &= -y + \mu x + x^2\end{aligned} \tag{8.49}$$

with a single parameter μ. Since the center manifold is tangent to $x=\mu=0$, the series for h begins with quadratic terms: $h(x,\mu)=ax^2+bx\mu+c\mu^2+O(3)$. Substitution into (8.48) gives coefficients,

$$-(ax^2+bx\mu+c\mu^2)+\mu x+x^2 = (2ax+b\mu)(\mu-x^2)+O(3)$$

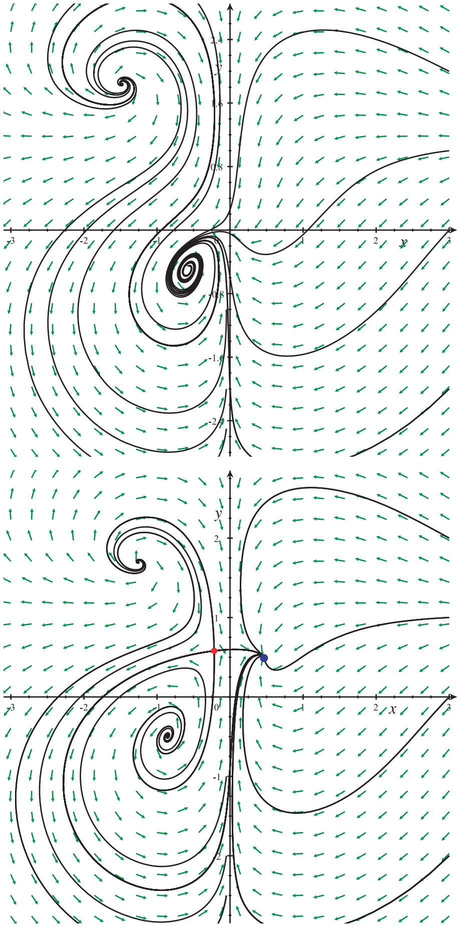

Figure 8.9. *Phase portraits of* (8.46) *for* $\mu = -0.1$ *and* $\lambda = 0$, *so that* $m > 0$ *(left), and* $\mu = 0.1$ *and* $\lambda = 0.6$ *so that* $m < 0$ *(right). In the right panel the newly created equilibria are a saddle and a stable node.*

through quadratic order. Comparing terms of a given order in both variables gives the three equations

$$(1-a)x^2 = 0,\ (-b+1-2a)x\mu = 0,\ (c+b)\mu^2 = 0$$

with the solutions $a = 1,\ b = -1$, and $c = 1$. Thus the center manifold is defined by $y = x^2 - \mu x + \mu^2$. Note that this is *not* the equilibrium equation, $y = \mu x + x^2$, because the invariance of the center manifold is a dynamical property. After more algebra we find, through cubic order,

$$h(x,\mu) = x^2 - \mu x + \mu^2 + 2x^3 - 7\mu x^2 + 14\mu^2 x - 14\mu^3 + O(4).$$

Substituting this into the ODE for x yields

$$\dot{x}|_{W^c} = \mu + \mu^2 x - (1-\mu)x^2 + x^3 + O(4)$$

for the dynamics on the center manifold. This equation is equivalent to the standard one-dimensional unfolding (8.17), and we can compute the extremal value $m = \mu +$

$\mu^4/4+O(\mu^5)$ and the curvature $c=-2$. Therefore, as μ crosses zero from below there is a saddle-node bifurcation that creates a pair of equilibria on the center manifold near $x=0$. ∎

8.7 ▪ Degenerate Saddle-Node Bifurcations

Theorem 8.10 showed that the one-dimensional normal form $f_o(x)=x^2$ has the miniversal unfolding $f(x;\mu)=\mu+x^2$. If the unfolding does not satisfy the transversality condition, $D_\mu f(0;0)\neq 0$, then the bifurcation can be somewhat different in character. For example, for the special unfolding

$$\dot{x}=\mu x+x^2, \tag{8.50}$$

we find $m(\mu)=-\mu^2/4\leq 0$. Thus m never changes sign. In fact, there are always two equilibria, at $x=0$ and $-\mu$. As we showed in §8.1, a bifurcation does occur at $\mu=0$ because the fixed points exchange stability types. This *transcritical* bifurcation is not a versal unfolding of the saddle-node singularity, since we have seen that more general unfoldings have parameter values for which there are no equilibria. However, one can observe this bifurcation in systems where the transversality condition, Corollary 8.16, is not satisfied, but $D_{\mu_1 x}g_1\neq 0$.

Another interesting bifurcation occurs when the nondegeneracy condition of Theorem 8.15 is violated, that is, when the quadratic term of f_o vanishes. The one-dimensional version of this corresponds to the vector field

$$f_o(x)=dx^3+g(x),$$

where $g=o(x^3)$. An unfolding of this system will in general contain all three of the lower-order terms,

$$f(x;\mu)=a(\mu)+b(\mu)x+c(\mu)x^2+d(\mu)x^3+g(x;\mu).$$

A special case of this system occurs when $a(\mu)=c(\mu)\equiv 0$. This gives rise to the *pitchfork* bifurcation, which corresponds to an equilibrium losing stability by the creation of two new equilibria. It is special because the constant term vanishes. (We will see in §8.9 that the quadratic term is not essential.)

Example 8.20. Consider the ODE

$$\dot{x}=\mu x-x^3. \tag{8.51}$$

There is always one equilibrium at $x=0$, and there are two others at $x=\pm\sqrt{\mu}$, provided $\mu>0$. Note that the origin is stable for $\mu<0$ but becomes unstable for $\mu>0$. The two new orbits have eigenvalues $Df(\pm\sqrt{\mu})=-2\mu$, which implies that they are stable when $\mu>0$. The resulting bifurcation diagram is shown in Figure 8.10. ∎

The form of the pitchfork bifurcation in (8.51) is called *supercritical*, because the new orbits that are created are stable. If, instead, the new orbits are unstable, the bifurcation is *subcritical.* A prototype for the subcritical case is simply

$$\dot{x}=\mu x+x^3.$$

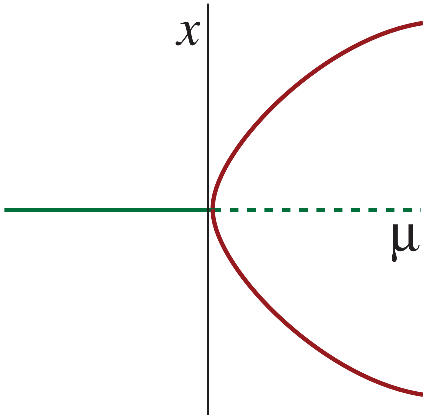

Figure 8.10. *Supercritical pitchfork bifurcation of* (8.51) *creates a pair of stable equilibria.*

Here the new orbits exists for $\mu < 0$ and are unstable, while when $\mu > 0$ the only equilibrium is unstable. That this subcritical bifurcation creates orbits as μ is decreased is not the essential point—we could replace $\mu \to -\mu$, and then the bifurcation would create orbits for an increasing parameter. The essential point is that the newly created orbits are unstable.

This bifurcation is common in systems with symmetry. The study of bifurcations in the presence of symmetries has received much recent attention (Golubitsky and Schaeffer 1985; Golubitsky, Stewart, and Schaeffer 1988). The pitchfork is also special case of a codimension-two bifurcation, the cusp—see §8.9.

8.8 ▪ Andronov–Hopf Bifurcation

In the 1890s Poincaré and Lyapunov studied what we now call the Andronov–Hopf bifurcation; however, Andronov proved the first theorem, for the two-dimensional case, in 1929. Hopf obtained the higher-dimensional result in 1942. This bifurcation typically occurs when an equilibrium has a pair of eigenvalues that cross the imaginary axis and corresponds to the creation or destruction of a periodic orbit. We will begin by studying the two-dimensional case, where the singular vector field $f_o(x)$ has a center at the origin and thus can be written as in (6.13),

$$\begin{aligned} \dot{x} &= -\omega y + p(x,y), \\ \dot{y} &= \omega x + q(x,y), \end{aligned} \tag{8.52}$$

where $p, q = o(x,y)$. The normal form for this case to cubic order was given in (8.43) after some prodigious linear algebra. Here we present an alternative and ultimately easier method for obtaining this form. The idea is to use complex coordinates:

$$z = x + iy, \ \bar{z} = x - iy. \tag{8.53}$$

The variables $(z, \bar{z})$ are to be thought of as independent. Indeed, if we were to allow the variables (x,y) to become complex, then the transformation $(x,y) \to (z,\bar{z})$ would be a diffeomorphism from $\mathbb{C}^2$ to $\mathbb{C}^2$. In this case, the transformation (8.53) applied to

(8.52) yields

$$\dot z = i\omega z + p\left(\frac{z+\bar z}{2}, \frac{z-\bar z}{2i}\right) + iq\left(\frac{z+\bar z}{2}, \frac{z-\bar z}{2i}\right),$$
$$\dot{\bar z} = -i\omega \bar z + p\left(\frac{z+\bar z}{2}, \frac{z-\bar z}{2i}\right) - iq\left(\frac{z+\bar z}{2}, \frac{z-\bar z}{2i}\right),$$

which could be thought of as a system of ODEs on $\mathbb{C}^2$. However, in our case we will assume that p and q are real-valued functions of their real arguments. In this case, the ODE for $\bar z$ is exactly the complex conjugate of that for z: $\overline{(\dot z)} = \dot{\bar z}$. If we keep this in mind, the ODE for $\bar z$ does not need to be considered. It is important to remember that $\bar z$ is an independent variable, however, and, moreover, since the equations depend on both z and $\bar z$, they are *not* analytic functions!

To compute the normal form, we expand the function $p + iq$ in a power series in z and $\bar z$ to obtain

$$\dot z = i\omega z + \sum_{\substack{m,n=0 \\ m+n>1}} a_{m,n} z^m \bar z^n. \tag{8.54}$$

The monomial basis consists of the functions $z^m \bar z^n$. The matrix for the linear system is diagonal $A = \mathrm{diag}(i\omega, -i\omega)$, and so the normal form results obtained for the diagonal case hold. In particular the monomials $z^m \bar z^n e_j$ are eigenfunctions of the homological operator L_A and the resonant terms in the z equation ($j = 1$) are those for which the eigenvalue (8.36) vanishes: $\mu_{(m,n),1} = (m,n)\cdot\lambda - \lambda_1 = 0$, or explicitly, $m(i\omega) + n(-i\omega) - i\omega = 0$. Thus the resonant terms are those for which $n = m-1$, i.e., those that have the form $z(z\bar z)^m = z\,|z|^{2m}$, where $|z|^2 \equiv z\bar z$ is the squared complex modulus. Since every other term can be eliminated, the general normal form can be written

$$\dot z = i\omega z + z\left(c\,|z|^2 + d\,|z|^4 + e\,|z|^6 + \cdots\right). \tag{8.55}$$

This is much easier than the process leading to (8.43)! We must, however, keep in mind that the coefficients c, d, etc., are complex. Indeed, upon comparison with the real normal form (8.43), $c = \alpha + i\beta$. The normal form (8.55) could be further simplified by scaling time to set $\omega = 1$ and scaling z so that c (if it is nonzero) has magnitude one, if we desire.

To unfold (8.55), we should add back in all the terms in (8.54) that have been eliminated and allow that coefficients in (8.55) to depend upon parameters $\mu \in \mathbb{R}^k$, thus obtaining a general power series of the form (8.54) again. In particular, the linear coefficient will become $\lambda(\mu)$, where $\lambda(0) = i\omega$. However, all the monomials $z^m \bar z^n$ with $n \neq m-1$ can still be eliminated by a further coordinate change, using the same argument that led to (8.55). Even the terms with $n = m-1$ could be eliminated by the coordinate transformation when $\lambda \neq i\omega$, since they are no longer resonant. However, this would lead to a change of variables that does not exist at $\mu = 0$. Therefore, to obtain a normal form that is valid in a neighborhood of $\mu = 0$, we must not try to eliminate these resonant terms. Consequently, a versal unfolding of (8.55) is

$$\dot z = \lambda(\mu) z + z\left(c(\mu)\,|z|^2 + d(\mu)\,|z|^4 + \cdots\right). \tag{8.56}$$

Equation (8.56) still has an equilibrium at the origin. We could have anticipated this using the persistence result, Corollary 8.4, since the normal form (8.55) has no zero eigenvalues.

Just as in the analysis in §6.3, it is easiest to understand the behavior of this system in polar coordinates. Defining $z = re^{i\theta}$, so that $r^2 = z\bar{z}$ and $\theta = \frac{1}{2i}\ln(z/\bar{z})$, gives

$$\begin{aligned}
\dot{r} &= \frac{1}{2r}\left(\dot{z}\bar{z} + z\dot{\bar{z}}\right) \\
&= \frac{1}{2r}\left(\lambda z\bar{z} + (c\,|z|^2 + d\,|z|^4)\,z\bar{z} + \bar{\lambda}\bar{z}z + \left(\bar{c}\,|z|^2 + \bar{d}\,|z|^4\right)\bar{z}z\right) \\
&= \mathrm{Re}(\lambda)r + r\left(\alpha r^2 + \gamma r^4 + O(r^6)\right), && (8.57) \\
\dot{\theta} &= \frac{1}{2ir^2}\left(\bar{z}\dot{z} - z\dot{\bar{z}}\right) \\
&= \frac{1}{2ir^2}\left(\lambda z\bar{z} + (c\,|z|^2 + d\,|z|^4)\,z\bar{z} - \bar{\lambda}\bar{z}z - \left(\bar{c}\,|z|^2 + \bar{d}\,|z|^4\right)\bar{z}z\right) \\
&= \mathrm{Im}(\lambda) + \beta r^2 + \delta r^4 + O(r^6),
\end{aligned}$$

where $c(\mu) = \alpha(\mu) + i\beta(\mu)$ and $d(\mu) = \gamma(\mu) + i\delta(\mu)$. Note that the r dynamics decouple from the θ dynamics. As promised, (8.57) shows that there is indeed an equilibrium at $r = 0$ that persists through the bifurcation. The remaining equilibria are given by

$$0 = F(r^2;\mu) = Re(\lambda) + \alpha(\mu)r^2 + \gamma(\mu)r^4 + O(r^6);$$

this equation can be thought of as a function of r^2 and μ. If we make the assumption that $\alpha(0) \neq 0$, F satisfies the hypotheses of the implicit function theorem: $F(0;0) = 0$, and $D_x F(0;0) = \alpha(0) \neq 0$. Since $r \geq 0$, there is a unique new equilibrium $r(\mu) \approx \sqrt{-\mathrm{Re}(\lambda)/\alpha}$, provided $\alpha\ \mathrm{Re}(\lambda) < 0$. Since $\mathrm{Im}(\lambda(0)) \neq 0$, there is a neighborhood of $\mu = 0$ such that $\theta(t)$ is monotone increasing for small r. Therefore, the equilibrium in r corresponds to a periodic orbit of the full system. Since the bifurcation set is determined by the single condition $\mathrm{Re}(\lambda) = 0$, the Andronov–Hopf bifurcation has codimension one.

The stability of the periodic orbit is easy to determine. At the equilibrium point, $r(\mu)$, the eigenvalue for the r dynamics is

$$D_r f(r;\mu) = \mathrm{Re}(\lambda) + 3\alpha r^2(\mu) + 5\gamma r^4(\mu) + \cdots = -2\mathrm{Re}(\lambda) + O(\mu^2).$$

Therefore when $\alpha < 0$, the new orbit exists when $\mathrm{Re}(\lambda) > 0$ and is asymptotically stable. When $\alpha > 0$, the new orbit exists when $\mathrm{Re}(\lambda) < 0$ and is unstable. These two cases, sketched in Figure 8.11, correspond to the supercritical and subcritical Andronov–Hopf bifurcations, respectively. A simple mnemonic for this bifurcation is that when both the periodic orbit and the equilibrium $r = 0$ exist, they have opposite stabilities.

Recall from §4.9 that a periodic orbit that is isolated is called a *limit cycle*. Thus the Andronov–Hopf bifurcation corresponds to the creation (or destruction) of a limit cycle.

This analysis also applies in higher dimensions—though the demonstration is considerably more complicated. In this case we assume that the singular system has a two-dimensional eigenspace with pure imaginary eigenvalues. The center manifold theorem of §5.6 can be used to obtain a reduction to the center space of the form (5.35), though some effort must be exerted to show that the reduction is smooth enough (Chow and Hale 1982; Chow, Li, and Wang 1994). The reduced system on W^c then takes the form (8.52).

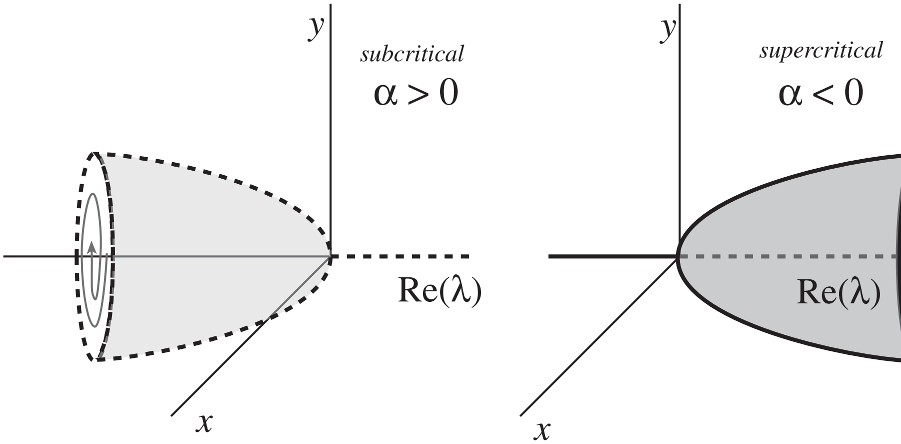

Figure 8.11. *Subcritical and supercritical Andronov–Hopf bifurcations.*

Theorem 8.21 (Andronov–Hopf bifurcation). *Let $f(x;\mu)$ be a C^3 vector field in $\mathbb{R}^n$ such that*

$$f(0;0)=0,$$

$$\operatorname{spec}(D_x f(0;0))=\{(i\omega,-i\omega,\lambda_3,\dots,\lambda_n):\operatorname{Re}(\lambda_k)\neq 0, k\geq 3\}. \qquad (\textit{singularity})$$

Theorem 8.3 *then implies there is a smooth curve of equilibria $x(\mu)$ with eigenvalues $\lambda_\pm(\mu)$ where $\lambda_\pm(0)=\pm i\omega$. The normal form on the center manifold of f_o has an unfolding of the form* (8.56)*. Assume that the normal form coefficient α satisfies*

$$\alpha(0)=\operatorname{Re}(c(0))\neq 0 \qquad (\textit{nondegeneracy})$$

and that a parameter μ causes the eigenvalue to cross the imaginary axis

$$\frac{d}{d\mu}\operatorname{Re}(\lambda(0))\neq 0 \qquad (\textit{transversality}). \tag{8.58}$$

Then there is an Andronov–Hopf bifurcation corresponding to the birth of a limit cycle that has a quadratic tangency with the eigenspace of $\pm i\omega$ at $\mu=0$. The limit cycle exists when $\alpha\operatorname{Re}(\lambda)<0$ and is stable if $\operatorname{Re}(\lambda)>0$ and unstable if $\operatorname{Re}(\lambda)<0$.

The main difficulty in the application of this theorem is verifying that $\alpha(0)\neq 0$. In general this is a tedious calculation, since we have to compute the normal form (8.55) through third order. For more than two dimensions this is especially hard because the center manifold must be computed to third order *before* the dynamics on the center subspace can be obtained. In two dimensions the calculation of α is manageable. Indeed, we can do the calculations to obtain a general formula for α in terms of the coefficients of p and q in (8.52) up to third order (Guckenheimer and Holmes 1983) to obtain

$$\begin{aligned}\alpha &= \frac{1}{16}\left(p_{xxx}+p_{xyy}+q_{xxy}+q_{yyy}\right)\\ &\quad -\frac{1}{16\omega}\left(q_{xy}(q_{xx}+q_{yy})-p_{xy}\left(p_{xx}+p_{yy}\right)+p_{xx}q_{xx}-p_{yy}q_{yy}\right).\end{aligned} \tag{8.59}$$

Here each subscript indicates a derivative, and all are evaluated at the origin.

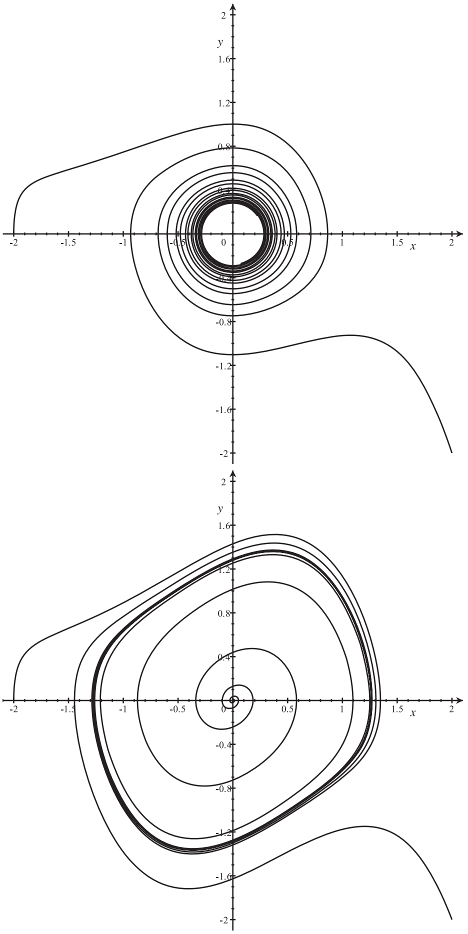

Figure 8.12. *Phase portrait of the van der Pol system* (8.60) *with* $\mu = 0$ *(left) and* $\mu = 0.2$ *(right). The origin is a topological sink in the left panel and an unstable focus in the right panel.*

Example 8.22 (van der Pol oscillator). The van der Pol oscillator

$$\ddot{x} - \left(2\mu - x^2\right)\dot{x} + x = 0$$

was derived in §1.4. It is a special case of Liénard's system, and according to Theorem 6.42, it is guaranteed to have a unique limit cycle when $\mu > 0$.

Turning this into a first-order system by defining $y = \dot{x}$ gives

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -x + 2\mu y - x^2 y. \end{aligned} \tag{8.60}$$

The origin is an equilibrium with eigenvalues $\lambda = \mu \pm i\sqrt{1-\mu^2}$, so it is a stable focus for $-1 < \mu < 0$, a linear center at $\mu = 0$, and an unstable focus for $0 < \mu < 1$. At $\mu = 0$, the matrix already has the normal form (8.11), though with $\omega = -1$. Thus α can be computed using (8.59); noting that $q_{xxy} = -2$ is the only nonzero coefficient yields $\alpha = -1/8$. Thus there is a stable periodic orbit created for $\mu > 0$ in a supercritical Andronov–Hopf bifurcation. Two phase portraits are shown in Figure 8.12. ■

8.9 ▪ The Cusp Bifurcation

There are five distinct codimension-two bifurcations for vector fields. These correspond to the various ways a vector field can be made singular or degenerate by varying two parameters. As we have seen, one singularity that arises upon varying a single parameter corresponds to a single eigenvalue with a zero real part. Since the eigenvalues of a matrix depend continuously on its elements, if we vary two parameters it should be possible to make two real eigenvalues zero. This corresponds to the *Takens–Bogdanov* bifurcation—we will study it in the next section. When there are complex eigenvalues they come in conjugate pairs—thus a single parameter can be used to change the real part of the pair, and by varying two parameters one could move two pairs of complex eigenvalues to the imaginary axis. This is called the *Hopf–Hopf* bifurcation. Since there must be at least two pairs of eigenvalues for this to occur, it requires a phase space with four or more dimensions. The final codimension-two arrangement of eigenvalues is a single real eigenvalue at zero and a single pair on the imaginary axis; this bifurcation is called the *fold-Hopf* or *Gavrilov–Guckenheimer* bifurcation; for this to occur the system must have at least a three-dimensional phase space.

The remaining two codimension-two bifurcations pertain to cases when the nondegeneracy assumptions in the codimension-one cases are not satisfied. For example, if the normal form coefficient, α, in the Hopf bifurcation case passes through zero, then a subcritical Hopf bifurcation is converted into a supercritical one. This situation is called a *degenerate Hopf* or *Bautin* bifurcation. Finally, for the saddle-node bifurcation, we assumed that the quadratic term in the center component of the vector field was nonzero. By varying a second parameter, it may be possible to make this coefficient vanish. This degenerate saddle-node gives rise to the *cusp* bifurcation. It is this final case that we study in this section.

Since codimension-two bifurcation theorems are considerably more complex than the codimension-one cases, we will study only the simplest cases in which these occur. For the cusp bifurcation, this corresponds to the degenerate saddle node in one dimension. We already looked at a special case of this in §8.7, the pitchfork bifurcation. As we remarked there, the pitchfork is not a versal unfolding of the singularity. Our goal here is to find a miniversal unfolding of this case.

Example 8.23. Consider the pitchfork normal form (8.51), but add a second parameter to represent the constant term

$$\dot{x} = \mu_1 + \mu_2 x - x^3. \tag{8.61}$$

Although an explicit form for the equilibrium solutions $x^*(\mu_1, \mu_2)$ to this system can be obtained since the vector field is a cubic polynomial, this form is not especially useful. It is more illuminating to graphically find the equilibria by plotting the two functions $y = \mu_2 x - x^3$ and $y = -\mu_1$ and looking for intersections. Since the cubic crosses each horizontal line either once or thrice, there will be either one or three equilibria.

The bifurcation set can be found by looking for places where there are degenerate equilibria:

$$\begin{aligned} f(x;\mu) &= \mu_1 + \mu_2 x - x^3 = 0, \\ D_x f(x;\mu) &= \mu_2 - 3x^2 = 0. \end{aligned}$$

The *resultant*, $R(\mu)$, is the equation obtained by eliminating x from this pair; its roots correspond to double roots of f. Since the second equation gives $x^2 = \mu_2/3$, we square

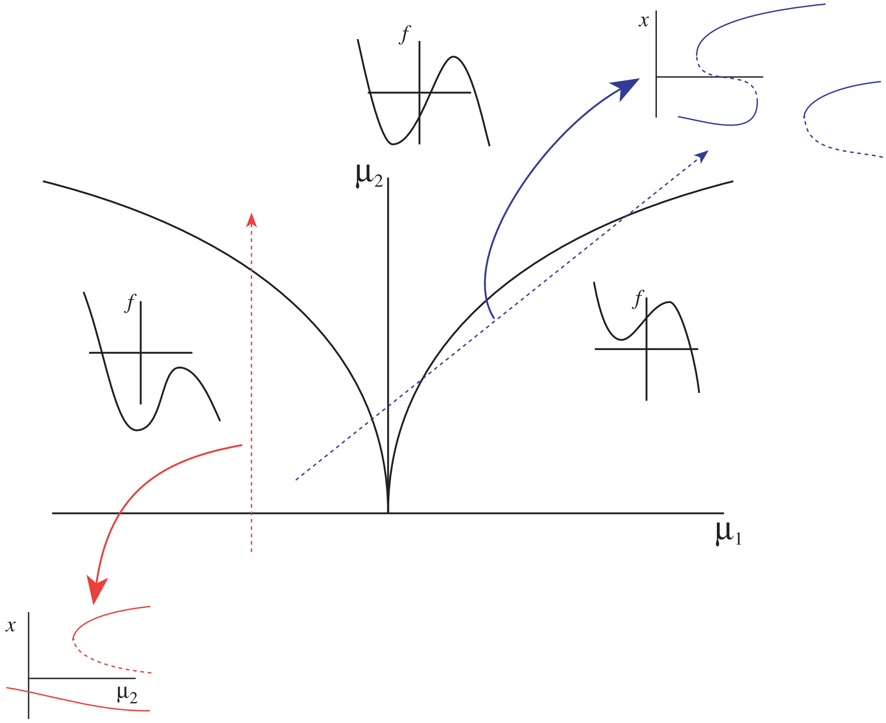

Figure 8.13. *Bifurcation parameter plane for* (8.61), *showing the bifurcation set* (8.62). *Also shown are two representative one-parameter sweeps through the bifurcation (vertical and diagonal dashed curves) and the resulting one-parameter bifurcation diagrams.*

the first equation to obtain an equation that contains only x^2: $\mu_1^2 = (\mu_2 x - x^3)^2 = x^2(\mu_2^2 - 2\mu_2 x^2 + x^4)$. Substituting for x^2 yields the resultant

$$R = 27\mu_1^2 - 4\mu_2^3 = 0. \tag{8.62}$$

This curve is called *Neile's semicubical parabola*. It has the form of a cusp in the μ plane—see Figure 8.13. For parameter values on the cusp there is a double root at $x = -\text{sgn}(\mu_1)\sqrt{\mu_2/3}$.

Inside the cusp region there are three equilibria, on the boundary there are two, and outside and at the cusp point there is one. Crossing the cusp at a fixed nonzero value of μ_1 results in a saddle-node bifurcation at a nonzero value for x. Setting $\mu_1 = 0$ and moving along the μ_2-axis gives the pitchfork case we considered before. These examples make it clear that the behavior in the neighborhood of the cusp point depends on the way that the parameters traverse Neile's parabola. This bifurcation requires varying two parameters, so it is codimension two. ■

We an also think of the cusp as an elementary form of *catastrophe*: there are paths through parameter space where the positions of the equilibria vary smoothly and paths that result in a saddle node, far from an existing equilibrium (catastrophe) (Arnold et al. 1999; Golubitsky et al. 1985).

More generally, the cusp corresponds to an unfolding of a singular vector field with $f_o(0) = Df_o(0) = D^2 f_o(0) = 0$, but $D^3 f_o(0) = d \neq 0$. A versal unfolding of f_o will have all of the low-order terms

$$f(x;\mu) = a(\mu) + b(\mu)x + c(\mu)x^2 + d(\mu)x^3 + g(x;\mu) \tag{8.63}$$

with $a(0) = b(0) = c(0) = 0$ and $d(0) \neq 0$.

It is always possible to eliminate the quadratic term from this function by translation of x. To see this, consider the equation

$$C(x;\mu) = \tfrac{1}{2}D_x^2 f(x;\mu) = c(\mu) + 3d(\mu)x + \tfrac{1}{2}D_x^2 g(x;\mu) = 0. \tag{8.64}$$

The implicit function theorem implies that (8.64) always has a solution $x_o(\mu)$ near $\mu = 0$, since $D_x C(0;0) = 3d(0) \neq 0$. Now define a new variable $y = x - x_o$ and rewrite (8.63) for y:

$$\dot{y} = f(x_o + y;\mu) = A(\mu) + B(\mu)y + D(\mu)y^3 + h(y;\mu). \tag{8.65}$$

Note that the quadratic coefficient would become $C(\mu) = ½D_{xx}f(x_o(\mu);\mu)$, but this is identically zero. The other coefficients are defined by

$$A(\mu) = f(x_o;\mu),\ B(\mu) = D_x f(x_o;\mu), \text{ and } D(\mu) = \tfrac{1}{6}D_{xxx}f(x_o;\mu). \tag{8.66}$$

By assumption D is nonzero when μ is small.

Next we rescale the variable y to eliminate the coefficient D by setting $z = \sqrt{|D|}y$. Finally we choose a new set of parameters $m_1 = A(\mu)\sqrt{|D|}$ and $m_2 = B(\mu)$ to obtain a vector field in the form

$$\dot{z} = F(z;m) = m_1 + m_2 z + sz^3 + h(z;m), \tag{8.67}$$

where $s = \operatorname{sgn}(D(0))$. This gives essentially the form (8.61) that we studied in the example, with the addition of a sign. We conclude with the formal statement.

Theorem 8.24 (cusp bifurcation). *Let $f \in C^3(\mathbb{R} \times \mathbb{R}^k, \mathbb{R})$ and*

$$f(0;0) = D_x f(0;0) = 0, \qquad \textit{(nonhyperbolic)}$$

$$D_x^2 f(0;0) = 0, \qquad \textit{(singularity)}$$

$$D_x^3 f(0;0) \neq 0. \qquad \textit{(nondegeneracy)}$$

Let N be the neighborhood of $\mu = 0$ for which $x_o(\mu)$ is the unique solution of $D_x^2 f(x_o(\mu);\mu) = 0$ such that $x_o(0) = 0$. Then f has a cusp bifurcation in N with bifurcation set

$$27A^2(\mu)D(\mu) = -4B^3(\mu), \tag{8.68}$$

where A, B, and D are defined by (8.66).

Proof. We leave the proof to the reader. See Exercise 17. □

To show that this bifurcation actually occurs—i.e., that the parameter plane in Figure 8.13 actually applies, it is necessary to have a condition of *transversality*. Since the bifurcation is unfolded by the parameters (A, B), the requirement is that the mapping $\mu \to (A, B)$ is *onto* a neighborhood of the origin; i.e., for each (small enough) (A, B), there is a μ that realizes this value. This occurs when we can solve the implicit system $F(\mu;m)=(A(\mu)-m_1,B(\mu)-m_2)=(0,0)$ for μ. There need to be at least two parameters, but there could be many more; of all the parameters, pick two—(μ_1,μ_2), say. Now since $F(0;0)=0$, the implicit function implies that if the Jacobian $D_\mu F$ is nonsingular at the origin, i.e., if the matrix

$$\begin{pmatrix} \frac{\partial A}{\partial \mu_1} & \frac{\partial A}{\partial \mu_2} \\ \frac{\partial B}{\partial \mu_1} & \frac{\partial B}{\partial \mu_2} \end{pmatrix}$$

is nonsingular, then there exists unique (μ_1,μ_2) for each (m_1,m_2). If this criterion is not satisfied for the two parameters we chose, then we can look for another pair. Thus the ultimate criterion is that the matrix $D_\mu(A,B)\big|_{\mu=0}$ has rank two. This becomes the assumption of transversality; when it is satisfied, the unfolding $f(x;\mu)$ is versal.

Corollary 8.25. *Given $f(x;\mu)$ as in Theorem* 8.24*, if*

$$\operatorname{Rank} D_\mu(f(0;\mu),D_x f(0;\mu))\big|_{\mu=0}=2, \qquad (\textit{transversality})$$

then there is cusp bifurcation at $\mu=0$.

Proof. Since $A(\mu)=f(x_o(\mu);\mu)$, then

$$D_\mu A(\mu)|_{\mu=0}=D_\mu f(0;0)+D_x f(0;0)\frac{dx_o}{d\mu},$$

but by assumption $D_x f(0;0)=0$. Similarly with $B(\mu)=D_x f(x(\mu);\mu)$, then

$$D_\mu B(\mu)|_{\mu=0}=D_\mu(D_x f(0;0))+D_x^2 f(0;0)\frac{dx_o}{d\mu},$$

but by assumption $D_x^2 f(0;0)=0$. Consequently, the Jacobian becomes $D_\mu(A,B)\big|_{\mu=0}=D_\mu(f(0;\mu),D_x f(0;\mu))\big|_{\mu=0}$. □

8.10 ▪ Takens–Bogdanov Bifurcation

As a second codimension-two bifurcation, we will study the case of a double zero eigenvalue. The simplest system for which this can occur is an ODE in $\mathbb{R}^2$. As remarked in §8.3, the linearization $Df(0)$ can be semisimple or not. In the former case, it is identically zero and this corresponds to the nonhyperbolic node studied in §6.2. We consider the latter case in this section. In §8.3, we argued that the normal form for this matrix is the Jordan form (8.13)

$$J=\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Following the normal form analysis in §8.5, we will obtain the nonlinear normal form by considering the homological operator for J,

$$L_J(h)=Jh-DhJx=\begin{pmatrix} y\partial_x h_x - h_y \\ y\partial_x h_y \end{pmatrix}.$$

For functions in $\mathbb{H}_2^2$, using the basis (8.23), the matrix representation for L_J is

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

The column space of L defines its range, $\mathrm{rng}(L_J) = \mathrm{span}\{p_2, p_3, p_1 - 2p_5, p_6\}$, which is four-dimensional. The resonant space $\mathbb{G}_2^2$, is any complementary subspace to $\mathrm{rng}(L_J)$. Since $p_4 \notin \mathrm{rng}(L_J)$, it certainly must be an element $\mathbb{G}_2^2$. The second vector can be any linear combination of p_1 and p_5 that is independent of $p_1 - 2p_5$. To be economical, it is nice to choose a resonant space with a minimal number of monomials. Using this criterion there are two choices for $\mathbb{G}_2^2$: $\mathrm{span}\{p_1, p_4\}$ or $\mathrm{span}\{p_4, p_5\}$. In either case, all the quadratic terms in the ODE can be eliminated except for the basis elements of $\mathbb{G}_2^2$, resulting in the quadratic normal forms

$$\begin{aligned} \dot{x} &= y + ax^2 \\ \dot{y} &= dx^2 \end{aligned} \qquad \text{or} \qquad \begin{aligned} \dot{x} &= y \\ \dot{y} &= dx^2 + exy. \end{aligned}$$

The second form has some advantages, so we will use it.[50] To unfold this bifurcation there must be at least two parameters that represent the two eigenvalues. It can be argued that one miniversal unfolding is given by the system (Guckenheimer et al. 1983)[51]

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= \mu_1 + \mu_2 y + dx^2 + exy. \end{aligned}$$

We assume that the parameters d and e are nonzero; it is not hard to see that by a suitable scaling of the x and y variables and time (possibly reversing its direction), these parameters can be scaled to $d = e = 1$, giving the normal form

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= \mu_1 + \mu_2 y + x^2 + xy. \end{aligned} \tag{8.69}$$

Equation (8.69) has exactly two equilibria at $(x_\pm, 0) = (\pm\sqrt{-\mu_1}, 0)$ when $\mu_1 < 0$ and one at $(0,0)$ when $\mu_1 = 0$. The Jacobian of (8.69) at an equilibrium is

$$Df = \begin{pmatrix} 0 & 1 \\ 2x_\pm & \mu_2 + x_\pm \end{pmatrix}. \tag{8.70}$$

Thus the characteristic polynomial is $\lambda^2 - (\mu_2 + x_\pm)\lambda - 2x_\pm$, so the point $(x_+, 0)$ is a saddle when it exists, and the point $(x_-, 0)$ is a source if $\mu_2 > \sqrt{-\mu_1}$ and is a sink otherwise. On the boundary curve $\mu_2 = \sqrt{-\mu_1}$ this latter equilibrium is a center. Thus we expect that an Andronov–Hopf bifurcation should occur there; indeed, after transforming to put the center at the origin, the computation of α from (8.59) gives $\alpha > 0$ so that the bifurcation is seen to be subcritical, creating an unstable periodic orbit around the sink when $\mu_2 < \sqrt{-\mu_1}$. A few distinguishing phase portraits are

[50] For example, note that the second normal form is equivalent to the nonlinear "oscillator" $\ddot{x} - dx^2 - ex\dot{x} = 0$.

[51] The original Bogdanov unfolding, replacing the term $\mu_2 y$ by $\mu_2 x$, is used by (Kuznetsov 1995).

shown in Figure 8.14. In particular, the left column shows three portraits for $\mu_1 \le 0$. The central portrait in this column, at $(\mu_1, \mu_2) = (-0.25, 0.48)$, exhibits the unstable limit cycle surrounding the stable focus at $(x, y) = (-0.5, 0)$.

We summarize our informal results with a theorem (Bogdanov 1975; Takens 2001).

Theorem 8.26 (Takens–Bogdanov). *Suppose that $f \in C^2(\mathbb{R}^2 \times \mathbb{R}^k, \mathbb{R}^2)$ and satisfies the conditions*

$$f(0,0;0) = 0, \qquad Df(0,0;0) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad (\textit{singularity})$$

$$(D_{xx}f_x + D_{xy}f_y)|_{(0,0;0)} \neq 0, \quad D_{xx}f_y|_{(0,0;0)} \neq 0, \quad (\textit{nondegeneracy})$$

$$\mathrm{Rank}\left(D_{(x,\mu)}(f, \mathrm{tr}(Df), \det(Df))\right)|_{(0,0;0)} = 4. \qquad (\textit{transversality})$$

Then there exists a neighborhood of the origin in which the dynamics of f is induced (up to a reversal of time) by the normal form (8.69).

As can be inferred from Figure 8.14, the leftward branches of the stable and unstable manifolds of the saddle of (8.69) appear to cross between the two portraits at $(-0.25, 0.48)$ and $(-0.5, 0.5)$. Indeed, as we will see in the next section, the Takens–Bogdanov normal form has a curve, emanating from the origin, of such “homoclinic bifurcations.”

8.11 ▪ Homoclinic Bifurcations

So far we have considered only bifurcations of equilibria. These are local bifurcations in the sense that they can be studied in a neighborhood by using local power series expansions. Other local bifurcations occur when a periodic orbit loses or gains stability. If one has some analytical information about the orbit (enough to compute the “Poincaré map” for the orbit; recall §4.12), then power series techniques suffice to understand these as well.

There are global bifurcations, however, for which no local information is sufficient. The simplest of these corresponds to the creation or destruction of a “homoclinic” or “heteroclinic” orbit. More complicated global bifurcations are responsible, in some cases, for the onset of “chaos.” We defined homoclinic and heteroclinic orbits in §5.2 and gave several examples of Hamiltonian systems that had such orbits.

