Nonlinear Systems of Differential Equations
Michael Taylor

Contents

Introduction

This final chapter brings to bear all the material presented before and pushes on to the heart of the subject, nonlinear systems of differential equations. Section 1 begins with a demonstration of existence and uniqueness (for t close to t_0) of solutions to

$$\frac{dx}{dt} = F(t,x), \quad x(t_0) = x_0. \tag{0.1}$$

Here $x(t)$ is a path in $\Omega \subset \mathbb{R}^n$ and F is bounded and continuous on $I \times \Omega$ (with $t_0 \in I$), and satisfies a Lipschitz condition in x. (See (1.2) for a definition.) We study the issue of global existence, including positive results when $F(t,x)$ is linear in x. Section 2 studies the smoothness of the solution to (0.1) as a function of x_0, given various additional hypotheses on F, and related issues.

Section 3 reveals a geometric flavor to (0.1), described in the language of vector fields and the flows they generate. A *vector field* on $\Omega \subset \mathbb{R}^n$ is a map $F : \Omega \to \mathbb{R}^n$. This is a special case of (0.1), where F is independent of t. The path $x(t)$ in Ω satisfying (0.1) for such F is called the *orbit* of F through x_0; denote it $\Phi^t(x_0)$. This gives rise to the family of maps Φ^t, called the flow generated by F. The *phase portrait* is introduced as a tool to understand the orbits and flow, from a visual perspective. We pay particular attention to how phase portraits look near critical points of a vector field F (which are points where F vanishes), including special types known as sources, sinks, saddles, and centers.

Section 4 discusses a particular class of vector fields, gradient vector fields, on a domain $\Omega \subset \mathbb{R}^n$. In case $n = 2$, this relates to the topic of exact equations, discussed in many texts early on. We have broken with tradition and moved the discussion of exactness to here, to see it in a broader context.

We move from generalities about nonlinear systems to settings in which they arise. Section 5 introduces a class of differential equations arising from Newton's law $F = ma$. This resumes the study introduced in §5 of Chapter 1. This time we are studying the interaction of several bodies, each moving in n-dimensional space. We concentrate on central force problems. We show how a two-body central force problem (for motion in $\mathbb{R}^n$) gives rise to a second order $n \times n$ system, in "center of mass coordinates." We look at this two-body problem in more detail in §6, and derive Newton's epoch-making analysis of the planetary motion problem.

In §7 we introduce another (though ultimately related) class of problems that lead to differential equations, namely variational problems. The general setup is to consider

$$I(u) = \int_a^b L(u(t), u'(t))\, dt, \tag{0.2}$$

for paths $u : [a, b] \to \Omega \subset \mathbb{R}^n$, given smooth L on $\Omega \times \mathbb{R}^n$, and find conditions under which I has a minimum, or maximum, or more generally a stationary point u. We produce a differential equation known as the Lagrange equation for u. This method has many important ramifications. One of the most important is to produce differential equations for physical problems, providing an alternative to the method discussed in §5. We illustrate this in §7 by obtaining a derivation of the pendulum equation, alternative to that given in §6 of Chapter 1. We proceed to more sophisticated uses of the variational method. In §8 we discuss the "brachistochrone problem," tossed about by the early leading lights of calculus, one of the foundational variational problems. In §9 we discuss the double pendulum, a physical problem that is confounding when one uses the $F = ma$ approach, and which well illustrates the "Lagrangian" approach. An alternative to Lagrangian differential equations is the class of Hamiltonian differential equations. The passage from Lagrangian to Hamiltonian equations is previewed (in special cases) in §§7 and 9, and developed further in §10.

The majority of the systems studied in this chapter are not amenable to solution in terms of explicit formulas. In §11 we introduce a tool that has revolutionized the study of these equations, namely numerical approximation. Behind this revolution is the availability of personal computers. In §11 we present several techniques that allow for accurate approximation of solutions to (0.1), the most important being Runge-Kutta difference schemes.

In §12 we return to the study of qualitative features of phase portraits, initiated in §3. We define limit sets of orbits, and establish a result known as the Poincaré-Bendixson theorem, which provides a condition under which a limit set for an orbit of a planar vector field can be shown to be a closed curve, called a limit cycle.

Sections 13–14 are devoted to some systems of differential equations arising to model the populations of interacting species. In §13 we study "predator-prey" equations. We study several models. In some, all the orbits are periodic, except for one critical point. In others, there is a limit cycle, arising via the mechanism examined in §12. In §14 we look at other interacting species equations, namely equations modeling competing species.

One phenomenon behind the Poincaré-Bendixson theorem is that an orbit of a vector field F in the plane locally divides the plane into two parts, one to the left of the orbit and one to the right. Since another orbit of F cannot cross it, this tends to separate the plane into pieces, in each of which the phase portrait has a fairly simple appearance. In dimension three and higher, this mechanism to enforce simplicity does not work, and far more complicated scenarios are possible. This leads to the occurrence of "chaos" for $n \times n$ systems of differential equations when $n \geq 3$. We explore some aspects of this in the last section of this chapter, §15.

This chapter has six appendices. In Appendix A we give basic information on the derivative of functions of several variables, reviewing material typically covered in third semester calculus and setting up notation that is used in the chapter. Appendix B discusses some basic results about convergence, including the notion of *compactness*. In Appendix C we show that if the linearization of a vector field F at a critical point behaves like a saddle, so does F. Appendix D discusses an approximation procedure for computing the periods of orbits, for a certain family of planar vector fields, with reference to how Einstein's correction of Newton's equations for planetary motion yields a calculation of the precession of the planet's perihelion. In Appendix E we show that a spherically symmetric planet produces the same gravitational field as if all its mass were concentrated at its center. In Appendix F we prove the Brouwer fixed-point theorem (in dimension 2), a use of which arises in §15. The proof we give makes use of material developed in §4.

1. Existence and uniqueness of solutions

We investigate existence and uniqueness of solutions to a first order nonlinear $n \times n$ system of differential equations,

$$\frac{dx}{dt} = F(t,x), \quad x(t_0) = x_0. \tag{1.1}$$

We assume F is bounded and continuous on $I \times \Omega$, where I is an open interval about t_0 and Ω is an open subset of $\mathbb{R}^n$, containing x_0. We also assume F satisfies a Lipschitz condition in x:

$$\|F(t,x) - F(t,y)\| \leq L\|x-y\|, \tag{1.2}$$

for all $t \in I,\ x,y \in \Omega$, with $L \in (0,\infty)$. Such an estimate holds if Ω is convex and F is C^1 in x and satisfies

$$\|D_x F(t,x)\| \leq L, \tag{1.3}$$

for all $t \in I,\ x \in \Omega$. At this point, the reader might want to review the concept of the derivative of a function of n variables, by looking in Appendix A. The estimate (1.3) follows readily from (A.9). Our first goal is to prove the following.

Proposition 1.1. *Assume $F : I \times \Omega \to \mathbb{R}^n$ is bounded and continuous and satisfies the Lipschitz condition (1.2), and let $x_0 \in \Omega$. Then there exists $T_0 > 0$ and a unique C^1 solution to (1.1) for $|t - t_0| < T_0$.*

The first step in proving this is to rewrite (1.1) as an integal equation:

$$x(t) = x_0 + \int_{t_0}^{t} F(s, x(s))\, ds. \tag{1.4}$$

The equivalence of (1.1) and (1.4) follows from the Fundamental Theorem of Calculus. It suffices to find a continuous solution x to (1.4) on $[t_0 - T_0, t_0 + T_0]$, since then the right side of (1.4) will be C^1 in t.

We will apply a technique known as Picard iteration to construct a solution to (1.4). We set $x_0(t) \equiv x_0$ and then define $x_n(t)$ inductively by

$$x_{n+1}(t) = x_0 + \int_{t_0}^{t} F(s, x_n(s))\, ds. \tag{1.5}$$

We show that this converges uniformly to a solution to (1.4), for $|t - t_0| \le T_0$, if T_0 is taken small enough. To get this, we quantify some hypotheses made above. We assume

$$\overline{B_R(x_0)} = \{x \in \mathbb{R}^n : \|x - x_0\| \le R\} \subset \Omega \tag{1.6}$$

and

$$\|F(s, x)\| \le M, \quad \forall\, s \in I,\ x \in \overline{B_R(x_0)}. \tag{1.7}$$

Clearly $x_0(t) \equiv x_0$ takes values in $\overline{B_R(x_0)}$ for all t. Suppose that $x_n(t)$ has been constructed, taking values in $\overline{B_R(x_0)}$, and $x_{n+1}(t)$ is defined by (1.5). We have

$$\|x_{n+1}(t) - x_0\| \le \int_{t_0}^{t} \|F(s, x_n(s))\|\, ds \le M|t - t_0|, \tag{1.8}$$

so $x_{n+1}(t)$ also takes values in $\overline{B_R(x_0)}$ provided $|t - t_0| \le T_0$ and

$$T_0 \le \frac{R}{M}. \tag{1.9}$$

As long as (1.9) holds and $[t_0 - T_0, t_0 + T_0] \subset I$, we get an infinite sequence $x_n(t)$ of functions, related by (1.5).