Fragility of Heteroclinic Orbits

It is much more difficult to explicitly construct heteroclinic orbits in general systems. Generically, the unstable manifold of one equilibrium does not coincide with the stable manifold of another, unless there is some special symmetry (like the conservation of energy (4.28) for planar systems).

Example 8.27. An illustrative example is

$$\begin{aligned} \dot{x} &= \mu + x^2 - xy, \\ \dot{y} &= y^2 - x^2 - 1. \end{aligned} \tag{8.71}$$

When $\mu = 0$ this system has two equilibria, $(0, \pm 1)$. Both are saddle points since the Jacobian matrix is

$$Df = \begin{pmatrix} 2x - y & -x \\ -2x & 2y \end{pmatrix} \underset{(x,y)=(0,\pm 1)}{\rightarrow} \begin{pmatrix} \mp 1 & 0 \\ 0 & \pm 2 \end{pmatrix}.$$

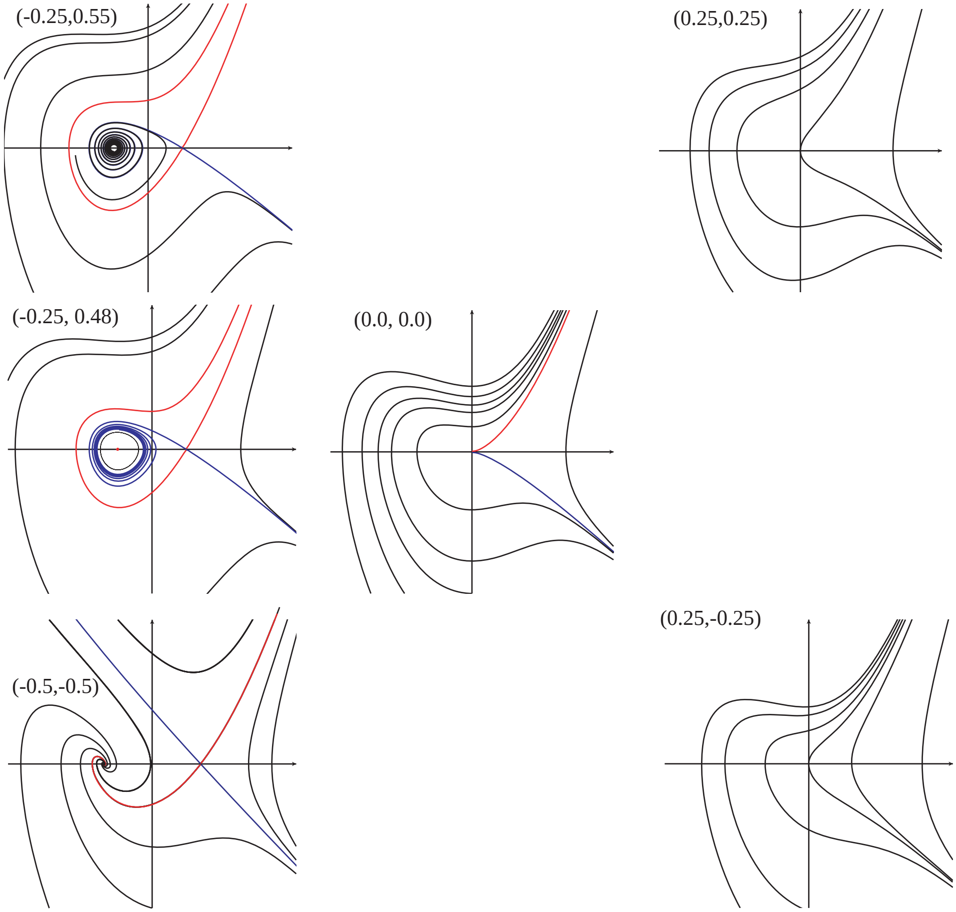

Figure 8.14. *Phase portraits for the Takens–Bogdanov unfolding* (8.69) *at six different sets of values of* (μ_1, μ_2).

The y-axis of (8.71) with $\mu = 0$ is invariant, and on this axis the y dynamics is simply $\dot{y} = y^2 - 1$, which has a stable equilibrium at -1 and an unstable equilibrium at $+1$. Thus the segment $\{(0, y) : -1 < y < 1\}$ is a heteroclinic orbit when $\mu = 0$.

The implicit function theorem guarantees that the saddle points persist when μ is small. They are easy to find by series expansion in μ: set $x = a\mu + b\mu^2 + \cdots, y = \pm 1 + c\mu + d\mu^2 + \cdots$, to obtain

$$
\begin{array}{rrl}
\dot{x} = & 0 = & \mu + (a\mu + b\mu^2)\big(a\mu + b\mu^2 \mp 1 - c\mu - d\mu^2\big) \\
\dot{y} = & 0 = & (\pm 1 + c\mu + d\mu^2)^2 - (a\mu + b\mu^2)^2 - 1 \\
& 0 = & (1 \mp a)\mu + (\mp b + a^2 - ca)\mu^2 + \cdots \\
\Rightarrow & & \\
& 0 = & \pm 2c\mu + (-a^2 + c^2 \pm 2d)\mu^2 + \cdots .
\end{array}
$$

This shows that $a = b = \pm 1$, $c = 0$, and $d = \pm 1/2$, so that $x = \pm(\mu + \mu^2 + \cdots)$ and $y = \pm(1 + 1/2\mu^2 + \cdots)$ are the new equilibria. Of course, the equations are simple

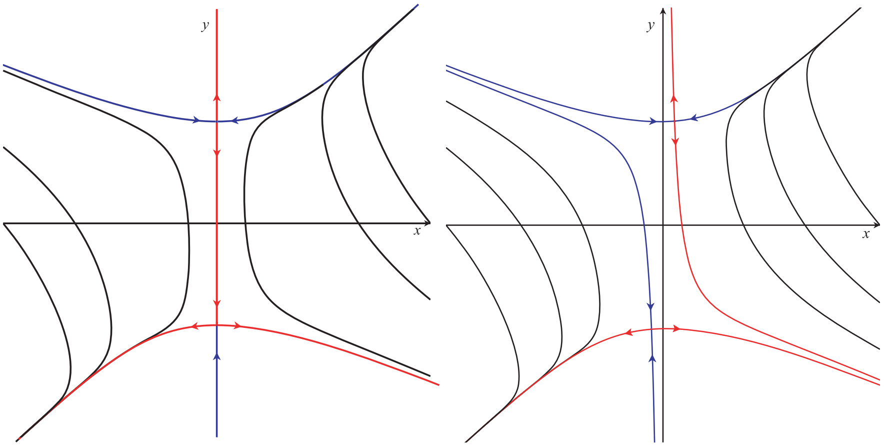

Figure 8.15. *Heteroclinic connection of* (8.71) *when* $\mu = 0$ *(left) is destroyed when* $\mu = 0.1$ *(right).*

enough here to find equilibria explicitly,

$$x = \pm\frac{\mu}{\sqrt{1-2\mu}}, \quad y = \pm\frac{1-\mu}{\sqrt{1-2\mu}},$$

which shows us that the equilibria persist until $\mu = 1/2$.

To study the change in the unstable manifolds with μ, note that when $x \approx 0$, we have $\dot{x} \approx \mu$, while if $|y| < 1$, then $\dot{y} < 0$. Thus if $\mu > 0$, the upper saddle is to the right of the y-axis, and its unstable manifold begins moving downward but necessarily moves to the right. The lower saddle, by contrast, is to the left of the y-axis, and its downward moving stable manifold comes from larger negative x. Thus the two manifolds no longer join, and there is no longer a saddle connection; see Figure 8.15.

It is easy to see that the heteroclinic connection is destroyed for $\mu < 0$ too, since the equilibria move to the opposite sides of the y-axis, but in this case $\dot{x} < 0$. ∎

The implication of this example is that if an ODE in the plane has a heteroclinic connection, then changing a single parameter can destroy it. In other words, homoclinic connections are singularities for planar ODEs and their unfolding gives rise to a codimension-one bifurcation.

Generic Homoclinic Bifurcations in $\mathbb{R}^2$

We will now develop the general theory for bifurcations of a homoclinic orbit in the plane. Our results will show that when the system is non-Hamiltonian, the destruction of a homoclinic orbit is a codimension-one bifurcation and is associated with the creation of a periodic orbit.

Suppose that $f(x;\mu)$ is an unfolding of a planar vector field $f_o(x)$ that has a saddle equilibrium $p_o = (0,0)$ with a homoclinic connection γ_o. The implicit function theorem implies there is an interval I of 0 such that for $\mu \in I$ the saddle equilibrium $p(\mu)$ persists and remains a saddle. Thus the stable manifold theorem in §5.4 implies that $p(\mu)$ has stable and unstable manifolds $W^s(p(\mu))$ and $W^u(p(\mu))$. To study the way in which these manifolds move with μ, we choose any point $q \in \gamma_o$ and let S be a local, one-dimensional section at q (recall §4.12). That is, S is a line segment through q that is perpendicular to the vector field $f_o(q)$; see Figure 8.16.

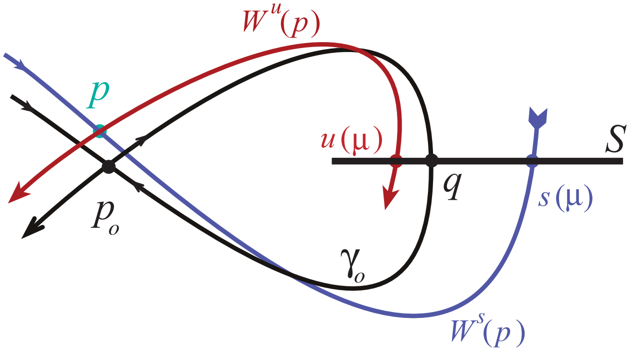

Figure 8.16. *Construction of the map for a homoclinic bifurcation.*

The points of intersection of $W^u(p)$ and $W^s(p)$with S are denoted $u(\mu)$ and $s(\mu)$, respectively. These functions exist for some interval I in μ because the stable and unstable manifolds move continuously with μ. Moreover, $s(0) = u(0) = q$. Note that there is no local way to compute s and u; to find them we must construct the manifolds of $p(\mu)$ over a macroscopic distance.

However, if the reader is willing to agree that s and u are, in principle, computable, then we can use them to state a bifurcation theorem.

Theorem 8.28 (homoclinic bifurcations). *Let $f \in C^2(\mathbb{R}^2 \times \mathbb{R}, \mathbb{R}^2)$ and suppose $f_o(x) = f(x;0)$ has a saddle equilibrium p_o such that*

$$p_o \text{ has a homoclinic orbit } \gamma_o,$$

$$\tau \equiv \mathrm{tr}(Df_o(p_o)) = \nabla \cdot f_o(p_o) \neq 0. \qquad (\textit{nondegeneracy})$$

Let $p(\mu)$ be the saddle equilibrium of f that continues from p_o and denote its manifolds by $W^u(p)$ and $W^s(p)$. Define a section S to f_o at a point $q \in \gamma_o$ and let $s(\mu) = S \cap W^s(p)$ and $u(\mu) = S \cap W^u(p)$ be the continuous functions of μ such that $s(0) = u(0) = q$. Suppose that

$$\Delta \equiv \frac{d}{d\mu}(s(\mu) - u(\mu))_{\mu=0} \neq 0. \qquad (\textit{transversality})$$

Then if $\tau < 0$ (> 0), there is a family $\gamma(\mu)$ of stable (unstable) periodic orbits that bifurcate from γ_o. The periods of these orbits are unbounded as $\mu \to 0$. Moreover, there is an ε (which may be negative) such that there is exactly one periodic orbit in a neighborhood of γ_o when $\mu \in (0, \varepsilon)$.

Proof (Sketch). We consider the stable case $\tau < 0$. Since S is transverse to $f(x;0)$, by continuity, it will be transverse to $f(x;\mu)$ for μ small enough and x close to q. Let $P_\mu : S \to S$ be the first return map for $f(x;\mu)$ to the section S (recall §4.12). As Figure 8.16 indicates, the return map is defined only for points on S that are closer to the equilibrium than $s(\mu)$; other points typically escape and do not return. As $x \to s(\mu)_-$ we have $P_\mu(x) \to u(\mu)_-$. Note that P_o is defined for all points in the interior of the homoclinic loop γ_o.

The transversality assumption $\Delta \neq 0$ implies s and u change at different rates with μ; suppose, for example, $\Delta > 0$. Then if $\mu < 0$, $u = P_\mu(s) > s$, and if $\mu > 0$, $u < s$. This means that the value of $P_\mu(s(\mu))$ is above the diagonal for the first case and below

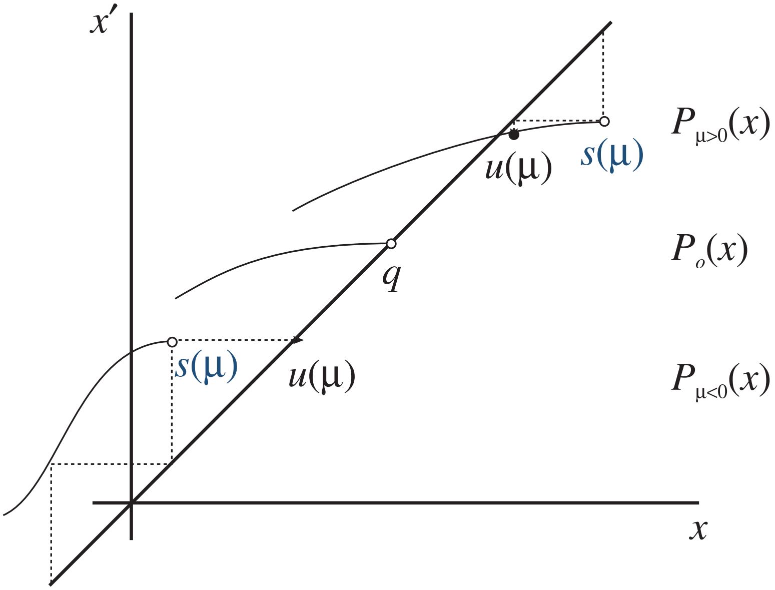

Figure 8.17. *Poincaré return map near the homoclinic loop assuming that* $\tau < 0$ *and* $\Delta > 0$.

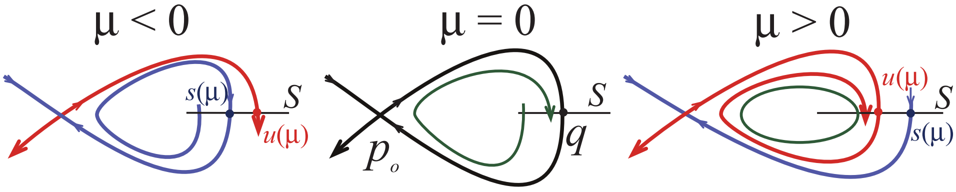

Figure 8.18. *Sketch of the phase space near a homoclinic bifurcation when* $\tau < 0$ *and* $\Delta > 0$.

the diagonal for the second case; see Figure 8.17. Therefore, since P_μ is defined for some interval to the left of s, the graph of P must intersect the diagonal at some point $x^* < u(\mu)$ when $\mu > 0$. At this point, where $P_\mu(x^*) = x^*$, the Poincaré map has a fixed point—this corresponds to a periodic orbit, $\gamma(\mu)$ of the flow, as sketched in Figure 8.18.

To study the stability of the periodic orbit, we must evaluate the slope of P_μ at x^*. The fact that $\nabla \cdot f_o(p_o) < 0$ implies that the stable eigenvalue of $Df_o(p_o)$ is stronger than the unstable eigenvalue, and this means that orbits inside the homoclinic loop at $\mu = 0$ are strongly attracted to the loop; indeed, we claim that $DP_o(q) = 0$. To see this, we look at the behavior of orbits near p_o. Choose local coordinates in the neighborhood of p_o such that the stable direction is the x-axis and the unstable direction is the y-axis. Denote the eigenvalues of $Df_o(p_o)$ by $-\alpha < 0 < \beta$; by assumption $\mathrm{tr}(Df_o(p_o)) = -\alpha + \beta < 0$, so $\alpha > \beta$.

To the extent that the linear approximation is valid, we have $x(t) = x_o e^{-\alpha t}$ and $y(t) = y_o e^{\beta t}$. Thus the trajectory that starts at $(\varepsilon, \Delta y)$ and ends at $(\Delta x, \varepsilon)$ takes a time

$t=\beta^{-1}\ln(\varepsilon/\Delta y)$, so that $\Delta x=\varepsilon\left(\frac{\varepsilon}{\Delta y}\right)^{-\alpha/\beta}$. This implies that

$$\frac{\Delta x}{\Delta y}=\left(\frac{\Delta y}{\varepsilon}\right)^{\alpha/\beta-1}\to 0 \ \text{ as } \Delta y\to 0,$$

since by assumption $\alpha/\beta>1$. Thus trajectories that start close to the x-axis approach much closer to the y-axis. This argument can be corrected by using Grönwall's inequality (Lemma 3.28) to take into account the nonlinear terms, at the expense of decreasing the exponent slightly. Moreover, the argument does not change much if we extend the trajectory backward and forward to crossing points on S.[52] The calculation implies that the point $x'=P_o(x)$ on S is much closer to q than x was, so that $DP_o(x)>0$, and the graph of P_o is monotone increasing. Moreover, since $(x'-q)/(x-q)\to 0$, as $x\to q$, $DP_o(q)=0$. A fixed point with zero multiplier is called "superstable." More generally, we can argue that $DP_\mu(s(\mu))=0$ for the same reasons.

The Floquet multiplier of the periodic orbit is the slope of the Poincaré map at x^* and must continuously move away from zero. Whenever the multiplier is less than one, the periodic orbit is stable. Similar techniques show that when $\Delta<0$, the periodic orbit exists for $\mu<0$ and is unstable. □

8.12 ▪ Melnikov's Method

As we have seen in §5.2, planar Hamiltonian systems often have homoclinic or heteroclinic solutions when there are saddle equilibria. We have also shown in the previous section that it is relatively easy to determine when a perturbation of the homoclinic orbit destroys it and gives birth to a nearby periodic orbit. Unfortunately, this theorem requires the computation of $s(\mu)$ and $u(\mu)$, which are difficult to obtain. In this section we develop a method, usually attributed to Melnikov—even though it was originally due to Poincaré—for finding their lowest-order behavior.

We start with a general question: When can a system in the plane have an orbit that is a closed loop? Suppose that a system with a C^1 vector field $f=(p,q)^T$ has an invariant loop $\gamma:\mathbb{S}^1\to\mathbb{R}^2$; it could be periodic, homoclinic, or a family of heteroclinic trajectories that form a loop (a separatrix cycle). Using the fact that $\dot{x}-p(x,y)\equiv 0$ and $\dot{y}-q(x,y)\equiv 0$ along any trajectory, consider the integral

$$\begin{aligned}0&=\oint_\gamma(-(\dot{y}-q)dx+(\dot{x}-p)dy)\\&=\oint_\gamma(\dot{x}dy-\dot{y}dx)+\oint_\gamma(qdx-pdy).\end{aligned}$$

The first integrand can be written $\dot{x}dy-\dot{y}dx=(\dot{x}\dot{y}-\dot{y}\dot{x})dt\equiv 0$, so that only the second term is possibly nonzero, giving

$$0=\oint_\gamma(qdx-pdy)=\int_{\mathrm{Int}(\gamma)}\nabla\cdot f. \tag{8.72}$$

Here we have used Green's theorem to convert this to the integral over the interior of γ. Thus we have obtained the next lemma.

[52] This is where our "proof" is only a sketch.

Lemma 8.29 (Bendixson's Criterion). *If $\dot{x} = f(x)$ has an invariant loop γ, then* $\int_{\mathrm{Int}(\gamma)} \nabla \cdot f = 0$.

Consider, for example, the perturbed Hamiltonian system

$$\begin{aligned} \dot{x} &= f_1(x,y) + \varepsilon g_1(x,y), \\ \dot{y} &= f_2(x,y) + \varepsilon g_2(x,y), \end{aligned} \qquad f = \left(\frac{\partial H}{\partial y}, -\frac{\partial H}{\partial x} \right)^T . \tag{8.73}$$

Suppose that there is an $\varepsilon_0 > 0$ such that (8.73) has an invariant closed loop γ_ε whenever $\varepsilon \in [-\varepsilon_0, \varepsilon_0]$. Since $\nabla \cdot f \equiv 0$ for a Hamiltonian system (see §9.2), the integral (8.72) becomes

$$0 = \varepsilon \oint_{\gamma_\varepsilon} (g_2 dx - g_1 dy) = \varepsilon \int_{\mathrm{Int}(\gamma_\varepsilon)} \nabla \cdot g .$$

Since the integral above vanishes identically as a function of ε on the interval $[-\varepsilon_0, \varepsilon_0]$, it must have a zero ε derivative. This implies in particular that

$$0 = M \equiv \left. \frac{d}{d\varepsilon} \right|_{\varepsilon=0} \varepsilon \oint_{\gamma_\varepsilon} (g_2 dx - g_1 dy) = \oint_{\gamma_0} (g_2 dx - g_1 dy) = \int_{\mathrm{Int}(\gamma_0)} \nabla \cdot g, \tag{8.74}$$

where the integrals are now taken along the "unperturbed" orbit γ_o.

When γ_o is a periodic orbit, $\gamma_o(t) = \gamma_o(t+T)$, we can use $dx = \dot{x} dt$ and $dy = \dot{y} dt$ to convert the integral (8.74) into a time integral over one period of the orbit. On the $\varepsilon = 0$ orbit $(\dot{x}, \dot{y}) = f$, so (8.74) becomes

$$0 = \int_0^T (g_2(x(t),y(t))\dot{x} - g_1(x(t),y(t))\dot{y}) dt = \int_0^T (g_2 f_1 - g_1 f_2)_{\gamma_o(t)} dt \equiv \int_0^T f \wedge g|_{\gamma_o(t)} dt,$$

where we have defined the wedge product: $f \wedge g \equiv f_1 g_2 - f_2 g_1$.

When γ_o is a homoclinic orbit, the "period" becomes infinity, and (x,y) are functions that limit to the saddle point in both directions of time. Since $f(x(t),y(t)) \to 0$ as the orbit approaches equilibrium, and does so exponentially fast, the integral converges as $t \to \pm\infty$. Thus (8.74) becomes

$$M = \int_{-\infty}^{\infty} f \wedge g|_{\gamma_o(t)} dt. \tag{8.75}$$

This integral is known as a *Melnikov integral*. Its vanishing is a necessary condition for the existence of a closed orbit γ near the original homoclinic orbit.

Lemma 8.30. *Suppose that* (8.73) *has a homoclinic loop γ_o when $\varepsilon = 0$. Then a necessary condition for the existence of an invariant loop when ε is small is that* (8.75) *vanishes.*

Example 8.31. Consider the non-Hamiltonian perturbation of the system (4.29):

$$\begin{aligned} \dot{x} &= y + \varepsilon x, \\ \dot{y} &= x - 3ax^2 + \varepsilon bxy, \end{aligned} \tag{8.76}$$

so that $g = (x, bxy)$. To apply the Melnikov criterion, we only need an expression for the unperturbed solution homoclinic to $(0,0)$, which is $y_\pm = \pm x\sqrt{1-2ax}$. Then the

Melnikov integral (8.74) becomes

$$
\begin{aligned}
M &= \oint_{\gamma_o} g_2 dx - g_1 dy = \int_{y_+} g_2 dx - g_1 dy + \int_{y_-} g_2 dx - g_1 dy, \\
&= \int_0^{\frac{1}{2a}} \left(bxy_+(x)dx - x dy_+(x)\right) - \int_0^{\frac{1}{2a}} \left(bxy_-(x)dx - x dy_-(x)\right), \\
&= \int_0^{\frac{1}{2a}} \left(bxy_+ - x\frac{dy_+}{dx}\right)dx - \int_0^{\frac{1}{2a}} \left(bxy_- - x\frac{dy_-}{dx}\right)dx.
\end{aligned}
$$

The two integrals can be combined since $y_- = -y_+$ to give[53]

$$
M = 2\int_0^{\frac{1}{2a}} x\left(by_+ - \frac{dy_+}{dx}\right)dx = 2\frac{2b+7a}{105a^2},
$$

which is nonzero unless $b = -7a/2$. Therefore, except on this curve, there are no nearby closed loop orbits. This result applies when $\varepsilon \ll 1$. As we see in Figure 8.19, there are indeed no nearby periodic orbits when $(a,b) = (1,0)$, but a nearby periodic orbit is created when $b \approx -7a/2$ even when ε is as large as 0.1. ∎

8.13 ▪ Melnikov's Method for Nonautonomous Perturbations

Nonautonomous perturbations to a planar system can also be treated by similar methods. In his Ph.D. thesis in 1963, Viktor Melnikov devised a perturbative technique to compute the motion of the stable and unstable manifolds. Begin, as before, with a Hamiltonian vector field in the plane that has a homoclinic loop. Upon adding a perturbation that could be non-Hamiltonian and periodically time dependent, the loop will typically be destroyed. We will compute the distance between the stable and unstable manifolds of the perturbed fixed point.

Letting $z = (x,y)$, consider the system

$$
\dot{z} = f(z) + \varepsilon\hat{g}(z,t), \quad \nabla\cdot f = 0, \quad \hat{g}(z,t+T) = \hat{g}(z,t), \tag{8.77}
$$

where f and $\hat{g}$ are C^2. Upon introducing a phase $\theta = \omega t$ with $\omega = 2\pi/T$, this system becomes autonomous on the extended phase space $\mathbb{R}^2 \times \mathbb{S}^1$:

$$
\begin{aligned}
\dot{z} &= f(z) + \varepsilon g(z,\theta), \\
\dot{\theta} &= \omega,
\end{aligned} \tag{8.78}
$$

with $g(z,\theta+2\pi) = g(z,\theta) = \hat{g}(z,\theta/\omega)$. Since the θ equation is independent of z, this is a skew-product system. The solution for $\varepsilon = 0$ is particularly simple: if $\varphi_t(z)$ is the flow for f, then

$$
(z(t),\theta(t)) = (\varphi_t(z), \theta + \omega t)
$$

is the flow for (8.78). Therefore, any equilibrium, p_o, of f becomes a periodic orbit of the extended system given by the closed loop $\gamma_o(t) = (p_o, \omega t \bmod 2\pi)$.

Moreover, if p_o is a hyperbolic equilibrium of f, then the implicit function theorem implies that in the extended phase space, the periodic orbit persists for $\varepsilon > 0$.

[53] The substitution $u = \sqrt{1-2ax}$ simplifies the integrals.

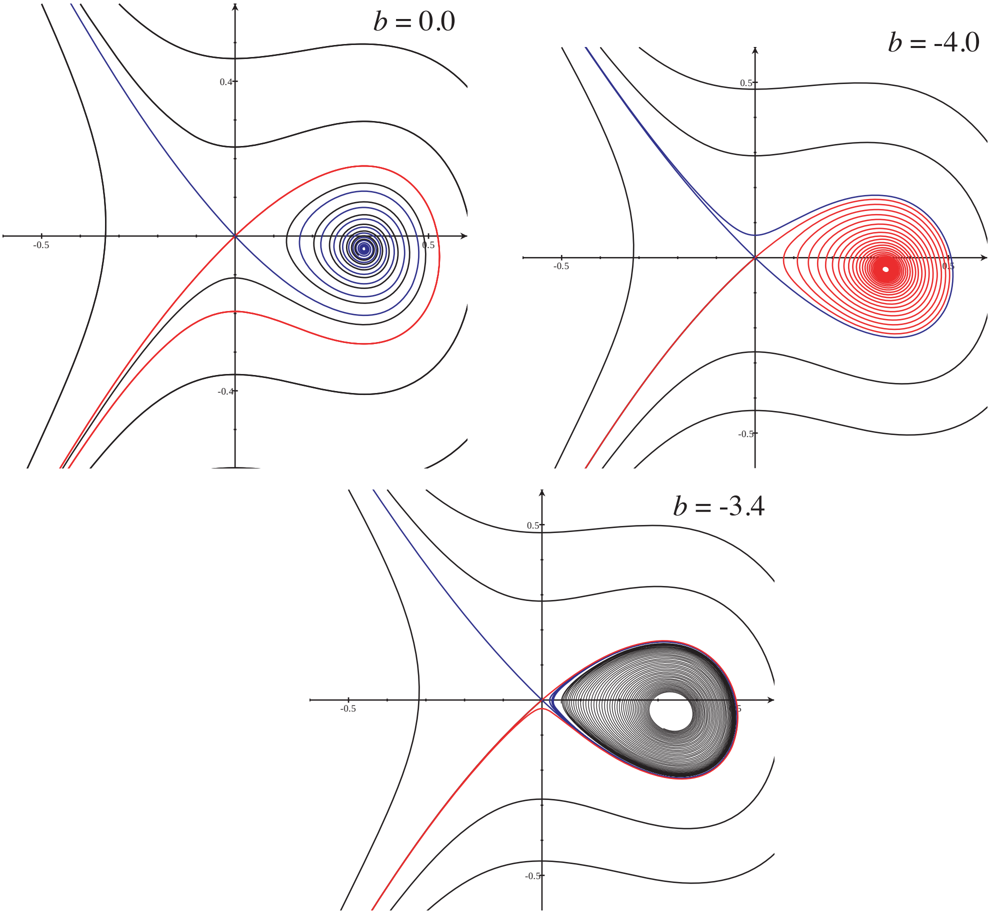

Figure 8.19. *Flows for the system* (8.76) *with* $\varepsilon = 0.1$ *and* $a = 1.0$. *The three figures show* $b = 0.0$, -4.0, *and* -3.4, *respectively. Between the upper two panels the stable and unstable manifolds must cross. They are nearly coincident in the third, where* b *is just above* $-7a/2$.

Theorem 8.32 (persistence of hyperbolic periodic orbits). *If* $\gamma_o(t) = (p_o, \omega t)$ *is a hyperbolic periodic orbit of* (8.78) *at* $\varepsilon = 0$, *then there is an* $\varepsilon_0 > 0$ *such that for any for* $|\varepsilon| < \varepsilon_0$ *there is unique periodic orbit* $\gamma_\varepsilon(t)$ *of period* T *that continues from* $\gamma_o(t)$.

Proof. Let $\varphi_t(z)$ denote the flow of $f(z)$ in $\mathbb{R}^2$. The surface $\bar{S} = \{(z, \theta) : \theta = \theta_o\}$ is a global section of (8.78) for any ε; let $P_\varepsilon(z)$ be the Poincaré map on $\bar{S}$: when $\varepsilon = 0$, $P_o(z) = \varphi_T(z)$. Thus $P_o(z)$ has a fixed-point $p_o = P_o(p_o)$. Moreover, since the flow linearized about $z = p_o$ is $D\varphi_t(p_o) = e^{tDf(p_o)}$, we have $DP_o(p_o) = D\varphi_T(p_o) = e^{TDf(p_o)}$. The multipliers of $DP_o(p_o)$ are $\mu_\pm = e^{\lambda_\pm T}$, where $\lambda_\pm$ are the eigenvalues of $Df(p_o)$. Thus since p_o is hyperbolic, $\mathrm{Re}(\lambda_\pm) \neq 0$ and therefore $\mu_\pm \neq 1$.[54] Thus the equation

$$F(z; \varepsilon) = P_\varepsilon(z) - z = 0$$

satisfies the following hypotheses of the implicit function theorem, Theorem 8.3: (a)

[54] By restricting the map to the section $\bar{S}$, we eliminate the tangent eigenvector $\dot{\gamma}$, which does have multiplier 1.

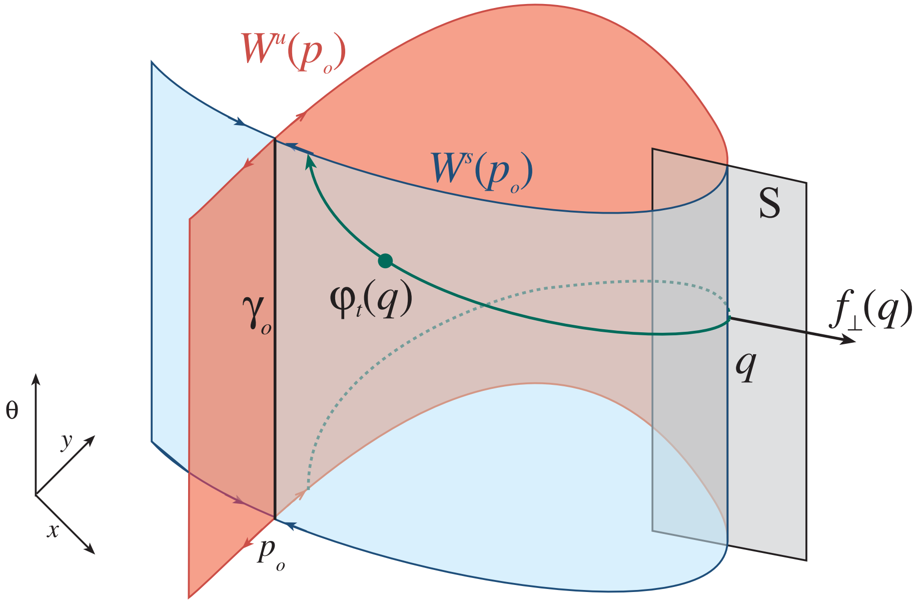

Figure 8.20. *The unperturbed flow of* (8.78) *and the homoclinic manifold* $H(\gamma_o)$.

$DP_o - I$ is nonsingular, and (b) $P_o(p_o) - p_o = 0$. Thus there exists a unique fixed point p_ε that continues from p_o for $|\varepsilon| < \varepsilon_0$. The orbit of p_ε is a periodic orbit of (8.78). □

Now suppose that when $\varepsilon = 0$, the equilibrium p_o of (8.77) has a homoclinic loop, $\gamma_o \subset W^s(p_o) \cup W^u(p_o)$. Then the corresponding periodic orbit γ_o of (8.78) has a two-dimensional homoclinic manifold

$$H(\gamma_o) = \left\{(z,\theta) : z \in \gamma_o, \theta \in \mathbb{S}^1\right\}, \tag{8.79}$$

as sketched in Figure 8.20. Every orbit on this manifold is both forward and backward asymptotic to γ_o.

We now wish to study the effect of the perturbations on $H(\gamma_o)$. To do this we develop an expression for the rate of change of the manifolds with ε. This expression gives rise to a vector field called the Melnikov vector field.

For any $q \in \Gamma_o$, the mapping $(t,\theta) \to (\varphi_t(q),\theta)$ uniquely represents any point on $H(\gamma_o)$. To measure the distance between the perturbed manifolds $W^s(\gamma_\varepsilon)$ and $W^u(\gamma_\varepsilon)$ we use the perpendicular vector to the unperturbed manifolds: since $f_\perp = (-f_2, f_1)$ is the two-dimensional vector perpendicular to the unperturbed flow, then at (t,θ), the three-dimensional vector perpendicular to the manifold is $(f_\perp(\varphi_t(q)), 0)$. Note that the dot product of a vector with $f_\perp$ is equal to the wedge product with f:

$$f_\perp \cdot v = -f_2 v_1 + f_1 v_2 = f \wedge v. \tag{8.80}$$

Define a local section at a point q on the homoclinic manifold by

$$S = \{(z,\theta) : z = q + \sigma f_\perp(q), \sigma \in (-\delta,\delta), \theta \in [0,2\pi)\}$$

for some $\delta > 0$; see Figure 8.20. Since S is transverse to the flow at $\varepsilon = 0$, it is still transverse to the perturbed flow for small enough ε.

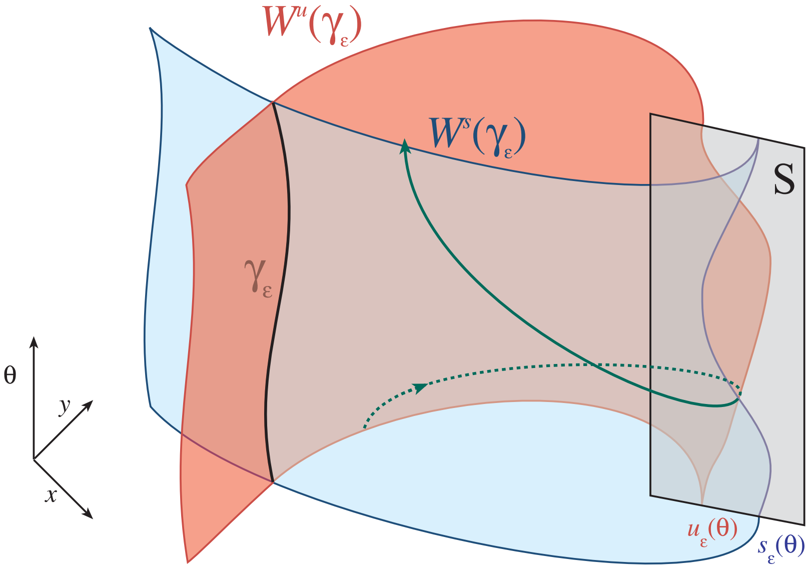

Figure 8.21. *Flow of* (8.78) *for* $\varepsilon \neq 0$ *and the perturbed stable and unstable manifolds of* γ_ε.

We denote the perturbed flow of (8.78) ($\varepsilon \neq 0$) by

$$(z(t),\theta(t)) = (\psi_t(z,\theta), \theta + \omega t). \tag{8.81}$$

The intersections of the manifolds with S are denoted

$$(s_\varepsilon(\theta),\theta) = W^s(\gamma_\varepsilon) \cap S \qquad (u_\varepsilon(\theta),\theta) = W^u(\gamma_\varepsilon) \cap S; \tag{8.82}$$

see Figure 8.21. We assume that these continue from the original intersections, so that $s_o(\theta) = u_o(\theta) = q$.[55]

Theorem 3.30, which guarantees smoothness of the flow with respect to parameters and with respect to initial conditions, implies that we can find the solution ψ_t as a power series expansion away from φ_t. For example, the solution on the stable manifold that starts at $(s_\varepsilon(\theta),\theta)$ can be expanded as

$$\psi_t(s_\varepsilon(\theta),\theta) = \varphi_t(q,\theta) + \varepsilon \xi_t^s(q,\theta) + O(\varepsilon^2). \tag{8.83}$$

Straightforward application of Grönwall's inequality, Lemma 3.28—varying both the initial condition and the parameter ε—shows that the difference, ξ_t^s, is bounded for any finite time.

Now since $\psi_t(s_\varepsilon(\theta),\theta) \to \gamma_\varepsilon$ as $t \to \infty$, there is some finite time T_δ such that ψ_t is within δ of γ_ε for any $\delta > 0$. Moreover, the implicit function theorem implies γ_ε is $O(\varepsilon)$ close to γ_o. Since the Grönwall inequality implies that, up to T_δ, the deviation is $O(\varepsilon)$, $\psi_t = \varphi_t + O(\varepsilon)$ for all $t > 0$. A similar argument leads to the conclusion that for points on the unstable manifold, the deviation is bounded for all $t < 0$.

A measure of the distance between the manifolds is the dot product of the difference between these points with the perpendicular vector field $f_\perp$, or equivalently the

[55] Note that there will probably be many intersections of W^s and W^u with S; indeed, as we will see, if the manifolds intersect transversely, Smale's horseshoe theorem implies there must be infinitely many. Right now we are interested in only the "first" intersections.

wedge product with f itself:[56]

$$\Delta_\varepsilon(t,\theta) \equiv f(\varphi_t(q)) \wedge (\psi_t(u_\varepsilon(\theta),\theta) - \psi_t(s_\varepsilon(\theta),\theta)) = \varepsilon M(t,\theta) + O(\varepsilon^2).$$

We have noted here that $\Delta_o(t,0) \equiv 0$, since the manifolds coincide at $\varepsilon = 0$, and we have defined the *Melnikov function* as the rate of change of this distance with ε:

$$M(t,\theta) \equiv \left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} \Delta_\varepsilon(t,\theta) = f(\varphi_t(q)) \wedge (\xi_t^u(q,\theta) - \xi_t^s(q,\theta)) = M^u(t,\theta) - M^s(t,\theta). \tag{8.84}$$

On the section S itself, the deviation is given by

$$\Delta_\varepsilon(\theta) \equiv \Delta_\varepsilon(0,\theta) = \varepsilon M(0,\theta) + O(\varepsilon^2). \tag{8.85}$$

To compute the terms in (8.84), we derive a differential equation for M. We consider the stable and unstable terms in (8.84) separately. For example, the stable part has derivative

$$\frac{d}{dt}M^s = Df(\varphi_t(q))\frac{d}{dt}\varphi_t(q) \wedge \xi_t^s(q,\theta) + f(\varphi_t(q)) \wedge \frac{d}{dt}\xi_t^s(q,\theta). \tag{8.86}$$

To evaluate this we need the differential equation for $\xi_t^s = d/d\varepsilon|_{\varepsilon=0}\, \psi_t(s_\varepsilon(\theta),\theta)$; this is obtained by differentiation of (8.78) with respect to ε:

$$\begin{aligned} \frac{d}{dt}\xi_t^s &= \frac{d}{d\varepsilon}\left(\frac{d}{dt}\psi_t\right)_{\varepsilon=0} = \frac{d}{d\varepsilon}\big(f(\psi_t(s_\varepsilon(\theta),\theta)) + \varepsilon g(\psi_t(s_\varepsilon(\theta),\theta),\theta+\omega t)\big)_{\varepsilon=0} \\ &= Df(\varphi_t(q))\xi_t^s + g(\varphi_t(q),\theta+\omega t). \end{aligned} \tag{8.87}$$

This linear equation is to be solved with the initial condition

$$\xi_0^s = \left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0} s_\varepsilon(\theta).$$

Substituting $\dot{\varphi}_t = f(\varphi_t)$ and (8.87) into (8.86) gives

$$\frac{d}{dt}M^s = Df(\varphi_t)f(\varphi_t) \wedge \xi_t^s + f(\varphi_t) \wedge (Df(\varphi_t)\xi_t^s + g).$$

The first two terms can be combined; it seems easiest to expand all the vector and matrix products to see this:

$$\begin{aligned} (Df(\varphi_t)f(\varphi_t)) \wedge \xi + f(\varphi_t) \wedge (Df(\varphi_t)\xi) &= \left(Df_{1j}f_j\right)\xi_2 - \left(Df_{2j}f_j\right)\xi_1 \\ &\quad + f_1\left(Df_{2j}\xi_j\right) - f_2\left(Df_{1j}\xi_j\right) \\ &= (-Df_{21}f_1 - Df_{22}f_2 + f_1Df_{21} - f_2Df_{11})\xi_1 \\ &\quad + (Df_{11}f_1 + Df_{12}f_2 + f_1Df_{22} - f_2Df_{12})\xi_2 \\ &= -(Df_{22} + Df_{11})f_2\xi_1 + (Df_{11} + Df_{22})f_1\xi_2 \\ &= -\mathrm{tr}(Df)f \wedge \xi. \end{aligned}$$

Now since $f \wedge \xi = M^s$ we obtain

$$\frac{d}{dt}M^s(t,\theta) = -\mathrm{tr}(Df(\varphi_t(q)))M^s(t,\theta) + f(\varphi_t(q)) \wedge g(\varphi_t(q),\theta+\omega t).$$

[56] To get the actual distance we could divide by the norm of f, but we are interested in only a measure of the distance and, in particular, whether the distance is zero.

Since f is by assumption a Hamiltonian vector field, $\mathrm{tr}(Df)\equiv 0$; this implies that the ODE is now trivial,[57]

$$\frac{d}{dt}M^s(t,\theta)=f(\varphi_t(q))\wedge g(\varphi_t(q),\theta+\omega t), \tag{8.88}$$

since everything on the right-hand side is now known. Thus we can simply integrate the equation to obtain M^s. Note that $M^s(t,\theta)$ vanishes exponentially fast as $t\to\infty$ because $\varphi_t(q)\to p_o$ and $f(p_o)=0$. Therefore if we integrate (8.88) from (t,∞), we have

$$M^s(t,\theta)=-\int_t^\infty f(\varphi_\tau(q))\wedge g(\varphi_\tau(q),\theta+\omega\tau)d\tau.$$

A similar calculation gives $M^u(t,\theta)=\int_{-\infty}^t f(\varphi_\tau(q))\wedge g(\varphi_\tau(q),\theta+\omega\tau)d\tau$. Putting these together in (8.84) gives the Melnikov function

$$M(\theta)=\int_{-\infty}^\infty f(\varphi_\tau(q))\wedge g(\varphi_\tau(q),\theta+\omega\tau)d\tau. \tag{8.89}$$

Here we have noted that this is independent of t (see Exercise 23). This expression is almost identical to (8.75) for the autonomous case! Note that

$$M(\theta+2\pi)=M(\theta) \tag{8.90}$$

since g is a periodic function.

By construction, when $M(\theta)=0$, the deviation in the manifolds is zero to first order in ε. Our expectation is that this means that there is a true crossing of the manifolds nearby. Indeed, this is true, provided that a nondegeneracy condition is satisfied, as we state in the next theorem.

Theorem 8.33 (Melnikov). *Suppose there is a point θ_o on the saddle connection such that $M(\theta_o)=0$ and $D_\theta M(\theta_o)\neq 0$. Then when ε is sufficiently small, $W^s(\gamma_\varepsilon)$ and $W^u(\gamma_\varepsilon)$ intersect transversely at a point within $O(\varepsilon)$ of (q,θ_o).*

Proof. This follows almost immediately from Theorem 8.3. The separation between the manifolds on the section S is measured by the formal expression (8.85). By definition $\Delta_o(\theta)=0$, and $d/d\varepsilon\Delta_\varepsilon(\theta)|_{\varepsilon=0}=M(\theta)$. The implicit function theorem cannot be applied to $\Delta_\varepsilon(\theta)$; however, the new function

$$F(\theta;\varepsilon)=\Delta_\varepsilon(\theta)/\varepsilon=M(\theta)+O(\varepsilon)$$

does satisfy the required conditions at the point θ_o: $F(\theta_o;0)=0$ and $D_\theta F(\theta_o;0)\neq 0$. Thus there is a unique curve $\theta(\varepsilon)$ such that $F(\theta(\varepsilon);\varepsilon)=0$, or equivalently $\Delta_\varepsilon(\theta(\varepsilon))=0$. Since M changes sign as θ traverses θ_o, the separation must change sign upon crossing $\theta(\varepsilon)$. □

When the manifolds cross at a point $s_\varepsilon(\theta)=u_\varepsilon(\theta)$, they cross on the orbit of this point as well. Since this orbit $(\psi_t(s_\varepsilon(\theta)),\theta+\omega t)$ moves periodically in θ, as it approaches the equilibrium, the crossing point will intersect each section

$$\bar{S}=\{(x,\theta):\theta=\theta_o\}$$

[57] The nonhomogeneous linear equation for M when $\mathrm{tr}(Df)\neq 0$ is also not hard to solve. See Exercise 21.

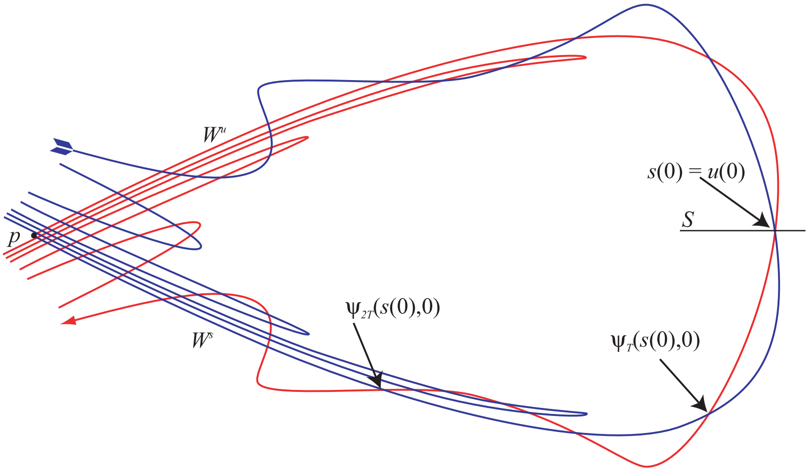

Figure 8.22. *Sketch of a cross section* $\bar{S}$ *for* $\theta_o = 0$ *of the stable and unstable manifolds. Here we suppose that* $s(0) = u(0) = 0$*, so that the crossing takes place on the section* $\bar{S}$*. The next crossing on the orbit of* $s(0)$ *occurs at time* T*, the period of* g*.*

infinitely many times. The resulting picture is extremely intricate, and only a brief indication of the complexity is sketched in Figure 8.22. Indeed, when Poincaré discovered the possibility of the transverse crossing of stable and unstable manifolds, he said the following (in our translation from the French):

> *When one tries to depict the figure formed by these two curves and their infinity of intersections, each of which corresponds to a doubly asymptotic solution, these intersections form a kind of net, web or infinitely tight mesh; neither of the two curves can ever cross itself, but must fold back on itself in a very complex way in order to cross the links of the web infinitely many times. One is struck by the complexity of this figure that I am not even attempting to draw.* (Henri Poincaré, *New Methods in Celestial Mechanics*, 1892, Vol. 3, §397)

Since Poincaré, there have been many attempts to sketch this figure—many are incorrect!

Example 8.34. To compute the Melnikov function for the example Hamiltonian (5.3), we first need to find the solution on the unperturbed homoclinic orbit, defined by

$$\dot{x} = y = \pm x\sqrt{1-2ax}. \tag{8.91}$$

Choosing a point on the unperturbed manifold, $q = (1/2a, 0)$, we must find $\varphi_t(q)$. Since (8.91) is separable, it can be integrated to obtain

$$\pm t + c = \int \frac{dx}{x\sqrt{1-2ax}} = -2\int \frac{du}{1-u^2} = -2\tanh^{-1}\left(\sqrt{1-2ax}\right),$$

where we used the substitution $u = \sqrt{1-2ax}$. Choosing $c = 0$ to give the proper initial condition and solving for x yields

$$x(t) = \frac{1}{2a}\operatorname{sech}^2\left(\frac{t}{2}\right).$$

This solution is valid for all $t \in \mathbb{R}$, even though the ± signs have disappeared. The solution for y is obtained using $y = \dot{x}$:

$$y(t) = -\frac{1}{2a}\operatorname{sech}^2\left(\frac{t}{2}\right)\tanh\left(\frac{t}{2}\right).$$

Evaluation of the Melnikov expression (8.89) requires finding f along the orbit, but since $\dot{z} = f(z)$, this is equivalent to taking time derivatives

$$\begin{aligned} M(\theta) &= \int_{-\infty}^{\infty} [\dot{x}(t)g_2(x(t),y(t),\omega t+\theta) - \dot{y}(t)g_1(x(t),y(t),\omega t+\theta)]dt, \\ &= -\frac{1}{4a}\int_{-\infty}^{\infty} \operatorname{sech}^2\left(\frac{t}{2}\right)\left(2\tanh\left(\frac{t}{2}\right)g_2 + \left(2-3\operatorname{sech}^2\left(\frac{t}{2}\right)\right)g_1\right)dt. \end{aligned}$$

This integral is well behaved near $\pm\infty$ since $\operatorname{sech}^2(\tau)$ is exponentially small there, provided only that g is bounded at the position of the unperturbed hyperbolic equilibrium, $(0,0)$.

Suppose, for example, that $g(x,y,\theta) = (0, \cos(\theta))$. The integral becomes

$$M(\theta) = -\frac{1}{2a}\int_{-\infty}^{\infty} \operatorname{sech}^2\left(\frac{t}{2}\right)\tanh\left(\frac{t}{2}\right)\cos(\omega t+\theta)dt,$$

which can be simplified by partial integration and expansion of the trig function:

$$M(\theta) = \frac{\omega}{2a}\int_{-\infty}^{\infty} \operatorname{sech}^2\left(\frac{t}{2}\right)\sin(\omega t+\theta)dt = \frac{\omega}{a}\sin\theta\int_{-\infty}^{\infty}\operatorname{sech}^2(\tau)e^{2i\omega\tau}d\tau.$$

Here we used the fact that sech^2 is even to convert the last integral into an exponential, and we changed variables to $\tau = t/2$. Note that $\operatorname{sech}^2\tau = (\cos(i\tau))^{-2}$ has double poles at the zeros of the cosine, or at $\tau_n = i\pi(2n+1)/2$. The integral is most easily done by the Cauchy residue method: close the integral in the upper half-plane and sum over the residues R_n at the poles τ_n:

$$M(\theta) = 2\pi i \sum_{n=0}^{\infty} R_n.$$

Near a pole,

$$\cos(i\tau) = 0 - i\sin(i\tau_n)(\tau-\tau_n) + O(\tau-\tau_n)^3 = i(-1)^n(\tau-\tau_n) + O(\tau-\tau_n)^3,$$

and the residue at τ_n is given by the $O(\tau-\tau_n)^{-1}$ term in the expansion:

$$\begin{aligned} \operatorname{sech}^2(\tau)e^{2i\omega\tau} &= \frac{-1}{(\tau-\tau_n)^2}e^{2i\omega\tau_n}e^{2i\omega(\tau-\tau_n)} + O(\tau-\tau_n)^0 \\ &= \frac{-e^{2i\omega\tau_n}}{(\tau-\tau_n)^2}(1+2i\omega(\tau-\tau_n)) + O(\tau-\tau_n)^0. \end{aligned}$$

Thus $R_n = -2i\frac{\omega^2}{a}\sin\theta e^{2i\omega\tau_n}$, and the Melnikov function becomes

$$\begin{aligned} M(\theta) &= \frac{4\pi\omega^2}{a}\sin\theta\sum_{n=0}^{\infty}e^{2i\omega\tau_n} = \frac{4\pi\omega^2}{a}\sin\theta e^{-\omega\pi}\sum_{n=0}^{\infty}e^{-2\pi\omega n} \\ &= \frac{4\pi\omega^2}{a}\sin\theta\frac{e^{-\pi\omega}}{1-e^{-2\pi\omega}} = \frac{2\pi\omega^2}{a}\sin\theta\operatorname{csch}(\pi\omega). \end{aligned} \tag{8.92}$$

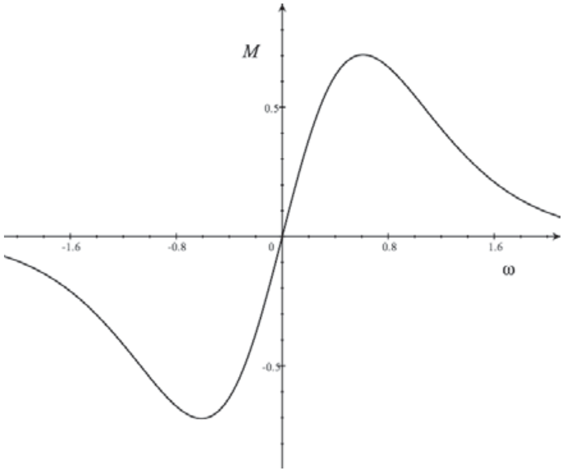

Figure 8.23. *Melnikov function* (8.92) *when* $\theta = \pi/2$ *and* $a = 1$, *as a function of* ω.

Note that $M(\theta)$ is periodic in θ and vanishes at $\theta = 0$ and π but is otherwise nonzero, as is shown in Figure 8.23. We therefore may conclude that the stable and unstable manifolds intersect transversely. ■

The calculation in the example points out one potential pitfall in the Melnikov formulation: Theorem 8.33 is valid only if the parameters in the system, other than the small parameter ε, are assumed to be $O(1)$. By contrast if $\omega \gg 1$, i.e., $\omega = O(1/\varepsilon)$, then the size of the Melnikov function is exponentially small in ε. This invalidates the ordering assumptions and the implicit function argument that we used. Proving that the manifolds intersect transversely in this case is much, much harder. This very interesting case of a rapidly oscillating perturbation requires a theory of "asymptotics beyond all orders" for its resolution (Delshams and Seara 1997; Holmes, Marsden, and Scheurle 1988; Segur 1993).

8.14 ▪ Shilnikov Bifurcation

Homoclinic orbits can also occur in higher-dimensional systems. For example, the intersections of the stable and unstable manifolds of a hyperbolic periodic orbit in three or more dimensions can exhibit complexity similar to that of the nonautonomous systems studied in §8.13. Another novel situation arises when a hyperbolic equilibrium in three or more dimensions has a homoclinic orbit. A bifurcation of such an orbit can give rise to infinitely many periodic orbits—Leonid Shilnikov extensively studied this case in the 1960s.

Consider a vector field $f_o(x)$ in $\mathbb{R}^3$ that has a hyperbolic equilibrium p_o with one unstable eigenvalue $\lambda_1 > 0$ and two stable eigenvalues λ_2 and λ_3 ordered so that $\mathrm{Re}(\lambda_3) \le \mathrm{Re}(\lambda_2) < 0$. The equilibrium p_o is a saddle when the stable eigenvalues are real and is a saddle-focus when $\lambda_{2,3} = \alpha \pm i\beta$ and $\beta \neq 0$. Just as in the planar case, Theorem 8.28, the sum of stable and unstable eigenvalues, τ, is important in determining the character of the bifurcation; it turns out that for the three-dimensional case only the "leading" stable eigenvalue is generically important. Consequently, the "leading trace" (or what Shilnikov calls the *saddle value*) is defined as

$$\tau \equiv \lambda_1 + \mathrm{Re}(\lambda_2). \tag{8.93}$$

Suppose that there exists an orbit γ_o that is homoclinic to p_o:

$$\gamma_o \subset W^u(p_o) \cap W^s(p_o).$$

Just as in Theorem 8.28, Shilnikov studied bifurcations in a neighborhood of γ_o. For this case the neighborhood, U, is a tube enclosing $\gamma_o \cup p_o$.

If $f(x;\mu)$ is an unfolding of f_o, then, just as for the two dimensions, varying any parameter μ will generically destroy the homoclinic orbit. The main question we address is, are there other orbits that are necessarily created in U when γ_o is destroyed? There are four cases depending on whether p_o is a saddle or saddle-focus and on the sign of τ.

Theorem 8.35 (Shilnikov homoclinic saddle). *Let $f(x;\mu)$ be a generic, one-parameter unfolding of a three-dimensional vector field $f(x;0)$ that has a saddle equilibrium p_o with real eigenvalues*

$$\lambda_3 < \lambda_2 < 0 < \lambda_1 \tag{8.94}$$

and $\tau = \lambda_1 + \lambda_2 \neq 0$. Suppose that p_o has a homoclinic orbit γ_o and that as $t \to \infty$, γ_o approaches p_o along its leading stable eigenvector (with eigenvalue λ_2). Then there exists a neighborhood U of $\gamma_o \cup p_o$ such that a unique limit cycle is created in U as μ crosses zero. Moreover,

(a) *if $\tau > 0$, then the limit cycle is unstable (it has one unstable multiplier), and*

(b) *if $\tau < 0$, the limit cycle is a sink.*

The hypothesis of a generic unfolding means that the parameter μ causes the stable and unstable manifolds of $p(\mu)$ to split, in other words, that there is a nonzero quantity analogous to Δ in Theorem 8.28. When $\tau < 0$, it can be seen that the limit cycle is created when the branch of the unstable manifold of the equilibrium $p(\mu)$ crosses from "below" to "above" $W^s(p(\mu))$ as sketched in Figure 8.24.

Note that when $\mu = 0$, the orbit $\gamma_o \subset W^s(p_o)$, so that the stable manifold must accumulate on itself as $t \to -\infty$, forming a ribbon as sketched in Figure 8.24. In the neighborhood of each point of γ_o for finite time, the stable manifold is a smooth two-dimensional surface. As $t \to -\infty$ one tangent vector to this surface limits on the unstable eigenvector v_1 of p_o. Generically the second tangent vector will limit on the direction of most rapid contraction, that is, the eigenvector v_3 corresponding to λ_3. There are two topologically distinct configurations for this ribbon: in one the ribbon is orientable, as shown in Figure 8.24, and in the other it acquires a half-twist so that the $W^s(p) \cap U$ is a Möbius band. This topology has an influence on the homoclinic bifurcation when $\tau > 0$: in the untwisted case, the limit cycle is created when $W^u(p)$ crosses from above $W^s(p)$, while in the twisted case it is created upon a crossing from below.

Finally, note that Theorem 8.35 can also be applied to a saddle equilibrium with a two-dimensional unstable eigenspace and a one-dimensional stable eigenspace simply by reversing the direction of time.

In either case, the creation of a limit cycle in this bifurcation could be expected from our study of the two-dimensional case. The same cannot be said for a homoclinic bifurcation when the equilibrium is a saddle-focus.

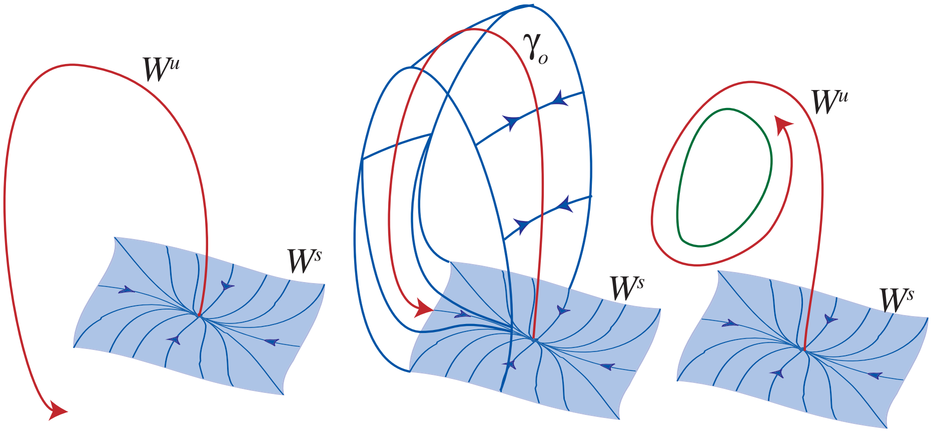

Figure 8.24. *Homoclinic bifurcation for a three-dimensional saddle with $\tau < 0$. The blue surfaces represent the stable manifold and the red curves the unstable manifold.*

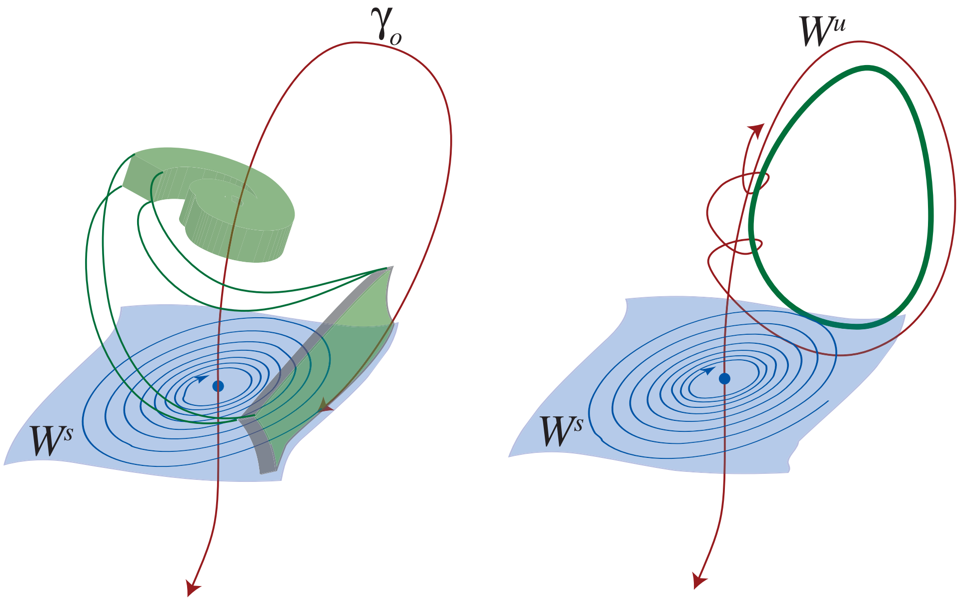

Figure 8.25. *Dynamics near a homoclinic orbit to a saddle-focus equilibrium with blue stable manifold and red unstable manifold. The left panel shows the spiral structure on a Poincaré section (gray) near the homoclinic trajectory γ_o, and the right the creation of a periodic orbit.*

Theorem 8.36 (saddle-focus homoclinic). *Let $f(x;\mu)$ be a generic, one-parameter unfolding of a three-dimensional vector field $f(x;0)$ that has a saddle-focus equilibrium p_o with eigenvalues $\lambda_1 > 0$ and $\lambda_{2,3} = \alpha \pm i\beta$ with $\alpha < 0$, $\beta \neq 0$, and $\tau = \lambda_1 + \alpha \neq 0$. Suppose that p_o has a homoclinic orbit γ_o. Then there exists a neighborhood U of $\gamma_o \cup p_o$ such that*

(a) *if $\tau > 0$, there is a $\mu_o > 0$ such that there are infinitely many saddle limit cycles in U for all $|0 < \mu| < \mu_o$; or*

(b) *if $\tau < 0$, a unique, stable limit cycle is created in U when μ passes through zero.*

The first case is remarkable in that an incredibly intricate structure, with infinitely many periodic orbits, occurs in a parameter region around the homoclinic orbit. The basic point is that since γ_o lies on $W^s(p_o)$, it must spiral infinitely many times as it approaches p_o. Similarly, a ball of orbits that passes near the equilibrium is twisted into a thin spiral that is spread along the unstable manifold; see Figure 8.25. This behavior persists for small μ and is reflected in the behavior of nearby orbits. When the leading trace is negative, this spiral is contracted rapidly toward the unstable manifold, and the resulting Poincaré map is a contraction on a cross section near γ_o when the homoclinic connection is broken to the "same side" of W^s as the homoclinic branch of W^u. This contraction mapping has a unique fixed point that corresponds to the newly created periodic orbit. When the leading trace is positive, contraction near the stable manifold is overwhelmed by expansion along the unstable manifold, and the image of the spiral crosses itself giving rise to an infinite number of fixed points of the Poincaré mapping. The theorem is proved by constructing a two-dimensional section transverse to $W^u(p_o)$ and showing that the Poincaré map is guaranteed to have infinitely many fixed points.

The proofs of the Shilnikov theorems can be found in (Kuznetsov 1995; Shilnikov et al. 1998).

8.15 ▪ Exercises

In each of these problems, where appropriate, use your favorite computer software to create phase portraits of these systems and compare with the theoretical results.

1. Find the equilibria and bifurcation points of the following one-dimensional ODEs. Draw the bifurcation diagram. Sketch the phase portraits for parameter values that represent the distinct classes of motion.

   (a) $\dot{x} = \mu + x^3$,
   (b) $\dot{x} = 1 + \mu x + x^3$,
   (c) $\dot{x} = \mu x + 2x^2 - x^3$,
   (d) $\dot{x} = \mu x + \sin(x)$,
   (e) $\dot{x} = \mu + 2x^2 - x^4$,
   (f) $\dot{x} = -4\mu^2 + 5\mu x^2 - x^4$.

2. Show that the flow of the vector field $\dot{y} = \nu y - y^2$ is diffeomorphic to the flow induced by the vector field $\dot{x} = \mu - x^2$. Thus the "transcritical bifurcation" of the first equation is nothing more than a disguised saddle-node bifurcation.

3. Show that if $A(c)$ is a matrix that depends continuously upon a parameter c, then the eigenvalues of A depend continuously on c as well. Now suppose that A depends smoothly on c. Show by example that the eigenvalues of $A(c)$ need not be smoothly dependent upon c. (*Hint*: Consider the solutions $\lambda(c)$ defined implicitly by the characteristic polynomial $p(\lambda(c), c) = \det(\lambda I - A(c))$.)

4. Here we will show that there is a neighborhood of the Takens–Bogdanov form (8.13) in which every 2×2 matrix is linearly conjugate to A_μ (8.14).

   (a) First suppose that M has eigenvalues $\lambda_1 \neq \lambda_2$. Note that M is linearly conjugate to the matrix $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$. Find an explicit linear conjugacy between Λ and A_μ. Consider the cases when the eigenvalues are real and when they are complex.

(b) Suppose that M has eigenvalues $\lambda_1 = \lambda_2 = \lambda \in \mathbb{R}$ and has geometric multiplicity one. Show that it is conjugate to the Jordan normal form $K = \left(\begin{smallmatrix} \lambda & 1 \\ 0 & \lambda \end{smallmatrix}\right)$. Find an explicit linear conjugacy between K and A_μ. (*Hint*: The generalized eigenvectors of M can be used to construct the first conjugacy.)

(c) The conjugacy in (a) fails when $\lambda_1 = \lambda_2$. Why? To see that semisimple matrices with a double eigenvalue are not near the Takens–Bogdanov form, we will use the Euclidean norm on $\mathbb{R}^4$ in the matrix components (a,b,c,d). Show that if M is semisimple and has a double eigenvalue, then there is an ε such that M is not in the ball of radius ε about J.

5. Consider the space $\mathbb{H}_k$ of homogeneous polynomials on $\mathbb{R}^n$.

(a) Show that $\mathbb{H}_k$ is a vector space with the monomial basis (8.22). (*Hint*: Recall that a vector space is closed under the operations of addition and scalar multiplication.)

(b) Show that the dimension of $\mathbb{H}_k$ is

$$\dim(\mathbb{H}_k) = \begin{pmatrix} k+n-1 \\ n-1 \end{pmatrix} = \frac{(k+n-1)!}{(n-1)!k!}.$$

(*Hint*: Recall that the binomial coefficient $\binom{m}{n}$ is "m choose n"—the number of ways of putting n identical balls into m boxes.)

(c) What is the dimension of $\mathbb{H}_k^n$?

6. Show that the homological operator L_A (8.28) is a linear operator on $\mathbb{H}_k^n$.

7. One possible inner product on $\mathbb{H}_k^n$ is a generalization of the Frobenius inner product on matrices. If $p(x) \in \mathbb{H}_k^n$, let $p(\partial)$ be the differential operator with each x_i replaced by $\partial/\partial x_i$. We define

$$\langle p, q \rangle = p(\partial) \cdot q(x)|_{x=0}\,. \tag{8.95}$$

(a) Compute the inner product $\left\langle p_{m,i}, p_{\hat{m},j} \right\rangle$ of two vector monomials (8.24). In particular show that the inner product vanishes unless $m = \hat{m}$ and $i = j$.

(b) Compute the inner product of two, degree-one vector fields $p = Ax$, $q = Bx$, and show that the result is the Frobenius inner product of the matrices, $\langle A, B \rangle_F = \operatorname{tr}(AB^T)$.

(c) Show that (8.95) is indeed an inner product, that is, that $\langle p, q \rangle = \langle q, p \rangle$ and $\langle p, p \rangle > 0$ unless $p = 0$.

(d) Show that the adjoint of the homological operator (8.28) with this inner product is $L_A^\dagger = L_{A^\dagger}$, where since A is real, $A^\dagger = A^T$. Thus one choice for the complement $\mathbb{G}$ of $\operatorname{rng}(L_A)$ is $\ker(L_{A^\dagger})$.

8. Verify the calculations leading to the normal form (8.43) of the center in $\mathbb{R}^2$. In particular derive the homological operator L_A, find its action on the standard bases of $\mathbb{H}_2^2$ and $\mathbb{H}_3^2$, and obtain the matrices L. Find the eigenvectors and eigenvalues. Show that the null, left eigenvectors in $\mathbb{H}_3^2$ are given by (8.42) but that the null, right eigenvectors, together with range of $L_A(\mathbb{H}_3^2)$, do indeed span $\mathbb{H}_3^2$.

9. Find a versal unfolding for the following system:

$$\begin{aligned}\dot{x} &= xy,\\ \dot{y} &= -y - x^2.\end{aligned}$$

Sketch the various types of phase portraits that are possible for nearby vector fields. (*Hint*: It may be helpful to find the center manifold for the degenerate system.)

10. Consider the following system:

$$\begin{aligned}\dot{x} &= \lambda x - x^2 + 2xy,\\ \dot{y} &= (\lambda - 1)y + x^2.\end{aligned}$$

(a) Verify that this system has the normal form (8.40) and satisfies the singularity and nondegeneracy conditions for a saddle-node bifurcation at $(x,y) = (0,0)$ when $\lambda = 0$.

(b) Using Theorem 8.15, compute the bifurcation function $F(x, y(x;\lambda);\lambda)$ and find the first two terms in the series expansion for its extremal value $m(\lambda)$ near $\lambda = 0$. What does this tell you about bifurcations of this point?

(c) Is the bifurcation in (b) a saddle-node bifurcation? If not, how could you change the parameter dependence to fix it?

(d) Analyze all of the fixed points and their stability as a function of λ.

11. Consider the system

$$\begin{aligned}\dot{x} &= x + 2y,\\ \dot{y} &= -x - y + xy.\end{aligned}$$

(a) Find the linear transformation $(x,y)^T = P(\xi,\eta)^T$ that transforms the linear part of this system into the real normal form $J = \left(\begin{smallmatrix}\mu & \omega\\ -\omega & \mu\end{smallmatrix}\right)$. (*Hint*: Recall §2.5—use the real and imaginary parts of the eigenvectors $v_\pm = u \pm iw$.)

(b) Transform the full system to the new coordinates (ξ,η).

(c) Use complex coordinates (8.53), and rewrite this as a system for $(z,\bar{z})$.

12. Consider the system

$$\begin{aligned}\dot{x} &= \mu x - y + (x^2+y^2)^2(\alpha x - \beta y),\\ \dot{y} &= x + \mu y + (x^2+y^2)^2(\alpha y + \beta y).\end{aligned}$$

Transform it into complex coordinates using (8.53) and show that this system is in the normal form (8.56). Show that when $\alpha \neq 0$, it has a degenerate Andronov–Hopf bifurcation at $\mu = 0$. Determine whether it is subcritical or supercritical.

13. Consider the system

$$\begin{aligned}\dot{x} &= \mu x - y + ay^2 + x^3,\\ \dot{y} &= x + \mu y + xy^2 + y^2.\end{aligned}$$

(a) Determine $\alpha(a)$ using (8.59).

(b) Find the set on which this system has an Andronov–Hopf bifurcation. Is the bifurcation subcritical or supercritical?

(c) Investigate numerically the behavior for values of a such that $\alpha > 0$, $\alpha < 0$, and $\alpha = 0$.

14. Show that the three-species food-chain model (1.11) has an invariant plane when the top predator, P, is extinct and that the carrying capacity K can be eliminated by scaling the resource population so that the model reduces to

$$\dot{R} = R(1-R) - x_c y_c \frac{CR}{R+R_o},$$
$$\dot{C} = -x_c C\left(1 - y_c \frac{R}{R+R_o}\right),$$

where all of the new parameters are assumed positive.

(a) Show that this system has an equilibrium in the physically relevant positive quadrant for certain ranges of the parameters y_c and R_o.

(b) Show that this equilibrium is a center when $R_o = \frac{y_c - 1}{y_c + 1}$ and that it satisfies the transversality requirement for an Andronov–Hopf bifurcation.

(c) Numerically investigate the Andronov–Hopf bifurcation for the parameters $x_c = 0.4$, $y_c = 2$, as R_o varies. Is it subcritical or supercritical?

15. Show that the following systems have an equilibrium that undergoes an Andronov–Hopf bifurcation for some parameter value μ. Find the bifurcation point and determine whether the bifurcation is subcritical or supercritical.

(a) $\begin{aligned} \dot{x} &= 1 - (\mu+1)x + x^2 y, \\ \dot{y} &= \mu x - x^2 y, \end{aligned}$ (the Brusselator)

(b) $\ddot{x} + \dot{x}^3 - 2\mu\dot{x} + x = 0$ (Rayleigh's oscillator)

(c) $\begin{aligned} \dot{x} &= y, \\ \dot{y} &= -x + \mu y + x^2 + xy + y^2 \end{aligned}$ (Bautin's model).

16. Two predators, with populations p_1 and p_2, hunt the same prey species, with population s. Using a saturating (recall (1.11) and Exercise 13) nonlinearity for the predators hunting efficiency, Farkas (1984) modeled this system by

$$\dot{s} = s\left[r\left(1 - \frac{s}{K}\right) - m_1 \frac{p_1}{a_1 + s} - m_2 \frac{p_2}{a_2 + s}\right],$$
$$\dot{p}_1 = p_1\left(m_1 \frac{s}{a_1 + s} - d_1\right),$$
$$\dot{p}_2 = p_2\left(m_2 \frac{s}{a_2 + s} - d_2\right).$$

As usual, all parameters are positive, and the biologically relevant phase space is the positive octant.

(a) By rescaling the dynamical variables and time, show that we can effectively set $K = r = 1$. Thus the effective parameter space is six-dimensional and labeled by $\mu = (a_1, a_2, m_1, m_2, d_1, d_2)$.

(b) Show that (in the newly scaled model), the ith predator necessarily goes extinct if $m_i < d_i$. We will assume that both predators can grow. Moreover, we will assume that predator 1 is relatively more a "K-strategist"; i.e., its efficiency saturates earlier, $a_1 < a_2$, while predator 2 is relatively more of an "r-strategist," that is, it has a higher relative maximal birthrate: $b_1 = m_1/d_1 < b_2 = m_2/d_2$.

(c) Assume that the critical prey populations, s^*, for the growth of each predator are equal (i.e., $\dot{p}_i > 0$ if $s > s^*$). Show that this implies that $a_1(b_2 - 1) = a_2(b_1 - 1)$ and does not contradict the assumptions in (b) on the predators' different strategies.

(d) There are two isolated equilibria and one line segment of equilibria for the model. Find them.

(e) Show that as s^* ranges over the interval $[\frac{1}{b_1+1}, \frac{1}{b_2+1}]$ the equilibria on the line segment successively undergo Andronov–Hopf bifurcations. Farkas calls this sequence of bifurcations a *zip bifurcation.*

(f) Investigate the dynamics near the zip bifurcation numerically.

17. Complete the proof of Theorem 8.24.

(a) Suppose that $f(x;\mu)$ is given by (8.63). Carry out the transformations leading to (8.65).

(b) Prove there is a unique solution $m_2(z;m_1)$ to $D_z F(z;m_1,m_2) = 0$ in a neighborhood N of the origin in $\mathbb{R} \times \mathbb{R}$.

(c) Show that there is a unique solution $m_1(z)$ to

$$G(z;m_1) = F(z;m_1,m_2(z;m_1)) = 0$$

in a neighborhood of the origin in $\mathbb{R}$.

(d) Show that the curve $(m_1(z), m_2(z;m_1(z)))$ is, to lowest order, Neile's parabola $27m_1^2 = -4sm_2^3$. Transforming back to (A,B,D), show that this becomes (8.68).

18. Continue the normal form transformation in §8.10 for the Takens–Bogdanov case to cubic order by finding the range and cokernel of the homological operator L_J on the eight cubic monomials in $\mathbb{H}_3^2$. Show that the normal form can be written

$$\begin{aligned}\dot{x} &= y,\\ \dot{y} &= dx^2 + exy + fx^3 + gx^2y.\end{aligned}$$

19. Taken's choice for the normal form for the Takens–Bogdanov bifurcation is

$$\begin{aligned}\dot{x} &= \nu_1 x + y + x^2,\\ \dot{y} &= \nu_2 + x^2.\end{aligned}$$

Study the phase portraits for this system as the parameters (ν_1,ν_2) vary, and show that the bifurcation diagram is equivalent to that of the normal form (8.69) under a transformation $\nu(\mu)$.

20. Consider the Hamiltonian system defined by

$$H(x,y) = \tfrac{1}{2}y^2 + \mu x + \tfrac{1}{3}x^3.$$

Write down the ODEs, find the equilibria, and demonstrate how the phase portraits change as μ varies. Identify the bifurcations, considering especially the behavior of the orbit homoclinic to the saddle equilibrium.

21. Consider the system

$$\begin{aligned}\dot{x} &= y + \varepsilon x,\\ \dot{y} &= x - xy - x^3.\end{aligned}$$

(a) Find the fixed points and characterize their stability.

(b) Show that when $\varepsilon = 0$, this system is time reversible but not Hamiltonian (recall §6.5).

(c) Prove that when $\varepsilon = 0$, the origin has a homoclinic orbit. (*Hint*: The nullclines $\dot{y} = 0$ and $\dot{x} = 0$ divide the plane into sectors with distinct directions. Use time reversal symmetry.)

(d) Study the system numerically for small ε. Is there a homoclinic bifurcation?

22. Find a formula that generalizes (8.89) for the Melnikov function $M(\theta)$ when $\mathrm{tr}(Df) \neq 0$.

23. Why is the Melnikov function (8.89) independent of t?

24. Study the bifurcations of your three-dimensional quadratic system of equations from Table 1.1 as you vary the reduced parameters.

Chapter 9

Hamiltonian Dynamics

> *The laws which we have explained abundantly serve to account for all the motions of the celestial bodies, and of our sea.* (Isaac Newton, *Principia Mathematica*, 1687)

In the earlier chapters we primarily studied dynamical systems without assuming any special structure. However, many physical systems do have a "geometric" structure, and this should be acknowledged by the model builder and safeguarded by the dynamicist. For example, a flow that conserves energy must lie on the surfaces defined by constant energy and is therefore geometrically restricted. In this chapter we will consider several of these special classes of dynamical systems. Perhaps the most useful are *Hamiltonian* systems, as virtually all the *fundamental* models in physics are described by such dynamics. We do not have the time to develop a complete course on the physics of these systems but will predominantly concentrate on their mathematical structure.[58]

9.1 ▪ Conservative Dynamics

A system $\dot{x} = f(x)$ is said to be conservative when there is a quantity that is invariant along the flow, i.e., if there is a function $I(x)$ such that[59]

$$0 = \frac{d}{dt} I(x) = DI \frac{dx}{dt} = \nabla I \cdot f. \tag{9.1}$$

Thus I is an *invariant* if it has a gradient everywhere normal to the vector field f. Invariants are also called *integrals* of motion, or *conserved quantities.* If the equations model a physical system, then the conserved quantity often has physical significance; for example, it could be the total momentum in a system of interacting particles (see Exercise 1). However, in some cases, the physical meaning of the invariant is obscure.

Example 9.1 (Lotka–Volterra dynamics). Sometimes the Lotka–Volterra systems of

[58] References include (Abraham and Marsden 1978; Arnold 1978; MacKay and Meiss 1987; Meyer and Hall 1992).

[59] As always, DI stands for the collection of first derivatives of the function I. We distinguish this from ∇I, which is the column vector formed from the first derivatives. Typically $DI = \nabla I^T$ is a row vector.