We produce one more constraint on T_0, which will guarantee convergence. Note that, for $n \geq 1$,

$$\begin{aligned} \|x_{n+1}(t) - x_n(t)\| &= \Bigl\| \int_{t_0}^{t} \bigl[F(s, x_n(s)) - F(s, x_{n-1}(s))\bigr]\, ds \Bigr\| \\ &\leq \int_{t_0}^{t} \|F(s, x_n(s)) - F(x, x_{n-1}(s))\|\, ds \\ &\leq L \int_{t_0}^{t} \|x_n(s) - x_{n-1}(s)\|\, ds, \end{aligned} \tag{1.10}$$

the last inequality by (1.2). Hence

$$\max_{|t-t_0|\leq T_0} \|x_{n+1}(t) - x_n(t)\| \leq LT_0 \max_{|s-t_0|\leq T_0} \|x_n(s) - x_{n-1}(s)\|. \tag{1.11}$$

The additional constraint we impose on T_0 is

$$T_0 \leq \frac{1}{2L}. \tag{1.12}$$

Noting that

$$\max_{|t-t_0|\leq T_0} \|x_1(t) - x_0\| \leq R, \tag{1.13}$$

we see that

$$\max_{|t-t_0|\leq T_0} \|x_{n+1}(t) - x_n(t)\| \leq 2^{-n} R. \tag{1.14}$$

Consequently, the infinite series

$$x(t) = x_0 + \sum_{n=0}^{\infty} \bigl(x_{n+1}(t) - x_n(t)\bigr) \tag{1.15}$$

is absolutely and uniformly convergent for $|t - t_0| \leq T_0$, with a continuous sum, satisfying

$$\max_{|t-t_0|\leq T_0} \|x(t) - x_n(t)\| \leq 2^{1-n} R. \tag{1.16}$$

It readily follows that

$$\int_{t_0}^{t} F(s, x_n(s))\, ds \longrightarrow \int_{t_0}^{t} F(s, x(s))\, ds, \tag{1.17}$$

so (1.4) follows from (1.5) in the limit $n \to \infty$.

To finish the proof of Proposition 1.1, we establish uniqueness. Suppose $y(t)$ also satisfies (1.4) for $|t-t_0| \le T_0$. Then

$$\begin{aligned} \|x(t)-y(t)\| &= \Bigl\| \int_{t_0}^{t} \bigl[F(s,x(s))-F(s,y(s))\bigr]\,ds \Bigr\| \\ &\le \int_{t_0}^{t} \|F(s,x(s))-F(s,y(s))\|\,ds \\ &\le L \int_{t_0}^{t} \|x(s)-y(s)\|\,ds, \end{aligned}$$

and hence

$$\max_{|t-t_0|\le T_0} \|x(t)-y(t)\| \le T_0 L \max_{|s-t_0|\le T_0} \|x(s)-y(s)\|. \tag{1.18}$$

As long as (1.12) holds, $T_0 L \le 1/2$, so (1.18) clearly implies $\max_{|t-t_0|\le T_0} \|x(t)-y(t)\| = 0$, which gives the asserted uniqueness.

Note that the Lipschitz hypothesis (1.2) was needed only for $x, y \in \overline{B_R(x_0)}$. Thus we can extend Proposition 1.1 to the following setting:

$$\begin{gathered} \text{For each closed, bounded } K \subset \Omega, \text{ there exists } L_K < \infty \text{ such that} \\ \|F(t,x)-F(t,y)\| \le L_K \|x-y\|, \ \forall\, x,y \in K, \ t \in I. \end{gathered} \tag{1.19}$$

We can also replace the bound on F by

$$\begin{gathered} \text{For each } K \text{ as above, there exists } M_K < \infty \text{ such that} \\ \|F(t,x)\| \le M_K, \ \forall\, x \in K, \ t \in I. \end{gathered} \tag{1.20}$$

Results of Appendix B imply that there exists $R_K > 0$ such that

$$\widetilde{K} = \overline{\bigcup_{x \in K} B_{R_K}(x)} \quad \text{is a compact subset of} \ \ \Omega.$$

It follows that for each $x_0 \in K$, the solution to (1.1) exists on the interval

$$\{t \in I : |t-t_0| \le \min(R_K/M_{\widetilde{K}}, 1/2L_{\widetilde{K}})\}. \tag{1.21}$$

Now that we have local solutions to (1.1), it is of interest to investigate when global solutions exist. Here is an example of breakdown:

$$\frac{dx}{dt} = x^2, \quad x(0) = 1. \tag{1.22}$$

Here $I = \mathbb{R}, \ n = 1, \ \Omega = \mathbb{R}$, and $F(x) = x^2$ is smooth, satisfying the local bounds (1.20)–(1.21). The equation (1.22) has the unique solution

$$x(t) = \frac{1}{1-t}, \quad t \in (-\infty, 1), \tag{1.23}$$

which blows up as $t \nearrow 1$. It is useful to know that "blowing up" is the only way a solution can fail to exist globally. We have the following result.

Proposition 1.2. *Let F be as in Proposition 1.1, but with the Lipschitz and boundedness hypotheses relaxed to (1.19)–(1.20). Assume $[a,b]$ is contained in the open interval I and assume $x(t)$ solves (1.1) for $t \in (a,b)$. Assume there exists a closed, bounded set $K \subset \Omega$ such that $x(t) \in K$ for all $t \in (a,b)$. Then there exist $a_1 < a$ and $b_1 > b$ such that $x(t)$ solves (1.1) for $t \in (a_1,b_1)$.*

Proof. We deduce from (1.21) that there exists $\delta > 0$ such that for each $x_1 \in K,\ t_1 \in [a,b]$, the solution to

$$\frac{dx}{dt} = F(t,x), \quad x(t_1) = x_1 \tag{1.24}$$

exists on the interval $[t_1-\delta, t_1+\delta]$. Now, under the current hypotheses, take $t_1 \in (b-\delta/2, b),\ x_1 = x(t_1)$, with $x(t)$ solving (1.1). Then solving (1.24) continues $x(t)$ past $t = b$. Similarly one can continue $x(t)$ past $t = a$.

Here is an example of a global existence result that can be deduced from Proposition 1.2. Consider the 2×2 system for $x = (y,v)$:

$$\begin{aligned} \frac{dy}{dt} &= v, \\ \frac{dv}{dt} &= -y^3. \end{aligned} \tag{1.25}$$

Here we have $\Omega = \mathbb{R}^2,\ F(t,x) = F(t,y,v) = (v,-y^3)$. If (1.25) holds for $t \in (a,b)$, we have

$$\frac{d}{dt}\Bigl(\frac{v^2}{2} + \frac{y^4}{4}\Bigr) = v\frac{dv}{dt} + y^3\frac{dy}{dt} = 0, \tag{1.26}$$

so each $x(t) = (y(t),v(t))$ solving (1.25) lies in a level curve $y^4/4 + v^2/2 = C$, hence is confined to a closed, bounded subset of $\mathbb{R}^2$, yielding global existence of solutions to (1.25).

We can also apply Proposition 1.2 to establish global existence of solutions to linear systems,

$$\frac{dx}{dt} = A(t)x, \quad x(0) = x_0, \tag{1.27}$$

given $A(t)$ continuous in $t \in I$ (an interval about 0), with values in $M(n,\mathbb{C})$. It suffices to establish the following.

Proposition 1.3. *If $\|A(t)\| \le K$ for $t \in I$, then the solution to (1.27) satisfies*

$$\|x(t)\| \le e^{K|t|}\|x_0\|. \tag{1.28}$$

Proof. It suffices to prove (1.28) for $t \geq 0$. Then $y(t) = e^{-Kt}x(t)$ satisfies

$$\frac{dy}{dt} = C(t)y, \quad y(0) = x_0, \tag{1.29}$$

with $C(t) = A(t) - K$. Hence $C(t)$ satisfies

$$\operatorname{Re}(C(t)u, u) \leq 0, \quad \forall\, u \in \mathbb{C}^n. \tag{1.30}$$

Then (1.28) is a consequence of the following estimate, of interest in its own right.

Lemma 1.4. *If $y(t)$ solves (1.29) and (1.30) holds for $C(t)$, then*

$$\|y(t)\| \leq \|y(0)\| \quad \textit{for} \quad t \geq 0. \tag{1.31}$$

Proof. We have

$$\begin{aligned} \frac{d}{dt}\|y(t)\|^2 &= (y'(t), y(t)) + (y(t), y'(t)) \\ &= 2\operatorname{Re}(C(t)y(t), y(t)) \\ &\leq 0. \end{aligned} \tag{1.32}$$

Thanks to Proposition 1.3, we have for $s, t \in I$, the solution operator for (1.27),

$$S(t, s) \in M(n, \mathbb{C}), \quad S(t, s)x(s) = x(t), \tag{1.33}$$

introduced in §8 of Chapter 3. As noted there, we have the Duhamel formula

$$x(t) = S(t, t_0) + \int_{t_0}^{t} S(t, s)f(s)\, ds, \tag{1.34}$$

for the solution to

$$\frac{dx}{dt} = A(t)x + f(t), \quad x(t_0) = x_0. \tag{1.35}$$

If $F(t, x)$ depends explicitly on t, we call (1.1) a non-autonomous system. If F does not depend explicitly on t, we say (1.1) is autonomous. The following device converts a non-autonomous system to an autonomous one. Take the $n \times n$ system (1.1). Then the $(n+1) \times (n+1)$ system

$$\frac{dx}{dt} = F(y, x), \ \frac{dy}{dt} = 1, \quad x(t_0) = x_0, \ y(t_0) = t_0 \tag{1.36}$$

has the autonomous form

$$\frac{dz}{dt} = G(z), \quad z(t_0) = (x_0, t_0), \tag{1.37}$$

for $z = (x, y)$, with $G(z) = (F(y, x), 1)$, and the solution to (1.36) is $(x(t), t)$, where $x(t)$ solves (1.1). Thus for many purposes it suffices to consider autonomous sytems.

To close this section, we note how a higher order $n \times n$ system, such as

$$\frac{d^k x}{dt^k} = F(t, x, \dots, x^{(k-1)}), \quad x(t_0) = x_0, \dots, x^{(k-1)}(t_0) = x_{k-1}, \tag{1.38}$$

can be converted to a first order $nk \times nk$ system, for

$$y = \begin{pmatrix} y_0 \\ \vdots \\ y_{k-1} \end{pmatrix}, \quad y_j(t) \in \mathbb{R}^n, \quad 0 \le j \le k-1. \tag{1.39}$$

The system is

$$\begin{aligned} \frac{dy_0}{dt} &= y_1, \\ &\vdots \\ \frac{dy_{k-2}}{dt} &= y_{k-1}, \\ \frac{dy_{k-1}}{dt} &= F(t, y_0, \dots, y_{k-1}), \end{aligned} \tag{1.40}$$

with initial data

$$y_j(t_0) = x_j, \quad 0 \le j \le k-1. \tag{1.41}$$

If $y(t)$ solves (1.40)–(1.41), then $x(t) = y_0(t)$ solves (1.38), and we have

$$x^{(j)}(t) = y_j(t), \quad 0 \le j \le k-1. \tag{1.42}$$

Note how this construction is parallel to that done in the linear case in Chapter 3, §3.

Exercises

1. Apply the Picard iteration method to

$$\frac{dx}{dt} = ax, \quad x(0) = 1,$$

given $a \in \mathbb{C}$. Taking $x_0(t) \equiv 1$, show that

$$x_n(t) = \sum_{k=0}^{n} \frac{a^k}{k!} t^k.$$

2. Discuss the matrix analogue of Exercise 1.

3. Consider the initial value problem

$$\frac{dx}{dt} = x^2, \quad x(0) = 1.$$

Take $x_0 \equiv 1$ and use the Picard iteration method (1.5) to write out

$$x_n(t), \quad n = 1, 2, 3.$$

Compare the results with the formula (1.23).

4. Given A_0, $A_1 \in M(n, \mathbb{C})$, consider the initial value problem

$$\frac{dx}{dt} = (A_0 + A_1 t)x, \quad x(0) = x_0.$$

Take $x_0(t) \equiv x_0$ and use the Picard iteration (1.5) to write out

$$x_n(t), \quad n = 1, 2, 3.$$

Compare and contrast the results with calculations from §10 of Chapter 3.

5. Let $x_n(t)$ be an approximate solution to (1.1), and assume that

$$\|x(t) - x_n(t)\| \le \delta_n |t - t_0|^n, \quad \text{for} \ \ t \in I.$$

Let $x_{n+1}(t)$ be defined by (1.5), and assume the Lipschitz condition (1.2) holds. Show that

$$\|x(t) - x_{n+1}(t)\| \le \frac{L\delta_n}{n+1} |t - t_0|^{n+1}, \quad t \in I.$$

6. Modify the system (1.25) to

$$\frac{dy}{dt} = v, \quad \frac{dv}{dt} = -y^3 - v.$$

Show that solutions satisfy

$$\frac{d}{dt}\Bigl(\frac{v^2}{2}+\frac{y^4}{4}\Bigr)\le 0,$$

and use this to establish global existence for $t \ge 0$.

7. Consider the initial value problem

$$\frac{dx}{dt} = |x|^{1/2}, \quad x(0) = 0.$$

Note that $x(t) \equiv 0$ is a solution, and

$$\begin{aligned} x(t) = \frac{1}{4}t^2, &\quad t \ge 0, \\ 0, &\quad t \le 0 \end{aligned}$$

is another solution, on $t \in (-\infty,\infty)$. Why does this not contradict the uniqueness part of Proposition 1.1? Can you produce other solutions to this initial value problem?

8. Take $\beta \in (0,\infty)$ and consider the initial value problem

$$\frac{dx}{dt} = x^\beta, \quad x(0) = 1.$$

Show that this has a solution for all $t \ge 0$ if and only if $\beta \le 1$.

9. Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be C^1 and suppose $x(t)$ solves

$$\frac{dx}{dt} = F(x), \quad x(t_0) = x_0, \tag{1.43}$$

for $t \in I$, an open interval containing t_0. Show that, for $t \in I$,

$$\frac{d}{dt}\|x(t)\|^2 = 2x(t)\cdot F(x(t)). \tag{1.44}$$

Show that, if $\alpha > 0$ and $x(t) \ne 0$,

$$\frac{d}{dt}\|x(t)\|^\alpha = \alpha\|x(t)\|^{\alpha-2}x(t)\cdot F(x(t)). \tag{1.45}$$

10. In the setting of Exercise 9, suppose F satisfies an estimate

$$\|F(x)\| \le C(1+\|x\|)^\beta, \quad \forall x \in \mathbb{R}^n, \quad C < \infty, \quad \beta < 1. \tag{1.46}$$

Show that there exists $\alpha > 0$ and $K < \infty$ such that, if $\|x(t)\| \ge 1$ for $t \in I$,

$$\frac{d}{dt}\|x(t)\|^\alpha \le K, \quad \forall\, t \in I.$$

Use this to establish that the solution to (1.43) exists for all $t \in \mathbb{R}$.

Exercises 11–13 below will extend the conclusion of Exercise 10 to the case $\beta = 1$ in (1.46). One approach is via the following result, known as *Gronwall's inequality.*

Proposition 1.5. *Assume*

$$g \in C^1(\mathbb{R}), \quad g' \geq 0. \tag{1.47}$$

Let u and v be real valued, continuous functions on I satisfying

$$\begin{aligned} u(t) &\leq A + \int_{t_0}^t g(u(s))\, ds, \\ v(t) &\geq A + \int_{t_0}^t g(v(s))\, ds. \end{aligned} \tag{1.48}$$

Then

$$u(t) \leq v(t), \quad \text{for } \ t \in I,\ t \geq t_0. \tag{1.49}$$

Proof. Set $w(t) = u(t) - v(t)$. Then

$$\begin{aligned} w(t) &\leq \int_{t_0}^t \bigl[g(u(s)) - g(v(s))\bigr]\, ds \\ &= \int_{t_0}^t M(s)w(s)\, ds, \end{aligned} \tag{1.50}$$

where

$$M(s) = \int_0^1 g'(\tau u(s) + (1-\tau)v(s))\, d\tau. \tag{1.51}$$

Hence we have

$$w(t) \leq \int_{t_0}^t M(s)w(s)\, ds, \quad M(s) \geq 0, \quad M \in C(I), \tag{1.52}$$

and we claim this implies

$$w(t) \leq 0, \quad \forall\, t \in I,\ t \geq t_0. \tag{1.53}$$

In other words, we claim that $w(t) \leq 0$ on $[t_0, b]$ whenever $[t_0, b] \subset I$. To see this, let t_1 be the largest number in $[t_0, b]$ with the property that $w \leq 0$ on $[t_0, t_1]$. We claim that $t_1 = b$.

Assume to the contrary that $t_1 < b$. Noting that $\int_{t_0}^{t_1} M(s)w(s)\,ds \le 0$, we deduce from (1.52) that

$$w(t) \le \int_{t_1}^{t} M(s)w(s)\,ds, \quad \forall\, t \in [t_1, b]. \tag{1.54}$$

Hence, with

$$K = \max_{[t_1,b]} M(s) < \infty, \tag{1.56}$$

we have, for $a \in (t_1, b)$,

$$\max_{[t_1,a]} w(t) \le (a - t_1)K \max_{[t_1,a]} w(s). \tag{1.57}$$

If we pick $a \in (t_1, b)$ such that $(a - t_1)K < 1$, this implies

$$w(t) \le 0, \quad \forall\, t \in [t_1, a], \tag{1.58}$$

contradicting the maximality of t_1. Hence actually $t_1 = b$, and we have the implication (1.52) $\Rightarrow$ (1.53), completing the proof of Proposition 1.5.