§1.4 have invariants. For example, the system (recall Exercise 6.4)

$$\begin{aligned}\dot{x} &= -dx(1-y),\\ \dot{y} &= by(1-x)\end{aligned}$$

represents the interaction between predators, whose population is x and isolated death rate is $d > 0$, with prey, whose population is y and isolated birthrate is $b > 0$. The populations have been normalized so that the carrying capacities are 1 for both x and y. This system has equilibria at $(0,0)$ and $(1,1)$. The origin is a saddle; the x-axis is its stable manifold and the y-axis is its unstable manifold. Thus the positive quadrant is an invariant set. It is easy to see that $(1,1)$ is a linear center with eigenvalues $\lambda = \pm i\sqrt{bd}$. To show that it is nonlinearly stable it would be desirable to construct a Lyapunov function; in fact, this system has an invariant. To see this we study the phase curve equation (1.22)

$$\frac{dy}{dx} = -\frac{b}{d}\frac{y(1-x)}{x(1-y)} \quad \Rightarrow \quad b\int \frac{1-x}{x}dx = -d\int \frac{1-y}{y}dy.$$

Integration of the separated equation gives a constant of integration that can be interpreted as an invariant function:

$$I(x,y) = b\,(x - \log x) + d\,(y - \log y).$$

It is easy to verify explicitly that I is constant along the trajectories; of course, it is also a weak Lyapunov function; recall §4.6. Expanding the function I near the point $(1,1)$ yields

$$I(x,y) = b + d + \frac{b}{2}(x-1)^2 + \frac{d}{2}(y-1)^2 + O(3)$$

so that the contours of I are ellipses near $(1,1)$. This implies that $(1,1)$ is a topological center; recall §6.3. ∎

Example 9.2 (wave–wave interactions). Propagating waves are typically represented by functions of the form $a(t)e^{ik\cdot x}$. Here $a(t)$ is the complex Fourier amplitude of the wave with wave vector k. In the linear approximation the amplitudes undergo a pure oscillation with a frequency ω: $a(t) = a(0)e^{i\omega t}$; often the frequency is a function of the wave vector—this function is called the dispersion relation, $\omega = \Omega(k)$. If the medium in which the waves are propagating is nonlinear, then waves with distinct wavenumbers interact; the lowest-order nonlinear terms couple waves in triplets that satisfy $k_3 = k_1 + k_2$ giving rise to *triad* interactions. For example, a single triad with amplitudes a_1, a_2, and a_3 obeys the equations

$$\begin{aligned}\dot{a}_1 &= -i\omega_1 a_1 + ic\bar{a}_2 a_3,\\ \dot{a}_2 &= -i\omega_2 a_2 + ic\bar{a}_1 a_3,\\ \dot{a}_3 &= -i\omega_3 a_3 + ica_1 a_2,\end{aligned} \tag{9.2}$$

where c is a real coupling constant. These equations describe, for example, water waves in a fluid with density gradients such as an ocean, the interaction of phonons in solids, as well as the nonlinear interaction of various plasma waves (Davidson 1972).

The system (9.2) has several invariant quantities. Defining the wave actions to be $J_i = \bar{a}_i a_i = |a_i|^2$, it is easy to see that this system has two invariants:

$$I_1 = J_1 + J_3, \quad I_2 = J_2 + J_3. \tag{9.3}$$

For example,

$$\frac{d}{dt}I_2 = \bar{a}_2(-i\omega_2 a_2 + ic\bar{a}_1 a_3) + a_2(i\omega_2\bar{a}_2 - ica_1\bar{a}_3)$$
$$+\bar{a}_3(-i\omega_3 a_3 + ica_1 a_2) + a_3(i\omega_3\bar{a}_3 - ic\bar{a}_1\bar{a}_2)$$
$$= ic(\bar{a}_2\bar{a}_1 a_3 - a_2 a_1\bar{a}_3 + \bar{a}_3 a_1 a_2 - a_3\bar{a}_1\bar{a}_2) = 0.$$

These invariants imply that for the amplitude of the first two waves to grow, the amplitude of the third one must decay. This interaction gives rise to the so-called "decay instability" of wave three into waves one and two. It is investigated further in Exercise 6. ∎

In general, the construction of an invariant requires the solution of the first-order partial differential equation (PDE) (9.1). In principle this could be done by the method of characteristics (Guenther and Lee 1996); however, the characteristic equations are $\dot{x} = f$, precisely the original ordinary differential equation (ODE)! Thus if one has no physical insight into a particular system (for example, knowledge of a symmetry—see §9.7), then it is difficult to find invariants. There are methods, based on Lie groups, that allow one to discover nonobvious symmetries and their associated invariants (Hydon 2000; Olver 1993).

9.2 ▪ Volume-Preserving Flows

Suppose $\varphi_t(x)$ is a flow with vector field $f(x)$ and let U_o represent a set of initial conditions. Each $x_o \in U_o$ moves to a point $x(t) = \varphi_t(x_o)$ under the flow, so that U_o is transformed into a domain $U_t = \varphi_t(U_o)$ at time t; see Figure 9.1. The theorem of calculus for transformation of volume integrals implies that the volume of U_t is

$$\mathrm{vol}(U_t) = \int_{U_t} dx = \int_{U_o} \det(D\varphi_t(x))dx. \tag{9.4}$$

The rate of change of this volume is easily computed from Abel's theorem, (2.50), upon recalling that $D_x\varphi_t(x) = \Phi(t;x)$ is the fundamental matrix of the linearization (7.7). Thus

$$\mathrm{vol}(U_t) = \int_{U_o} \exp\left(\int_0^t \mathrm{tr}Df(x(s))\,ds\right)dx,$$

so that

$$\frac{d}{dt}\mathrm{vol}(U_t) = \int_{U_o} \mathrm{tr}(Df(x(t)))\det(D\varphi_t(x))\,dx = \int_{U_t} \mathrm{tr}(Df(x))dx. \tag{9.5}$$

One immediate conclusion of (9.5) is that the rate of change of the volume vanishes for any region U_o if and only if $\mathrm{tr}(Df) \equiv 0$. Such flows are called *volume preserving*. Volume-preserving flows commonly occur in applications.

Example 9.3 (passive tracers). One common example of a volume-preserving flow arises in fluid mechanics; recall §1.4. Suppose that $v(x,t)$ is the *Eulerian* velocity field of a fluid; i.e., $v(x,t)$ is the velocity of the fluid seen by an observer at a fixed point x in space at time t. A drop of dye put in the fluid at x will then be swept along with

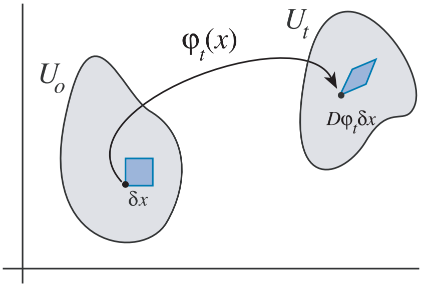

Figure 9.1. *Volume-preserving flow.*

the fluid—to the extent that the inertia of the dye is the same as that of the fluid—and thus obey the *Lagrangian* ODE (1.15),

$$\dot{x} = v(x,t). \tag{9.6}$$

The solution $x(t) = \varphi_t(x_o)$ of this system gives the position of an observer moving with the flow.

When $\nabla \cdot v = 0$ the fluid is said to be incompressible. This is a good approximation when the fluid velocity is far below the speed of sound (subsonic). Note that for (9.6),

$$\mathrm{tr}(Df) = \sum_{i=1}^{3} \frac{\partial}{\partial x_i} v_i = \nabla \cdot v = 0,$$

so that the flow of a passive tracer in an incompressible fluid is volume preserving. One example of this is the ABC flow of §1.4. Two-dimensional fluid flows, like the flow within a soap film, give another class of simple examples. An incompressible vector field in two dimensions can be written as the curl of a *stream function* $\psi(x,y,t)$,

$$v = \hat{z} \times \nabla\psi = \left(-\frac{\partial\psi}{\partial y}, \frac{\partial\psi}{\partial x}\right),$$

so that $\nabla \cdot v = -\frac{\partial^2\psi}{\partial x \partial y} + \frac{\partial^2\psi}{\partial y \partial x} = 0$. Consequently, this system is also Hamiltonian (see §9.3) with $H(x,y,t) = -\psi(x,y,t)$. ∎

Hamiltonian flows are examples of volume-preserving flows—but not every volume-preserving flow is Hamiltonian. Indeed, as we will soon see, Hamiltonian systems have additional geometrical structure.

9.3 ▪ Hamiltonian Systems

William Rowan Hamilton conceived in 1834 what is now called *Hamiltonian* dynamics as a reformulation of Newton's equation $F = ma$ for a set of point particles in a force field (Hamilton 1834). When the force, F, is conservative, it can be written as the gradient of a *potential energy* function, V; by convention $F = -\nabla V$, so that the force is in the direction of decreasing potential energy. Using the standard technique to convert second-order differential equations into a system of first-order equations (recall §1.2) gives

$$\frac{dx_i}{dt} = v_i, \quad m_i \frac{dv_i}{dt} = -\nabla_i V. \tag{9.7}$$

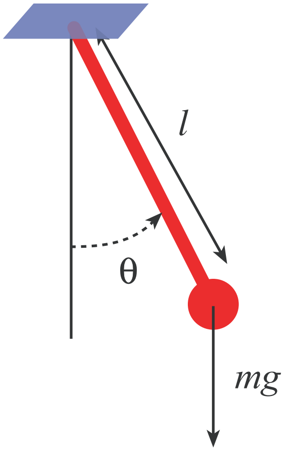

Figure 9.2. *Planar pendulum.*

Equations of the form (9.7) also hold for collections of interacting particles or mechanical components as in (1.12). Denote the collection of position variables of all the components by the single vector q; these are the *configuration* variables. Similarly p denotes the collection of the kinetic momenta, $p_i = m_i v_i$. Hamilton noticed that these equations could be obtained through differentiation of a scalar quantity that he called the characteristic function,

$$H(q,p) = \sum_{i=1}^{n} \frac{p_i^2}{2m_i} + V(q). \tag{9.8}$$

Today H is called the *Hamiltonian*. By direct comparison of (9.7) with (9.8), we see that the equations of motion are

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \quad \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}. \tag{9.9}$$

Physically, H is the total energy of the system; it is the sum of *kinetic*, $T = \sum p_i^2/2m_i$, and *potential*, $V(q)$, energies. The total energy is an invariant of the system, by (9.1),

$$\frac{dH}{dt} = \sum_{i=1}^{n} \frac{\partial H}{\partial q_i}\frac{dq_i}{dt} + \frac{\partial H}{\partial p_i}\frac{dp_i}{dt} = \sum_{i=1}^{n} \frac{\partial H}{\partial q_i}\frac{\partial H}{\partial p_i} - \frac{\partial H}{\partial p_i}\frac{\partial H}{\partial q_i} = 0. \tag{9.10}$$

This generalizes the calculation of (4.28) for one configuration and momentum variable.

Example 9.4 (The Planar Pendulum). Newton's equations for a planar pendulum of length l in a gravitational field of strength g, as shown in Figure 9.2, are

$$ml\ddot{\theta} = -mg\sin(\theta) = -\frac{\partial}{\partial\theta}(-mg\cos(\theta)),$$

where θ is the angle measured from the bottom. The angular momentum is $p = ml^2\dot{\theta} = I\dot{\theta}$, where I is the moment of inertia. The Hamiltonian is $H(\theta,p) = \frac{p^2}{2I} -$

$mgl\cos\theta$, and Hamilton's equations are

$$\begin{aligned}\dot{\theta} &= \frac{p}{I},\\ \dot{p} &= -mgl\sin\theta.\end{aligned} \tag{9.11}$$

Since the energy of the pendulum is conserved, the motion is along contours of $H = E$ in the phase space (θ, p). ∎

The Hamiltonian formulation is not limited to functions having the form "kinetic plus potential" energies. More generally a Hamiltonian is any smooth (C^1) function $H : M \to \mathbb{R}$, where M is a $2n$-dimensional manifold with coordinates $z = (q, p)$. We call M the phase space, q the configuration, and p the *canonical momenta*; each is n-dimensional. Such a Hamiltonian is said to have *n degrees of freedom*, even though it defines motion on a $2n$-dimensional manifold. Thus the planar pendulum has one degree of freedom.

A Hamiltonian can also depend explicitly on time, $H : M \times \mathbb{R} \to \mathbb{R}$. It then defines a system of differential equations for $(q, p, t) \in M \times \mathbb{R}$ in the *extended phase space*. In this case the flow is nonautonomous, but the equations are still given by (9.9). It is conventional to say that $H(q, p, t)$ has $n + 1/2$ degrees of freedom, since the time variable effectively increases the dimension of the phase space by one.

Example 9.5. One physical situation in which a more general Hamiltonian function arises corresponds to the motion of a charged particle in an electromagnetic field. In the nonrelativistic limit $v \ll c$ (where c is the speed of light), a point particle with charge e obeys the Lorentz force law:

$$\dot{x} = v, \quad m\dot{v} = e\left(E + \frac{v}{c} \times B\right). \tag{9.12}$$

This equation is not of the form (9.7) unless the magnetic field B vanishes. However, the Lorentz law does arise from a Hamiltonian system. Since the magnetic field is source free, $\nabla \cdot B = 0$, it can be written as the curl of a vector potential A: $B(x, t) = \nabla \times A(x, t)$. The canonical momentum conjugate to the particle position x is defined to be

$$p = mv + \frac{e}{c}A,$$

and the corresponding Hamiltonian is still the total energy of the system, $H = 1/2 mv^2 + e\phi$, where ϕ is the scalar potential, and $E = -\nabla\phi - \frac{1}{c}\frac{\partial A}{\partial t}$. So that (9.9) applies, the Hamiltonian must be written in its canonical coordinates:

$$H(q, p, t) = \frac{1}{2m}\left|p - \frac{e}{c}A(q, t)\right|^2 + e\phi(q, t). \tag{9.13}$$

It is an exercise in vector identities to show that the Hamiltonian equations for (9.13) reproduce the Lorentz force; see Exercise 3.

If the product in (9.13) is expanded, then H has a term that looks like the conventional kinetic energy $p^2/2m$ (though it does not have that interpretation physically). The remaining terms include the factor $p \cdot A$ that depends linearly upon the momentum and yields the Lorenz force. ∎

It is often convenient to write Hamilton's equations (9.9) in a more compact, matrix form

$$\frac{dz}{dt} = J\nabla H, \quad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}. \tag{9.14}$$

Here $z = (q,p)^T$ represents a point in phase space; J, the *Poisson matrix*, is the $2n \times 2n$ antisymmetric matrix shown, and I is the $n \times n$ identity matrix.[60] Note that J is nondegenerate; indeed, $\det(J) = 1$. Moreover, $J^2 = -I$ (here I is the $2n \times 2n$ identity), so that $J^{-1} = -J$.

The time rate of change of any scalar function $F \in C^1(M \times \mathbb{R}, \mathbb{R})$ on the extended phase space can be computed from (9.14) by using the chain rule $\frac{dF}{dt} = \frac{\partial F}{\partial t} + \frac{\partial F}{\partial z}\dot{z}$. This can be compactly written

$$\frac{dF}{dt} = \frac{\partial F}{\partial t} + \{F, H\}.$$

Here the expression $\{F,H\}$ is called the (canonical) *Poisson bracket*, defined as

$$\{F,H\} \equiv \nabla F^T J \nabla H = \sum_{i=1}^{n} \left(\frac{\partial F}{\partial q_i}\frac{\partial H}{\partial p_i} - \frac{\partial F}{\partial p_i}\frac{\partial H}{\partial q_i} \right). \tag{9.15}$$

For example, the equations of motion (9.14) can be rewritten in Poisson bracket form

$$\dot{z} = \{z, H\}. \tag{9.16}$$

The result (9.10) can now be generalized to show that any time-independent Hamiltonian is *conservative*.

Lemma 9.6 (conservation of energy). *If H is time independent, then energy is preserved along trajectories: $H(q(t), p(t)) = E$.*

Proof. $dH/dt = \{H,H\} = \nabla H^T J \nabla H = 0$ because J is antisymmetric. □

For example, if the system has one degree of freedom, then the motion is along the curves defined by the contours of H. Since these contours determine the phase portrait, motion in a one-degree-of-freedom Hamiltonian is not very interesting.[61] As we will see below, the motion of 1.5- and two-degree-of-freedom Hamiltonian systems can be much more complicated.

Joseph Liouville showed that Hamiltonian flows preserve volume.

Lemma 9.7 (Liouville). *If H is C^2, then its flow is volume preserving.*

Proof. According to (9.5) we need to show that $\mathrm{tr}(Df) = 0$. Here

$$f = (\partial H/\partial p, -\partial H/\partial q).$$

Thus

$$\mathrm{tr}(Df) = \sum_{i=1}^{n} \frac{\partial}{\partial q_i}\frac{\partial H}{\partial p_i} - \frac{\partial}{\partial p_i}\frac{\partial H}{\partial q_i} = 0,$$

[60] The Poisson matrix Jshould be distinguished from another matrix, ω, the symplectic form, though these two are sometimes identified. Take care to note that various authors use different sign conventions!

[61] However, it does give a good algorithm for plotting the contours of a function!

since the partial derivatives commute. Note that this is valid for the nonautonomous case as well. □

Another simple consequence of the Hamiltonian form of the equations is that fixed points are equivalent to critical points of H.

Lemma 9.8 (equilibria). *A point z^* is an equilibrium point of an autonomous Hamiltonian flow if and only if it is a critical point of H.*

Proof. So that $\dot{z} = 0$ in (9.14), we must have $\nabla H = 0$ because the matrix J is nondegenerate. □

Example 9.9. The equilibria of the planar pendulum (9.11) are points where $\partial H/\partial\theta = mg\sin\theta = 0$ and $\partial H/\partial p = p/I = 0$. Thus the equilibria are $(n\pi, 0)$ for any integer n. ■

Since equilibria are critical points of H, their stability can be determined by examination of the Hessian matrix of H. We will do this in §9.10. Meanwhile, a simple implication of Lemma 9.8 is the following.

Lemma 9.10. *A nondegenerate minimum or maximum point of an autonomous Hamiltonian H is a Lyapunov stable equilibrium (recall* §4.5*).*

This follows because, near such a point, the contours $H = E$ are topological spheres enclosing the equilibrium.

9.4 ▪ Poisson Dynamics

While Hamiltonian dynamics provides a useful description of particles interacting through mechanical, electromagnetic, and gravitational forces, there are some systems that do not fit this form. *Poisson systems* are one such generalization. These are defined on a smooth manifold M with any dimension d—in particular d need not be even. Given a generalized Hamiltonian function $H : M \to \mathbb{R}$, Poisson dynamics is defined by (9.16) as

$$\dot{z} = \{z, H\};$$

however, in this case $\{\ ,\ \}$ is not necessarily given by the canonical form (9.15) but is a generalized

> ▷ *Poisson bracket*: A Poisson bracket $\{\ ,\ \}$ is a bilinear operator on a pair of functions in $C^2(M,\mathbb{R})$ that is antisymmetric, is a derivation, and satisfies the Jacobi identity.

These terms are defined as follows. Let $F, G, H \in C^2(M,\mathbb{R})$.

- Antisymmetry: $\{F,G\} = -\{G,F\}$;
- Bilinearity: $\{F+G,H\} = \{F,H\} + \{G,H\}$ and $\{aF, bG\} = ab\,\{F,G\}$ for any constants a and b;
- Derivation: $\{FH,G\} = F\,\{H,G\} + H\,\{F,G\}$;
- Jacobi identity:

$$\{F,\{G,H\}\} + \{G,\{H,F\}\} + \{H,\{F,G\}\} = 0. \tag{9.17}$$

The derivation property is equivalent to the product rule for derivatives. Indeed, we have the following.

Lemma 9.11. *Suppose L is a linear operator on $C^1(\mathbb{R},\mathbb{R})$ that obeys the derivation property, $L(fg) = fL(g) + gL(f)$, and that $L(x) = 1$. Then $L = d/dx$.*

Thus, the derivation property gives an alternative way of defining the derivative. The proof of Lemma 9.11 is left to the reader; see Exercise 4.

It is relatively easy to verify that the standard Poisson bracket (9.15) satisfies each of these properties; see Exercise 5. Moreover, every Poisson bracket has a form similar to (9.15).

Lemma 9.12. *Suppose $\{\ ,\ \}$ is a Poisson bracket on $C^2(\mathbb{R}^d,\mathbb{R})$. Then there exists an antisymmetric matrix $J(z)$ such that*

$$\{F,G\} = \nabla F^T J(z) \nabla G. \tag{9.18}$$

Proof (Sketch: B). Bilinearity implies that the bracket must be linear in each slot. The main assertion is that the derivation property implies that the Poisson bracket acts as a first derivative on each of its arguments; this follows from Lemma 9.11 and antisymmetry. Moreover, antisymmetry implies that the coefficients of the derivatives must be an antisymmetric matrix. □

Note that, in general, $J(z)$ is a function of z and, unlike the standard Poisson matrix, it need not be nonsingular. The converse of Lemma 9.12 is not true: a bracket defined by (9.18) for a general, antisymmetric matrix $J(z)$ does satisfy the antisymmetry, bilinearity, and derivation properties; however, it does not necessarily satisfy the Jacobi identity.

Lemma 9.12 implies that the equations of motion for a Poisson system have the same form as (9.14), $\dot{z} = J\nabla H$. Moreover, the time derivative of any function of z can be obtained using

$$\frac{dF}{dt} = DF\dot{z} = \nabla F^T J \nabla H = \{F,H\}. \tag{9.19}$$

The Jacobi identity implies that the time derivative of the Poisson bracket obeys the expected relationship:

$$\frac{d}{dt}\{F,G\} = \{\{F,G\},H\} = \{F,\{G,H\}\} + \{G,\{H,F\}\} = \{F,\dot{G}\} + \{\dot{F},G\}.$$

Autonomous Poisson systems are always conservative. In particular, the energy, H, is a conserved quantity, since $\dot{H} = \{H,H\} = 0$ by antisymmetry.

Example 9.13 (rigid body dynamics). The Euler equations for a free rigid body are most easily expressed as a Poisson system. Let $\omega \in \mathbb{R}^3$ represent the angular velocities of the body in body-fixed coordinates such that the moment of inertia tensor, I, is diagonal. The coordinate axes are the principal axes of the body, and the diagonal components of the moment of inertia are I_i. If the angular momenta are denoted $L_i = I_i\omega_i$ then the kinetic energy is $H = \sum_{i=1}^{3} L_i^2/2I_i$. The Euler equations describe the evolution of the angular momenta:

$$\begin{aligned}
\frac{d}{dt}L_1 &= \left(\frac{1}{I_3}-\frac{1}{I_2}\right)L_2L_3 \\
\frac{d}{dt}L_2 &= \left(\frac{1}{I_1}-\frac{1}{I_3}\right)L_3L_1 \quad \Rightarrow \quad \frac{d}{dt}L = L\times\nabla H. \\
\frac{d}{dt}L_3 &= \left(\frac{1}{I_2}-\frac{1}{I_1}\right)L_1L_2
\end{aligned} \tag{9.20}$$

It is obvious that these three equations cannot be written in the canonical Hamiltonian form since the phase space is odd-dimensional. However, this system is a Poisson system, as we can see by defining the generalized Poisson matrix,

$$J=\begin{pmatrix} 0 & -L_3 & L_2 \\ L_3 & 0 & -L_1 \\ -L_2 & L_1 & 0 \end{pmatrix}. \tag{9.21}$$

It is easy to see that the Euler equations (9.20) become $\dot{L}=J\nabla H$. To show that this is a Poisson bracket requires the verification of the Jacobi identity; see Exercise 5. The dynamics of (9.20) is explored further in Exercise 7. ∎

Poisson dynamics has many applications in fluid and plasma physics (Morrison 1998).

9.5 ▪ The Action Principle

The basic equations of physics are all derivable from variational principles. For example, Fermat's principle asserts that light going from point a to point b takes the path of shortest time. A path in space is a smooth, parameterized curve $\gamma:[a,b]\to\mathbb{R}^3$ so that $\gamma(s)$ represents a point in space and such that $\gamma(a)$ and $\gamma(b)$ are given endpoints. The travel time from a to b along γ is then

$$T[\gamma]=\int_a^b \frac{1}{c}\left|\frac{d\gamma}{ds}\right| ds,$$

where c is the speed of light. In empty space c is a constant, and in Euclidean space the paths of shortest time are straight lines; the corresponding paths on other manifolds are "geodesics." That light travels along geodesics at a constant speed is one of the principles of general relativity.

The *stationary action principle*[62] is an extension of this idea to general Hamiltonian systems. Let γ be a path in phase space parameterized with the distinguished parameter time: $\gamma:[a,b]\to M$, i.e., $\gamma=\{(q(t),p(t)), a\le t\le b\}$. The *action* of γ is the line integral

$$S[\gamma]\equiv\int_\gamma p\,dq-H\,dt=\int_a^b\left(p(t)\frac{dq}{dt}-H(q(t),p(t),t)\right)dt. \tag{9.22}$$

[62] Sometimes this is called the minimum action or extremal action principle; however, any critical point corresponds to an orbit, and while this could possibly be an extremum, it could also be a saddle.

The action is a functional: for each curve γ, it gives a real number analogous to the travel time in Fermat's principle.

The stationary action principle asserts that the realized paths are those whose action is stationary with respect to variation of the path: nearby paths have the same action. The stationary paths can be obtained from (9.22) using the calculus of variations. The basic idea is that when γ is a stationary point of $S[\gamma]$, the action of a nearby path $\hat{\gamma}$ should be, to first order, the same as that of γ.

Recall that the critical points x of a function f are obtained by setting $\hat{x} = x + \delta x$ and expanding to obtain $f(x+\delta x) = f(x) + Df(x)\delta x + o(\delta x)$. Thus the *first variation* is $\delta f = Df(x)\delta x$. The critical points of f are points where δf is zero, i.e., where the Jacobian vanishes. For (9.22), we will set $\hat{\gamma} = \gamma + \delta\gamma$ and find that $S[\gamma+\delta\gamma] = S[\gamma] + \delta S + o(\delta\gamma)$. Upon demanding that the first variation $\delta S = 0$, we will find that

$$\delta S = \int_a^b \frac{\delta S}{\delta\gamma}(\gamma(t))\,\delta\gamma(t)dt,$$

where $\delta S/\delta\gamma$ is called the *Fréchet* or *functional derivative* of S.[63] Since the variation $\delta\gamma$ is arbitrary, under appropriate assumptions of continuity, δS will vanish only if the Fréchet derivative vanishes for each $t \in (a,b)$. This will convert the global variational statement into a set of local differential equations.

> ▷ *Action principle*: The curves of stationary action (9.22) are the Hamiltonian trajectories.

More formally, we have the next lemma.

Lemma 9.14. *Suppose that $H \in C^1(M,\mathbb{R})$ and $\gamma \in C^1([a,b],M)$ such that the configurations $q(a) = q_a$ and $q(b) = q_b$ are fixed. Then, if γ is a stationary point of the action* (9.22), *it satisfies Hamilton's equations* (9.9).

Proof. Let $\gamma = \{(q(t),p(t)), a \le t \le b\}$ be the original path and $\hat{\gamma} = \{(q(t)+\delta q(t), p(t)+\delta p(t)), a \le t \le b\}$, where δq and δp are smooth and formally "small." Since the configurations of γ are fixed at the endpoints, $\delta q(a) = \delta q(b) = 0$. Substitute the perturbed path into (9.22) and expand to find

$$S[\hat{\gamma}] = \int_a^b \left[(p+\delta p)\frac{d}{dt}(q+\delta q) - \left(H(q,p,t) + \frac{\partial H}{\partial q}\delta q + \frac{\partial H}{\partial p}\delta p + \cdots\right)\right]dt.$$

Rearranging these terms and keeping only those that are of the first order in the small quantities $(\delta q, \delta p)$ gives the first variation in S:

$$\delta S = \int_a^b \left[\delta p\left(\frac{dq}{dt} - \frac{\partial H}{\partial p}\right) + p\frac{d}{dt}\delta q - \frac{\partial H}{\partial q}\delta q\right]dt.$$

To isolate δq, integrate the term $p\delta\dot{q}$ by parts using $p\delta\dot{q} = \frac{d}{dt}(p\delta q) - \delta q\dot{p}$. The integral of the total derivative term vanishes since δq is zero at the endpoints:

$$\int_a^b \frac{d}{dt}(p\delta q)dt = p(b)\delta q(b) - p(a)\delta q(a) = 0.$$

[63] The Fréchet derivative is a special case of the *Gâteaux derivative* defined on more general spaces.

Consequently,

$$\delta S[\gamma]=\int_a^b\left[\delta p\left(\frac{dq}{dt}-\frac{\partial H}{\partial p}\right)-\delta q\left(\frac{dp}{dt}+\frac{\partial H}{\partial q}\right)\right]dt.$$

Since $S[\gamma]$ is stationary, δS must vanish. Since $(\delta q,\delta p)$ are arbitrary and independent functions, and the integrand is continuous, each of the parenthesized terms above must vanish at each time $t\in(a,b)$. Thus the path γ satisfies Hamilton's equations (9.9). □

It is no doubt profound that we did not need to assume that the variations of p at a and b vanish, but only those of q. On a prosaic level this happened because the only parts integration was with respect to q; there is, however, a more philosophical level, as discussed by (Lanczos 1962).

The fact that δS vanishes for solutions of Hamilton's equations means that these solutions are points of stationary action. The action need not be minimal and often is not even a local minimum.

Example 9.15 (billiards). An ideal billiard is a point particle with unit mass on a billiard table, defined to be a closed region $D\subset\mathbb{R}^2$. For simplicity, the boundary of the table, ∂D, is assumed to be C^1 and a perfect reflector: the cushions do not absorb any energy. Since the billiard is a point particle, there are no effects from its spin (no *english* or *follow*). The Hamiltonian of such a system is

$$H=\tfrac{1}{2}p^2+V(q),\text{ where }V(q)=\begin{cases}0, & q\in D,\\ \infty, & q\in\partial D.\end{cases}$$

When the particle is in the interior of the domain, the equations are trivial: $\dot{p}=0$. Thus the trajectory will consist of a sequence of straight-line segments connected on the boundary. We will show that action is stationary when the angle of incidence equals the angle of reflection at a collision with the boundary. Consider a path γ that consists of straight-line segments connecting a sequence q_i, $i=0,1,\ldots,n$, of points on ∂D; see Figure 9.3. Since the speed $|p|=|\dot{q}|$ is constant on the path, the integrand of the action reduces to $(p\cdot\dot{q}-H)dt={}^1\!/\!_2p^2dt={}^1\!/\!_2p\cdot dl$, where dl is the length increment along the path. The action of γ is then

$$S[\gamma]=S[q_o,\ldots,q_n]=\frac{1}{2}p\sum_{j=0}^{n-1}L(q_j,q_{j+1}),$$

where $L(q_j,q_{j+1})=\left|q_j-q_{j+1}\right|$ is the length of the straight segment from q_j to q_{j+1}. Thus the action is simply a constant multiple of the total path length along the broken curve γ. To find a stationary path, fix q_0 and q_n and vary the intermediate points q_j, $j=1,\ldots,n-1$. The first variation in the action is then

$$\delta S[q_0,q_1,\ldots,q_n]=\frac{1}{2}p\sum_{i=1}^{n-1}\frac{\partial}{\partial q_i}\big(L(q_{i-1},q_i)+L(q_i,q_{i+1})\big)\delta q_i.$$

Since the points q_i must lie on the boundary, their first variations must be tangent to the boundary: $\delta q_i\propto\hat{t}$ where $\hat{t}$ is the tangent to ∂D at q_i. Thus δS vanishes for arbitrary variations when

$$\hat{t}\cdot\frac{\partial S}{\partial q_i}=0\ \Rightarrow\hat{t}\cdot\frac{\partial}{\partial q_i}L(q_{i-1},q_i)+\hat{t}\cdot\frac{\partial}{\partial q_i}L(q_i,q_{i+1})=0. \qquad (9.23)$$

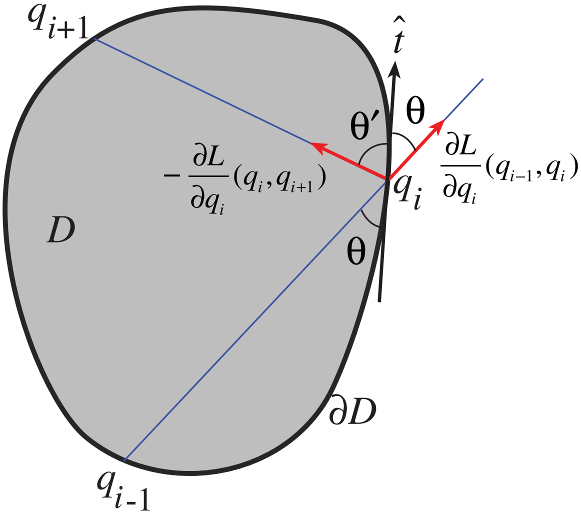
q_{i+1}
$\hat{t}$
θ'
θ
$-\frac{\partial L}{\partial q_i}(q_i, q_{i+1})$
$\frac{\partial L}{\partial q_i}(q_{i-1}, q_i)$
q_i
D
θ
∂D
q_{i-1}

Figure 9.3. *Orbits of the billiard defined by* (9.23).

Geometrically, the derivative

$$\frac{\partial}{\partial q'} L(q,q') = \frac{1}{L}\left((x'-x)\hat{x} + (y'-y)\hat{y}\right)$$

is the unit vector pointing along the segment from $q = (x,y)$ to $q' = (x',y')$. Thus the dot product with $\hat{t}$ is the cosine of the angle with the boundary. Therefore, stationary action implies the incoming angle is equal to the outgoing angle as shown in Figure 9.3.

The transformation $(q_{i-1}, q_i) \to (q_i, q_{i+1})$ implicitly defined by (9.23) is an example of a discrete dynamical system, or map, like the Poincaré map defined in §4.12. The Hamiltonian nature of billiard dynamics implies that this map is *symplectic* (Meiss 1992).

Note that any broken trajectory is certainly not a global minimum of the action; for example, the three-point trajectory $\gamma = [a, q_1, b]$ is certainly longer than the two-point trajectory $[a,b]$. ∎

9.6 ▪ Poincaré Invariant

One geometrical consequence of the action principle for Hamilton's equations is the existence of what is called the Poincaré invariant. This *invariant* is not an invariant *function* in the sense of §9.1, but is rather an action, the

▷ *loop action*: Let $\mathscr{L} : \mathbb{S}^1 \to M \times \mathbb{R}$ be any closed loop in the extended phase space; then the loop action is defined as

$$S(\mathscr{L}) = \oint_{\mathscr{L}} p\,dq - H\,dt. \qquad (9.24)$$

If the loop is parameterized as $\mathscr{L} = \{(q(s), p(s), t(s)) : s \in [0,1]\}$ with $q(1) = q(0)$, etc., then the integral (9.24) explicitly becomes

$$S(\mathscr{L}) = \int_0^1 \left(p(s)\frac{dq}{ds}(s) - H(q(s), p(s), t(s))\frac{dt}{ds}(s) \right) ds.$$

Poincaré discovered that for any $\mathscr{L}$, $S[\mathscr{L}]$ remains constant as $\mathscr{L}$ evolves with the Hamiltonian flow: we say that $S[\mathscr{L}]$ is an *integral invariant* for the flow. One implication is that Hamiltonian flows are volume preserving, though the preservation of the Poincaré invariant is much stronger than this.

Theorem 9.16 (Poincaré invariant). *The loop action is invariant under a Hamiltonian flow.*

Proof. We give a proof that relies upon the n-dimensional version of Stokes's theorem[64]—if you do not know this theorem, then the proof will seem clear only for the one-degree-of-freedom case. Denote the line element $dl = (dq, dp, dt)$, and define a vector $A = (p, 0, -H)$ in (q, p, t) coordinates. The action of a loop, $\mathscr{L}$, is the integral of the *one form* $A \cdot dl$ around the loop. Allow the loop to deform by letting each point on the loop move under the flow to form a two-dimensional tube $\mathscr{T}$ in the extended phase space; see Figure 9.4. Now consider *any* other loop $\mathscr{L}'$ on the tube $\mathscr{T}$ that is contractible to $\mathscr{L}$. Since these two loops bound a piece of the tube $\mathscr{T}$, Stokes's theorem implies that the difference between the loop actions can be written as a surface integral over this piece of $\mathscr{T}$,

$$S(\mathscr{L}) - S(\mathscr{L}') = \oint_{\mathscr{L}} A \cdot dl - \oint_{\mathscr{L}'} A \cdot dl = \oint_{\mathscr{T}} dA \wedge dl,$$

where $\wedge$ is the wedge product, which we will discuss in more detail below. In three dimensions, $dA \wedge dl = \nabla \times A \cdot ds^2$, where ds^2 is the surface element on $\mathscr{T}$ and the curl of A is

$$\nabla \times A = \left(-\frac{\partial H}{\partial p}, \frac{\partial H}{\partial q}, -1 \right) = -v,$$

the negative of the velocity vector of the Hamiltonian flow. Since v lies along the tube $\mathscr{T}$ and the surface element ds^2 is normal to the tube, $\nabla \times A \cdot ds^2 = 0$. Consequently the loop action is invariant along the flow.

In more dimensions, the calculation of the surface integrand relies upon exterior calculus to calculate the wedge product. Formally, we have

$$d(A \cdot dl) = dA \wedge dl = d\left(\sum_{i=1}^{n} p_i dq_i - H dt \right) = dp \wedge dq - dH \wedge dt.$$

Let (t, s) be coordinates on $\mathscr{T}$, where t, time, parameterizes the orbits and s is any transverse coordinate. Denoting the Hamiltonian flow by φ, we see that any point on

[64] Generally, Stokes's theorem states that the integral of an $(n-1)$ form α over the boundary of an n-dimensional surface is the integral of $d\alpha$ over the surface.

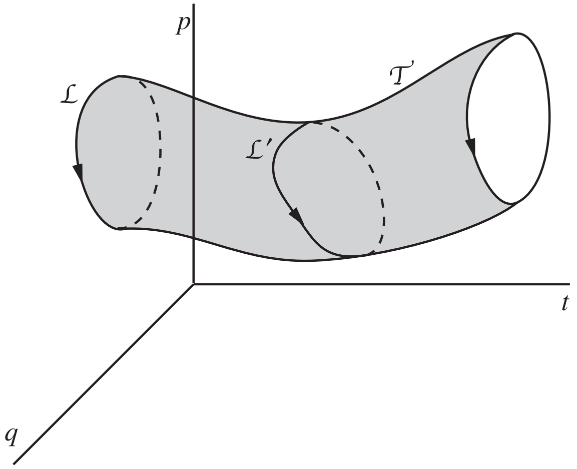

Figure 9.4. *Preservation of the loop action.*

$\mathscr{T}$ has the representation $(q(t,s), p(t,s)) = \varphi_t(q(0,s), p(0,s))$. Consequently,

$$\begin{aligned} dA \wedge dl &= dp \wedge dq - dH \wedge dt \\ &= \left[\sum_{i=1}^{n} \left(\frac{\partial p_i}{\partial t} \frac{\partial q_i}{\partial s} - \frac{\partial p_i}{\partial s} \frac{\partial q_i}{\partial t} \right) + \frac{\partial H}{\partial s} \right] dt \wedge ds \\ &= \left(\frac{\partial p}{\partial t} \frac{\partial q}{\partial s} - \frac{\partial p}{\partial s} \frac{\partial q}{\partial t} + \frac{\partial H}{\partial p} \frac{\partial p}{\partial s} + \frac{\partial H}{\partial q} \frac{\partial q}{\partial s} \right) dt \wedge ds. \end{aligned}$$

The term in the last set of parentheses is manifestly zero by Hamilton's equations (9.9). Thus, $dA \wedge dl = 0$ on the tube. □

The invariance of the loop action under a Hamiltonian flow will be used in §9.14 to construct a Poincaré section for a two-degree-of-freedom system. The loop action is also extensively used in the computation of fluxes in the theory of chaotic transport for Hamiltonian systems (Meiss 1992).

9.7 ▪ Lagrangian Systems

Historically, Hamiltonian dynamics arose from an earlier variational formulation due to Joseph-Louis Lagrange and Leonhard Euler. As we will see, the Lagrangian variational principle is equivalent to the Hamiltonian action principle when a special condition is satisfied—the Legendre condition. For complicated mechanical systems and especially when there are constraints, the Lagrangian is easier to obtain than the Hamiltonian because of coordinate independence.

The Lagrangian is a C^2 real valued function of three arguments: a set of configuration variables, x, in some configuration space M, the corresponding velocities, v, which are vectors in the tangent space T_xM, and time. The Lagrangian is denoted by

$$L : TM \times \mathbb{R} \to \mathbb{R}$$

or in coordinates by $L(x, v, t)$. Let $\gamma \in C^2([a,b], M)$ be a temporally parameterized curve in M. For a given Lagrangian, the action of γ is defined by

$$A[\gamma]=\int_a^b L(\gamma(t),\dot{\gamma}(t),t)dt. \tag{9.25}$$

The Lagrangian action A is a functional like the Hamiltonian action (9.22): it maps the space of curves to $\mathbb{R}$. In distinction to the Hamiltonian case, the curve γ in (9.25) is a curve in *configuration space only*, not phase space. This is also seen by the fact that the quantity $\dot{\gamma}$ is defined as the time derivative of γ, while in the Hamiltonian formulation, p is a variable independent of q.

To make matters more confusing, the Lagrangian analogue of the action variational principle is called *Hamilton's principle*!