11. Assume $v \ge 0$ is a C^1 function on $I = (a, b)$, satisfying

$$\frac{dv}{dt} \le Cv, \quad v(t_0) = v_0, \tag{1.59}$$

where $C \in (0, \infty)$ and $t_0 \in I$. Using Proposition 1.5, show that

$$v(t) \le e^{C(t-t_0)} v_0, \quad \forall\, t \in [t_0, b). \tag{1.60}$$

12. In the setting of Exercise 11, avoid use of Proposition 1.5 as follows. Write (1.59) as

$$\frac{dv}{dt} = Cv - g(t), \quad v(t_0) = v_0, \ \ g \ge 0, \tag{1.61}$$

with solution

$$v(t) = e^{C(t-t_0)} v_0 - \int_{t_0}^{t} e^{C(t-s)} g(s)\,ds. \tag{1.62}$$

Deduce (1.60) from this.

13. Return to the setting of Exercise 9, and replace the hypothesis (1.46) by

$$\|F(x)\| \le C(1 + \|x\|), \quad \forall\, x \in \mathbb{R}^n. \tag{1.63}$$

Show that the solution to (1.43) exists for all $t \in \mathbb{R}$.
Hint. Take $v(t) = 1 + \|x(t)\|^2$ and use (1.44). Show that Exercise 11 (or 12) applies.

2. Dependence of solutions on initial data and other parameters

We study how the solution to a system of differential equations

$$\frac{dx}{dt} = F(x), \quad x(0) = y \tag{2.1}$$

depends on the initial condition y. As shown in §1, there is no loss of generality in considering the autonomous system (2.1). We will assume $F : \Omega \to \mathbb{R}^n$ is smooth, $\Omega \subset \mathbb{R}^n$ open and convex, and denote the solution to (2.1) by $x = x(t, y)$. We want to examine smoothness in y. Let $DF(x)$ denote the $n \times n$ matrix valued function of partial derivatives of F. (See Appendix A for more on this derivative.)

To start, we assume F is of class C^1, i.e., DF is continuous on Ω, and we want to show $x(t, y)$ is differentiable in y. Let us recall what this means. Take $y \in \Omega$ and pick $R > 0$ such that $\overline{B_R(y)}$, defined as in (1.6), is contained in Ω. We seek an $n \times n$ matrix $W(t, y)$ such that, for $w_0 \in \mathbb{R}^n$, $\|w_0\| \le R$,

$$x(t, y + w_0) = x(t, y) + W(t, y)w_0 + r(t, y, w_0), \tag{2.2}$$

where

$$r(t, y, w_0) = o(\|w_0\|), \tag{2.3}$$

which means

$$\lim_{w_0 \to 0} \frac{r(t, y, w_0)}{\|w_0\|} = 0. \tag{2.4}$$

When this holds, $x(t, y)$ is differentiable in y, and

$$D_y x(t, y) = W(t, y). \tag{2.5}$$

In other words,

$$x(t, y + w_0) = x(t, y) + D_y x(t, y)w_0 + o(\|w_0\|). \tag{2.6}$$

In the course of proving this differentiability, we also want to produce an equation for $W(t, y) = D_y x(t, y)$. This can be done as follows. Suppose $x(t, y)$ were differentiable in y. (We do not yet know that it is, but that is okay.) Then $F(x(t, y))$ is differentiable in y, so we can apply D_y to (2.1). Using the chain rule, we get the following equation,

$$\frac{dW}{dt} = DF(x)W, \quad W(0, y) = I, \tag{2.7}$$

called the linearization of (2.1). Here, I is the $n \times n$ identity matrix. Equivalently, given $w_0 \in \mathbb{R}^n$,

$$w(t,y) = W(t,y)w_0 \tag{2.8}$$

is expected to solve

$$\frac{dw}{dt} = DF(x)w, \quad w(0) = w_0. \tag{2.9}$$

Now, we do not yet know that $x(t,y)$ is differentiable, but we do know from results of §1 that (2.7) and (2.9) are uniquely solvable. It remains to show that, with such a choice of $W(t,y)$, (2.2)–(2.3) hold.

To rephrase the task, set

$$x(t) = x(t,y), \quad x_1(t) = x(t, y + w_0), \quad z(t) = x_1(t) - x(t), \tag{2.10}$$

and let $w(t)$ solve (2.9). The task of verifying (2.2)–(2.3) is equivalent to the task of verifying

$$\|z(t) - w(t)\| = o(\|w_0\|). \tag{2.11}$$

To show this, we will obtain for $z(t)$ an equation similar to (2.9). To begin, (2.10) implies

$$\frac{dz}{dt} = F(x_1) - F(x), \quad z(0) = w_0. \tag{2.12}$$

Now the fundamental theorem of calculus gives

$$F(x_1) - F(x) = G(x_1, x)(x_1 - x), \tag{2.13}$$

with

$$G(x_1, x) = \int_0^1 DF\big(\tau x_1 + (1-\tau)x\big)\, d\tau. \tag{2.14}$$

If F is C^1, then G is continuous. Then (2.12)–(2.13) yield

$$\frac{dz}{dt} = G(x_1, x)z, \quad z(0) = w_0. \tag{2.15}$$

Given that

$$\|DF(u)\| \le L, \quad \forall\, u \in \Omega, \tag{2.16}$$

which we have by continuity of DF, after possibly shrinking Ω slightly, we deduce from Proposition 1.3 that

$$\|z(t)\| \le e^{|t|L}\|w_0\|, \tag{2.17}$$

that is,

$$\|x(t,y) - x(t,y+w_0)\| \le e^{|t|L}\|w_0\|. \tag{2.18}$$

This establishes that $x(t,y)$ is *Lipschitz* in y.

To proceed, since G is continuous and $G(x,x) = DF(x)$, we can rewrite (2.15) as

$$\frac{dz}{dt} = G(x+z,x)z = DF(x)z + R(x,z), \quad z(0) = w_0, \tag{2.19}$$

where

$$F \in C^1(\Omega) \Longrightarrow \|R(x,z)\| = o(\|z\|) = o(\|w_0\|). \tag{2.20}$$

Now comparing (2.19) with (2.9), we have

$$\frac{d}{dt}(z-w) = DF(x)(z-w) + R(x,z), \quad (z-w)(0) = 0. \tag{2.21}$$

Then Duhamel's formula gives

$$z(t) - w(t) = \int_0^t S(t,s)R(x(s),z(s))\,ds, \tag{2.22}$$

where $S(t,s)$ is the solution operator for $d/dt - B(t)$, with $B(t) = DF(x(t))$, which as in (2.17), satisfies

$$\|S(t,s)\| \le e^{|t-s|L}. \tag{2.23}$$

We hence have (2.11), i.e.,

$$\|z(t) - w(t)\| = o(\|w_0\|). \tag{2.24}$$

This is precisely what is required to show that $x(t,y)$ is differentiable with respect to y, with derivative $W = D_y x(t,y)$ satisfying (2.7). Hence we have:

Proposition 2.1. *If $F \in C^1(\Omega)$ and if solutions to (2.1) exist for $t \in (-T_0, T_1)$, then, for each such t, $x(t,y)$ is C^1 in y, with derivative $D_y x(t,y)$ satisfying (2.7).*

We have shown that $x(t,y)$ is both Lipschitz and differentiable in y. The continuity of $W(t,y)$ in y follows easily by comparing the differential equations of the form (2.7) for $W(t,y)$ and $W(t,y+w_0)$, in the spirit of the analysis of $z(t)$ done above.

If F possesses further smoothness, we can establish higher differentiability of $x(t,y)$ in y by the following trick. Couple (2.1) and (2.7), to get a system of differential equations for (x,W):

$$\begin{aligned}\frac{dx}{dt} &= F(x),\\ \frac{dW}{dt} &= DF(x)W,\end{aligned} \tag{2.25}$$

with initial conditions

$$x(0)=y, \quad W(0)=I. \tag{2.26}$$

We can reiterate the preceding argument, getting results on $D_y(x,W)$, hence on $D_y^2 x(t,y)$, and continue, proving:

Proposition 2.2. *If $F \in C^k(\Omega)$, then $x(t,y)$ is C^k in y.*

Similarly, we can consider dependence of the solution to

$$\frac{dx}{dt} = F(\tau,x), \quad x(0)=y \tag{2.27}$$

on a parameter τ, assuming F smooth jointly in (τ,x). This result can be deduced from the previous one by the following trick. Consider the system

$$\frac{dx}{dt} = F(z,x), \ \frac{dz}{dt}=0, \quad x(0)=y, \ z(0)=\tau. \tag{2.28}$$

Then we get smoothness of $x(t,\tau,y)$ jointly in (τ,y). As a special case, let $F(\tau,x)=\tau F(x)$. In this case $x(t_0,\tau,y)=x(\tau t_0,y)$, so we can improve the conclusion in Proposition 2.2 to the following:

$$F \in C^k(\Omega) \Longrightarrow x \in C^k \ \text{ jointly in } \ (t,y). \tag{2.29}$$

Exercises

1. Suppose $\tau \in \mathbb{R}$ in (2.27). Show that $\xi = \partial x/\partial\tau$ satisfies

$$\frac{d\xi}{dt} = D_x F(\tau,x)\xi + \frac{\partial}{\partial\tau}F(\tau,x), \quad \xi(0)=0.$$

2. Consider the family of differential equations for $x_\tau(t)$,

$$\frac{dx}{dt} = x + \tau x^2, \quad x(0) = 1.$$

Write down the differential equations satisfied by $\xi = \partial x/\partial\tau$ and by $\eta = \partial^2 x/\partial\tau^2$.

3. Let $x = x_\tau(t),\ y = y_\tau(t)$ solve

$$\frac{dx}{dt} = -y + \tau(x^2 + y^2), \quad \frac{dy}{dt} = x, \quad x(0) = 1,\ \ y(0) = 0. \tag{2.30}$$

Knowing smooth dependence on τ, find differential equations for the coefficients $X_j(t), Y_j(t)$ in power series expansions

$$\begin{aligned} x_\tau(t) &= X_0(t) + \tau X_1(t) + \tau^2 X_2(t) + \cdots, \\ y_\tau(t) &= Y_0(t) + \tau Y_1(t) + \tau^2 Y_2(t) + \cdots. \end{aligned} \tag{2.31}$$

Note that $X_0(t) = \cos t,\ Y_0(t) = \sin t$.

4. Using the substitution $\xi(t) = -x(-t),\ \eta(t) = y(-t)$, show that, for τ sufficiently small, solutions to (2.30) are periodic in t.

5. Let $p(\tau)$ denote the period of the solution to (2.30). Using (2.31), show that $p(\tau)$ is smooth in τ for $|\tau|$ small. Note that $p(0) = 2\pi$. Compute $p'(0)$. Compare results in Appendix C.

6. Suppose y in (2.1) is a critical point of F, i.e., $F(y) = 0$. Show that (2.7) becomes

$$\frac{dW}{dt} = LW, \quad W(0) = I, \quad \text{where} \ \ L = DF(y),$$

hence

$$F(y) = 0 \Longrightarrow D_y x(t, y) = e^{tL}.$$

3. Vector fields, orbits, and flows

Let $\Omega \subset \mathbb{R}^n$ be an open set. A vector field on Ω is simply a map

$$F : \Omega \longrightarrow \mathbb{R}^n, \tag{3.1}$$

such as encountered in (2.1). We say F is a C^k vector field if F is a C^k map. A C^∞ vector field is said to be smooth. By convention, if we simply call F a vector field, we mean it is a smooth vector field. In this section we always assume F is at least C^1.

One can also look at time-dependent vector fields (cf. (1.1)), but in this section we restrict attention to the autonomous case.

The solution to (2.1), i.e., to

$$\frac{dx}{dt} = F(x), \quad x(0) = y, \tag{3.2}$$

will be denoted

$$x(t) = \Phi_F^t(y). \tag{3.3}$$

Results of §§1–2 imply that for each closed bounded $K \subset \Omega$ there exists an interval $I = (-T_0, T_1)$ about 0 such that, for each $t \in I$,

$$\Phi_F^t : K \longrightarrow \Omega, \tag{3.4}$$

and this is a C^k map if F is a C^k vector field. The family of maps Φ_F^t from K to Ω is called the *flow* generated by F. We have

$$\Phi_F^0(y) \equiv y, \tag{3.5}$$

i.e., Φ_F^0 is the identity map. We also have

$$\Phi_F^{s+t}(y) = \Phi_F^t \circ \Phi_F^s(y), \tag{3.6}$$

provided all these maps are well defined. Given $y \in \Omega$, the path

$$t \mapsto \Phi_F^t(y) \tag{3.7}$$

is called the *orbit* through y.

Another way to state the defining property of Φ_F^t is that (3.5) holds and

$$\frac{d}{dt}\Phi_F^t(x) = F(\Phi_F^t(x)). \tag{3.8}$$

We next obtain interesting information on the t-derivative of

$$v^t(x) = v(\Phi_F^t(x)), \tag{3.9}$$

given $v \in C_0^1(\Omega)$, i.e., v is of class C^1 and vanishes outside some closed bounded $K \subset \Omega$. The chain rule (cf. Appendix A, especially (A.8)) plus (3.8) yields

$$\frac{d}{dt}v^t(x) = F(\Phi_F^t(x)) \cdot \nabla v(\Phi_F^t(x)). \tag{3.10}$$

In particular,

$$\frac{d}{ds} v(\Phi_F^s(x))\Big|_{s=0} = F(x)\cdot \nabla v(x). \tag{3.11}$$

Here ∇v is the gradient of v, given by $\nabla v = (\partial v/\partial x_1, \dots, \partial v/\partial x_n)^t$. A useful alternative formula to (3.10) is

$$\begin{aligned} \frac{d}{dt} v^t(x) &= \frac{d}{ds} v^t(\Phi_F^s(x))\Big|_{s=0} \\ &= F(x)\cdot \nabla v^t(x), \end{aligned} \tag{3.12}$$

the first equality following from (3.6) and the second from (3.11), with v replaced by v^t.

One significant consequence of (3.12), which will lead to the important result (3.17) below, is that, for $v \in C_0^1(\Omega)$,

$$\begin{aligned} \frac{d}{dt}\int_\Omega v(\Phi_F^t(x))\,dx &= \int_\Omega F(x)\cdot\nabla v^t(x)\,dx \\ &= -\int_\Omega \operatorname{div} F(x)\, v(\Phi_F^t(x))\,dx. \end{aligned} \tag{3.13}$$

Here $\operatorname{div} F(x)$ is the *divergence* of the vector field $F(x) = (F_1(x), \dots, F_n(x))^t$, defined by

$$\operatorname{div} F(x) = \frac{\partial F_1}{\partial x_1}(x) + \cdots + \frac{\partial F_n}{\partial x_n}(x). \tag{3.14}$$

The last equality in (3.13) follows by integration by parts,

$$\int_\Omega F_k(x)\frac{\partial v^t}{\partial x_k}\,dx = -\int_\Omega \frac{\partial F_k}{\partial x_k} v^t(x)\,dx,$$

followed by summation over k. We reiterate the content of (3.13):

$$\frac{d}{dt}\int_\Omega v(\Phi_F^t(x))\,dx = -\int_\Omega \operatorname{div} F(x)\, v(\Phi_F^t(x))\,dx. \tag{3.15}$$

So far, we have (3.15) for $v \in C_0^1(\Omega)$. We can extend this by noting that (3.15) implies

$$\begin{aligned} &\int_\Omega v(\Phi_F^t(x))\,dx - \int_\Omega v(x)\,dx \\ &\quad = -\int_0^t \int_\Omega \operatorname{div} F(x) v(\Phi_F^s(x))\,dx\,ds. \end{aligned} \tag{3.16}$$

Basic results on the integral allow one to pass from $v \in C_0^1(\Omega)$ in (3.16) to more general v, including $v = \chi_B$ (the characteristic function of B, defined to be equal to 1 on B and 0 on $\Omega \setminus B$), for smoothly bounded closed $B \subset \Omega$, amongst other functions.

In more detail, if $B \subset \Omega$ is a smoothly bounded, closed set, let $B_\delta = \{x \in \mathbb{R}^n : \operatorname{dist}(x, B) \le \delta\}$. There exists $\delta_0 > 0$ such that $B_\delta \subset \Omega$ for $\delta \in (0, \delta_0]$. For such δ, one can produce $v_\delta \in C_0^1(\Omega)$ such that

$$v_\delta = 1 \ \text{ on } \ B, \quad 0 \le v_\delta \le 1, \quad v_\delta = 0 \ \text{ on } \ \mathbb{R}^n \setminus B_\delta.$$

Then

$$\Bigl| \int \chi_B(x)\, dx - \int v_\delta(x)\, dx \Bigr| \le \operatorname{vol}(B_\delta \setminus B) \to 0, \ \text{ as } \ \delta \to 0,$$

so, as $\delta \to 0$,

$$\int\limits_\Omega v_\delta(x)\, dx \longrightarrow \int\limits_\Omega \chi_B(x)\, dx.$$

Similar arguments give

$$\int\limits_\Omega v_\delta(\Phi_F^t(x))\, dx \longrightarrow \int\limits_\Omega \chi_B(\Phi_F^t(x))\, dx,$$

and

$$\int_0^t \int\limits_\Omega \operatorname{div} F(x)\, v_\delta(\Phi_F^s(x))\, dx\, ds \longrightarrow \int_0^t \int\limits_\Omega \operatorname{div} F(x)\, \chi_B(\Phi_F^s(x))\, dx\, ds.$$

These results allow one to take $v = \chi_B$ in (3.16).

Now one can pass from (3.16) back to (3.15), via the fundamental theorem of calculus. Note that

$$\operatorname{Vol} \Phi_F^t(B) = \int \chi_B(\Phi_F^{-t}(x))\, dx.$$

We can apply (3.15) with t replaced by $-t$, and v by χ_B, and deduce the following.