> ▹ *Hamilton's principle*: The paths that are realized by the dynamical system represented by L are those for which the action (9.25) is stationary for fixed endpoints $\gamma(a)=x_o$ and $\gamma(b)=x_1$.

Stationary points of this action are computed using the calculus of variations as in §9.5. The action $A[\gamma]$ is stationary if it does not vary when the curve is slightly changed, $\gamma(t)\to\gamma(t)+\delta\gamma(t)$. The change in the action upon doing this can be formally expanded in $\delta\gamma$,

$$A[\gamma+\delta\gamma]-A[\gamma]=\int_a^b \frac{\delta A}{\delta\gamma}\delta\gamma(t)dt+o(\delta\gamma), \tag{9.26}$$

where $\delta A/\delta\gamma$ is the Fréchet derivative as before. To compute the Fréchet derivative, expand L in a Taylor series, and integrate by parts to isolate $\delta\gamma$:

$$\begin{aligned}A[\gamma+\delta\gamma]-A[\gamma]&=\int_a^b (L(\gamma+\delta\gamma,\dot{\gamma}+\delta\dot{\gamma},t)-L(\gamma,\dot{\gamma},t))dt\\&=\int_a^b \Big(\frac{\partial}{\partial x}L(\gamma,\dot{\gamma},t)\delta\gamma(t)+\frac{\partial}{\partial v}L(\gamma,\dot{\gamma},t)\delta\dot{\gamma}(t)+o(\delta\gamma)\Big)dt\\&=\int_a^b \delta\gamma(t)\Big(\frac{\partial}{\partial x}L(\gamma,\dot{\gamma},t)-\frac{d}{dt}\frac{\partial}{\partial v}L(\gamma,\dot{\gamma},t)\Big)dt\\&\qquad+\frac{\partial}{\partial v}L(\gamma,\dot{\gamma},t)\delta\gamma(t)\Big|_{t=a}^{t=b}+o(\delta\gamma).\end{aligned}$$

The boundary terms vanish because the endpoints are fixed: $\delta\gamma(a)=\delta\gamma(b)=0$. Consequently, the Fréchet derivative of the action is

$$\frac{\delta A}{\delta\gamma}=\frac{\partial}{\partial x}L(\gamma,\dot{\gamma},t)-\frac{d}{dt}\frac{\partial}{\partial v}L(\gamma,\dot{\gamma},t). \tag{9.27}$$

A sufficient condition[65] for the action to be stationary is that the Fréchet derivative (9.27) vanish, so that $\gamma=x(t)$ is a solution of the *Euler–Lagrange* differential equations:

$$\frac{d}{dt}\frac{\partial}{\partial v}L=\frac{\partial}{\partial x}L,\quad x(a)=x_o,\quad x(b)=x_1. \tag{9.28}$$

[65] This is not a necessary condition, as we will see in §9.8.

Note that this is a boundary value problem and not a traditional initial value problem. Thus, it is not obvious, nor even necessarily true, that a solution exists for every pair(x_o, x_1) and every time interval $[a,b]$. Moreover, (9.28) is typically a system of second-order ODEs and not, according to our usual definition, a dynamical system.

Nevertheless, we have so far demonstrated the following.

Lemma 9.17. *A C^2 curve γ that satisfies the Euler–Lagrange equations* (9.28) *is a stationary curve of the action* (9.25).

The converse of this result, discussed in §9.8, is not true without additional conditions on L.

Lagrangian mechanics includes all "frictionless" Newtonian mechanics. For example, the equations of motion for a system of point particles interacting under conservative forces can be written as

$$m_i \frac{d^2}{dt^2} x_i = -\frac{\partial}{\partial x_i} V(x),$$

where the coordinates of the particles are listed sequentially to give a vector $(x_1, x_2, \dots, x_n)$, m_i are the particle masses, and $V(x)$ is the potential energy. For example (x_1, x_2, x_3) may represent the (x, y, z) coordinates of the first particle, (x_4, x_5, x_6) those of the second, etc. It is easy to see that the Lagrangian for this system is

$$L(x, \dot{x}) = \frac{1}{2} \sum_{i=1}^{n} m_i \dot{x}_i^2 - V(x) = \frac{1}{2} \dot{x}^T \rho \dot{x} - V(x),$$

where $\rho = \operatorname{diag}(m_i)$ is the mass matrix. Hence the Lagrangian for a mechanical system is of the form "kinetic minus potential" energy.

Coordinate Independence of the Action

One of the nicest properties of the Lagrangian formulation is that it is independent of coordinates. This implies that the modeler is free to choose whatever coordinate system is most convenient.

Theorem 9.18. *Suppose $h : N \to M$ is a C^2 embedding and $L(x, \dot{x}, t)$ is a Lagrangian for $x \in M$. Then the dynamics of $y \in N$, where $x = h(y)$, is given by the Euler–Lagrange equations of the Lagrangian*

$$\tilde{L}(y, \dot{y}, t) = L(h(y), Dh(y)\dot{y}, t). \tag{9.29}$$

Proof. Consider the Euler–Lagrange equations for $\tilde{L}$:

$$\begin{aligned}
\frac{d}{dt}\frac{\partial}{\partial \dot{y}} \tilde{L}(y, \dot{y}, t) - \frac{\partial}{\partial y} \tilde{L}(y, \dot{y}, t) &= \frac{d}{dt}\frac{\partial}{\partial \dot{y}} L(h(y), Dh(y)\dot{y}, t) - \frac{\partial}{\partial y} L(h(y), Dh(y)\dot{y}, t) \\
&= \frac{d}{dt}\left(Dh^T(y) \frac{\partial}{\partial \dot{x}} L(x, \dot{x}, t)\right) - Dh^T(y) \frac{\partial}{\partial x} L(x, \dot{x}, t) - (D^2 h(y)\dot{y})^T \frac{\partial}{\partial \dot{x}} L(x, \dot{x}, t) \\
&= Dh^T(y)\left(\frac{d}{dt}\frac{\partial}{\partial \dot{x}} L(x, \dot{x}, t) - \frac{\partial}{\partial x} L(x, \dot{x}, t)\right).
\end{aligned}$$

Since $Dh(y)$ is nonsingular, the Euler–Lagrange equations for $\tilde{L}$ are satisfied precisely when those for L are satisfied. □

The meaning of Theorem 9.18 is this: to find the proper Lagrangian in a general coordinate system we can simply substitute for $x = h(y)$ and $\dot{x} = Dh(y)\dot{y}$ in L to get the new Lagrangian. This is often much easier than transforming the ODEs themselves.

Example 9.19. The dynamical equations for a system that is constrained are often quite difficult to write down. An example of this is the motion of a particle sliding on a surface, discussed in §3.3. There we had to compute the force that constrained the particle to the surface. Lagrangian mechanics allows us to completely avoid this difficulty. For example, consider a bead sliding on a wire defined parametrically by a curve $c : \mathbb{R} \to \mathbb{R}^3$. Suppose the bead has mass m and there is no friction, but there is an external force, for example, gravity, represented by potential energy $V(x)$. The Lagrangian for the system is $L = T - V$, where T is the kinetic energy of the bead $T = {}^1\!/\!{}_2 m\dot{x}^2$. One way to model this system is to write out Newton's equations for the three components of x and then impose the constraints that restrict the bead to follow the wire by introducing external forces.[66] It is much easier to use the parametric representation $c(s)$ for the curve and the one-dimensional coordinate s to describe the motion. The kinetic energy should be expressed as a function of s and $\dot{s}$; this can easily be done using the parametric form of the constraining curve, $c(s)$, for then $\dot{x} = Dc(s)\dot{s}$ and $T = {}^1\!/\!{}_2 m\,|Dc|^2\,\dot{s}^2$. According to Theorem 9.18, the new Lagrangian is $\tilde{L} = L(c(s), Dc(s)\dot{s}, t)$, or

$$\tilde{L}(s,\dot{s},t) = \tfrac{1}{2} m\,|Dc|^2\,\dot{s}^2 - V(c(s)). \tag{9.30}$$

For example, suppose the curve is a vertically aligned ellipse, $c = \{(x(s), 0, z(s)) = (a\sin s, 0, b\cos s) : s \in [0, 2\pi)\}$ and the external force is a constant gravity field, so that $V(x,y,z) = mgz$. Then $Dc = (a\cos s, 0, -b\sin s)^T$ and

$$\tilde{L}(s,\dot{s},t) = \tfrac{1}{2} m\left(a^2\cos^2 s + b^2\sin^2 s\right)\dot{s}^2 - mgb\cos s. \tag{9.31}$$

This gives the Euler–Lagrange equation

$$\begin{aligned} 0 = \frac{d}{dt}\frac{\partial}{\partial \dot{s}}\tilde{L}(s,\dot{s},t) - \frac{\partial}{\partial s}\tilde{L}(s,\dot{s},t) &= m\frac{d}{dt}\left[\left(a^2\cos^2 s + b^2\sin^2 s\right)\dot{s}\right] \\ &\quad - m\cos s\sin s\left(b^2 - a^2\right)\dot{s}^2 - mgb\sin s, \end{aligned}$$

or equivalently

$$0 = \left(a^2\cos^2 s + b^2\sin^2 s\right)\ddot{s} + \cos s\sin s\left(b^2 - a^2\right)\dot{s}^2 - gb\sin s.$$

The second term, which was obtained simply as a result of transforming the coordinates, is physically the result of the forces that constrain the bead to the wire. As usual, this second-order equation can be converted to a first-order system by defining $v = \dot{s}$,

$$\begin{aligned} \dot{s} &= v, \\ \dot{v} &= \sin s\,\frac{gb - (b^2 - a^2)v^2\cos s}{a^2 + (b^2 - a^2)\sin^2 s}. \end{aligned} \tag{9.32}$$

[66] Another is to impose the constraints by adding a Lagrange multiplier to the variational principle.

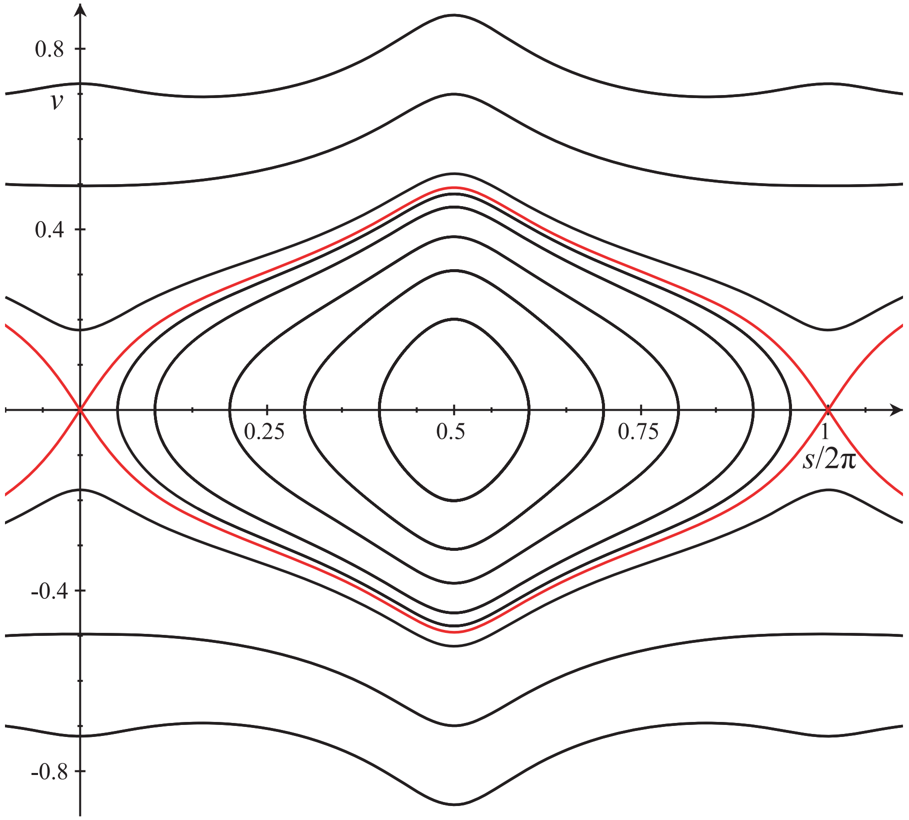

Figure 9.5. *Phase portrait for the system* (9.32) *with* $a = 1$, $b = \sqrt{5}$, *and* $g\,b = 1$. *There is a center equilibrium at* $s = \pi$ *and saddles at* $s = 0$ *and* 2π.

Note that when the ellipse degenerates into a circle, $a = b$, the equations reduce to those of the planar pendulum (9.11) with $s = \theta - \pi$. A more general example is shown in Figure 9.5. This system can be transformed to a Hamiltonian system by finding the canonical momentum; see §9.9 and Exercise 11.

The dynamics of the bead can become chaotic if the elliptical wire is allowed to rotate (Bollt and Klebanoff 2002). ■

Example 9.20. The ideal spherical pendulum is a mass on the end of a rigid, massless rod of length l under the force of gravity; see Figure 9.6. Again the Lagrangian is $\frac{1}{2} m\dot{x}^2 - mgz$. It is natural to use spherical coordinates (r, θ, ϕ) centered at the attachment point of the pendulum. The transformation $x = h(r, \theta, \phi)$ is given by

$$x = r\sin\theta\cos\phi, \quad y = r\sin\theta\sin\phi, \quad z = -r\cos\theta.$$

After some algebra, the kinetic energy can be transformed into the new coordinate system,

$$\tfrac{1}{2} m\dot{x}^2 = \tfrac{1}{2} m\left(\dot{r}^2 + r^2\sin^2\theta\dot{\phi}^2 + r^2\dot{\theta}^2\right).$$

The rigid pendulum has the constraint $r = l$. Thus we set $\dot{r} = 0$ and $r = l$ to obtain

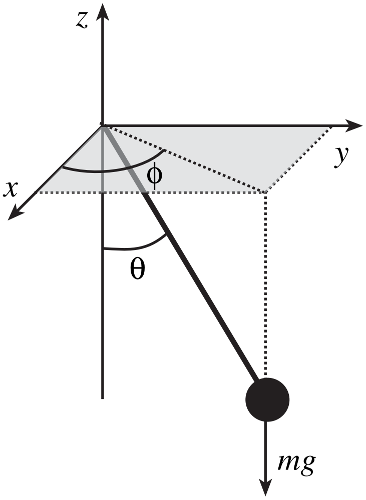

Figure 9.6. *Spherical pendulum.*

the new Lagrangian

$$\tilde{L}(\theta,\phi,\dot{\theta},\dot{\phi}) = \tfrac{1}{2}ml^2\left(\sin^2\theta\dot{\phi}^2+\dot{\theta}^2\right)+mgl\cos\theta. \tag{9.33}$$

Note that the new Lagrangian is independent of ϕ, though it does depend upon $\dot{\phi}$. This is due to the rotational symmetry of our system. The implication is that the equation of motion for ϕ is especially simple:

$$0 = \frac{d}{dt}\frac{\partial L}{\partial\dot{\phi}} = \frac{d}{dt}(ml^2\sin^2\theta\dot{\phi}) = \frac{d}{dt}p_\phi.$$

Thus, $p_\phi(\theta,\dot{\phi})$ is an invariant; physically, it is the vertical or axial component of the angular momentum. More generally, if a Lagrangian does not depend upon one of the coordinates, say q_s, then the momentum corresponding to that coordinate is an invariant, since

$$\frac{d}{dt}p_s = \frac{d}{dt}\frac{\partial L}{\partial\dot{q}_s} = \frac{\partial L}{\partial q_s} = 0.$$

Such a coordinate is called *cyclic*.

The equation of motion for θ is

$$\frac{d}{dt}\frac{\partial L}{\partial\dot{\theta}} - \frac{\partial L}{\partial\theta} = ml^2\ddot{\theta} - \frac{p_\phi^2}{ml^2}\frac{\cos\theta}{\sin^3\theta} + mgl\sin\theta$$

and solving for the highest derivative gives

$$ml^2\ddot{\theta} = \frac{p_\phi^2}{ml^2}\frac{\cos\theta}{\sin^3\theta} - mgl\sin\theta. \tag{9.34}$$

Note that this equation has a new force-like term, the "centrifugal force" caused by the angular motion; see Figure 9.7. The centrifugal force is singular at $\theta = 0$ because to maintain a fixed angular momentum as $\theta \to 0$, $\dot{\phi}$ would have to go to ∞.

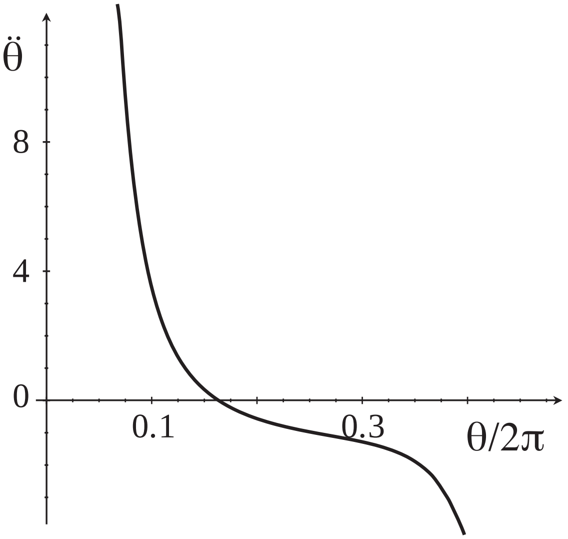

Figure 9.7. *Balance of the centrifugal and gravitational forces* (9.34) *with* $p_\phi = ml^2$ *and* $g = l$.

The phase portrait for (9.34) is shown in Figure 9.8. There is an equilibrium solution when the gravitational and centrifugal forces balance—corresponding to the zero value shown in Figure 9.7. For this choice of θ, the pendulum rotates at a constant angular speed $\dot{\phi}$; otherwise, since θ is oscillatory, the angular speed must also oscillate to keep p_ϕ constant. ■

Symmetries and Invariants

The invariance of the axial component of the angular momentum in the spherical pendulum arises because the potential and kinetic energies are independent of the spherical angle, ϕ. Equivalently, the rotational *symmetry* of the Lagrangian gives rise to an invariant. Recall from §6.4 that a symmetry corresponds to a diffeomorphism, S, that conjugates a flow to itself, $S \circ \varphi_t = \varphi_t \circ S$, (6.24). Since the new Lagrangian under a coordinate transformation is simply (9.29), the Euler–Lagrange equations in the new variables, $y = S(x)$, are identical to those in the old coordinates when $\tilde{L}(x, v, t) = L(x, v, t)$. We then say that the Lagrangian is *equivariant*:

$$L(S(x), DS(x)v, t) = L(x, v, t).$$

Emmy Noether discovered in 1915 that when the symmetry depends smoothly upon a parameter, $S(x) \to h_s(x)$ for $s \in \mathbb{R}$, then equivariance implies that the Euler–Lagrange equations have an invariant.[67]

Theorem 9.21 (Noether). *Suppose* $L(x, v, t)$ *is* C^2, $h_s : M \to M$ *is a* C^2 *diffeomorphism depending smoothly on a parameter* s, *and* L *is equivariant under* h_s*:*

$$L(h_s(x), Dh_s(x)v, t) = L(x, v, t). \tag{9.35}$$

[67] This result was obtained by Noether just as she arrived in Göttingen at the invitation of David Hilbert and began a long and ultimately successful battle with the university administration to be allowed to receive the *Habilitation* and join the faculty—an honor that at that time was not open to women.

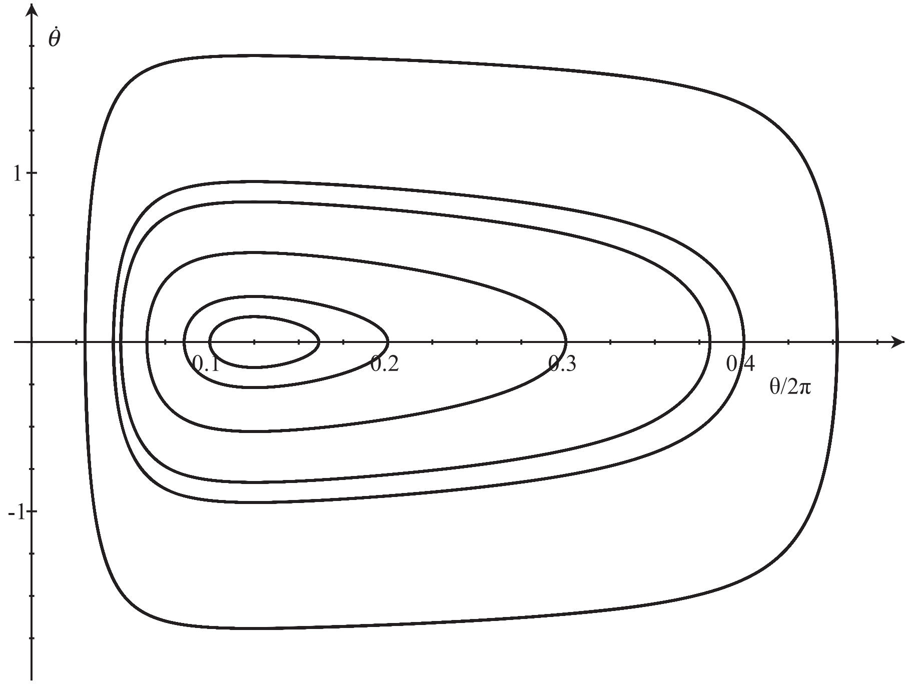

Figure 9.8. *Phase space of the spherical pendulum* (9.34) *with the parameters of Figure* 9.7.

Then the Euler–Lagrange equations for L have an invariant

$$I(x,\dot{x}) = \frac{\partial L}{\partial \dot{x}}(x,\dot{x},t)\,\frac{\partial h_s(x)}{\partial s}\bigg|_{s=0}. \tag{9.36}$$

Proof. Note that Euler–Lagrange equations for $y = h_s(x)$ are, by the assumption of equivariance of L and Theorem 9.18, the same as those for x. Differentiation of (9.35) with respect to s and use of the Euler–Lagrange equations (9.28) gives

$$\begin{aligned} 0 &= \frac{\partial}{\partial s} L(h_s(x), Dh_s(x)v, t) = \frac{\partial L}{\partial x}\frac{\partial h_s}{\partial s} + \frac{\partial L}{\partial v}\frac{\partial Dh_s}{\partial s}v \\ &= \left(\frac{d}{dt}\frac{\partial L}{\partial v}\right)\frac{\partial h_s}{\partial s} + \frac{\partial L}{\partial v}\frac{\partial Dh_s}{\partial s}\frac{dx}{dt} = \frac{d}{dt}\left(\frac{\partial L}{\partial v}\frac{\partial h_s}{\partial s}\right). \end{aligned}$$

Consequently I in (9.36) is independent of time along the trajectory. □

Theorem 9.21 directly applies to the case of rotational symmetry of the spherical pendulum upon choosing $h_s(r,\theta,\phi) = (r,\theta,\phi+s)$ since (9.33) does not depend upon ϕ. Many other applications of symmetry can be found in any text on classical mechanics (Arnold 1978; Barger and Olsson 1973; Goldstein, Poole, and Safko 2002).

9.8 ▪ The Calculus of Variations

I, Johann Bernoulli, address the most brilliant mathematicians in the world. Nothing is more attractive to intelligent people than an honest, challenging

> *problem, whose possible solution will bestow fame and remain as a lasting monument. Following the example set by Pascal, Fermat, etc., I hope to gain the gratitude of the whole scientific community by placing before the finest mathematicians of our time a problem which will test their methods and the strength of their intellect. If someone communicates to me the solution of the proposed problem, I shall publicly declare him worthy of praise.* (Johann Bernoulli, *Acta Eruditorum*, 1696)

The calculus of variations is concerned with finding the stationary values of a functional such as the action. Historically, it arose from the problem of computing the path a particle would take to minimize the travel time between two given points—the *brachistochrone* problem posed by Johann Bernoulli in the quote above. Euler showed in 1744 that this problem could be solved using the Euler–Lagrange equations (9.28) for a functional of the form (9.25). This theory has become known as the *classical* calculus of variations.

Although we have shown that smooth solutions of the Euler–Lagrange equations are stationary points of the action, it is not necessarily true that every stationary point satisfies the ODEs.

Example 9.22 (Weierstrass). Consider the problem of finding the curve $\gamma : [0,1] \to \mathbb{R}$ so that $\gamma = x(t)$ minimizes the functional

$$A[\gamma] = \int_{-1}^{1} t^2 \dot{x}^2 dt$$

with the endpoint conditions $x(-1) = -1$ and $x(1) = 1$. If γ were C^2, it would satisfy the Euler–Lagrange equations

$$\frac{d}{dt}(2t^2\dot{x}) = 0,$$

implying that $x(t) = t^{-1}$, which is certainly not C^2 at $t = 0$, violating the assumption.

This action obeys the inequality $A[\gamma] \geq 0$. Moreover, there is a sequence of smooth functions whose action limits to zero:

$$x_n(t) = \frac{\arctan(nt)}{\arctan(n)}.$$

Note that x_n satisfies the required endpoint conditions. Furthermore,

$$A[x_n] = \frac{1}{\arctan^2(n)} \int_{-1}^{1} \left(\frac{nt}{1+(nt)^2}\right)^2 dt < \frac{1}{\arctan^2(n)} \int_{-1}^{1} \frac{dt}{1+(nt)^2} = \frac{2}{n\arctan(n)}.$$

As $n \to \infty$ the right-hand side approaches zero, and thus $A[x_n] \to 0$. The sequence $x_n(t)$ limits to the discontinuous curve $x_\infty(t) = \mathrm{sgn}(x)$. Thus the minimum of A is achieved, but not on a smooth function. ∎

Techniques for studying the existence of stationary points and, in particular, *minima* of the action are called *direct methods* in the calculus of variations. Leonida Tonelli did important early work on this problem in the 1920s.

Tonelli showed that with two additional conditions on L, there are smooth curves that actually minimize the action. The first assumption is a growth or *coercivity* condition in its dependence upon the velocity, namely, that there is a constant $p > 1$ and constants $\alpha > 0$ and β such that

$$L(x,v,t) \geq \alpha |v|^p + \beta$$

for all $t \in [a,b]$ and $x \in M$. Under this assumption the action has a lower bound, and there exists a sequence of absolutely continuous functions $x_n(t)$ whose actions converge to this infimum. Moreover, the sequence $x_n(t)$ converges uniformly to a limit $x_\infty(t)$. However, it is not guaranteed that this curve is smooth nor that $A[x_\infty]$ is the minimal value (Akhiezer 1962).

These latter properties follow from a second assumption that is often satisfied by physical systems whose kinetic energies are proportional to v^2. Let $\rho = D_v^2 L$ be the Hessian of L with respect to the velocities,

$$\rho_{ij}(x,v,t) \equiv \frac{\partial^2 L}{\partial v_i \partial v_j}. \tag{9.37}$$

The crucial requirement is the

> ▷ *Legendre condition*: The Hessian ρ is a uniformly, positive-definite matrix. That is, there is a $c > 0$ such that for all (x,v,t) and all vectors $w \in \mathbb{R}^n$,
>
> $$w^T \rho w \geq c|w|^2. \tag{9.38}$$

This is commonly satisfied when the Lagrangian corresponds to a mechanical system. In this case,

$$L(x,v,t) = \tfrac{1}{2} v^T \rho(x) v - V(x), \tag{9.39}$$

where ρ is the *mass* matrix; it is often uniformly positive definite from physical considerations. We found a form like this both for the elliptical loop (9.31) and the spherical pendulum (9.33) examples in §9.7, although in the latter case the mass matrix is only semi-definite.

When the Lagrangian satisfies the Legendre condition, Tonelli showed that a minimizing trajectory exists and that it satisfies the Euler–Lagrange equations. One version of this theorem is as follows (Mather 1991).

Theorem 9.23 (Tonelli). *Suppose $L(x,v,t)$ is C^2 on $M \times \mathbb{R}^n \times \mathbb{R}$, where M is a compact n-dimensional manifold, and satisfies the following conditions:*

(i) *the Legendre condition;*

(ii) *depends periodically on time, $L(x,v,t+T) = L(x,v,t)$;*

(iii) *grows superlinearly in the velocity,*

$$\frac{L(x,v,t)}{|v|} \to \infty \text{ as } v \to \infty;$$

(iv) *and has a complete flow (recall §4.2), $\varphi_t(x,v)$, $t \in \mathbb{R}$.*

Then for any points $a, b \in M$, there is a C^2 curve $\gamma(t)$ that satisfies the Euler–Lagrange equations and is a minimum of the action.

The completeness condition (that the flow exists for every initial condition for all time) is satisfied only when the velocity on every trajectory remains bounded. If condition (iv) is not satisfied, then the minimal curve still exists, but it may only be C^1 and thus not satisfy the Euler–Lagrange equations. Condition (iv), as well as the requirement that L be C^2, can be violated in common systems, for example, for gravitational forces between point particles, where the potential has a $1/r$ singularity.

9.9 ▪ Equivalence of Hamiltonian and Lagrangian Mechanics

When the Legendre condition (9.38) is satisfied, Lagrangian mechanics is equivalent to Hamiltonian mechanics. To convert the second-order Euler–Lagrange equations (9.28) to a first-order system, like the Hamiltonian system (9.9), it is natural to define an auxiliary variable

$$p \equiv \frac{\partial L}{\partial v}(q, v, t), \tag{9.40}$$

since then the Euler–Lagrange equation becomes$\dot{p} = \partial L(q, v, t)/\partial q$. This is not a complete dynamical system since v is not specified. However, since $v = \dot{q}$ on the Euler–Lagrange trajectory, if v can be given as a function of (q, p, t), the system can be closed with a second equation of the form $\dot{q} = v(q, p, t)$. Equation (9.40) implicitly defines the required function.

Lemma 9.24. *When $L(q, \cdot, t) : \mathbb{R}^n \to \mathbb{R}$ is a C^2 function for each $(q, t) \in M \times \mathbb{R}$ and satisfies the Legendre condition* (9.38)*, then* (9.40) *defines a unique implicit function $v(q, p, t)$.*

Proof. Define $F(v; p, q, t) = p - D_v L(q, v, t)$. Since $D_v F = -D_v^2 L$ is nonsingular by the Legendre condition, the implicit function theorem implies that near any pair v_0, p_0, where $F(v_0; p_0, q, t) = 0$, there is a unique C^1 solution $v(q, p, t)$ to $F = 0$. We now claim that the map $P_{q,t} : v \to p$ given by $p = P_{q,t} = D_v L(q, v, t)$ is bijective and so has a unique inverse. To see this, given any v_1, let $p_1 = P(v_1)$ and define the line segment $v(s) = v_0 + s(v_1 - v_0)$, $s \in [0, 1]$. On this segment, $\frac{\partial P}{\partial s} = D_v^2 L \frac{\partial v}{\partial s}(s) = \rho(q, v(s), t) \cdot (v_1 - v_0)$; thus the fundamental theorem of calculus implies that

$$p_1 - p_0 = \int_0^1 \rho(q, v(s), t) \cdot (v_1 - v_0)\, ds.$$

Taking the dot product of this with $(v_1 - v_0)$ and using (9.38) gives

$$(v_1 - v_0) \cdot (p_1 - p_0) \geq c\, |v_1 - v_0|^2. \tag{9.41}$$

Thus whenever $v_1 \neq v_0$, $p_1 \neq p_0$, so $P_{q,t}$ is injective. Moreover, $P_{q,t}$ is surjective since as $v_1 \to \infty$, then $p_1 \to \infty$ as well, and must do so in nearly the same direction according to (9.41). Thus the inverse of $P_{q,t}$ is globally unique. □

The first-order system $\dot{p} = \partial L/\partial q$, $\dot{q} = v$, can be made explicit by defining a Hamiltonian using the *Legendre transformation*

$$H(q, p, t) = p \cdot v(q, p, t) - L(q, v(q, p, t), t). \tag{9.42}$$

Indeed the equations of motion are Hamiltonian with this function H:

$$\begin{aligned} \frac{\partial H}{\partial p}(q, p, t) &= v + \left(p - \frac{\partial L}{\partial v}\right) \cdot \frac{\partial v}{\partial p} = v = \dot{q}, \\ -\frac{\partial H}{\partial q}(q, p, t) &= \frac{\partial L}{\partial q} - \left(p - \frac{\partial L}{\partial v}\right) \cdot \frac{\partial v}{\partial q} = \frac{\partial L}{\partial q} = \dot{p}. \end{aligned}$$

Geometrically, for each (q, p, t) the value H is the maximum distance between the plane $y = p \cdot v$ and the graph $y = L(q, v, t)$. This leads to the geometrical definition

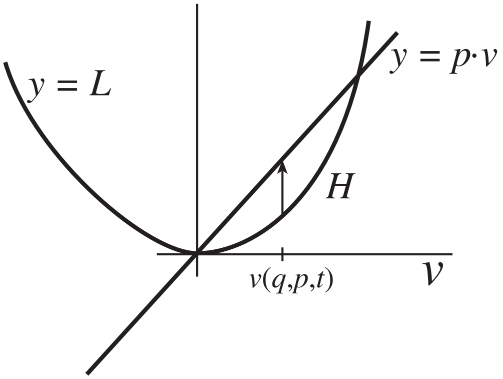

Figure 9.9. *Legendre transformation* (9.43).

▷ *Legendre transformation*: For each q, t the Legendre transformation $L \to H$ is defined by

$$H(q,p,t) = \max_v [p \cdot v - L(q,v,t)]. \tag{9.43}$$

Note that the first derivative condition for an extremum leads precisely to (9.40); moreover, the extremum is a maximum by the Legendre condition; see Figure 9.9.

The inverse of the Legendre transformation can be used to obtain the Lagrangian from the Hamiltonian. In fact, when the Lagrangian satisfies the Legendre condition, the Hamiltonian also satisfies a convexity condition, as can be seen by a direct calculation:

$$\frac{\partial^2 H}{\partial p^2} = \frac{\partial}{\partial p}\left[v + \left(p - \frac{\partial L}{\partial v}\right)\frac{\partial v}{\partial p}\right] = \frac{\partial v}{\partial p} = \rho^{-1},$$

which is positive definite. If ρ^{-1} is also uniformly positive definite, then we can define

$$L(x,v,t) = \max_p (p \cdot v - H(x,p,t)), \tag{9.44}$$

which implies that $v = \partial H / \partial p$. Consequently, the Legendre transformation is an involution.

Finally, the action (9.25) can be written in terms of the Hamiltonian as

$$A[\gamma] = \int_a^b L dt = \max_p \int_a^b (p\dot{q} - H(q,p,t))dt. \tag{9.45}$$

Apart from the "max," this is identical to the Hamiltonian action (9.22). Moreover, along any solution curve, $\dot{q} = \partial H / \partial p$ which is the extremal value. Thus $A[\gamma] = S[\gamma]$ along solutions.

Example 9.25. The mechanical system (9.39) has a momentum

$$p = \frac{\partial L}{\partial v} = \rho(q)v \quad \Rightarrow v = \rho^{-1}(q)p$$

if the symmetric mass matrix, $\rho(q)$, is nonsingular. Thus the Hamiltonian is given by

$$\begin{aligned} H(q,p) &= p^T v - L = p^T \rho^{-1}(q)p - \tfrac{1}{2}\left(p^T \rho^{-1} p - V(q)\right) \\ &= \tfrac{1}{2} p^T \rho^{-1}(q)p + V(q). \end{aligned}$$

Since the Hamiltonian is autonomous, the energy is constant along the trajectories: $H(q,p)=E$. Explicit examples are given in Exercises 10, 11, and 14. ∎

Note that the Hamiltonian is autonomous whenever L is independent of time. This is another manifestation of Noether's Theorem 9.21: "time-translation invariance" of L: $L(q,v,t)=L(q,v,t+s)$ for all s implies the conservation of energy; see Exercise 13.

9.10 ▪ Linearized Hamiltonian Systems

As we saw in §9.3, any equilibrium z^* of a Hamiltonian system (9.14) is a critical point of H. In a neighborhood of this point, the terms in H that are linear in $\delta z = z - z^*$ will vanish, giving

$$H(z^*+\delta z)=H(z^*)+\tfrac{1}{2}\delta z^T D^2 H(z^*)\delta z+O(\delta z^3).$$

Here D^2H is the Hessian matrix of H,

$$(D^2H)_{ij}\equiv\frac{\partial^2 H}{\partial z_i \partial z_j};$$

it is necessarily symmetric. Since the constant $H(z^*)$ will not appear in the equations of motion, the Hamiltonian for the linearized system about z^* can simply be taken to be the quadratic form $H = {}^1\!/_2 \delta z^T S \delta z$, where $S = D^2H(z^*)$. For this system, the equations of motion are

$$\delta\dot{z}=JS\delta z=K\delta z, \tag{9.46}$$

where J is the Poisson matrix (9.14).

A matrix of the form $K=JS$, where $S=S^T$, is called a *Hamiltonian* matrix. Just as the symmetric matrices form a group under addition, $Sym(2n)$, so do the Hamiltonian matrices: if K_1 and K_2 are Hamiltonian matrices, then so is K_1+K_2; this group is called $sp(2n)$.[68] Note that whenever $K=JS$, then $JK=-S$ since $J^2=-I$. Also, since $J^T=-J$, then $S=-JK=J^TK=(J^TK)^T=K^TJ$, so $sp(2n)$ is characterized by

▷ *Hamiltonian matrices*: $sp(2n)=\left\{K\in\mathbb{R}^{2n\times 2n} : JK+K^TJ=0\right\}$.

Since there is a one-to-one correspondence between matrices in $Sym(2n)$ and those in $sp(2n)$, they have the same number of independent elements; thus $\dim(sp(2n)) = (2n+1)n$.

We know from Chapter 2 that the formal solution to the linear system (9.46) is $\delta z(t) = \exp(tK)\delta z(0)$. The matrix $\exp(tK)$ is called a symplectic matrix. The set of symplectic matrices is also a group—though the group operation is now matrix multiplication—the symplectic group, $Sp(2n)$.

The product of two Hamiltonian matrices is not necessarily a Hamiltonian matrix. However, there is a kind of product that can be defined that does map onto the group. This additional structure is related to the Poisson bracket. Suppose $H_1 = {}^1\!/_2 z^T S_1 z$ and $H_2 = {}^1\!/_2 z^T S_2 z$ are two quadratic Hamiltonians. Consider a third function defined as

$$H_3=\{H_1,H_2\}, \tag{9.47}$$

[68] This is the Lie algebra of the symplectic group. Thus, Hamiltonian matrices are also called *infinitesimally symplectic* matrices.

where { , } is the Poisson bracket (9.15). A simple calculation shows that H_3 is also a quadratic Hamiltonian:

$$H_3 = \nabla H_1^T J \nabla H_2 = (z^T S_1) J (S_2 z) = \tfrac{1}{2}\left(z^T S_1 J S_2 z + z^T S_2 J^T S_1 z\right)$$
$$= \tfrac{1}{2} z^T (S_1 J S_2 - S_2 J S_1) z = \tfrac{1}{2} z^T S_3 z,$$

where $S_3 = S_1 J S_2 — S_2 J S_1$. Note that $S_3^T = S_3$, since S_1 and S_2 are symmetric. Therefore, the matrix $K_3 = J S_3$ is a Hamiltonian matrix, and

$$K_3 = J S_3 = J S_1 J S_2 - J S_2 J S_1$$
$$= [K_1, K_2],$$

where [,] is the commutator, $[A,B] = AB — BA$; recall §2.6. This additional structure means that the group $sp(2n)$ is a *Lie algebra*.[69] In general, a Lie algebra is an additive group with an additional operation, a Lie bracket denoted [,]. Lie brackets must satisfy the Jacobi identity (9.17). For a matrix algebra the bracket is simply the commutator. The point is this: if A and B are in the Lie algebra, then so is $C = [A,B]$. There are many interesting Lie algebras that arise in applications; for example, $su(d)$ is the Lie algebra of $d \times d$ Hermitian matrices with zero trace (Hall 2003; Isham 1999). By contrast, the symmetric matrices do not form a Lie algebra since the commutator of two symmetric matrices is antisymmetric.

We summarize this discussion as a lemma.

Lemma 9.26. *$sp(2n)$ is a $(2n+1)n$-dimensional Lie algebra.*

Eigenvalues of Hamiltonian Matrices

The stability of a Hamiltonian equilibrium is governed by the eigenvalues of its Hamiltonian matrix. According to a theorem first proved by Poincaré and Lyapunov, the eigenvalues are restricted due to the condition that JK is symmetric.

Theorem 9.27. *If λ is an eigenvalue of a Hamiltonian matrix, K, then so is $—\lambda$. Moreover, the characteristic polynomial of K is even.*

Proof. Recall that for any two square matrices A and B, $\det(AB) = \det(A)\det(B)$ and $\det(A^T) = \det(A)$. Moreover, if I is the $2n \times 2n$ identity matrix, then $\det(—I) = (—1)^{2n} = 1$.

Suppose $K \in sp(2n)$. Since $JK = —K^T J$, and $\det(J) = 1$, the characteristic polynomial $p(x) = \det(xI - K)$ obeys

$$p(x) = \det(J)\det(xI - K) = \det(xJ - JK) = \det(xJ + K^T J)$$
$$= \det(xI + K^T)\det(J) = \det(xI + K)^T$$
$$= p(-x).$$

In particular, whenever $p(\lambda) = 0$, so is $p(—\lambda)$. □

Theorem 9.27 implies that $p(x)$ has only n nonzero coefficients,

$$p(x) = x^{2n} + a_1 x^{2n-2} + \cdots + a_{n-1} x^2 + a_n = 0. \tag{9.48}$$

[69] Named after Sophus Lie, 1842–1899, a Norwegian mathematician. *Lie* is pronounced *Lee*.