Proposition 3.1. *If F is a C^1 vector field, generating the flow Φ_F^t, well defined on Ω for $t \in I$, and $B \subset \Omega$ is smoothly bounded, then, for $t \in I$,*

$$\frac{d}{dt} \operatorname{Vol} \Phi_F^t(B) = \int\limits_{\Phi_F^t(B)} \operatorname{div} F(x)\, dx. \tag{3.17}$$

This result is behind the notation div F, i.e., the *divergence* of F. Vector fields F with positive divergence generate flows Φ_F^t that magnify volumes as t increases, while vector fields with negative divergence generate flows that shrink volumes as t increases.

We say the flow generated by a vector field F is *complete* provided $\Phi_F^t(y)$ is defined for all $t \in \mathbb{R}$, $y \in \Omega$. We say it is forward complete if $\Phi_F^t(y)$ is defined for all $t \in [0, \infty)$, $y \in \Omega$. The flow is backward complete if $\Phi_F^t(y)$ is defined for all $t \in (-\infty, 0]$, $y \in \Omega$. Here is an occasionally useful criterion for forward completeness.

Proposition 3.2. *Let F be a C^1 vector field on $\Omega = \mathbb{R}^n$. Assume there exists $R < \infty$ and a function $V \in C^1(\mathbb{R}^n)$ such that*

$$V(x) \to +\infty \quad \textit{as} \quad \|x\| \to \infty \tag{3.18}$$

and

$$\|x\| \geq R \Longrightarrow \nabla V(x) \cdot F(x) \leq 0. \tag{3.19}$$

Then the flow Φ_F^t is forward complete.

Proof. Let $x(t) = \Phi_F^t(x_0)$ be an orbit, defined for $t \in I$, some interval about 0. Then

$$\|x(t)\| \geq R \Longrightarrow \frac{d}{dt} V(x(t)) = \nabla V(x(t)) \cdot F(x(t)) \leq 0. \tag{3.20}$$

Hence, for $t \in I,\ t \geq 0,\ x(t)$ is confined to the closed bounded set

$$\Bigl\{ x \in \mathbb{R}^n : V(x) \leq \max V(y),\ y \in B_R(0) \cup \{x_0\} \Bigr\}. \tag{3.21}$$

From here, Proposition 1.2 yields forward completeness.

One way to display the behavior of the flow generated by a vector field F on a domain Ω is to draw a "phase portrait." This consists of graphs of selected integral curves of F, with arrows indicating the direction of F along each integral curve. Such portraits are particularly revealing when $\dim \Omega = 2$, and also of considerable use when $\dim \Omega = 3$. As an example, consider Fig. 3.1, the phase portrait of the flow associated to the 2×2 system

$$\begin{aligned} \frac{d\theta}{dt} &= \psi, \\ \frac{d\psi}{dt} &= -\frac{g}{\ell} \sin \theta, \end{aligned} \tag{3.22}$$

Figure 3.1

which arises from the pendulum equation (cf. Chapter 1, (6.6))

$$\frac{d^2\theta}{dt^2} + \frac{g}{\ell}\sin\theta = 0, \tag{3.23}$$

by adding the variable $\psi = d\theta/dt$. Here $g, \ell > 0$. The system (3.22) has the form (3.2) with $x = (\theta, \psi)$ and

$$F(\theta, \psi) = \begin{pmatrix} \psi \\ -\frac{g}{\ell}\sin\theta \end{pmatrix}. \tag{3.24}$$

Note that Fig. 3.1 looks like Fig. 6.2 of Chapter 1, except that here we have added arrows, to indicate the direction of the flow. As noted in Chapter 1, the orbits of this flow are level curves of the function

$$\mathcal{E}(\theta, \psi) = \frac{\psi^2}{2} - \frac{g}{\ell}\cos\theta, \tag{3.25}$$

since if $(\theta(t), \psi(t))$ solves (3.22),

$$\frac{d}{dt}\mathcal{E}(\theta, \psi) = \psi\psi' + \frac{g}{\ell}(\sin\theta)\theta' = 0. \tag{3.26}$$

It is instructive to expand on this last calculation. In general, if $(\theta', \psi') = F(\theta, \psi)$,

$$\frac{d}{dt}\mathcal{E}(\theta, \psi) = \nabla\mathcal{E}(\theta, \psi) \cdot F(\theta, \psi), \quad \text{where} \quad \nabla\mathcal{E}(\theta, \psi) = \begin{pmatrix} \partial\mathcal{E}/\partial\theta \\ \partial\mathcal{E}/\partial\psi \end{pmatrix}. \tag{3.27}$$

Figure 3.2

Now the formula (3.24) gives

$$F(\theta, \psi) = -J\,\nabla\mathcal{E}(\theta, \psi), \tag{3.28}$$

where

$$J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \tag{3.29}$$

so the vanishing of $d\mathcal{E}(\theta, \psi)/dt$ follows from (3.27)–(3.28) and the skew-symmetry of J, which implies

$$v \cdot Jv = 0, \quad \forall\, v \in \mathbb{R}^2. \tag{3.30}$$

A vector field of the form (3.28) is a special case of a *Hamiltonian* vector field, a class of vector fields that will be discussed further in §§5, 7, and 10.

We mention some noteworthy features of the phase portrait in Fig. 3.1, features to look for in other such portraits. First, there are the *critical* points of F, i.e., the points where F vanishes. In case (3.24), the set of critical points in

$$\{(k\pi, 0) : k \in \mathbb{Z}\}.$$

Fig. 3.1 indicates different natures of the orbits near these critical points, depending on whether k is even or odd. For k even, the orbits near $(k\pi, 0)$ consist of closed curves. We say these critical points are *centers*; cf. Fig. 3.2.

For k odd, the orbits near $p = (k\pi, 0)$ consist of curves of the following

Figure 3.3

nature:

$$
\text{(3.31)} \qquad \begin{aligned}
&\text{(a)} \quad \text{two orbits that approach } p \text{ as } t \to +\infty, \\
&\text{(b)} \quad \text{two orbits that approach } p \text{ as } t \to -\infty, \\
&\text{(c)} \quad \text{orbits that miss } p, \text{ looking like saddles.}
\end{aligned}
$$

We say these critical points are *saddles.* Cf. Fig. 3.3. Sometimes one calls them hyperbolic critical points.

Considerable insight is obtained from the study of the *linearization* of F at each critical point. Generally, if F is a C^1 vector field on $\Omega \subset \mathbb{R}^n$, $x_0 \in \Omega$, and $F(x_0) = 0$, the linearization of F at x_0 is given by

$$
L = DF(x_0) \in \mathcal{L}(\mathbb{R}^n). \tag{3.32}
$$

This construction extends the notion of linearization given in §8 of Chapter 1. We expect that

$$
\Phi_F^t(x_0 + y) \approx x_0 + e^{tL} y,
$$

for $\|y\|$ small. Cf. Exercise 6 of §2 (but mind the change in notation). Going further, we expect some important qualitative features of the flow Φ_F^t near x_0 to be captured by the behavior of e^{tL}, and this is born out, with some exceptions. If $DF(x_0)$ has zero as an eigenvalue (we say x_0 is a degenerate critical point) this approximation is not typically useful. It has a better chance if $\det DF(x_0) \neq 0$. We then say x_0 is a nondegenerate critical point for F.

In case F is given by (3.24), with critical points at $p_k = (k\pi, 0)$, we have

$$L_0 = DF(0,0) = \begin{pmatrix} 0 & 1 \\ -\frac{g}{\ell} & 0 \end{pmatrix}. \tag{3.33}$$

The eigenvalues of this matrix are $\pm i\sqrt{g/\ell}$, and the orbits of e^{tL_0} are ellipses, with qualitative features like Fig. 3.2, a center. Meanwhile,

$$L_1 = DF(\pm\pi,0) = \begin{pmatrix} 0 & 1 \\ \frac{g}{\ell} & 0 \end{pmatrix}. \tag{3.34}$$

The eigenvalues of this matrix are $\pm\sqrt{g/\ell}$, with corresponding eigenvectors $(1, \pm\sqrt{g/\ell})^t$, and the orbit structure for e^{tL_1} has qualitative features like Fig. 3.3, a saddle.

In general, if F is a planar vector field with a nondegenerate critical point at x_0, and if all the eigenvalues of $DF(x_0)$ are purely imaginary, F itself might not have a center at x_0, i.e., the orbits of F near x_0 might not be closed orbits surrounding x_0. Here is an example. Take

$$F(x) = Jx - \|x\|^2 x, \quad x \in \mathbb{R}^2, \tag{3.35}$$

with J as in (3.29). Then $x_0 = 0$ is a critical point, and $DF(0) = J$. Thus the linearization has a center. However, if $x(t)$ is an orbit for this vector field, then

$$\begin{aligned} \frac{d}{dt}\|x(t)\|^2 &= 2x \cdot x' \\ &= 2x \cdot (Jx - \|x\|^2 x) \\ &= -2\|x\|^4, \end{aligned} \tag{3.36}$$

i.e., $\rho(t) = \|x(t)\|^2$ satisfies

$$\frac{d\rho}{dt} = -2\rho^2. \tag{3.37}$$

This is separable and we have

$$\rho(0) = \rho_0 \Longrightarrow \rho(t) = \frac{\rho_0}{1 + 2t\rho_0} \to 0 \ \text{ as } \ t \nearrow +\infty, \tag{3.38}$$

so the orbits of this vector field spiral into the origin as $t \nearrow +\infty$, though much more slowly than they do in the case of spiral sinks, a type of critical point that we will encounter shortly.