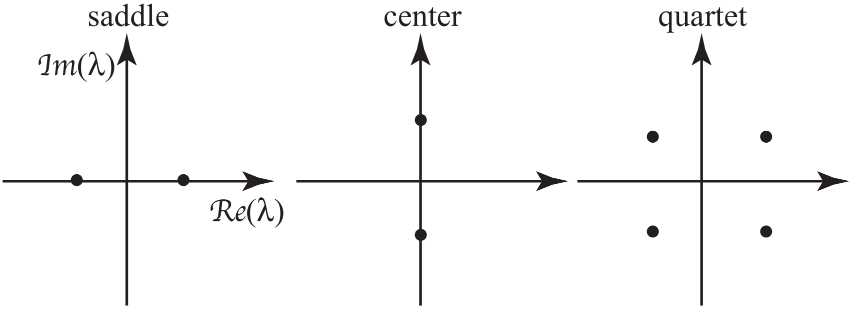

Figure 9.10. *Hamiltonian eigenvalue configurations in the complex λ-plane.*

Since the coefficient of x^{2n-1} is $-\mathrm{tr}(K)$, a simple consequence of (9.48) is the vanishing of the trace of a Hamiltonian matrix.

In addition, if a Hamiltonian matrix is real, then its characteristic polynomial is real, so that whenever it has a complex eigenvalue, λ, then its conjugate $\bar{\lambda}$ is also an eigenvalue.

One consequence of Theorem 9.27 is that it is impossible for a Hamiltonian equilibrium to be asymptotically stable, since this would require that all the eigenvalues have negative real parts. When H is real, there are four possible groupings of the eigenvalues:

(a) *Hyperbolic* (*saddle*): λ is real. Then there is a pair of eigenvalues $(\lambda, -\lambda)$.

(b) *Elliptic* (*center*): $\lambda = i\omega$ is imaginary. Then $-\lambda = \bar{\lambda}$ and $(i\omega, -i\omega)$ form a pair.

(c) *Krein quartet*: λ is complex, and $\mathrm{Re}(\lambda) \neq 0$. Then there is a quartet of eigenvalues $(\lambda, -\lambda, \bar{\lambda}, -\bar{\lambda})$.

(d) *Parabolic*: A double eigenvalue $\lambda = 0$.

The first three configurations are shown in Figure 9.10.

The assertion that the parabolic case corresponds to a multiplicity-two eigenvalue is not obvious. However, the previous theorem can be generalized to show this.

Theorem 9.28 (Hamiltonian eigenvalues). *If $K \in sp(2n)$ has an eigenvalue λ of multiplicity k, then $-\lambda$ is an eigenvalue of multiplicity k. Moreover, the multiplicity of the eigenvalue 0, if it occurs, is even.*

Proof. Since $JK + K^T J = 0$, $K = J^{-1}(-K^T)J$. Therefore, K is similar to $-K^T$. Similar matrices have the same eigenvalues counted with multiplicity (and the same Jordan normal forms). Since eigenvalues and multiplicities of K^T are the same as those of K, the multiplicity of λ is the same as that of $-\lambda$.

Since $\mathrm{tr}(K) = 0 = \sum_{i=1}^{2n} \lambda_i$, if zero is an eigenvalue then it must have even multiplicity, because the remaining $\lambda_i \neq 0$ come in opposite pairs. □

Example 9.29. A set of uncoupled, simple harmonic oscillators is defined by the quadrati Hamiltonian

$$H = \frac{1}{2}\sum_{j=1}^{n} \omega_j \left(p_j^2 + q_j^2\right). \tag{9.49}$$

The Hamiltonian matrix, in block form, is

$$K=JS=\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}\begin{pmatrix} \operatorname{diag}(\omega_j) & 0 \\ 0 & \operatorname{diag}(\omega_j) \end{pmatrix}=\begin{pmatrix} 0 & \operatorname{diag}(\omega_j) \\ -\operatorname{diag}(\omega_j) & 0 \end{pmatrix}.$$

Note that K has zero trace as required. The characteristic polynomial $\det(\lambda I-K)$ can be expanded by rows, and each row has only two nonzero elements. Expanding along the first row gives

$$p(\lambda)=\det\begin{pmatrix} \operatorname{diag}(\lambda) & -\operatorname{diag}(\omega_i) \\ \operatorname{diag}(\omega_i) & \operatorname{diag}(\lambda) \end{pmatrix}$$

$$=\lambda\det\begin{pmatrix} \lambda & & & 0 & -\omega_2 & & \\ & \ddots & & \vdots & & \ddots & \\ & & \lambda & 0 & & & -\omega_n \\ 0 & \cdots & 0 & \lambda & 0 & \cdots & 0 \\ \omega_2 & & & 0 & \lambda & & \\ & \ddots & & \vdots & & \ddots & \\ & & \omega_n & 0 & & & \lambda \end{pmatrix}$$

$$+(-1)^n\omega_1\det\begin{pmatrix} 0 & \lambda & & 0 & -\omega_2 & & \\ \vdots & \ddots & \ddots & 0 & & \ddots & \\ 0 & 0 & 0 & \lambda & & & -\omega_n \\ \omega_1 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \ddots & & 0 & \lambda & & \\ \vdots & & \ddots & 0 & & \ddots & \\ 0 & & & \omega_n & & & \lambda \end{pmatrix}.$$

Now each of the two subdeterminants has only one nonzero element in its nth row. Expanding along these rows shows that the remaining determinants are the same:

$$p(\lambda)=(-1)^{2n}\left(\lambda^2+\omega_1^2\right)\det\begin{pmatrix} \lambda & & & -\omega_2 & & \\ & \ddots & & & \ddots & \\ & & \lambda & & & -\omega_n \\ \omega_2 & & & \lambda & 0 & \cdots \\ & \ddots & & & \ddots & \\ & & \omega_n & & & \lambda \end{pmatrix}.$$

The new $2(n-1)\times 2(n-1)$ matrix has the same form as the initial one; thus this process can be repeated to finally obtain

$$p(\lambda)=\prod_{j=1}^{n}(\lambda^2+\omega_j^2),$$

showing that K has the $2n$ eigenvalues $\pm i\omega_j$. Thus this Hamiltonian has all its eigenvalues on the imaginary axis, and its motion corresponds to a center. ∎

Recall from Theorem 2.20 that the domain of a square matrix can be decomposed into a direct sum of the generalized eigenspaces E_{λ_i} that correspond to the eigenvalue

λ_i. A special property of the Hamiltonian case is that the spaces corresponding to eigenvalues that are not in a ± pair or a Krein quartet are "skew" orthogonal.

Theorem 9.30. *If K is a Hamiltonian matrix with eigenvectors $\xi_i, i = 1, 2,$ and corresponding eigenvalues λ_i such that $\lambda_1 + \lambda_2 \neq 0$, then ξ_1 and ξ_2 are skew orthogonal:*

$$\xi_1^T J \xi_2 = 0. \tag{9.50}$$

More generally, the generalized eigenspaces E_{λ_i} are skew orthogonal.

Proof. Recall that if K is a Hamiltonian matrix, then $S = JK$ is symmetric. Multiply the eigenvalue equation $K\xi_1 = \lambda_1 \xi_1$ on the left by $\xi_2^T J$ to obtain $\xi_2^T JK\xi_1 = \lambda_1 \xi_2^T J \xi_1$. Subtracting the corresponding equation for ξ_2 gives

$$\begin{aligned} \xi_2^T JK\xi_1 &= \lambda_1 \xi_2^T J \xi_1, \\ \xi_1^T JK\xi_2 &= \lambda_2 \xi_1^T J \xi_2, \\ 0 &= (\lambda_1 + \lambda_2)\xi_2^T J \xi_1. \end{aligned}$$

Since $\lambda_1 + \lambda_2 \neq 0$, this implies (9.50). The result for generalized eigenspaces can be proved inductively from this; see (Meyer and Hall 1992, p. 51 et seq.). □

9.11 ▪ Krein Collisions

In 1950 the Russian mathematician Mark Krein obtained another interesting result about Hamiltonian eigenvalues during a turbulent period in his life when he was twice dismissed from Odessa University. His result concerns the possible changes of stability of Hamiltonian equilibria as a parameter is varied. Suppose that the equilibrium is linearly stable, so that all eigenvalues start on the imaginary axis. As a parameter is varied, these eigenvalues will change continuously; recall Exercise 8.3. Moreover, they cannot leave the axis unless unless a pair of eigenvalues collide—creating a multiplicity-two eigenvalue—because that would violate Theorem 9.27. Consequently, there are two ways that an elliptic point can lose stability. One is for a pair of eigenvalues to collide at 0, the parabolic case, and continue to real eigenvalues, the hyperbolic case. A second is for a pair $\pm i\omega_1$ to collide with a pair $\pm i\omega_2$, called a *Krein collision*, and split off the imaginary axis—giving rise to a Krein quartet. The question that Krein addressed is, can every Krein collision lead to instability?

The answer is no. Instability is possible only if certain conditions on the *Krein signature* are satisfied. To any Hamiltonian matrix, K, there corresponds a linear Hamiltonian

$$H(\xi) = -\tfrac{1}{2}\xi^T JK\xi. \tag{9.51}$$

The value of H is independent of time along its flow, and it can be used to define the

▷ *Krein signature.* Suppose $K \in sp(2n)$ has a nonzero eigenvalue pair $\pm i\omega$ with corresponding eigenvectors $v = u \pm iw$. Let $\xi \in E_{\pm i\omega} = \text{span}(u, w)$ be any vector in the invariant subspace for $\pm i\omega$. The Krein signature of $E_{\pm i\omega}$ is

$$\sigma_\omega = \text{sgn} H(\xi). \tag{9.52}$$

It is not hard to see that σ_ω is independent of the choice of $\xi \in E_{\pm i\omega}$. Note that $K(u+iw) = i\omega(u+iw)$, so that $Ku = -\omega w$, and $Kw = \omega u$. Thus for $\xi = \alpha u + \beta w$,

$$H(\xi) = -\tfrac{1}{2}(\alpha u + \beta w)^T JK(\alpha u + \beta w) = -\tfrac{1}{2}\omega(\alpha u + \beta w)^T(-\alpha Jw + \beta Ju)$$

$$= \tfrac{1}{2}\omega(\alpha^2 + \beta^2)u^T Jw = -\tfrac{1}{2}(\alpha^2 + \beta^2)u^T JKu = (\alpha^2 + \beta^2)H(u),$$

where we have used $u^T Ju = 0$ since J is antisymmetric. Thus the sign of $H(\xi)$ is independent of the choice vector $\xi \in E_{\pm i\omega}$.

Example 9.31. The Krein signature is essentially the direction of rotation in the canonical plane. For example, let $K = \left(\begin{smallmatrix} 0 & \omega \\ -\omega & 0 \end{smallmatrix}\right)$ so that $H(z) = 1/2\omega(x^2 + y^2)$. The eigenvector for $i\omega$ is $v = (1, i)^T$, so that $u = (1, 0)^T$, and $H(u) = 1/2\omega$. Thus $\sigma_\omega = \operatorname{sgn}(\omega)$, which corresponds to the direction of rotation. ■

Since the two-by-two block of K corresponding to $\pm i\omega$ can always be written in the antidiagonal form of the example, we see that $H(\xi)$ is nonzero for any $\xi \in E_{\pm i\omega}$, and thus its sign is well defined.

Krein's theorem concerns a one-parameter family of Hamiltonian matrices, $K(s)$, with eigenvalues on the imaginary axis that collide for some value of s. The theorem shows that K cannot lose stability if the signatures of its colliding eigenvalues are the same (Arnold and Avez 1968, Appendix 29; Yakubovitch and Starzhinskii 1975).

Theorem 9.32 (Krein collisions). *Let $K(s)$ be a Hamiltonian matrix that depends upon a parameter s. Suppose that for $s < 0$, K has $2d \le 2n$ distinct, imaginary eigenvalues that are nonzero for $s \le 0$. Suppose that these eigenvalues collide at $s = 0$. If all the colliding eigenvalues have the same Krein signature, then there exists an $\varepsilon > 0$ such that the $2d$ eigenvalues remain on the imaginary axis for $0 < s < \varepsilon$.*

Proof. Let $E(s) = E_1 \oplus E_2 \oplus \ldots E_d$ be the invariant subspace of dimension $2d$ containing the eigenvectors of the imaginary eigenvalues that collide at $s = 0$. Suppose (without loss of generality) that when $s < 0$ all the Krein signatures are positive, $\sigma_{\omega_k} > 0$. Our first goal is to show that the signature of any vector $\eta \in E$ is positive for all $s < 0$. Any vector in E can be written as a linear combination $\eta = \sum_{k=1}^d c_k \xi_k$ of vectors $\xi_k \in E_k$. By (9.50), when $s < 0$, the eigenvectors ξ_k are skew orthogonal, $\xi_j^T J \xi_k = 0$ when $j \ne k$; moreover, since E_k is invariant, $K\xi_k \in E_k$, so that $\xi_j^T JK\xi_k = 0$ as well. Thus the quadratic form

$$H(\eta) = -\frac{1}{2}\sum_{i,j=1}^{d} c_i c_j \xi_i^T JK\xi_j = \sum_{i=1}^{d} c_i^2 H(\xi_i)$$

is positive for any $\eta \ne 0$. Consequently, the quadratic form H is positive definite when restricted to $E(s)$ for any $s < 0$. Thus the set $\{\eta \in E(s) : H(\eta) = 1\}$ is an ellipsoid.

By continuity, H is still positive definite at $s = 0$, and must remain positive definite up to some positive value ε. Thus $H(\eta) = 1$ defines an ellipsoid for $s < \varepsilon$. Since H is invariant, the vector $\eta(t) = e^{-tK}\eta(0)$ still belongs to the ellipsoid. Thus the flow restricted to the space $E(s)$ is stable for $s < \varepsilon$. □

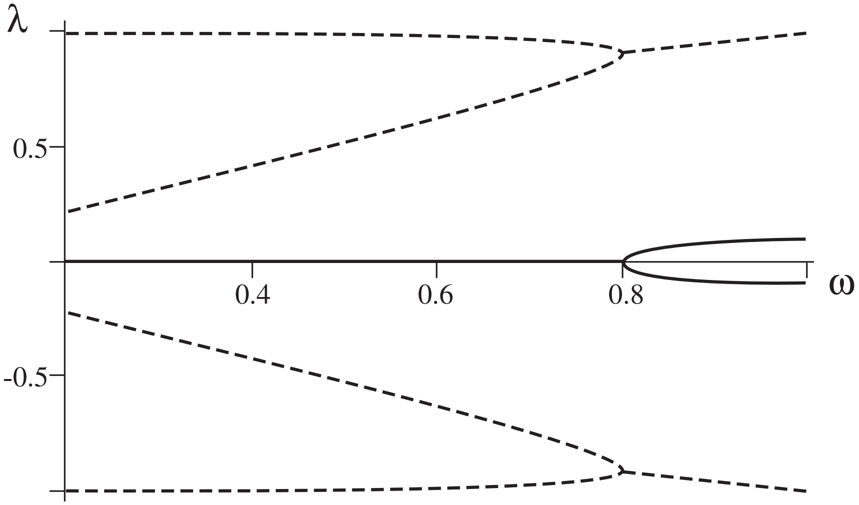

Figure 9.11. *Krein collision for* (9.53) *with* $\varepsilon = 0.2$*. The imaginary part of the four eigenvalues is dashed and the real part is solid. When* $0 < \omega < 0.8012$ *the eigenvalues are imaginary. At* $\omega \approx 0.8012$*, they collide and split off to form a Krein quartet.*

Example 9.33. Consider the Hamiltonian

$$H = \frac{1}{2}\left(p_1^2 + q_1^2\right) - \frac{\omega}{2}\left(p_2^2 + q_2^2\right) + \varepsilon p_1 p_2. \tag{9.53}$$

When $\varepsilon = 0$ and $\omega > 0$, the signatures of the two independent oscillators in H are opposite. The characteristic polynomial of K is $p(\lambda) = (\lambda^2 + 1)(\lambda^2 + \omega^2) + \omega\varepsilon^2$. Thus

$$\lambda^2 = \frac{1}{2}\left(-1 - \omega^2 \pm \sqrt{(\omega^2-1)^2 - 4\omega\varepsilon^2}\right) = \frac{1}{2}\left(-1 - \omega^2 \pm (\omega^2 - 1) \mp \frac{2\omega\varepsilon^2}{(\omega^2-1)}\right) + O(\varepsilon^4).$$

When $\omega \neq 1$ and $\varepsilon \ll 1$, these eigenvalues are purely imaginary and close to the values $\lambda = \pm i$ and $\pm i\omega$. However, when $\omega = 1$, we have $\lambda^2 = -1 \pm i\varepsilon$, giving a Krein quartet. Thus (9.53) becomes unstable when $\omega \to 1$. More generally, the Krein collision occurs when the discriminant $\Delta = (\omega^2 - 1)^2 - 4\omega\varepsilon^2$ vanishes. A sketch of the variation of the eigenvalues with ω is shown in Figure 9.11.

By contrast, when ω is negative the two oscillators have the same signature at $\varepsilon = 0$. Moreover, the discriminant Δ no longer changes sign and all four eigenvalues remain on the imaginary axis until a pair collides at 0 for some large enough ε. ■

Nonlinearly, an equilibrium whose eigenvalues undergo a mixed-signature Krein collision often gives rise to a periodic orbit. By analogy with the corresponding generic bifurcation, this is called the Hamiltonian–Hopf bifurcation (van der Meer 1985).

9.12 ▪ Integrability

There are many definitions of "integrability," but all are based on some notion of finding explicit solutions for the orbits (Zakharov 1991). One way to help find such an explicit form is to construct an integral or invariant (9.1). For an autonomous Hamiltonian system, this means finding a function F on phase space that is constant along the orbits (9.15),

$$\frac{d}{dt}F = \{F,H\} = 0,$$

so that it "Poisson commutes" with H. Of course, the Hamiltonian itself is an integral and restricts the motion to an energy surface. One would expect that each integral would allow us to restrict the motion to a surface of one fewer dimension, so that $2n-1$ integrals (including H) would seem to be needed to solve the system. This, however, is not necessary, as was first noticed by Liouville:

> ▷ *Liouville integrable*: An n degree-of-freedom Hamiltonian system is integrable if there exist n integrals F_i that are almost everywhere independent and in *involution*, $\{F_i, F_j\} = 0$.

The integrals are independent at a point if the n gradient vectors, ∇F_i, are linearly independent vectors. The integrals need not be independent everywhere. For example, the gradient of the energy vanishes at every equilibrium, but since equilibria are typically isolated, this should not be an obstruction to integrability.

The fact that the equations can be effectively solved when there are only n integrals is due to the canonical structure. When the integrals are in involution, they can be used as new momentum coordinates in the Hamiltonian. In fact, each invariant can be thought of as a Hamiltonian in its own right, with the set of differential equations

$$\frac{d}{ds_i}z = \{z, F_i\} \tag{9.54}$$

for a "time" s_i. By virtue of the involution property each of these flows has n integrals: each of the functions F_j is invariant under the time-s_i flow generated by F_i. This fact leads to a dramatic restriction of the motion of an integrable Hamiltonian system.

Theorem 9.34 (Liouville–Arnold). *Suppose H is Liouville integrable. For $c \in \mathbb{R}^n$, let $M_c = \{z : F_i = c, i = 1,2,\dots,n\}$ be a level set of the integrals on which the gradients are linearly independent. Then M_c is a smooth, invariant submanifold. If M_c is compact and connected, it is diffeomorphic to the n-torus. In this case, there are n angle coordinates, θ_i, on M_c such that the Hamiltonian flow on M_c is conjugate to*

$$\frac{d}{dt}\theta = \Omega(F) \tag{9.55}$$

for some frequency vector Ω.

Proof (Ideas). This theorem is proved in (Arnold 1978, pp. 271–274). We give only some of the ideas. That M_c is a smooth submanifold (recall §5.5) follows from the fact that the functions F_i are independent. Define the flow of (9.54) to be $\varphi_{s_i}(z)$. Then the involution property implies that

$$\varphi_{s_i} \circ \varphi_{s_j} = \varphi_{s_j} \circ \varphi_{s_i},$$

i.e., the flows commute. That M_c is an n-torus follows from group theory: the only compact, connected manifold that admits n independent, commuting flows is the n-torus. The angle coordinates on each level set are found by looking for directions in the space of times, s_i, that give closed loops on the torus. There are n such independent loops on an n-torus, and each loop corresponds to one direction. The paths traced out by these loops define the angle variables θ. □

A further consequence of this theorem is that there is a choice of angle variables that have conjugate momenta defined in a neighborhood of a regular level set M_c (Arnold 1978). These are commonly called *action* variables and are denoted I_i. An integrable Hamiltonian is said to be in *action-angle* form if it depends only upon the momentum variables—in this case when

$$H(\theta,I)=H(I).$$

Note that for action-angle variables, the equations of motion are

$$\begin{aligned} \dot{I} &= -\frac{\partial H}{\partial \theta} = 0, \\ \dot{\theta} &= \frac{\partial H}{\partial I} = \Omega(I), \end{aligned} \tag{9.56}$$

showing that the actions are themselves invariant, and thus they must be functions of the n invariants F_i. Moreover, comparing (9.55) with (9.56) shows that the frequencies Ω, when written as a function of the actions, are the gradient of the scalar $H(I)$. In particular this means that

$$\frac{\partial \Omega_i}{\partial I_k} = \frac{\partial \Omega_k}{\partial I_i}.$$

One strategy for understanding the dynamics of an integrable system is to construct its action-angle coordinates (Goldstein et al. 2002). This, however, can be nontrivial even when the integrals are known. A famous example is the Kovalevskaya top; see Exercise 14 (Kovalevskaya 1889). Although the three integrals of this top have been known since 1888, construction of the action variables is highly nontrivial (Dubrovin, Krichever, and Novikov 1985; Dullin, Juhnke, and Richter 1994).

9.13 ▪ Nearly Integrable Dynamics

Even though an n-degree-of-freedom Hamiltonian system can be integrable, it is typically not (when $n > 1$)—many of the orbits are chaotic. However, a Hamiltonian that is near to an integrable one still exhibits some of the features of integrable systems. In this section we will consider the dynamics of the system

$$H(\theta,I) = H_o(I) + \varepsilon H_1(\theta,I) \tag{9.57}$$

that is integrable when $\varepsilon = 0$. Our goal is to understand what features of the integrable dynamics persist for small ε.

Invariant Tori

In the integrable case, the trajectories lie on the *invariant tori* defined as levels sets $M_c = \{(\theta,I) : I_i = c_i\}$ of the action variables. On each torus the dynamics is given by (9.56), which has the flow

$$\varphi_t(\theta,I) = (\theta + \omega t \bmod 2\pi, I), \tag{9.58}$$

with $\omega = \Omega(I)$. The resulting orbits depend in an intricate way on the relationship between the components of the frequency vector. The simplest case occurs when $\omega = \alpha m$ for $m \in \mathbb{Z}^n$, an integer vector, and any $\alpha \in \mathbb{R}$. In this case the orbit is periodic with period $T = 2\pi/\alpha$ (or an equilibrium if $\alpha = 0$).

The opposite of the periodic case was considered in an example in §7.1 and Exercise 7.1, that is, the case that ω is *incommensurate*,

$$\omega \cdot m \neq 0 \; \forall m \in \mathbb{Z}^n \setminus \{0\}. \tag{9.59}$$

The flow (9.58) is then quasiperiodic and transitive: every orbit is dense on $\mathbb{T}^n$.

Between these two extreme cases are the commensurate frequency vectors: ω values such that there exists at least one nonzero integer vector m for which $\omega \cdot m = 0$. The set of integer solutions of this equation is called the *resonance module*; see Exercise 17. In this case the orbits are dense on lower-dimensional tori.

Example 9.35. Suppose $n = 3$ and $\omega = (1, \gamma, 2\gamma)$, where γ is irrational, for example, the golden mean $\gamma = (1+\sqrt{5})/2$. The only integer solutions to $\omega \cdot m = 0$ are then $m = (0, 2k, -k)$ for some $k \in \mathbb{Z}$. Thus there is, up to normalization, precisely one vector of commensurability. The flow (9.58) with this frequency is not dense on $\mathbb{T}^3$, but it does cover a two-dimensional surface. Indeed, the flow restricted to two components (θ_1, θ_2) is dense on the two-torus because the vector $(1, \gamma)$ is incommensurate. This holds as well for the components (θ_1, θ_3). However, because of the rational ratio ω_2/ω_3, there is always a simple relation between θ_2 and θ_3, namely,

$$\theta_3(t) - 2\theta_2(t) \bmod 2\pi = \theta_3(0) - \theta_2(0) \bmod 2\pi,$$

which is constant. Thus the orbits densely cover two-dimensional tori, and the collection of these tori covers the three-torus. ■

The motion on incommensurate tori is also called *nonresonant*, as opposed to the *resonant* motion on tori that are commensurate.

The n-dimensional, nonresonant invariant tori of the n-degree-of-freedom integrable system are prevalent when the frequencies $\Omega(I)$ vary nontrivially as I changes—that is, when the oscillators represented by H_o are *anharmonic*. So that this is true, it is necessary that Ω satisfies a *nondegeneracy* or *twist* condition. The function $\Omega : \mathbb{R}^n \to \mathbb{R}^n$ can be thought of as a *frequency map*: for each action value it gives a frequency vector. This map is nondegenerate if it is a local diffeomorphism, that is, its Jacobian is nonzero,

$$\det(D\Omega) = \det(D^2 H_o) \neq 0. \tag{9.60}$$

The implication of this condition is that a neighborhood of an action variable I_o is mapped one-to-one onto a neighborhood of the frequency $\Omega(I_o)$. In this case, incommensurate frequencies will occur for "almost all" actions: a set of full measure in frequency space. This is true even though commensurate values are still dense. (Recall that the rationals are dense in the reals, even though almost all reals are irrational.)

The nondegeneracy condition is stronger than that needed for many purposes because autonomous Hamiltonian systems are conservative, so their orbits lie on a particular $(2n-1)$-dimensional energy surface, $E = \{(q,p) : H(q,p) = E\}$. In this case the variation of the frequency in the direction transverse to the energy surface, ∇H, is irrelevant. Thus instead of the nondegeneracy condition (9.60), it is sufficient to require that the columns of the matrix $D\Omega$ span the $n-1$ vectors tangent to the energy surface, or equivalently that the $n \times (n+1)$ matrix $(D^2H_o, \nabla H_o)$ has rank n; this is called *isoenergetic nondegeneracy*. An alternative statement of this is

$$\det \begin{pmatrix} D^2H_o & \nabla H_o \\ DH_o & 0 \end{pmatrix} \neq 0. \tag{9.61}$$

In a system that satisfies (9.61) both resonant and nonresonant tori are dense on each energy surface.

KAM Theory

As we will see in §9.17, the resonant tori of an integrable system can be strongly affected by a perturbation of the form (9.57)—namely, they are often immediately destroyed upon perturbation. One of the most profound advances in Hamiltonian dynamics was the discovery by Andrei Kolmogorov in 1953 that "sufficiently" incommensurate tori are preserved upon perturbation. This result was formalized later by Vladimir Arnold (Arnold 1963) for analytic systems and by Jürgen Moser for sufficiently smooth systems (Moser 1962). The results of Kolmogorov, Arnold, and Moser now go by the name of *KAM theory*. For nice expositions of this complex theory see (de la Llave 2001; Pöschel 2001).

There are only two important concepts of KAM theory that we will focus upon. The first is that while the theory asserts that "many" incommensurate tori are preserved, it actually requires that neighborhoods of all commensurate tori be excluded. However, since rationals are dense, it is a delicate matter to exclude the neighborhood of every rational from consideration and not exclude everything! Luckily the neighborhood that KAM excludes has a width that decreases with the magnitude of the integer vector m: $|m \cdot \omega| < c\,|m|^{-\tau}$. Vectors that are not in these resonant neighborhoods are called

> ▷ *Diophantine frequencies.* A vector ω is Diophantine if there is a $c > 0$ and $\tau > n{-}1$ such that $\omega \in \mathscr{D}_{c,\tau} = \left\{\omega \in \mathbb{R}^n : |m \cdot \omega| \geq c\,|m|^{-\tau} \;\; \forall m \in \mathbb{Z}^n \setminus \{0\}\right\}$.

The set $\mathscr{D}_{c,\tau} \cap \mathbb{S}^{n-1}$ is a Cantor set when $\tau > n-1$, and $c \neq 0$. Moreover, as $c \to 0$, the measure of this Cantor set approaches one (Cassels 1957).

Example 9.36. Consider the case $n = 2$ and a vector (ω_1, ω_2) with $0 < \omega_1 < \omega_2$. Let $|m| = \max(|m_1|, |m_2|)$ be the sup-norm of m. Since we wish to find vectors nearly perpendicular to ω, we can assume that $m = (q, -p)$ with $0 < p < q$. Then the Diophantine condition becomes a condition on the frequency ratio:

$$\left|\frac{\omega_1}{\omega_2} - \frac{p}{q}\right| > \frac{d}{2|q|^{\tau+1}}$$

with $d = 2c/\omega_2$. Thus we are looking for irrational numbers $x \in (0,1)$ that are bounded away from rationals p/q. The set $\mathscr{D}_{c,\tau}$ corresponds to what remains after excluding an interval of width $d\big/|q|^{\tau+1}$ about each rational in $[0,1]$. Start by excluding the intervals $[0, d/2)$ and $(1-d/2, 1]$ and then the interval of width $d\big/2^{\tau+1}$ about $1/2$, etc. The total length of the excluded intervals in $\mathscr{D}$ is then

$$L = d + \sum_{q=2}^{\infty} \phi(q) \frac{d}{|q|^{\tau+1}},$$

where $\phi(q)$ is the number of integers in $[1, q]$ that are coprime with q—Euler's totient function. This function cannot be computed explicitly, but certainly $\phi(q) \le q-1$ (where equality occurs only when q is prime). Thus L is bounded by

$$L < d \sum_{q=1}^{\infty} \frac{1}{q^{\tau}}.$$

The sum is convergent whenever $\tau > 1$. Thus as $c \to 0$, the excluded length goes to zero and the measure of the Diophantine frequency ratios in $[0, 1]$ approaches 1. ■

A key concept in Kolmogorov's theory is the identification of invariant tori by their frequencies: instead of trying to understand the orbit of a particular point in phase space, he chooses a Diophantine frequency vector ω and follows the torus with that frequency as ε grows. The question then becomes, does the flow of (9.57) have an invariant n-dimensional torus with a given frequency for some $\varepsilon \neq 0$? One version of the answer is as follows.

Theorem 9.37 (KAM (Pöschel 1982)). *Suppose that $H_0(I)$ is real analytic and nondegenerate and suppose that $H_1(\theta, I)$ is C^r with $r > 2n$. Then there is a constant $\alpha > 0$ such that if $\varepsilon < \alpha c^2$ the system* (9.57) *has invariant tori for all $\omega \in \mathcal{D}_{c,n}$ in the range of the frequency map Ω. Similarly if an energy surface $\mathcal{E}$ is isoenergetically nondegenerate, then there are tori whose frequencies are proportional to each $\omega \in \mathcal{D}_{c,n}$ in the range of Ω.*[70] *In both cases, the dynamics on each torus is smoothly conjugate to the flow* (9.58).

The preservation of n-dimensional tori with Diophantine frequencies (often called *KAM tori*) is more than a mathematical curiosity; these tori can be easily observed in many examples; see Figure 9.12. This came as a surprise to many scientists and mathematicians. Indeed, one of the earliest of numerical experiments—by Enrico Fermi, John Pasta, and Stanislaw Ulam (FPU) on the MANIAC-I computer at Los Alamos in 1954—was an attempt to measure the "thermalization" of energy in the modes of a nonlinear string. They believed that nonlinear coupling of the linear normal modes would lead to the spread of energy to all the modes on the energy surface, and thus the motion of a perturbed integrable Hamiltonian should be topologically transitive or ergodic on the energy surface; recall §7.1. While it is difficult to directly apply the KAM theorem to the FPU system,[71] Theorem 9.37 certainly implies that "thermalization" is not a typical property of weakly perturbed, integrable Hamiltonian systems: most of the trajectories are confined to n-dimensional tori and do not wander densely through the energy surface.

However, KAM theory applies only for "small enough" ε, and explicit estimates of the necessary bound are difficult to obtain and often are very small indeed. As we will see next, numerical experiments show that some invariant tori are preserved for quite large values of ε for specific choices of H_1.

9.14 ▪ Onset of Chaos in Two Degrees of Freedom

The dynamics of a one-degree-of-freedom Hamiltonian system are easy to visualize since the phase space is two-dimensional. Of course, this case is also rather trivial because of the conservation of energy. The study of the motion of Hamiltonian systems

[70] That is, the frequency *ratio* is fixed.

[71] For a fascinating account of the history, see (Weissert 1997).

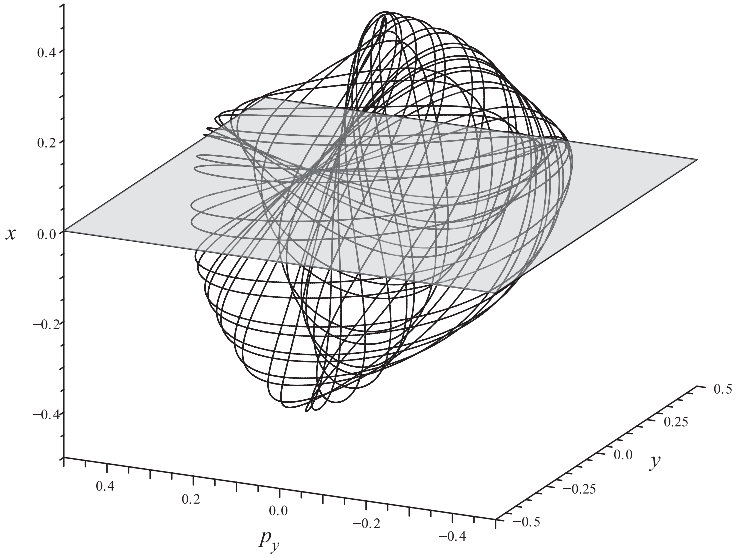

Figure 9.12. *Three-dimensional projection onto (x, y, p_y) of an invariant torus for the two-degree-of-freedom Hénon–Heiles system* (9.66) *with initial condition* $(0, -0.15, 0.376, 0.0)$ *so that $E = 1/12$. Also shown is a section at $x = 0$. This plot is obtained using Maple; see the appendix.*

with more degrees of freedom is difficult because their phase spaces have four or more dimensions. The case of two degrees of freedom can be quite effectively visualized using a Poincaré section; recall §4.12. This works because each orbit lies on a three-dimensional energy surface, $\mathscr{E} = \{z : H(z) = E\}$, and so restricting to a set of orbits on $\mathscr{E}$, a cross section to the Hamiltonian flow $\varphi_t(z)$ is a two-dimensional surface; see Figure 9.12.

Recall that for a section S the Poincaré return map is defined as

$$z' = P(z) = \varphi_{\tau(z)}(z), \tag{9.62}$$

where $\tau(z)$ is the first time that the orbit of $z \in S$ returns to S. When S is a two-dimensional surface, the dynamics of P is easy to visualize numerically.

However, the construction of a section is complicated by the fact that $\mathscr{E}$ is typically not Euclidean but rather is a manifold.

Example 9.38. For the harmonic oscillator Hamiltonian (9.49), $\mathscr{E}$ is the set

$$\sum_{j=1}^{n} \omega_j \left(p_j^2 + q_j^2\right) = 2E,$$

when $\omega_j > 0$, which is topologically the sphere $\mathbb{S}^{2n-1}$. More generally, consider the Hamiltonian

$$H(q,p) = \tfrac{1}{2}p^2 + V(q),$$

where $V(q)$ is a periodic function of q: $V(q + 2\pi m) = V(q)$ for each $m \in \mathbb{Z}^n$. Thus the phase space for H is $\mathbb{T}^n \times \mathbb{R}^n$. Suppose that V has a global minimum V_m at some point q_m and a global maximum V_M at q_M. When E is near its minimum value, $E = H(0, q_m) = V_m$, q is confined to a neighborhood of q_m. For this neighborhood V is approximately quadratic, $V(q) \approx V_m + q^T W q$, where W is a positive-definite matrix. This implies that when $E = V_m + \varepsilon$, $\mathscr{E}$ is (topologically) the sphere $\mathbb{S}^{2n-1}$, just as for the harmonic oscillator. The topology of the energy surface will typically change when E reaches the next lowest critical point of V because a larger range of the configuration variables becomes accessible. (This is especially true when the periodicity of the configuration comes into play.) When $E > V_M$, all configurations are accessible, so the energy surface includes all the configuration space $\mathbb{T}^n$. However, not all momenta are possible: indeed for each value of q, conservation of energy implies that

$$p^2 = 2(E - V(q)).$$

Consequently, for each q, the momentum is confined to the sphere $\mathbb{S}^{n-1}$. The radius of this sphere varies with q but is always nonzero when $E > V_M$. Consequently, $\mathcal{E} \cong \mathbb{T}^n \times \mathbb{S}^{n-1}$. The topology of $\mathcal{E}$ for values of E below V_M depends in detail on the critical points of V; see Exercise 18. ■

To construct a Poincaré section for a two degree-of-freedom system, $H(q_1, q_2, p_1, p_2)$, we would like to choose a two-dimensional surface in $\mathcal{E}$ that is a global section for the flow. Recall from §4.12 that a surface S is a section if the vector field is nowhere tangent to S and is a global section if the orbit of every point crosses S and returns.

It is often difficult to prove that a particular section is global, so in the interest of expediency, we will appeal to physical intuition. For example, in a system based on nonlinear oscillators, it is often true that the configuration variables oscillate about zero, and so one possible choice for a section is the surface for which one of these configurations vanishes, say,

$$\mathcal{Q} = \{(q, p) : q_2 = 0\}. \tag{9.63}$$

As we focus on an energy surface, $\mathcal{E}$, a candidate for the section is the intersection $S = \mathcal{E} \cap \mathcal{Q}$. However, this surface is typically not a section because the vector field is not everywhere transverse to S.

Example 9.39. Let

$$H = \tfrac{1}{2}(p^2 + q^2) + q_1^2 q_2^2 \tag{9.64}$$

and let $\mathcal{Q}$ be given by (9.63). For any $E > 0$, $S = \{(q_1, 0, p_1, p_2) : p_1^2 + p_2^2 + q_1^2 = 2E\}$ is a two sphere. Convenient coordinates on S could be (q_1, p_1), but both hemispheres $p_2 > 0$ and $p_2 < 0$ project onto the same disk. Indeed since $\dot{q}_2 = p_2$ the vector field is not transverse to the section at $p_2 = 0$. ■

To ameliorate this problem the section must be a subset of $\mathcal{E} \cap \mathcal{Q}$ to which the vector field is transverse. Since $\dot{q}_2 = -\partial H / \partial p_2$, the vector field will be transverse to $\mathcal{Q}$ whenever $\partial H / \partial p_2 \neq 0$. For example, we can choose as a section

$$S = \mathcal{E} \cap \mathcal{Q} \cap \{\partial H / \partial p_2 > 0\}. \tag{9.65}$$

Typically this set is an open disk and is bounded by a loop on which the vector field is tangent to the section.

Example 9.40. For (9.64), the section (9.65) is

$$S = \left\{(q_1, 0, p_1, p_2) : p_1^2 + q_1^2 = 2E - p_2^2, p_2 > 0\right\}.$$

This set is the "northern hemisphere" of a two-sphere, and it projects into the interior of the disk of radius $\sqrt{2E}$ in the (q_1, p_1) plane. Note that for each (q_1, p_1) in this disk, there is a unique $p_2 > 0$ in S, and since $q_2 = 0$ we know the full initial condition of the trajectory. Thus (q_1, p_1) is a good set of coordinates on S.

The boundary of the section is the circle $\{p_1^2 + q_1^2 = 2E, p_2 = q_2 = 0\}$. This is an invariant set of (9.64); indeed it corresponds to a periodic orbit. ∎

This example illustrates the best scenario we can expect: the section (9.65) is a disk whose boundary is an invariant set. Indeed, Birkhoff showed that when there is no globally transverse section, a necessary condition for the Poincaré map to be smooth is that the boundary of S be an invariant set (Dullin and Wittek 1995).

Example 9.41 (Hénon–Heiles Hamiltonian). In 1964 Michael Hénon and Carl Heiles were studying the motion of individual stars in the collective gravitational potential of the remaining stars in a galaxy (Hénon and Heiles 1964). On the scale of a galaxy, a star can be treated as a point mass and its motion is governed by a Hamiltonian of the form $H = p^2/2m + V(x, y, z)$ with $p \in \mathbb{R}^3$. Many galaxies have an (approximate) axisymmetry so that $V = V(\rho, z)$, where $\rho = x^2 + y^2$ is the cylindrical radius. This symmetry implies the conservation of the z-component of angular momentum; recall (9.36). Thus the three degree-of-freedom model has two invariants, energy, and angular momentum. Hénon and Heiles wished to address the question of existence of a third invariant. To study this they noted that the conserved angular momentum can be used to reduce the three degree-of-freedom model to one with two degrees of freedom. They studied the simplified two degree-of-freedom model

$$H = \tfrac{1}{2}(p_x^2 + p_y^2 + x^2 + y^2) + x^2 y - \tfrac{1}{3}y^3, \tag{9.66}$$

which does not have direct astronomical origin but could be thought of as a typical model for motion near an elliptic equilibrium; however, it has the special feature that the linear oscillators have the same frequency: it is in one-to-one resonance (Rod and Churchill 1985).

The ODEs for (9.66) are

$$\begin{aligned} \dot{x} &= p_x, & \dot{y} &= p_y, \\ \dot{p}_x &= -x - 2xy, & \dot{p}_y &= -y - x^2 + y^2. \end{aligned} \tag{9.67}$$

There are equilibria at the origin, where $E = 0$, and at $(0, 1, 0, 0)$ and $(\pm\sqrt{3}/2, -1/2, 0, 0)$, where $E = 1/6$. The origin is elliptic, and the remaining three points are saddles. Note that this system is reversible (recall §6.4) with the reversor $R(q, p) = (q, -p)$.

Hénon and Heiles used the section

$$S = \mathscr{E} \cap \{x = 0\} \cap \{p_x > 0\}. \tag{9.68}$$

When $0 < E < 1/6$, the energy surfaces have two components, one bounded and the other unbounded; the bounded component of (9.68) is a disk whose boundary is the curve $3p_y^2 + 3y^2 - 2y^3 = 6E$. This curve is invariant since if we set $x(0) = p_x(0) = 0$ in (9.67), then they remain zero. When E is small the bounding curve is nearly circular and most of the orbits appear to lie on invariant tori. An example of a three-dimensional projection of one trajectory was already shown in Figure 9.12. This orbit intersects plane $x = 0$ in the figure with both $p_x > 0$ and $p_x < 0$; only the former intersections correspond to the section (9.68). However, by reversibility the trajectories with initial conditions $(0, y, p_x, p_y)$ with $p_x < 0$ are equivalent to trajectories that start at a point $(0, y, -p_x, -p_y)$ integrated backward in time. Thus there is no need to restrict the intersections to those with positive p_x. The intersections of the trajectory

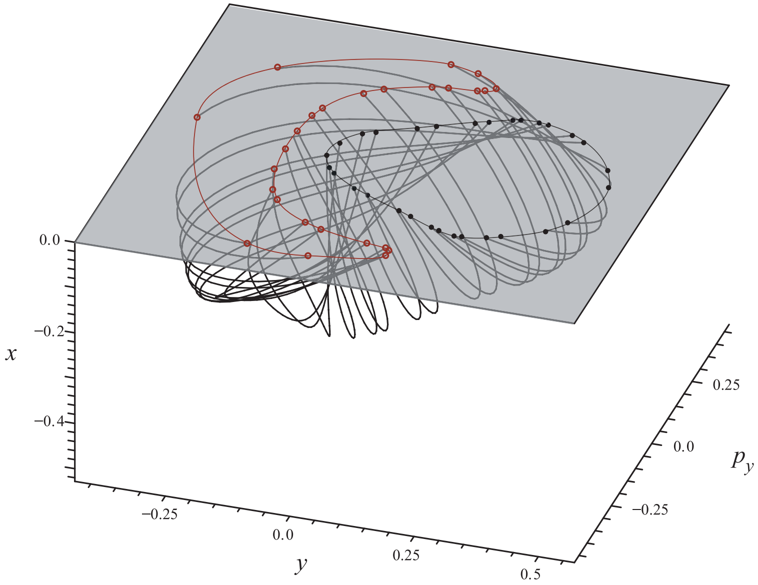

Figure 9.13. *The intersections of the trajectory of Figure* 9.12 *with the plane* $\{x = 0\}$ *appear to trace out two curves. The black dots correspond to the points with* $p_x > 0$ *and the red dots to* $p_x < 0$.

of Figure 9.12 with the plane $\{x = 0\}$ are shown in Figure 9.13. The set of intersections appears to fall on two curves, one for which the crossing is from above $x = 0$ to below ($p_x < 0$) and the other from below to above. These two curves illustrate the intersection of an invariant two-torus with S.

The Poincaré map for the section (9.68) can be easily plotted using computer algebra tools; see the appendix. When $E = 1/12$, many of the orbits on the Poincaré section in Figure 9.14 appear to cover circles, indicating that the orbits lie on invariant two-tori. There also appears to be a "period-three" saddle on the section, and a few points on its stable and unstable manifolds are shown. These manifolds enclose various invariant tori. When the energy is increased to $E = 1/8$ in Figure 9.15, the trajectories near this saddle appear to no longer lie on smooth curves. This is an indication that the KAM tori have been destroyed and are replaced by chaotic dynamics. Indeed, computations of the Lyapunov exponents for these trajectories show that they exhibit sensitive dependence; see Exercise 19. ■

Poincaré sections of the form (9.65) are usually visualized by projecting them onto the canonical pair of variables (q_1, p_1). One reason that this is a nice coordinate system for S is that the resulting return map is *area preserving*. This follows from the conservation of the Poincaré loop action (9.24). Let $\mathscr{L}$ be a loop on the section so that every point on $\mathscr{L}$ has energy E. Then, since $\oint_{\mathscr{L}} H dt = E \oint_{\mathscr{L}} dt$, and the integral of a perfect differential around a closed loop is zero, this term in the action vanishes. For the section (9.63), $q_2 \equiv 0$ on $\mathscr{L}$, and pdq reduces to $p_1 dq_1$ so that the loop action becomes

$$S[L] = \oint_{\mathscr{L}} p_1 dq_1,$$

which is simply the area enclosed by $\mathscr{L}$ in the (q_1, p_1) plane. Since $S[\mathscr{L}]$ is preserved

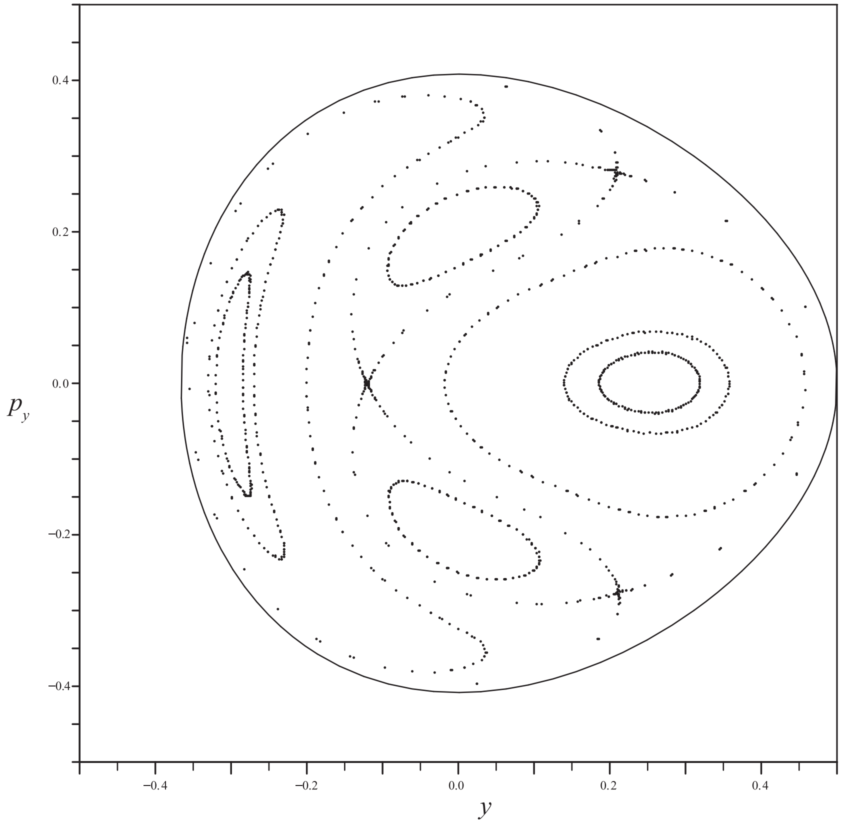

Figure 9.14. *Poincaré section of the Hénon–Heiles Hamiltonian* (9.66) *with* $E = 1/12$, *plotted using the code in the appendix.*

along trajectories, then the transformed loop $\mathscr{L}' = P(\mathscr{L})$ also lies on the section and has the same action; thus the Poincaré map preserves area. Consequently, the study of the dynamics of two degree-of-freedom Hamiltonian systems on energy surfaces (without critical points) essentially reduces to the study of area-preserving mappings—a beautiful subject in its own right (Meiss 1992).

9.15 ▪ Resonances: Single Wave Model

The planar pendulum provides a typical example of a locally integrable Hamiltonian system. Here we generalize this slightly and consider the nonautonomous system

$$H(q,p,t) = H_o(p) + a\cos(kq - \omega t), \tag{9.69}$$

where $(q,p) \in \mathbb{R}^2$. This system corresponds to a particle with kinetic energy $H_o(p)$ in a potential corresponding to a traveling wave with wavenumber k and frequency ω. Equation (9.69) describes physical systems such as the motion of a charged particle in an electrostatic wave.

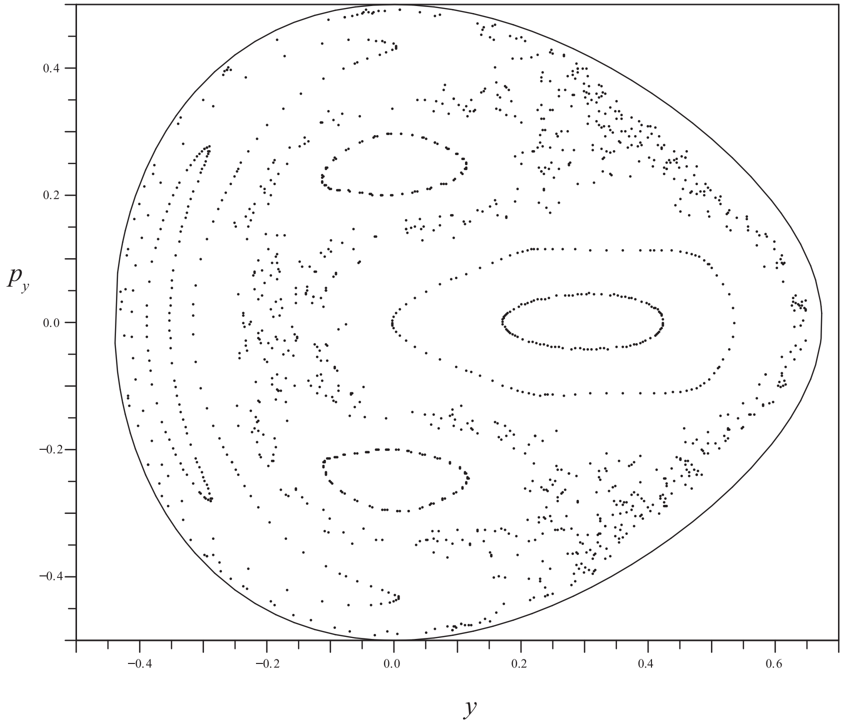

Figure 9.15. *Poincaré section of the Hénon–Heiles Hamiltonian* (9.66) *for* $E = 1/8$.

The Hamiltonian ODEs for (9.69) are

$$\begin{aligned} \dot{q} &= \frac{\partial H}{\partial p} = DH_o(p), \\ \dot{p} &= -\frac{\partial H}{\partial q} = ak\sin(kq - \omega t). \end{aligned} \tag{9.70}$$

First consider the case $a = 0$ where the momentum is constant. In this case the solution for the position is $q(t) = vt + q_o$, where $v = DH_o(p_o)$ is the constant velocity of the particle. Note that the Legendre condition $D^2H(p) \neq 0$ requires that $\partial v/\partial p \neq 0$, i.e., that the velocity changes with the momentum. Such systems are said to have *twist* or *shear*.

Now suppose that a is very small. The ODE for p leads to the expectation that p will change by $O(a)$, and then $q \approx vt + q_o + O(a)$. If this is correct to first order in a, this approximation for $q(t)$ can be substituted into the p equation to obtain

$$\dot{p} \approx ak\sin((kv - \omega)t + kq_o) + O(a^2).$$

The solution of this is

$$p(t) \approx c + \frac{ak}{\omega - kv}\cos((\omega - kv)t + kq_o), \tag{9.71}$$

providing $\omega - kv \neq 0$. In this case the assumption that p varies by a quantity $O(a)$ is valid. However, when the denominator vanishes, our approximation is not valid. The case that $v = \omega/k$ corresponds to the particle moving at the phase speed of the wave. It is known as *resonance*. If D^2H is nonzero, then the resonance equation

$$v = DH_o(p_R) = \frac{\omega}{k} \tag{9.72}$$

can be solved, according to the implicit function theorem, Theorem 8.3, to obtain the resonant momentum p_R. When the system is nearly resonant, i.e., when $\omega - kv = O(a)$, then the ordering that led to the assumption that p changes by $O(a)$ breaks down, and we must begin again.

To study the resonant case, we return to the nonautonomous system (9.70). The time dependence is simple enough that it can easily be transformed away by a Galilean coordinate transformation, i.e., by going to a frame moving with the phase speed ω/k of the wave. Upon defining $x = kq - \omega t$, (9.70) becomes

$$\begin{aligned} \dot{x} &= kDH_o(p) - \omega, \\ \dot{p} &= ak\sin(x). \end{aligned}$$

This system is also Hamiltonian[72] with a new Hamiltonian function

$$\hat{H}(x,p) = kH_o(p) - \omega p + ak\cos x. \tag{9.73}$$

Note that the new Hamiltonian is independent of time and so is conserved; physically it is related to the particle energy in the moving frame. Thus the motion can be completely characterized by graphing the contours of $\hat{H}$. The details of these depend upon $H_o(p)$; however, the contours near the resonant case can by understood by expanding for p near p_R. Upon defining a new momentum $y = p - p_R$, the first two terms in (9.73) expand to

$$kH(p_R+y) - \omega(p_R+y) = kH_o(p_R) - \omega p_R + (kDH(p_R) - \omega)y + \frac{k}{2}D^2H(p_R)y^2 + o(y^2).$$

The terms proportional to y vanish because of the resonance condition (9.72). If we drop the constant terms, since they do not affect the equations of motion, we obtain

$$\tilde{H}(x,y) = \frac{1}{2M}y^2 + ak\cos x, \tag{9.74}$$

where we assume that $D^2H_o(p_R) \neq 0$ so that the effective mass

$$M = \frac{1}{kD^2H_o(p_R)} \tag{9.75}$$

is finite. Equation (9.74) is *exactly* the pendulum Hamiltonian (recall (9.11))! The dynamics follows contours of the new energy $\tilde{H}$, as sketched in Figure 9.16.

The pendulum has two equilibria, the two critical points of $\tilde{H}(x,y)$, at $(0,0)$ and $(\pi,0)$. The Hessian matrix of $\tilde{H}$ and corresponding Hamiltonian matrix are

$$S = D^2\tilde{H} = \begin{pmatrix} -ak\cos x & 0 \\ 0 & M^{-1} \end{pmatrix}, \quad K = JS = \begin{pmatrix} 0 & M^{-1} \\ ak\cos x & 0 \end{pmatrix}.$$

[72] The theory of canonical transformations shows how to do this transformation on the Hamiltonian itself; see (Goldstein et al. 2002).

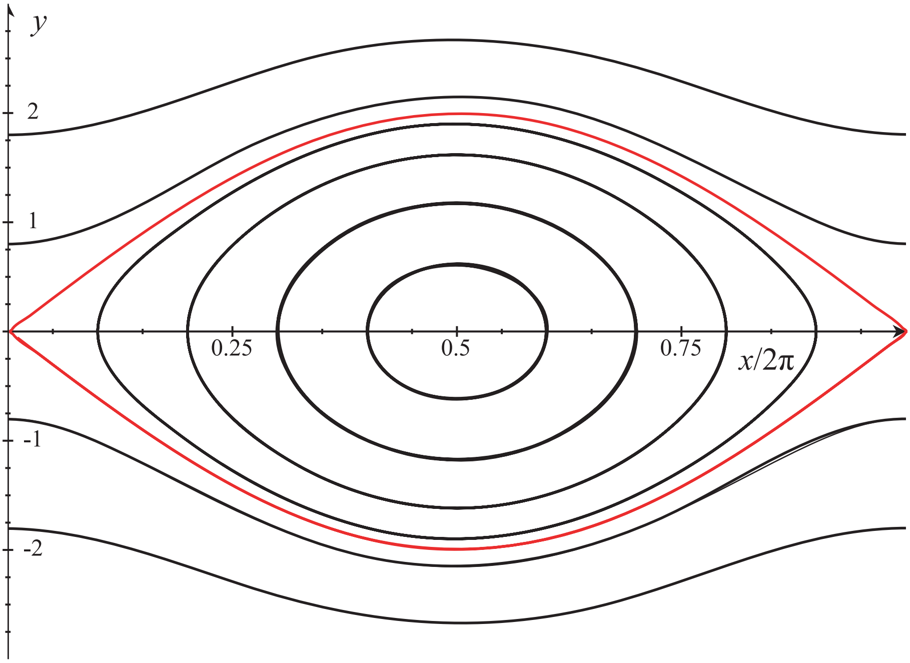

Figure 9.16. *Orbits of the pendulum Hamiltonian* (9.74) *for* $M = ka = 1$.

Thus at $x = 0$ the eigenvalues of K are $\lambda = \pm\sqrt{\frac{ka}{M}}$ and so $(0,0)$ is a saddle, while at $x = \pi$, the eigenvalues are $\lambda = \pm i\sqrt{\frac{ka}{M}}$ and so $(\pi, 0)$ is a center.

The stable and unstable manifolds of the saddle correspond to the contours of the level set $\tilde{H}(x,y) = \tilde{H}(0,0) = ak$. For this energy (9.74) can be solved for $y(x)$ to give

$$y_\pm(x) = \pm\sqrt{2Mak(1-\cos x)} = \pm 2\sqrt{Mak}\sin(x/2). \tag{9.76}$$

Thus the maximum and minimum of y on these contours occur at $x = \pi$, where $y_\pm(\pi) = \pm 2\sqrt{Mka}$. The stable and unstable manifolds form the *separatrix*, the red curve in Figure 9.16: it separates the contours that correspond to *trapped* (or *librating*) motion near the center from those that correspond to untrapped (or *rotating*) motion, where x is monotone increasing or monotone decreasing. The set of trapped trajectories is also known as a *resonance*.

Note that the width of the resonance, $y_+(\pi) - y_-(\pi)$, is proportional to $\sqrt{a}$. When a is small, $\sqrt{a} \gg a$. This is what was wrong with our assumption that p varies by $O(a)$ in the derivation of (9.71) near to the resonance.

Transforming back to the original coordinates, (q, p), the resonance now corresponds to a region centered at $p = p_R$ with upper and lower bounds $p_\pm = p_R + y_\pm(\pi)$; see Figure 9.17. The resonance width is

$$\Delta p = y_+(\pi) - y_-(\pi) = 4\sqrt{\frac{a}{D^2 H_o(p_R)}}. \tag{9.77}$$

Particles trapped in the resonance stay near $x = 0$, so they have a mean motion $q(t) \approx \omega t/k$ plus an oscillation about this line. Hence these particles are trapped in the traveling wave.

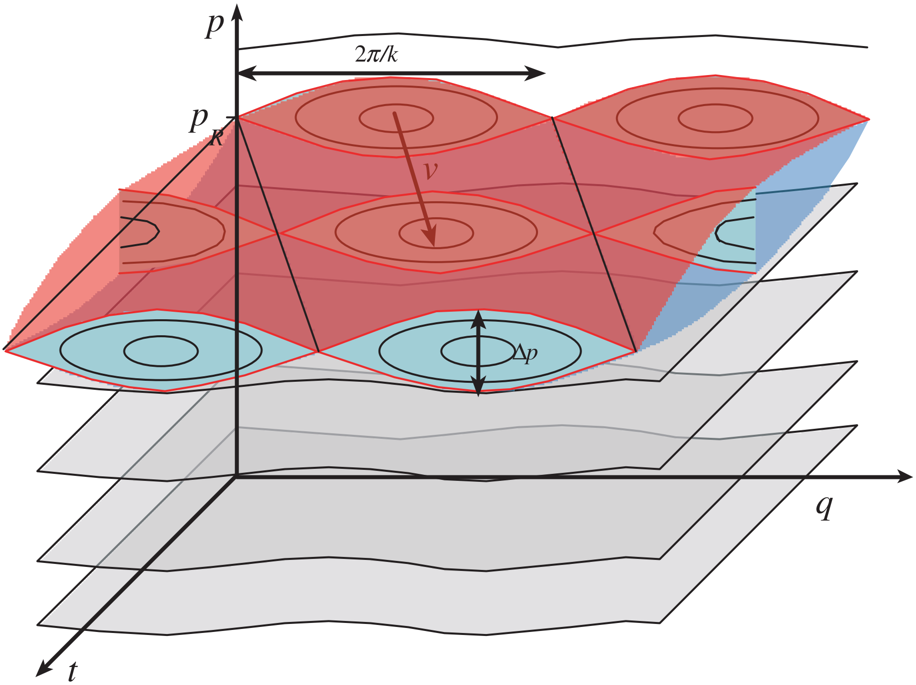

Figure 9.17. *Extended phase space for the Hamiltonian* (9.78) *when there is only one resonance.*

9.16 ▪ Resonances: Multiple Waves

More generally, suppose that $H(q,p,t)$ is C^2 and a periodic function of q with spatial period $L = 2\pi/k$ and a periodic function of t with temporal period $T = 2\pi/\omega$. The Hamiltonian can then be expanded in a double Fourier series, allowing for an arbitrary dependence upon p:

$$H(q,p,t) = H_o(p) + \sum_{l,m \neq 0} a_{lm}(p)\cos(k_l q - \omega_m t + \theta_{lm}). \tag{9.78}$$

Here $a_{lm}(p)$ is the (l,m)th Fourier amplitude, $\theta_{lm}(p)$ is its phase, $k_l = lk$, and $\omega_m = m\omega$. Each of the terms in the sum corresponds to a "wave" and can cause a resonance at the appropriate momentum value. For example, if only one of the waves, that of mode numbers l and m, is nearly resonant, then

$$H_{lm}(p,q,t) = H_o(p) + a_{lm}\cos(k_l q - \omega_m t - \theta_{lm})$$

is the Hamiltonian that dominates the dynamics. As in §9.15, this system can be transformed into a moving frame to eliminate the time dependence, setting $x = k_l q - \omega_m t$. The biggest excursions in p occur near the resonant momentum, p_{lm}, where the function $k_l H_o(p) - \omega_m p$ has a critical point. Expanding about the resonant momentum and assuming that a_{lm} is small, it is appropriate to keep only the lowest-order terms $a_{lm}(p_{lm})$ and $\theta_{lm}(p_{lm})$. In this case, the system reduces to the pendulum in the neighborhood of the resonance,

$$\tilde{H}(x,y) \approx \frac{1}{2M}y^2 + a_{lm}k_l\cos(x + \theta_{lm}), \tag{9.79}$$

where $M = (k_l D^2 H_o(p_{lm}))^{-1}$, $x = k_l q - \omega_m t$, and $y = p - p_{lm}$.

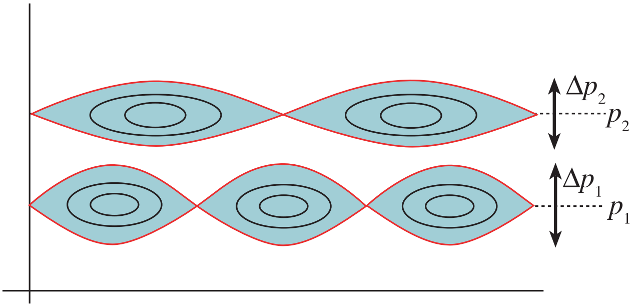

Figure 9.18. *The overlap criterion.*

9.17 ▪ Resonance Overlap and Chaos

The Russian physicist Boris Chirikov realized that the single resonance approximation that was developed above might be reasonable when resonances are far apart, but that it must break down when neighboring resonances overlap (Chirikov 1979). In his view, this phenomenon is responsible for the onset of chaos in Hamiltonian systems—or at least in typical cases, since exceptions in both directions to this statement can be found.

Consider, for example, $H_o(p) = p^2/2m$. The resonant momenta are the solutions of

$$DH_o(p_R) = \frac{p_R}{m} = \frac{\omega_m}{k_l}.$$

If there are finitely many Fourier modes, then the resonances have a nonzero spacing. Suppose that p_1 and p_2 correspond to the locations of two neighboring resonances and that their corresponding widths are Δp_1 and Δp_2; see Figure 9.18. The resonances are relatively independent of one another when they are far apart compared to the sum of their half-widths. Chirikov defined the overlap parameter

$$s_{12} = \frac{1}{2}\frac{\Delta p_2 + \Delta p_1}{p_2 - p_1}. \tag{9.80}$$

When $s_{12} \ll 1$, the resonances are far apart, and the momentum should vary by $O(a)$ when away from the resonances and by $O(\sqrt{a})$ when trapped in one resonance.

The single-resonance approximation should break down when $s_{12} \approx 1$. In this case a particle trapped in the first resonance could make a transition to the second resonance. In this way the average velocity of the particle could change from the phase speed of the first wave to that of the second. If a set of neighboring resonances overlap, then p could drift by an amount that is large compared to the typical resonance amplitude, $O(\sqrt{a})$. This is indeed observed numerically. Moreover, this drift can even look like a diffusive process—the motion switches from one resonance to the other in a seemingly random fashion (Meiss 1992). The overlap criterion, while crude, typically gives an estimate that is within a factor of three of the onset of "connected" chaos in these systems, i.e., chaos that allows the momentum to drift from resonance to resonance. The estimate works best when the resonances have comparable amplitudes; indeed, the overlap criterion fails completely when one of the resonance amplitudes is zero, since there is no second resonance in the system at all.

There are many refinements to the resonance overlap criterion. For example, with perturbation theory one can compute the "secondary" resonances that arise from the nonlinear beating between primary resonances. These secondary resonances reduce

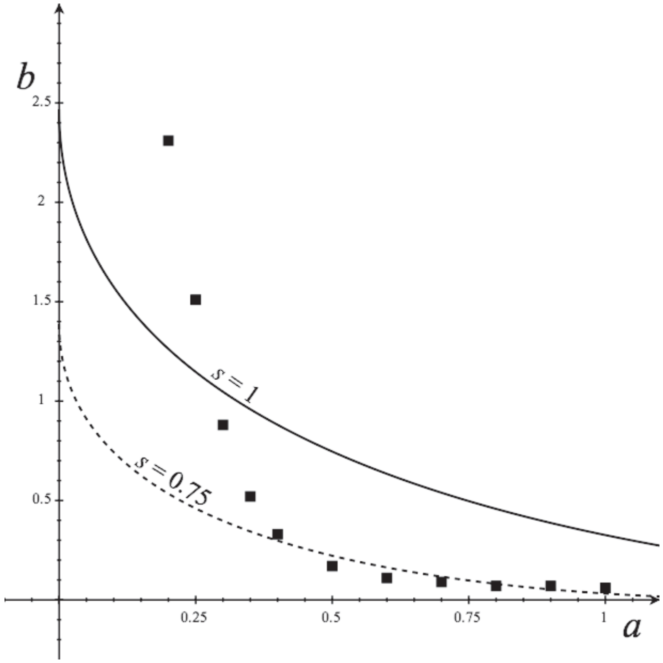

Figure 9.19. *Overlap criterion for the two-wave Hamiltonian* (9.81). *The two curves show* $s=1$ *and* $s=0.75$ *from* (9.82). *The boxes are numerical thresholds for connected chaos.*

the effective distance between resonances (Lichtenberg and Lieberman 1992). Indeed, a rule of thumb is that connected chaos occurs when $s_{12} \approx 2/3$ instead of 1, due to this mechanism. A more sophisticated version of this is "renormalization theory" that takes into account the creation of infinitely many secondary resonances (Escande 1985; MacKay 1993).

Example 9.42. The two-wave model is given by the Hamiltonian

$$H(q,p,t)=\frac{p^2}{2}-\frac{1}{4\pi^2}\big(a\cos(2\pi q)+b\cos(2\pi(2q-t))\big). \tag{9.81}$$

The first resonance is at $p_1 = 0$ and has a resonance width, from (9.77), of $\Delta p_1 = 4\sqrt{a/(2\pi)^2} = \frac{2}{\pi}\sqrt{a}$. The resonant momentum of the second resonance is $p_2 = \omega/k = 2\pi/4\pi = 1/2$, and its width is $\Delta p_2 = \frac{2}{\pi}\sqrt{b}$. Thus the resonance overlap parameter is

$$s=\frac{1}{2}\frac{\Delta p_1+\Delta p_2}{p_2-p_1}=\frac{2}{\pi}\left(\sqrt{a}+\sqrt{b}\right). \tag{9.82}$$

Empirically, it is found that $s = 1$ gives only a rough estimate of the onset of connected chaos; see Figure 9.19. The boxes in the figure are computed numerically by looking for an orbit that begins near the hyperbolic periodic orbit corresponding to one resonance and moves to a region near the hyperbolic orbit corresponding to the second resonance. Simulations of this system for several parameter values are shown in Figure 9.20. These figures are Poincaré sections of the extended phase space of the nonautonomous system with the section defined by $t = 0 \bmod 1$; in other words, a

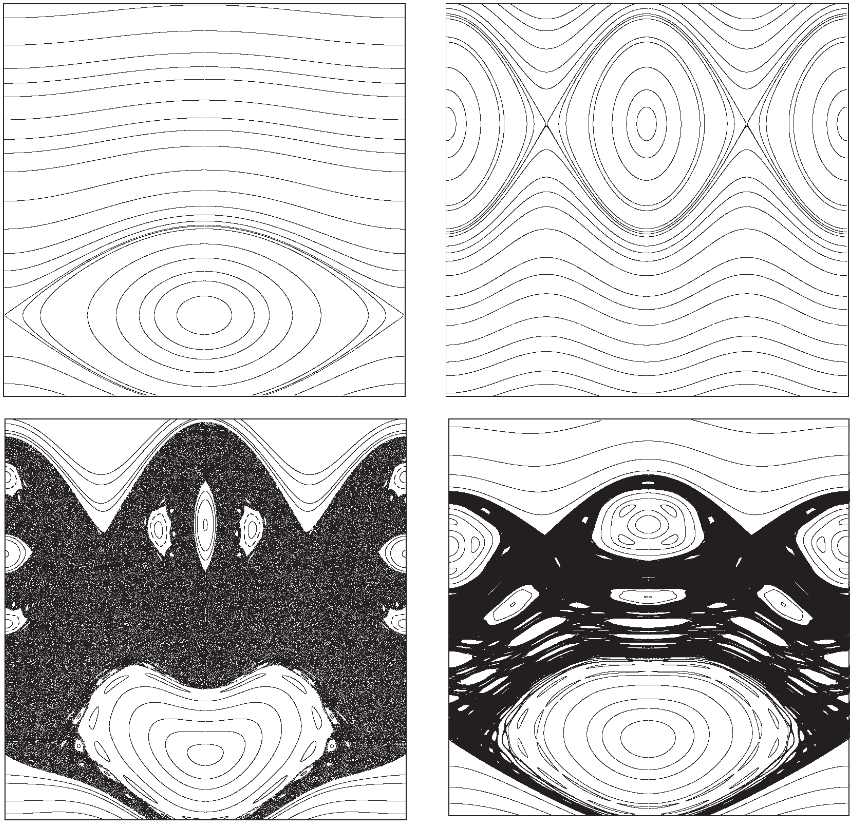

Figure 9.20. *Four stroboscopic plots in the two-resonance system* (9.81) *with* $(a,b) = (0.5,0)$, $(0,0.75)$ *on the top row and* $(0.5,0.75)$ *and* $(0.5,0.17)$ *on the bottom row. The overlap parameter* (9.82) *is one in the bottom left panel; however, connected chaos occurs at smaller parameter values due to resonance islands that are caused by nonlinear beating, as one in the bottom right, where* $s = 0.71$.

point is plotted on a trajectory at each integer time. The top two plots show the individual resonances with amplitudes chosen so that if the two plots were overlaid, the individual resonances would be just touching, thus giving $s = 1$. Both resonances are active in the bottom plots; the left plot for $s = 1$ shows a large chaotic region encompassing both resonances. In the bottom right plot, $s = 0.71$ and the chaotic region has just connected: for any smaller value of b there is a barrier to motion between the two resonances. ∎

9.18 ▪ Exercises

1. Let $x_i \in \mathbb{R}^3$ represent the positions of a system of N interacting particles with masses m_i and forces that depend only upon the interparticle distances $x_i - x_j$:

$$m_i \ddot{x}_i = \sum_{\substack{j=1\\ j\neq i}}^{N} f(x_j - x_i),$$

where $f : \mathbb{R}^3 \to \mathbb{R}^3$ is the force.

   (a) Show that the total momentum, $P = \sum_{i=1}^{N} m_i \dot{x}_i$, is an invariant if the force is odd: $f(-x) = -f(x)$.

   (b) Show that the total angular momentum, $L = \sum_{i=1}^{N} m_i \dot{x}_i \times x_i$, is an invariant if the force is directed along the interparticle separation: $f(x) = x g(|x|)$.

2. Show that the system of equations, defining Arnold's ABC flow, (1.16),

$$\begin{aligned}\dot{x} &= A\sin z + C\cos y,\\ \dot{y} &= B\sin x + A\cos z,\\ \dot{z} &= C\sin y + B\cos x,\end{aligned}$$

is volume preserving. Show that when $A = 0$ it has an invariant, $\psi(x,y,z) = B\cos x + C\sin y$. Discuss the phase portrait for this case.

3. Show that the equations of motion for the Hamiltonian (9.13) for a charged particle in an electromagnetic field are equivalent to the Lorentz force law (9.12). (*Hints*: Let the ith component of the curl be denoted

$$B_i = (\nabla \times A)_i = \sum_{j,k=1}^{3} \varepsilon_{ijk} \partial A_j / \partial x_k,$$

where ε_{ijk} is the completely antisymmetric symbol. Use the identity

$$\sum_{i=1}^{3} \varepsilon_{ijk}\varepsilon_{ilm} = \delta_{jl}\delta_{km} - \delta_{jm}\delta_{kl}.)$$

4. Prove Lemma 9.11. (*Hint*: Consider the action of L on the monomial basis x^n. Use the fact that every function in $C^1([a,b],\mathbb{R})$ can be approximated arbitrarily closely in the sup-norm by a polynomial.)

5. Verify that the standard Poisson bracket (9.15) and the generalized Poisson bracket (9.21) for the rigid body satisfy the Jacobi identity.

6. (*Wave–wave interactions.*) The system (9.2) is a Poisson dynamical system but can be transformed into a one-degree-of-freedom Hamiltonian system using the invariants (9.3).

   (a) Show that the bracket

$$\{F, G\} = i\sum_{i=1}^{3} \frac{\partial F}{\partial a_i}\frac{\partial G}{\partial \bar{a}_i} - \frac{\partial F}{\partial \bar{a}_i}\frac{\partial G}{\partial a_i} \tag{9.83}$$

is a nondegenerate Poisson bracket for functions of $z = (a_1, a_2, a_3, \bar{a}_1, \bar{a}_2, \bar{a}_3)$, where these six variables are thought of as independent.

(b) Show that (9.2), with the complex conjugate equations for the amplitudes $\bar{a}_k$, can be written as a Poisson system, (9.16), using the bracket (9.83), for some $H(z)$.

(c) Show that the transformation $(a_k, \bar{a}_k) \to (\theta_k, J_k)$ defined by

$$a_j = \sqrt{J_j} e^{i\theta_k}, \quad \bar{a}_k = \sqrt{J_k} e^{-i\theta_k}$$

converts the bracket (9.83) into the standard, canonical bracket (9.15) for the canonical coordinates θ_k and momenta J_k. Thus show that the system (9.2) is Hamiltonian with a new Hamiltonian $\tilde{H}(\theta, J) = H(z(\theta, J))$.

(d) Show that the transformation $(\theta, J) \to (\psi, I)$, defined by

$$I = (J_1 + J_3, J_2 + J_3, J_3), \quad \psi = (\theta_1, \theta_2, \theta_3 - \theta_1 - \theta_2),$$

is *canonical* in the sense that the bracket in the new coordinates is still the canonical bracket. Thus show that the system is Hamiltonian with the new Hamiltonian $\hat{H}(\psi, I) = \tilde{H}(\theta(\psi), J(I))$.

(e) Show that the Hamiltonian $\hat{H}$ does not depend upon ψ_1 and ψ_2 (they are *ignorable* variables), thus verifying the invariance of I_1 and I_2 shown in (9.3).

(f) The system $\hat{H}$ effectively has only one degree of freedom (ψ_3, I_3), with parameters I_1, I_2, c, and $\Omega = \omega_3 - \omega_2 - \omega_1$. Use your favorite software to sketch contours of $\hat{H}$ to investigate the orbits and discuss the implications for the dynamics of wave–wave interactions.

7. Poisson systems often have a special type of invariant, called a *Casimir*: a nonconstant function $C \in C^1(M, \mathbb{R})$ such that $\{C, F\} = 0$ for any $F \in C^1(M, \mathbb{R})$. Thus Casimir invariants are associated with the Poisson bracket instead of the Hamiltonian.

   (a) Show that if C is a Casimir for the bracket (9.18), then the matrix $J(z)$ is singular.

   (b) Show that if $\dim(M)$ is odd, then, since a Poisson bracket is antisymmetric, it must have at least one local Casimir.

   (c) Show that the canonical bracket (9.15) does not have a Casimir.

   (d) The rigid body bracket, (9.21), has one Casimir, C. Find it.

   (e) Since both H and C are invariants for the Euler equations (9.20), the flow is restricted to lie on the curves of an intersection of an energy surface $H = E$ and a Casimir surface $C = c$. Assuming that $I_1 > I_2 > I_3 > 0$, describe the dynamics of the Euler equations.

8. (*Ignorable coordinates.*) Consider a Lagrangian for a mechanical system of the form $L(x, \dot{x}) = {}^1\!/\!_2 \dot{x}^T \rho(x) \dot{x} - V(x)$, where $\rho(x)$ is a positive-definite symmetric matrix.

   (a) Show that the energy $E = {}^1\!/\!_2 \dot{x}^T \rho(x) \dot{x} + V(x)$ is conserved.

(b) Suppose that E is independent of one of the coordinates, say, x_1. Show that the corresponding canonical momentum p_1 is conserved.

(c) Suppose $V(x) = V(r)$, and $\rho(x) = \rho(r)$, where r is the polar radius in $\mathbb{R}^3$. Convert the Lagrangian to polar coordinates (r, θ, ϕ). Show that the canonical momenta p_θ and p_ϕ are conserved.

9. (*Spring-pendulum.*) Consider a harmonic spring, with potential energy $V(x) = \frac{1}{2}k(|x| - L)^2$, hanging from a frictionless support in a constant gravitational field. Allow the spring to move in a two-dimensional, vertical plane. Assume that the spring can extend and compress, but not bend.

   (a) Obtain the Lagrangian for this system in a Cartesian coordinate system. Derive the Lagrangian equations of motion.

   (b) Find the Hamiltonian and the Hamiltonian equations of motion.

   (c) Transform the Lagrangian into polar coordinates. Show that the resulting Euler–Lagrange equations are (1.35).

   (d) Find the two equilibria of this system, and the eigenvalues of the linearization about each equilibrium. One of the equilibrium is elliptic (a center). Show that if the equilibrium length of the spring, L^*, is $4L/3$, then the two oscillation frequencies have the ratio 1:2.

   (e) Expand the Hamiltonian found in (b) about the linearly stable equilibrium, keeping terms through cubic order. Use this system to study the stability of the periodic, vertically oscillating solution $x = (0, L^* + a\cos\omega t)$. You should find that the linearized equations can be reduced to the Mathieu equation; see (Rusbridge 1979).

   (f) Get a spring and study this system experimentally. The dynamics is particularly interesting when the frequencies have the 1:2 ratio found in (d), for then the Mathieu equation has a positive Floquet exponent (recall §2.8).

10. The equations for the double spring system of Figure 1.4 were derived in §1.4 when the natural length of the springs was assumed to be zero. More generally, the potential energy of a harmonic spring is $V(x) = \frac{1}{2}k(|x| - L)^2$.

    (a) Obtain the Lagrangian and Hamiltonian for the system depicted in Figure 1.4 when the two springs have differing spring constants and natural lengths.

    (b) Find the equilibria of this system and study their eigenvalues as the parameters vary.

11. Consider the system (9.32) that represents the dynamics of a bead on an elliptical wire.

    (a) Show that the points $(s, v) = (n\pi, 0)$ are equilibria and determine the linearization of the dynamics about each.

    (b) Use the Legendre transformation to transform the Lagrangian (9.31) into Hamilton's form. Obtain the Hamiltonian equations. Check that these reduce to those of the planar pendulum when $a = b$.

    (c) Show that the equilibrium at $(0, 0)$ is a topological center and that the saddle points $(\pm\pi, 0)$ have two heteroclinic connections.

12. Consider a particle of mass m moving without friction that is constrained to lie on a two-dimensional surface specified by $z = Z(x,y)$.