Despite the existence of such examples as (3.35), the fact that $(0,0)$ is a center for F, given by (3.24), is no accident, but rather a consequence of the fact that F has the form (3.28),

$$F(x) = -J\nabla\mathcal{E}(x), \tag{3.39}$$

so that, as derived in (3.27)–(3.30), orbits of F lie on level curves of $\mathcal{E}$. Generally, if $\mathcal{E}$ is a smooth real-valued function on a planar domain $\Omega \subset \mathbb{R}^2$ and the vector field F is given by (3.39), (nondegenerate) critical points of F and (nondegenerate) critical points of $\mathcal{E}$ coincide. If $x_0 \in \Omega$ is such a point

$$DF(x_0) = -JD^2\mathcal{E}(x_0), \tag{3.40}$$

where $D^2\mathcal{E}(x_0)$ is the matrix of second-order partial derivatives of $\mathcal{E}$ at x_0, i.e.,

$$D^2\mathcal{E} = \begin{pmatrix} \partial^2\mathcal{E}/\partial\theta^2 & \partial^2\mathcal{E}/\partial\psi\partial\theta \\ \partial^2\mathcal{E}/\partial\theta\partial\psi & \partial^2\mathcal{E}/\partial\psi^2 \end{pmatrix}.$$

We recall the following result, established in basic multivariable calculus. Let x_0 be a nondegenerate critical point of $\mathcal{E}$, so $D^2\mathcal{E}(x_0)$ is an invertible, real symmetric matrix. Then

$$\begin{aligned} D^2\mathcal{E}(x_0) \text{ positive definite } &\Leftrightarrow \mathcal{E} \text{ has a local minimum at } x_0, \\ D^2\mathcal{E}(x_0) \text{ negative definite } &\Leftrightarrow \mathcal{E} \text{ has a local maximum at } x_0, \\ D^2\mathcal{E}(x_0) \text{ indefinite } &\Leftrightarrow \mathcal{E} \text{ has a saddle at } x_0, \end{aligned} \tag{3.41}$$

We also note that, whenever $A \in M(2,\mathbb{R})$ is symmetric and invertible,

$$\begin{aligned} A \text{ positive definite } &\Leftrightarrow \det A > 0 \text{ and } \operatorname{Tr} A > 0, \\ A \text{ negative definite } &\Leftrightarrow \det A > 0 \text{ and } \operatorname{Tr} A < 0, \\ A \text{ indefinite } &\Leftrightarrow \det A < 0. \end{aligned} \tag{3.42}$$

Furthermore, if A is such a matrix and

$$B = -JA, \tag{3.43}$$

then

$$\det B = \det A, \tag{3.44}$$

and, for such $B \in M(2,\mathbb{R})$,

$$\begin{aligned} B \text{ has 2 real eigenvalues of opposite signs } &\Leftrightarrow \det B < 0, \\ B \text{ has 2 purely imaginary eigenvalues } &\Leftrightarrow \det B > 0 \text{ and } \operatorname{Tr} B = 0, \end{aligned} \tag{3.45}$$

Figure 3.4

Putting these observations together (cf. also Exercise 8 below), we have:

Proposition 3.3. *Let $\mathcal{E}$ be a smooth function on $\Omega \subset \mathbb{R}^2$, with a nondegenerate critical point at x_0. Let F be given by (3.39). Then*

$$\begin{aligned} &DF(x_0) \text{ has 2 purely imaginary eigenvalues} \\ &\Leftrightarrow \mathcal{E} \text{ has a local max or local min at } x_0, \end{aligned} \tag{3.46}$$

and

$$\begin{aligned} &DF(x_0) \text{ has 2 real eigenvalues of opposite sign} \\ &\Leftrightarrow \mathcal{E} \text{ has a saddle at } x_0. \end{aligned} \tag{3.47}$$

We move on to the 2×2 system

$$\begin{aligned} \frac{d\theta}{dt} &= \psi, \\ \frac{d\psi}{dt} &= -\frac{\alpha}{m}\psi - \frac{g}{\ell}\sin\theta, \end{aligned} \tag{3.48}$$

which arises from the damped pendulum equation (cf. Chapter 1, (7.6)),

$$\frac{d^2\theta}{dt^2} + \frac{\alpha}{m}\frac{d\theta}{dt} + \frac{g}{\ell}\sin\theta = 0, \tag{3.49}$$

by adding the variable $\psi = d\theta/dt$. Here $g, \ell, \alpha, m > 0$. The system (3.48) has the form (3.2) with $x = (\theta, \psi)$ and

$$F(\theta, \psi) = \begin{pmatrix} \psi \\ -\frac{\alpha}{m}\psi - \frac{g}{\ell}\sin\theta \end{pmatrix}. \tag{3.50}$$

The phase portrait for this system is illustrated in Fig. 3.4. We compare and contrast this portrait with that depicted in Fig. 3.1.

To start, the vector field (3.50) has the same critical points as the field given by (3.24), namely $\{(k\pi, 0) : k \in \mathbb{Z}\}$. The first striking difference is in the behavior near the critical points $(k\pi, 0)$ with k even. Fig. 3.4 depicts orbits spiraling into these critical points, as opposed to the picture in Fig. 3.1 of closed orbits circling these critical points. Let us consider the linearizations about these critical points. For F as in (3.50), we have

$$L = DF(0,0) = \begin{pmatrix} 0 & 1 \\ -\frac{g}{\ell} & -\frac{\alpha}{m} \end{pmatrix}, \tag{3.51}$$

with characteristic polynomial $\lambda(\lambda + \alpha/m) + g/\ell$, hence with eigenvalues

$$\lambda_\pm = -\frac{\alpha}{2m} \pm \sqrt{\frac{\alpha^2}{m^2} - \frac{4g}{\ell}}. \tag{3.52}$$

There are three cases:

Case I. $\alpha^2/m^2 < 4g/\ell$. Then $\lambda_\pm$ are complex conjugates, each with real part $-\alpha/2m$.

Case II. $\alpha^2/m^2 = 4g/\ell$. Then $\lambda_+ = \lambda_- = -\alpha/2m$.

Case III. $\alpha^2/m^2 > 4g/\ell$. Then λ_+ and λ_- are distinct real numbers, each negative.

In all three cases, we have $e^{tL}v \to 0$ as $t \nearrow +\infty$, for each $v \in \mathbb{R}^2$. In Case I, there is also spiraling, and the orbits look like those in Fig. 3.5(a). Fig. 3.4 depicts such behavior. In Case III, the orbits look like those in Fig. 3.5(c). In Case II, the orbits look like a cross between Fig. 3.5(b) and Fig. 3.5(c). These critical points are all called *sinks.* (Reverse the sign on F, and the associated orbits are called *sources*; cf. Fig. 3.6.) The three cases described above correspond to damped oscillatory, critically damped, and overdamped motion, as discussed in §9 of Chapter 1.

Further information on the nature of these orbits spiraling in toward these sinks can be obtained from a computation of the rate of change along the orbits of $\mathcal{E}(\theta, \psi)$, given by (3.25), i.e.,

$$\mathcal{E}(\theta, \psi) = \frac{\psi^2}{2} - \frac{g}{\ell}\cos\theta. \tag{3.53}$$

Figure 3.5

Figure 3.6

This time, instead of (3.26), we have

$$\begin{aligned}\frac{d}{dt}\mathcal{E}(\theta,\psi) &= \psi\psi' + \frac{g}{\ell}(\sin\theta)\theta' \\ &= -\psi\Bigl(\frac{\alpha}{m}\psi + \frac{g}{\ell}\sin\theta\Bigr) + \frac{g}{\ell}(\sin\theta)\psi \\ &= -\frac{\alpha}{m}\psi^2.\end{aligned} \tag{3.54}$$

While this calculation applies nicely to the problem at hand, it is useful to note the following general phenomenon.

Proposition 3.4. *Let F be a smooth vector field on $\Omega \subset \mathbb{R}^n$, with a critical point at $x_0 \in \Omega$. Assume*

$$\text{all the eigenvalues of } DF(x_0) \text{ have negative real part.} \tag{3.55}$$

Then there exists $\delta > 0$ such that

$$\|x - x_0\| \le \delta \Longrightarrow \lim_{t\to+\infty} \Phi^t_F x = x_0. \tag{3.56}$$

To prove this, we bring in the following linear algebra result.