    (a) Obtain the Lagrangian. Suppose that the kinetic energy is $T = {}^1\!/\!_2 m(\dot{x}^2 + \dot{y}^2 + \dot{z}^2)$ and the gravitational potential is mgz. Write the Lagrangian in the form of (9.39).

    (b) Derive Lagrange's equations. Solve for $(\ddot{x},\ddot{y})$ as functions of $(\dot{x},\dot{y},x,y)$. Note how complicated this is!

    (c) Suppose that $Z(x,y) = f(x-y)$. Show that the momentum $p = p_x + p_y$ is conserved. Here the momenta are defined by $p_x = \partial L/\partial \dot{x}$, etc. (*Hint*: It is much easier to use the fact that the Lagrangian has the form $L(x-y,\dot{x},\dot{y})$ and the basic Euler–Lagrange equations than to do this calculation with the equations you found in b.)

    (d) Specializing to the egg carton surface, $Z(x,y) = \cos x \;\cos y$, and using the energy in Exercise 8, argue that when $E < 0$, the particle is trapped forever in one cell of the carton.

13. A generalization of Noether's theorem, Theorem 9.21, implies that a Lagrangian that is independent of time has a conserved energy. Compute the total time derivative of $L(q(t),\dot{q}(t),t)$ along an orbit and use the Euler–Lagrange equations to show that if $\partial L/\partial t = 0$, then the quantity $E(q,\dot{q}) = \partial L/\partial \dot{q} \cdot \dot{q} - L$ is independent of time. For the case that E satisfies the Legendre condition, show that E takes the same value as H.

14. (*Kovalevskaya top.*) The Kovalevskaya top is a rigid body with moments of inertia $I = I_1 = I_2 = 2I_3$, supported at the origin with its center of mass at the point $(-a,0,0)$ in the plane of the equal moments. It can be described, using as coordinates the Euler angles (θ,φ,ψ), by a Lagrangian of the form $L = T - V$ with

$$T = \tfrac{1}{2}I\left[\dot{\theta}^2 + \sin^2\theta\,\dot{\varphi}^2 + \tfrac{1}{2}\left(\dot{\psi} + \cos\theta\,\dot{\varphi}\right)^2\right],$$
$$V = -mga\sin\theta\cos\psi.$$

    (a) Find the Hamiltonian for the Kovalevskaya top.

    (b) Show that there are two "obvious" invariants of this system due to the symmetry with respect to rotation in φ and to time translation.

    (c) Kovalevskaya found that there is a third invariant, given by

$$K = \left|\left(\sin\theta\,\dot{\varphi} - i\dot{\theta}\right)^2 + 2\frac{mga}{I}\sin\theta e^{i\psi}\right|^2.$$

    Use the equations of motion to show explicitly that K is an invariant.

15. (*Small oscillations.*) Consider the quadratic Hamiltonian

$$H(q,p) = \tfrac{1}{2}\left(p^T M^{-1} p + q^T V q\right), \tag{9.84}$$

where M and V are $n \times n$, symmetric matrices, and M is positive definite. The goal is to show that the equilibrium at $(0,0)$ has eigenvalues that come in pairs and are

    (i) $\pm i\omega_j$, pure imaginary for each positive eigenvalue of V, or

(ii) $\pm\mu_j$, real for each negative eigenvalue of V.

(a) Write down the Hamiltonian system of ODEs for (9.84) and the corresponding Hamiltonian matrix K.

(b) Look for solutions of the form $(q(t), p(t)) = e^{i\omega t}(\hat{q}, \hat{p})$. Show that the constant vector $\hat{q}$ must solve the equations

$$V\hat{q} = \omega^2 M\hat{q}. \tag{9.85}$$

Let $\lambda = \omega^2$. Show that λ is real. (*Hint*: Suppose λ were complex. Take the complex conjugate of (9.85). Multiply the original equation by $\bar{\hat{q}}^T$ and the new equation by $\hat{q}^T$ and subtract. Let $\hat{q} = \alpha + i\beta$, and show that positive definiteness of M implies that $\lambda = \bar{\lambda}$. This implies that $\omega = \pm\sqrt{\lambda}$ is either real—case (i), or pure imaginary—case (ii).)

(c) Show that since M is symmetric, its eigenvectors v_i are orthogonal, i.e., $v_i \cdot v_j = 0$ if $i \neq j$. Argue that since the eigenvalues m_i of M are positive, you can choose the norm of v_i such that $v_i^T M v_i = 1$. Let $A = (v_1, v_2, \dots, v_n)$. Show that by construction $A^T M A = I$. What is $A^T A$?

(d) Now convert (9.85) to a standard eigenvalue problem. Let $\hat{q} = Av$, and show that (9.85) becomes

$$A^T V A v = \lambda v.$$

Thus defining $W = A^T V A$ gives a standard eigenvalue problem for W. The spectrum of W determines the stability of the equilibrium.

(e) Since W is symmetric, there exists an orthogonal matrix O that diagonalizes $W, O^T W O = \Lambda$. Show that this implies that the matrix AO diagonalizes V (and also diagonalizes M).

(f) Thus for any q if $q = AOv$, then $q^T V q = v^T \Lambda v$. This is the "principal axis coordinate system." Show that this implies that Λ has the same number of positive elements as V has positive eigenvalues. Thus for example, if V is positive definite, then the equilibrium is stable.

16. Investigate the dynamics of the linear Hamiltonian system

$$H(q,p) = \tfrac{1}{2}\left(p_1^2 + q_1^2 - \omega p_2^2 - \omega q_2^2\right) + \varepsilon\left(p_1 q_2 - p_2 q_1\right)$$

as ε increases from zero. Consider especially the points $\omega = \pm 1$ where the eigenvalues collide on the imaginary axis. How does the behavior of this system correlate with the predictions from Theorem 9.32?

17. (*Resonance modules.*) Let $\omega \in \mathbb{R}^n$ be a frequency vector.

(a) Prove that the set $M = \{m \in \mathbb{Z}^n : \omega \cdot m = 0\} \subset \mathbb{Z}^n$ is a *module*, that is, a set that is closed under addition and multiplication by scalars $k \in \mathbb{Z}$. A module is basically a vector space, except that the scalars are taken from a *ring* rather than a *field*. This set is called the *resonance module*.

(b) Show that each module in $\mathbb{Z}^n$ has a basis consisting of $d \leq n$ integer vectors. Thus each resonance module has a dimension d, called the *multiplicity* of the resonance.

(c) Find the modules, a basis, and the multiplicity for the following frequency vectors:

$$(1,\sqrt{2},1), \quad (1,\sqrt{2},\sqrt{5}), \quad (\sqrt{2},2+\sqrt{2},5), \quad (\sqrt{2},2\sqrt{2},3\sqrt{2}).$$

(d) Discuss the flow (9.58) on invariant tori with the frequency vectors of (c).

18. Discuss the topology of the energy surfaces of the two degree-of-freedom Hamiltonian

$$H(q,p)=\tfrac{1}{2}p^2+a\cos(q_1)+b\cos(q_2+q_1)$$

on $\mathbb{T}^2 \times \mathbb{R}^2$ depending upon the values of the amplitudes a and b, as well as the energy, E. Plotting contours of the potential $V(q)$ may be helpful. Each energy surface on which there is a critical point of H may give rise to a change in the topology of E.

19. Using the methods of §7.2, compute the Lyapunov exponents for several trajectories of the Hénon–Heiles Hamiltonian that start on the section $x=0$ with $E=1/12$ and $1/8$. You should find that two of the Lyapunov exponents are zero, and the other two are paired, $\mu_3=-\mu_4$. Why?

20. (*Chaotic tumbling of Hyperion.*) Saturn's moon Hyperion is an irregularly shaped body with diameters of about $C=360$, $B=280$, and $A=215$ km. Moreover, its orbit has a relatively high eccentricity, $e=0.104$, in comparison with most of the larger bodies of the solar system. The combination of these two effects has been shown to lead to chaotic tumbling of Hyperion (Wisdom, Peale, and Mignard 1983). A simple model of the dynamics of the angle of orientation, x, of the semimajor axis of the satellite in a fixed elliptical orbit is

$$H(x,y)=\tfrac{1}{2}y^2-\tfrac{1}{2}\alpha\cos(2x-2t)+\tfrac{1}{4}e\alpha[\cos(2x-t)-7\cos(2x-3t)].$$

Here $\alpha=3(B-A)/2C$, and t is measured in units of Hyperion's orbital period, 21.2 days.

(a) Use the resonance overlap criterion to find the chaotic zones for Hyperion.

(b) Study this system numerically, using a stroboscopic plot.

(c) In 2005, NASA's Cassini mission confirmed that Hyperion is chaotically tumbling; see the Web site http://www.nasa.gov/mission_pages/cassini/main/. Compare the observations to your numerical solutions.

Mathematical Software

There are a number of excellent references on the use of mathematical software in dynamical systems—for example (Abell and Braselton 2004; Baumann 2004; Gander and Hrebícek 2004; Lee and Schiesser 2003; Lynch 2001, 2004). From these it is apparent that it would take hundreds of pages to comprehensively discuss the algorithms in any one language; consequently, we make no attempt to do that. However, a few simple commands can still be very helpful. In this appendix we give some examples in Mathematica, Maple, and MATLAB that can be used to make phase space portraits, solve linear systems, plot bifurcation diagrams, compute Lyapunov exponents, and draw Poincaré maps.

A.1 ▪ Vector Fields

It is quite easy to use a computer algebra system such as Mathematica, Maple, or MATLAB to create a plot that represents a vector field. For example, consider the vector field (1.5). In Mathematica this vector field is defined by

```
f = {Sin[x y] - y, y + x}
```

We then load the appropriate package and create the plot using "VectorPlot"

```
VectorPlot[f, {x, -Pi, Pi}, {y, -Pi, Pi},
          VectorScale -> Small, Axes -> True,
          AxesOrigin -> {-Pi, Pi}, Frame -> False]
```

This generates a plot of the vector field f as a 20×20 grid of arrows whose maximum length is scaled to one; see Figure 1.1. Mathematica puts the tail of each arrow on the appropriate grid point.

The corresponding commands in Maple are

```
>f:=[sin(x*y)-y,y+x];
>with(plots);
>fieldplot(f,x=-Pi..Pi, y=-Pi..Pi);
```

In MATLAB the grid of points is generated first, and then the command "quiver" plots the vector field:

```
[x,y]=meshgrid(-pi:pi/10:pi);
quiver(x,y,sin(x*y)-y,y+x)
```

A.2 ▪ Matrix Exponentials

All computer algebra programs have commands for diagonalizing and exponentiating matrices. Here, we give an example using Maple. First, load the linear algebra package, then define a matrix A and compute its characteristic polynomial and eigenvectors:

```
>with(LinearAlgebra):
>A := <<9,17,10>|<-8,-16,-10>|<4,9,7>>;
```

$$A=\begin{pmatrix} 9 & -8 & 4 \\ 17 & -16 & 9 \\ 10 & -10 & 7 \end{pmatrix}.$$

```
>p := CharacteristicPolynomial(A, lambda);
 factor(p);
```

$$p=\lambda^3-7\lambda+6,$$
$$(\lambda-1)(\lambda-2)(\lambda+3).$$

Therefore, the three eigenvalues are $(1,2,-3)$ and each has multiplicity one.

```
>(v,P) := Eigenvectors(A);
```

$$v,P:=\begin{pmatrix} 2 \\ -3 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 1 \\ \frac{1}{2} & 2 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

The output shows that the set of eigenvalues is $(2,-3,1)$, but this we already knew. The eigenvectors are given as the columns of the matrix P. To compute the exponential, define the diagonal matrix $e^{t\Lambda}$, and compute $Pe^{t\Lambda}P^{-1}$. The matrix multiplication operator is “.”, distinguishing it from scalar multiplication, “*”.

```
>etL := DiagonalMatrix([exp(2*t), exp(-3*t), exp(t)]):
>exptA := P . etL . MatrixInverse(P);
```

$$\exp tA:=\begin{pmatrix} -2e^{-3t}+3e^{t} & 2e^{-3t}-2e^{-t} & -e^{-3t}+e^{t} \\ e^{2t}-4e^{-3t}+3e^{t} & -e^{2t}+4e^{-3t}-2e^{t} & e^{2t}-2e^{-3t}+e^{t} \\ 2e^{2t}-2e^{-3t} & -2e^{2t}+2e^{-3t} & 2e^{2t}-e^{-3t} \end{pmatrix}.$$

Of course, it is much easier to use the built-in command `MatrixExponential(A, t)` to compute the exponential of tA. In MATLAB, a matrix is given in the notation `A = [a b; c d]` and the matrix exponential is computed numerically using the command `expm(A)`. Finally, in Mathematica a matrix is specified using braces, e.g., `A = {{a,b},{c,d}}`, and its exponential computed by `MatrixExp[t A]`.

A.3 ▪ Lyapunov Exponents

To compute Lyapunov exponents, we must solve both the differential equation and its linearization. Consider, for example, the Lorenz system (4.26), and its linearization (7.21). Using MATLAB’s Runge–Kutta routine, `ode45`, it is easy to write an integrator to solve the six-dimensional system for $w=(x,y,z,v_1,v_2,v_3)\in T\mathbb{R}^3$. This requires a function that returns the vector field for w:

```
function wdot=LorenzJet(t,w,r,sigma,b)
wdot=zeros(6,1);
wdot(1) = sigma*(w(2)-w(1));
wdot(2) = r*w(1)-w(2)-w(1)*w(3);
wdot(3) = w(1)*w(2)-b*w(3);
wdot(4) = sigma*(w(5)-w(4));
wdot(5) = (r-w(3))*w(4)-w(5)-w(1)*w(6);
wdot(6) = w(2)*w(4)+w(1)*w(5)-b*w(6);
```

A second function solves this system using ode45. Care must be taken, however, that the components of v do not get too large: indeed, they are expected to grow exponentially. As v grows the accuracy of the computation decreases and the ode45 routine will attempt to reduce the time step to compensate, causing it to eventually fail. Since the system (7.21) is linear, the length of v is irrelevant, and it can be rescaled at any time without affecting the trajectory. Since only the logarithm of $|v(t)|$ is needed to compute the Lyapunov exponents, if at some time we rescale, setting $v' = v/N$, it is only necessary to add $\ln N$ to $\ln |v'|$ to compute the original norm. The following simple function computes $\mu_{\max}$ using the vector field LorenzJet.

```
function mu = Lyapunov(tmax, r)

tstep = 1;
sigma = 10.0; b = 8/3;
x0 = [1,1,1];
v0 = [.1,.1,.1];
w0 = horzcat(x0,v0);
scalefactor = -log(norm(v0));
T=[]; L=[];

for time = 1:tstep:tmax
[t,w] = ode45(@LorenzJet,[time,time+tstep],w0,[],r,sigma,b);
Lyp = (scalefactor + 0.5*log(sum(w(:,4:6).^2,2)))./t;

T=[T; t]; % Store integration output
L=[L; Lyp];

nm = norm(w(end,4:6));
scalefactor = scalefactor + log(nm);
w0 = horzcat(w(end,1:3), w(end,4:6)/nm);
end
plot(T,L);
mu = Lyp(end);
```

Here the integration is done for a time tstep and then the vector v is renormalized by dividing by its norm, nm. At each rescaling the logarithm of nm is accumulated into scalefactor. Plots of the output of this routine are shown in Figure 7.5.

A.4 ▪ Bifurcation Diagrams

One way to plot bifurcation diagrams is to simply solve the equations for the equilibria as a function of the parameters and plot the results. However, since there typically are several equilibria for some parameter ranges, we must take care to get all of the solutions.

In some cases, however, the inverse function $\mu(x)$ has a single solution. For example, the vector field on $\mathbb{R}^1 \times \mathbb{R}^1$, defined in Maple by

```
>f := mu+x+mu*x^2-x^3;
```

$$f(x;\mu) = \mu + x + \mu x^2 - 3x^3, \tag{A.1}$$

has up to three equilibria $x_i^*(\mu)$. However, there is a single solution for μ:

```
>m := solve(f, mu):
>p1 := plot(m, x = -2 .. 2, mu = -1 .. 1);
```

However, this produces a plot with x vertical. To plot the normal bifurcation diagram we reflect about the diagonal:

```
>with(plots); with(plottools);
>display(reflect(p1, [[0, 0], [1, 1]]),
labels = ['mu', 'x']);
```

Alternatively, we can use a numerical solution to make the plot. For example, the equilibria of (8.3) can be easily plotted in Maple by

```
>xp := mu->fsolve(mu+x-ln(1+x),
     x,0..10);
>xm:= mu->fsolve(mu+x-ln(1+x),
     x,-1..0);
>plot({xm,xp},-3..0);
```

These commands create Figure 8.2.

In some cases, when the number of branches is uncertain, it is easier to use an implicit plotting routine. The vector field

```
>f := mu+x^2+(x-mu)^3;
```

$$f = \mu + x^2 + (x-\mu)^3 \tag{A.2}$$

can be solved for either x or μ; however, the expressions are not particularly elucidating. It is easier to visualize the equilibria by plotting the zero contour of$f(x;\mu)$:

```
>implicitplot(f, mu = -5 .. 5, x = -3 .. 3,
grid = [100, 100]);
```

This generates Figure A.1, showing a pair of saddle-node bifurcations.

In Mathematica, this implicit plot is made by plotting the zero contour:

```
ContourPlot[f,{mu,-5,5},{x,-3,3},Contours->{0.0},
     ContourShading->False]
```

In MATLAB, the relevant command is `Contour`.

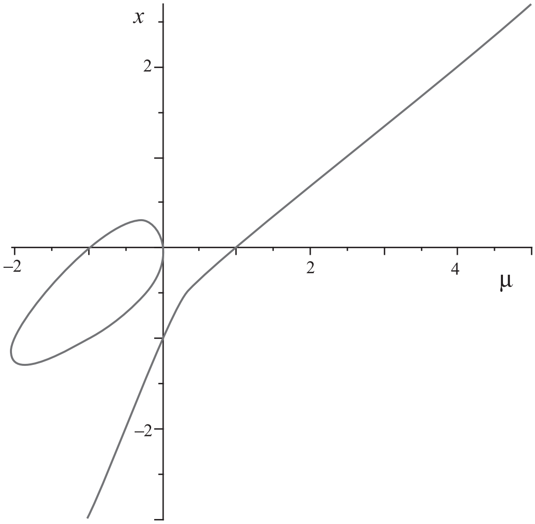

Figure A.1. *Equilibria of the vector field* (A.2).

A.5 ▪ Poincaré Maps

One easy way to plot the Poincaré map for the Hénon–Heiles Hamiltonian (9.66) is to use the Maple function "poincare." We start by defining the Hamiltonian:

```
>with(plots):
>with(DETools):
>H := (p1^2+p2^2+q1^2+q2^2)/2 + q1^2*q2-q2^3/3;
```

A three-dimensional projection of the dynamics, Figure 9.12, for a single initial condition is obtained using

```
>ic := {[0., .3759703843, 0., 0., -.15]};
>poincare(H,t=0..200,ic,stepsize = 0.1,scene=
[q2=-0.5..0.5,p2=-0.5..0.5,q1=-0.5..0.5],3);
```

Here the initial condition is of the form (t,p,q) with $t=0$, $(p_2=0, q_1=0, q_2=-0.15)$, and $p_1 \approx 0.376$ so the energy is 1/12. The command `poincare` integrates the Hamiltonian equations over the range of time $[0,200]$ using, by default, a fourth-order Runge–Kutta routine with the time step 0.1. The `scene` argument sets the projection to the three variables (q_2,p_2,q_1) and defines the section variable, q_1, which is by default at the value zero.

A two-dimensional section for a set of initial conditions obtained from

```
>ics := generate_ic(H,{t=0,p2=0, q2=-0.2..0.1,q1=0.0,
energy = 1/12},5);
>poincare(H,t=0..800,ics, stepsize=0.01,
iterations = 5,scene=[q2,p2,q1]);
```

The command "`generate_ic`" creates a list of five initial conditions on the section by computing the required value of p_1 for the given energy and $(q_1=0, q_2 \in -[0.2,0.1], p_2=0)$. The `scene` argument sets the projection to (q_2,p_2) and the section to $q_1=0$. These commands generate Figures 9.14 and 9.15, though the choice of initial conditions for the figures was different.

Bibliography

Abell, M. L., and J. P. Braselton (2004). *Differential Equations with Mathematica*. London, Academic Press.

Abraham, R., and J. E. Marsden (1978). *Foundations of Mechanics*. Reading, Benjamin.

Adrianova, L. Y. (1995). *Introduction to Linear Systems of Differential Equations*. Providence, AMS.

Akhiezer, N. I. (1962). *The Calculus of Variations*, New York, Blaisdell Publishing.

Alexander, J. C., B. R. Hunt, I. Kan, and J. A. Yorke (1996). "Intermingled Basins for the Triangle Map." *Ergodic Theory and Dynam. Systems* 16: 651–662.

Allee, W. C., A. E. Emerson, O. Park, T. Park, and K. P. Schmidt (1949). *Principles of Animal Ecology*. Philadelphia, Saunders.

Alligood, K. T., T. D. Sauer, and J. A. Yorke (1997). *Chaos*. New York, Springer-Verlag.

Arnold, V. I. (1963). "Proof of a Theorem of A.N. Kolmogorov on the Invariance of Quasiperiodic Motions Under Small Perturbations of the Hamiltonian." *Russ. Math. Surveys* 18(5): 9–36.

Arnold, V. I. (1978). *Mathematical Methods of Classical Mechanics*. New York, Springer.

Arnold, V. I. (1983). *Geometrical Methods in the Theory of Ordinary Differential Equations*. New York, Springer-Verlag.

Arnold, V. I., and A. Avez (1968). *Ergodic Problems of Classical Mechanics*. New York, Benjamin.

Arnold, V. I., V. S. Afrajmovich, Y. S. Ilyashenko, and L. P. Shilnikov (1999). *Bifurcation Theory and Catastrophe Theory*. Berlin, Springer-Verlag.

Arrowsmith, D. K., and C. M. Place (1992). *Dynamical Systems: Differential Equations, Maps and Chaotic Behavior*. London, Chapman and Hall.

Auslander, J. and J. A. Yorke (1980). "Interval Maps, Factors of Maps, and Chaos." Thoku Math. J. 32(2): 177-188.

Barger, V. D., and M. G. Olsson (1973). *Classical Mechanics*. New York, McGraw-Hill.

Baumann, G. (2004). *Mathematica for Theoretical Physics: Classical Mechanics and Nonlinear Dynamics*. New York, Springer Science.

Bielecki, A. (1956). "Une Remarque Sur la Méthode De Banach-Cacciopoli-Tikhonov Dans la Théorie Des Equations Différentiel les Ordinaires." *Bull. Acad. Pol. Sci.* 4: 261–264.

Blanchard, F., E. Glasner, S. Kolyada and A. Maass (2002). "On Li-Yorke Pairs." J. Reine Angew. Math. 547: 51-68.

Bogdanov, R. I. (1975). “Versal Deformation of a Singular Point of a Vector Field on the Plane in the Case of Zero Eigenvalues.” *Funkcional Anal. i Priložen.* 9(2): 63.

Bollt, E., and A. Klebanoff (2002). “A New and Simple Chaos Toy.” *Internat. J. Bifur. Chaos* 12(8): 1843–1857.

Bora, M. P., and D. Sarmah (2008). “Sawtooth Disruptions and Limit Cycle Oscillations.” *Comm. Nonlinear Sci. Numer. Simul.*, 13 (2): 296–313.

Carr, J. (1981). *Applications of Centre Manifold Theory*. New York, Springer-Verlag.

Cartwright, M. L. (1952). “Non-linear Vibrations: A Chapter in Mathematical History.” *Math. Gaz.* 26 (316):81–88.

Cassels, J. W. S. (1957). *An Introduction to Diophantine Approximation*. Cambridge, UK, Cambridge University Press.

Chicone, C. (1999). *Ordinary Differential Equations with Applications*. New York, Springer-Verlag.

Chirikov, B. V. (1979). “A Universal Instability of Many-Dimensional Oscillator Systems.” *Phys. Rep.* 52: 265–379.

Chow, S. H., and J. K. Hale (1982). *Methods of Bifurcation Theory*. New York, Springer-Verlag.

Chow, S. N., C. Li, and D. Wang (1994). *Normal Forms and Bifurcations of Planar Vector Fields*. Cambridge, UK, Cambridge University Press.

Coddington, E. A., and N. Levinson (1955). *Theory of Ordinary Differential Equations*. New York, McGraw-Hill.

Conley, C. (1978). *Isolated Invariant Sets and the Morse Index*. Providence, AMS.

Cvitanovic, P. (1995). “Dynamical Averaging in Terms of Periodic Orbits.” *Phys. D*, 83(1–3): 109–123.

Davidson, R. C. (1972). *Methods in Nonlinear Plasma Theory*. New York, Academic Press.

de la Llave, R. (2001). “A Tutorial on KAM Theory.” In *Smooth Ergodic Theory and Its Applications (Seattle, WA, 1999). Proc. Sympos. Pure Math.* 69. Providence, AMS 69: 175–292.

Delshams, A., and T. M. Seara (1997). “Splitting of Separatrices in Hamiltonian Systems with One and a Half Degrees of Freedom.” *Math. Phys. Electron*. J. 3: Paper 4 (electronic).

Devaney, R. L. (1986). *An Introduction to Chaotic Dynamical Systems*. Menlo Park, NJ, Benjamin/Cummings.

Diacu, F., and P. J. Holmes (1996). *Celestial Encounters: The Origins of Chaos and Stability*. Princeton, Princeton University Press.

Dieci, L., and E. S. van Vleck (2002). “Lyapunov Spectral Intervals: Theory and Computation.” *SIAM J. Numer. Anal.* 40(2): 516–542.

Dobson, A. P., A. D. Bradshaw, and J. M. Baker (1997). “Hopes for the Future: Restoration Ecology and Conservation Biology.” *Science* 277: 515–522.

Dombre, T., U. Frisch, J. M. Greene, M. Hénon, A. Mehr, and A. M. Soward (1986). “Chaotic Streamlines in the ABC Flows.” J. Fluid Mech. 167: 353–391.

Dubrovin, B. A., I. M. Krichever, and S. P. Novikov (1985). “Integrable Systems. I.” *Current Prob. Math. Fund. Dir.*, 4: 179–284, 291.

Dullin, H. R., and A. Wittek (1995). "Complete Poincaré Sections and Tangent Sets." *J. Phys. A* 28: 7157–7180.

Dullin, H. R., M. Juhnke, and P. H. Richter (1994). "Action Integrals and Energy Surfaces of the Kovalevskaya Top." *Internat. J. Bifur. Chaos* 4(6): 1535–1562.

Easton, R. W. (1998). *Geometric Methods for Discrete Dynamical Systems*. Cambridge, UK, Cambridge University Press.

Eden, A., C. Foias, B. Nicolaenko, and R. Temam (1994). *Exponential Attractors for Dissipative Evolution Equations*. Paris, Masson.

Enciso, G. A., and E. D. Sontag (2006). "Global Attractivity, I/O Monotone Small-Gain Theorems, and Biological Delay Systems." *Discrete Contin. Dynam. Syst.* 14(3): 549–578.

Escande, D. F. (1985). "Stochasticity in Hamiltonian Systems: Universal Aspects." *Phys. Rep.* 121: 165–261.

Falconer, K. J. (1990). *Fractal Geometry: Mathematical Foundations and Applications*. New York, Wiley.

Farkas, M. (1984). "Zip Bifurcation in a Competition Model." *Nonlinear Anal.* 8(11): 1295–1309.

Field, M. (1996). *Lectures on Bifurcations, Dynamics and Symmetry*. Harlow, UK, Longman.

Field, M., and M. Golubitsky (1995). *Symmetry in Chaos: A Search for Patterns in Mathematics, Art and Nature*. New York, Oxford University Press.

Floquet, G. (1883). "Sur les Équations Différentielles Linéaires à Coefficients Périodiques." *Ann. Sci. École Norm. Sup.* 12(2): 47–88.

Friedman, A. (1982). *Foundations of Modern Analysis*, New York, Dover Publications.

Gander, W., and J. Hrebícek (2004). *Solving Problems in Scientific Computing Using Maple and MATLAB*. Berlin, Springer-Verlag.

Glendinning, P., T.H. Jäger and G. Keller (2006). "How Chaotic Are Strange Non-Chaotic Attractors?" *Nonlinearity* 19: 2005–2022.

Goldstein, H., C. P. Poole, and J. L. Safko (2002). *Classical Mechanics*. Reading, MA, Addison-Wesley.

Golubitsky, M., and D. G. Schaeffer (1985). *Singularities and Groups in Bifurcation Theory* I. New York, Springer-Verlag.

Golubitsky, M., and I. Stewart (2002). *The Symmetry Perspective: From Equilibrium to Chaos in Phase Space and Physical Space*. Basel, Birkhäuser.

Golubitsky, M., I. Stewart, and D. G. Schaeffer (1988). *Singularities and Groups in Bifurcation Theory* II. New York, Springer-Verlag.

Guckenheimer, J., and P. Holmes (1983). *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. New York, Springer-Verlag.

Guenther, R. B., and J. W. Lee (1996). *Partial Differential Equations of Mathematical Physics and Integral Equations*. New York, Dover Publications.

Hall, B. C. (2003). *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction* New York, Springer-Verlag.

Hamilton, W. R. (1834). “On a General Method in Dynamics; by Which the Study of the Motions of All Free Systems of Attracting or Repelling Points Is Reduced to the Search and Differentiation of One Central Relation, or Characteristic Function.” *Phil. Trans. Roy. Soc.*, II: 247–308.

Hardy, G. H., and E. M. Wright (1979). *An Introduction to the Theory of Numbers*. Oxford, UK, Oxford University Press.

Harris, Jr., W. A., J. P. Fillmore, and D.R. Smith (2001). “Matrix Exponentials—Another Approach.” *SIAM Rev.* 43: 694–706.

Hénon, M., and C. Heiles (1964). “The Applicability of the Third Integral of Motion: Some Numerical Experiments.” *Astron. J.* 69: 73–79.

Hilbert, D. (1900). “Mathematische Probleme.” *Göttinger Nachr.* 253–297.

Hirsch, M. W. (1976). *Differential Topology*. New York, Springer-Verlag.

Hirsch, M. W., and S. Smale (1974). *Differential Equations, Dynamical Systems and Linear Algebra*, New York, Academic Press.

Hirsch, M. W., C. Pugh, and M. Shub (1977). *Invariant Manifolds*. New York, Springer-Verlag.

Hocking, J. G., and G. S. Young (1961). *Topology*. Mineola, Dover.

Holmes, P., J. Marsden, and J. Scheurle (1988). “Exponentially Small Splittings of Separatrices with Applications to KAM Theory and Degenerate Bifurcations.” In *Hamiltonian Dynamical Systems (Boulder, CO, 1987)* Contemp. Math. 81. Providence, AMS. 213–244.

Hydon, P. E. (2000). *Symmetry Methods for Differential Equations: A Beginner’s Guide*. Cambridge, UK, Cambridge University Press.

Ilyashenko, Y., and S. Yakovenko, Eds. (1995). *Concerning the Hilbert* 16*th Problem*. Providence, AMS.

Ince, E. L. (1956). *Ordinary Differential Equations*. New York, Dover.

Isham, C. J. (1999). *Modern Differential Geometry for Physicists*. Singapore, World Scientific.

Katok, A. B., and B. Hasselblatt (1999). *Introduction to the Modern Theory of Dynamical Systems*. Cambridge, UK, Cambridge University Press.

Kim, J. W., S. Y. Kim, B. Hunt, and E. Ott (2003). “Fractal Properties of Robust Strange Nonchaotic Attractors in Maps of Two or More Dimensions.” *Phy. Rev. E* (3). 67: 036211.

Kovalevskaya, S. (1889). “Sur le Problème De la Rotation D’un Corps Solide D’un Point Fixe.” *Acta Math*. 12: 177–232.

Kuznetsov, Y. A. (1995). *Elements of Bifurcation Theory*. New York, Springer-Verlag.

Lanczos, C. (1962). *The Variational Principles of Mechanics*. Toronto, University of Toronto.

Lee, H. J., and W. E. Schiesser (2003). *Ordinary and Partial Differential Equation Routines in C, C++, Fortran, Java, Maple, and MATLAB*. Boca Raton, FL, Chapman and Hall.

Li, T. Y., and J. A. Yorke (1975). “Period Three Implies Chaos.” *Amer. Math. Monthly* 82: 985–992.

Lichtenberg, A. J., and M. A. Lieberman (1992). *Regular and Chaotic Motion*. New York, Springer-Verlag.

Lorenz, E. N. (1963). “Deterministic Nonperiodic Flow.” *J. Atmos. Sci.* 20: 130–141.

Lynch, S. (2001). *Dynamical Systems with Applications using Maple*. Boston, Birkhäuser.

Lynch, S. (2004). *Dynamical Systems with Applications using Matlab*. Boston, Birkhäuser.

MacDonald, N. (1978). *Time Lags in Biological Models*. New York, Springer-Verlag.

MacKay, R. S. (1993). *Renormalisation in Area-Preserving Maps*. Singapore, World Scientific.

MacKay, R. S., and J. D. Meiss, Eds. (1987). *Hamiltonian Dynamical Systems: A Reprint Selection*. London, Adam-Hilgar Press.

Markley, N. G. (2004). *Principles of Differential Equations*. Hoboken, NJ, John Wiley and Sons.

Markus, L., and H. Yamabe (1960). “Global Stability Criteria for Differential Systems.” *Osaka Math. J.* 12: 305–317.

Mather, J. N. (1991). “Action Minimizing Invariant Measures for Positive Definite Lagrangian Systems.” *Math. Z.* 207: 169–207.

McGehee, R. (1974). “Triple Collision in the Collinear Three-Body Problem.” *Invent. Math.* 27: 191–227.

Meiss, J. D. (1992). “Symplectic Maps, Variational Principles, and Transport.” *Rev. Modern Phys.* 64(3): 795–848.

Meyer, K. R., and G. R. Hall (1992). *Introduction to the Theory of Hamiltonian Systems*. New York, Springer-Verlag.

Michaelis, L., and M. L. Menten (1913). “Die Kinetic Der Invertinwirkung.” *Biochem. Z.* 49: 333–369.

Milnor, J. (1985a). “On the Concept of Attractor.” *Comm. Math. Phys.* 99: 177–195.

Milnor, J. (1985b). “On the Concept of Attractor: Correction and Remarks.” *Comm. Math. Phys.* 102(3): 517–519.

Moler, C., and C. van Loan (1978). “Nineteen Dubious Ways to Compute the Exponential of a Matrix.” *SIAM Rev.* 20: 801–836.

Morrison, P. J. (1998). “Hamiltonian Description of the Ideal Fluid.” *Rev. Mod. Phys.* 70(2): 467–521.

Moser, J. K. (1962). “On Invariant Curves of Area-Preserving Mappings of an Annulus.” *Nachr. Akad. Wiss. Göttingen* II *Math. Phys.* 1: 1–20.

Murray, J. D. (1993). *Mathematical Biology*. New York, Springer-Verlag.

Nayfeh, A. H., and D. T. Mook (1979). *Nonlinear Oscillations*. New York, John Wiley and Sons.

Olver, P. J. (1993). *Applications of Lie Groups to Differential Equations*. New York, Springer-Verlag.

Olver, P. J., and C. Shakiban (2006). *Applied Linear Algebra*. Upper Saddle River, NJ, Pearson Prentice–Hall.

Perko, L. (2000). *Differential Equations and Dynamical Systems*. New York, Springer-Verlag.

Poincaré, H. (1890). “Sur le Problème Des Trois Corps Et les Équations De la Dynamique.” *Acta Math.*: 1–270.

Poincaré, H. (1892). *Les Methodes Nouvelles de la Mechanique Celeste*. Three vols. Paris, Gauthier-Villars. (Translated as (1992) New Methods in Celestial Mechanics. New York. Springer-Verlag.)

Poincaré, H. (1908). *Science et Méthode*. Paris, Flammarion (Translated as (1952) Science and Method. New York, Dover).

Poincaré, H. (1914). *La Valeur de la Science*. Paris, Flammarion (Translated as (2001) The Value of Science: Essential Writings of Henri Poincaré. New York, Modern Library.

Pöschel, J. (1982). “Integrability of Hamiltonian Systems on Cantor Sets.” *Comm. Pure Appl. Math.* 35(5): 653–696.

Pöschel, J. (2001). “A Lecture on the Classical KAM Theorem.” In *Smooth Ergodic Theory and Its Applications (Seattle, WA 1999). Proc. Sympos. Pure Math.* 69. Providence, AMS 69: 707–732.

Robinson, C. (1999). *Dynamical Systems: Stability, Symbolic Dynamics, and Chaos*. Boca Raton, FL, CRC Press.

Rod, D. L., and R. C. Churchill (1985). “A Guide to the Hénon-Heiles Hamiltonian.” In *Singularities and Dynamical Systems (Iráklion, 1983).* North–Holland Math. Stud. 103. Amsterdam, North–Holland 385–395.

Romeiras, F. J., and E. Ott (1987). “Strange Nonchaotic Attractors of the Damped Pendulum with Quasiperiodic Forcing.” *Phys. Rev. A* 35(10): 4404–4412.

Rosenweig, M. L. (1973). “Exploitation in Three Trophic Levels.” *Amer. Naturalist* 107: 275–294.

Rössler, O. E. (1976). “An Equation for Continuous Chaos.” *Phys. Lett. A* 57: 397–398.

Ruelle, D. (1981). “Small Random Perturbations of Dynamical Systems and the Definition of Attractors.” *Comm. Math. Phys.* 82: 137–151.

Rusbridge, M. G. (1979). “Motion of the Sprung Pendulum.” *Amer. J. Phys.* 48: 146–151.

Segur, H. (1993). “Asymptotics Beyond All Orders—-A Survey.” In *Chaos in Australia (Sydney, 1990).* River Edge, NJ, World Sci. Publ. 150–172.

Shi, S. L. (1980). “A Concrete Example of the Existence of Four Limit Cycles for Plane Quadratic Systems.” *Sci. Sinica* 23(2): 153–158.

Shilnikov, A. L. (1993). “On Bifurcations of the Lorenz Attractor in the Shimizu-Morioka Model.” *Phys. D* 62: 338–346.

Shilnikov, L. P., A. L. Shilnikov, D. V. Turaev, and L. O. Chua (1998). *Methods of Qualitative Theory in Nonlinear Dynamics, Part* I. Singapore, World Scientific.

Siegel, C. L., and J. K. Moser (1971). *Lectures on Celestial Mechanics*. New York, Springer-Verlag.

Smale, S. (1998). “Mathematical Problems for the Next Century.” *Math. Intell.* 20(2): 7–15.

Sprott, J. C. (1994). “Some Simple Chaotic Flows.” *Phys. Rev. E* 50: R647–R650.

Sternberg, S. (1958). “On the Structure of Local Homeomorphisms of Euclidean Space, II.” *Amer. J. Math.* 81: 623–631.

Strang, G. (1988). *Linear Algebra and Its Applications*. Boca Raton, FL, Brooks/Cole.

Strogatz, S. H. (1994). *Nonlinear Dynamics and Chaos: With Applications in Physics, Biology, Chemistry, and Engineering*. Reading, MA, Addison–Wesley.

Takens, F. (2001). "Forced Oscillations and Bifurcations." In *Global Analysis of Dynamical Systems*. Bristol, Inst. Phys. 1–61.

Taylor, A. E., and W. R. Mann (1983). *Advanced Calculus*. New York, John Wiley and Sons.

Toral, R., M. San Miguel, and R. Gallego (2000). "Period Stabilization in the Busse-Heikes Model of the Küppers-Lortz Instability." *Phys. A* 280: 315–336.

Tucker, W. (2002). "A Rigorous ODE Solver and Smale's 14th Problem." *Found. Comput. Math.* 2(1): 53–117.

van der Meer, J.-C. (1985). *The Hamiltonian Hopf Bifurcation*. Berlin, Springer-Verlag.

van der Pol, B. (1922). "On Oscillation Hysteresis in a Simple Triode Generator." *Phil. Mag.* 43: 700–719.

Vinograd, R. E. (1957). "The Inadequacy of the Method of Characteristic Exponents for the Study of Nonlinear Differential Equations." *Mat. Sb.* 41: 431–438.

Viswanath, D. (2004). "The Fractal Property of the Lorenz Attractor." *Phys. D* 190: 115–128.

Weissert, T. P. (1997). *The Genesis of Simulation in Dynamics: Pursing the Fermi-Pasta-Ulam Problem*. New York, Springer-Verlag.

Wiggins, S. (2003). *Introduction to Applied Nonlinear Dynamical Systems and Chaos*. New York, Springer-Verlag.

Wisdom, J., S. J. Peale, and F. Mignard (1983). "The Chaotic Rotation of Hyperion." *Icarus* 58: 137–152.

Wolfram, S. (1983). "Statistical Mechanics of Cellular Automata." *Rev. Modern Phys.* 55(3): 601–644.

Yakubovitch, V. A., and V. M. Starzhinskii (1975). *Linear Differential Equations with Periodic Coefficients*. New York, John Wiley and Sons.

Zakharov, V. E., Ed. (1991). *What Is Integrability?* Springer Series in Nonlinear Dynamics. Berlin, Springer-Verlag.

Index