Lemma 3.5. *Let $L \in M(n, \mathbb{R})$ and assume all the eigenvalues of L have real part < 0. Then there exists a (symmetric, positive definite) inner product $\langle \ , \ \rangle$ on $\mathbb{R}^n$ and a positive constant K such that*

$$\langle Lv, v\rangle \le -K\langle v, v\rangle, \quad \forall\, v \in \mathbb{R}^n. \tag{3.57}$$

We show how Lemma 3.5 allows us to prove Proposition 3.4. Apply the lemma to $L = DF(x_0)$. Note that there exist $a, b \in (0, \infty)$ such that

$$a\|v\|^2 \le \langle v, v\rangle \le b\|v\|^2, \quad \forall\, v \in \mathbb{R}^n, \tag{3.58}$$

where $\langle v, v\rangle$ is as in (3.57) and, as usual, $\|v\|^2 = v \cdot v$. Since F is smooth,

$$F(x_0 + y) = Ly + R(y), \tag{3.59}$$

with R smooth on a ball about 0 and $DR(0) = 0$. Hence

$$\|R(y)\| \le C\|y\|^2 \le C'\langle y, y\rangle. \tag{3.60}$$

For $y(t) = \Phi_F^t(x_0 + y_0) - x_0$, we have

$$\begin{aligned} \frac{d}{dt}\langle y(t), y(t)\rangle &= 2\langle y'(t), y(t)\rangle \\ &= 2\langle F(x_0 + y), y\rangle \\ &= 2\langle Ly, y\rangle + 2\langle R(y), y\rangle. \end{aligned} \tag{3.61}$$

Now (3.57) applies to the first term in the last line of (3.61), while Cauchy's inequality plus (3.60) yields

$$\begin{aligned} |\langle R(y), y\rangle| &\le \langle R(y), R(y)\rangle^{1/2}\langle y, y\rangle^{1/2} \\ &\le C\langle y, y\rangle^{3/2}. \end{aligned} \tag{3.62}$$

Hence

$$\begin{aligned} \frac{d}{dt}\langle y, y\rangle &\le -2K\langle y, y\rangle + C\langle y, y\rangle^{3/2} \\ &\le -K\langle y, y\rangle, \end{aligned} \tag{3.63}$$

the last inequality holding provided $\langle y, y\rangle^{1/2} \le K/C$. As long as δ in (3.56) is small enough that $\{x \in \mathbb{R}^n : \|x - x_0\| \le \delta\}$ is contained in Ω and $\|v\| \le$

$\delta \Rightarrow \langle v, v\rangle^{1/2} \le K/C$, if $x = x_0 + y_0$ and $\|y_0\| \le \delta$, then (3.63) holds for $y(t) = \Phi_F^t(x_0 + y_0) - x_0$ for all $t \ge 0$, and yields

$$\langle y(t), y(t)\rangle \le e^{-Kt}\langle y_0, y_0\rangle, \tag{3.64}$$

which in turn gives (3.56).

We now prove Lemma 3.5. As shown in §8 of Chapter 2, $\mathbb{C}^n$ has a basis $\{v_1, \dots, v_n\}$ with respect to which L is upper triangular, i.e.,

$$Lv_j = \lambda_j v_j + \sum_{k<j} a_{jk} v_k. \tag{3.65}$$

Alternatively, Appendix B of Chapter 2 shows that $\mathbb{C}^n$ has an orthonormal basis $\{v_j\}$ for which (3.65) holds. The eigenvalues of L are λ_j, so by hypothesis there exists $K_1 \in (0, \infty)$ such that $\operatorname{Re}\lambda_j \le -K_1$ for all j. Now if we take $\varepsilon > 0$ and set $w_j = \varepsilon^j v_j$, we get

$$Lw_j = \lambda_j w_j + \sum_{k<j} \varepsilon^{j-k} a_{jk} w_k. \tag{3.66}$$

Then setting

$$\Big\langle \sum a_j w_j, \sum b_k w_k \Big\rangle = \operatorname{Re} \sum a_j \bar{b}_j \tag{3.67}$$

defines a positive definite inner product (depending on $\varepsilon > 0$) on $\mathbb{C}^n$, hence by restriction on $\mathbb{R}^n$, and if $\varepsilon > 0$ is taken sufficiently small, the desired conclusion (3.57) follows, with $K = K_1/2$, from (3.66).

Having discussed the critical points of the vector field (3.50) at $(0, 0)$ and related issues, we now consider the critical points at $(\pm\pi, 0)$. We have

$$DF(\pm\pi, 0) = \begin{pmatrix} 0 & 1 \\ \frac{g}{\ell} & -\frac{\alpha}{m} \end{pmatrix}. \tag{3.68}$$

This matrix has eigenvalues

$$\lambda_\pm = -\frac{\alpha}{2m} \pm \sqrt{\frac{\alpha^2}{m^2} + \frac{4g}{\ell}}, \tag{3.69}$$

one positive and one negative. These critical points are saddles. The orbits near these critical points have a behavior such as described in (3.31). Unlike the case of F given by (3.24), where the orbits are level curves of $\mathcal{E}$, the proof of this is more subtle in the present situation. See Appendix C for a proof.

Having studied the various critical points depicted in Figs. 3.1 and 3.4, we point out some special orbits that appear in these phase portraits, namely orbits connecting two critical points. Generally, if F is a C^1 vector field on $\Omega \subset \mathbb{R}^n$ with critical points $p_1, p_2 \in \Omega$, an orbit $x(t)$ of Φ_F^t satisfying

$$\lim_{t\to-\infty} x(t) = p_1, \quad \lim_{t\to+\infty} x(t) = p_2 \tag{3.70}$$

is called a *heteroclinic orbit*, from p_1 to p_2, if $p_1 \neq p_2$. If $p_1 = p_2$, such an orbit is called a *homoclinic orbit.* In Fig. 3.1, we see heteroclinic orbits connecting $p_1 = (-\pi, 0)$ and $p_2 = (\pi, 0)$, one from p_1 to p_2 and one from p_2 to p_1. These lie on level curves where $\mathcal{E}(\theta, \psi) = g/\ell$.

Such a heteroclinic orbit describes the motion of a pendulum that is heading towards pointing vertically upward. As time goes on, the pendulum ascends more and more slowly, never quite reaching the vertical position. With a little less energy, the pendulum would stop a bit short of vertical and fall back, swinging back and forth. With a little more energy, the pendulum would swing past the vertical position. Recall that Fig. 3.1 portrays the motion of an idealized pendulum, without friction. The motion of a pendulum with friction is portrayed in Fig. 3.4.

In Fig. 3.4, we see a heteroclinic orbit from $(-\pi, 0)$ to $(0, 0)$, another from $(-\pi, 0)$ to $(-2\pi, 0)$, another from $(\pi, 0)$ to $(0, 0)$, another from $(\pi, 0)$ to $(2\pi, 0)$, etc. Given that there is an orbit $x(t) = (\theta(t), \psi(t))$ here such that $\lim_{t\to-\infty} x(t) = (-\pi, 0)$ and $\psi(t) > 0$ for large negative t, the fact that $\lim_{t\to+\infty} x(t) = (0, 0)$ can be deduced from (3.54), i.e.,

$$\frac{d}{dt}\mathcal{E}(\theta, \psi) = -\frac{\alpha}{m}\psi^2. \tag{3.71}$$

We end this section with a look at the phase portrait for one more vector field, namely

$$F(\theta, \psi) = \begin{pmatrix} \frac{g}{\ell}\sin\theta \\ \psi \end{pmatrix}. \tag{3.72}$$

See Fig. 3.7. In this case,

$$F = \nabla\mathcal{E} = \begin{pmatrix} \partial\mathcal{E}/\partial\theta \\ \partial\mathcal{E}/\partial\psi \end{pmatrix}, \tag{3.73}$$

with $\mathcal{E}$ given by (3.25), i.e.,

$$\mathcal{E}(\theta, \psi) = \frac{\psi^2}{2} - \frac{g}{\ell}\cos\theta. \tag{3.74}$$

Figure 3.7

Such a vector field is called a *gradient vector field*, and its flow Φ_F^t is called a *gradient flow*. Note that if $F(x) = \nabla\mathcal{E}(x)$ and $x(t)$ is an orbit of Φ_F^t, then

$$\frac{d}{dt}\mathcal{E}(x(t)) = \nabla\mathcal{E}(x(t)) \cdot \nabla\mathcal{E}(x(t)) = \|\nabla\mathcal{E}(x(t))\|^2. \tag{3.75}$$

The critical points of F again consist of $\{(k\pi, 0) : k \in \mathbb{Z}\}$, and again they behave differently for even k than for odd k. This time

$$DF(0,0) = \begin{pmatrix} \frac{g}{\ell} & 0 \\ 0 & 1 \end{pmatrix}, \tag{3.76}$$

which is positive definite. The origin is a (non-spiraling) *source*; cf. Fig. 3.6. In particular, if $x(t) = (\theta(t), \psi(t))$ is an orbit and $x(0)$ is close to $(0,0)$, then

$$\lim_{t\to-\infty} x(t) = (0,0). \tag{3.77}$$

This can be deduced from Proposition 3.4 by reversing time. It also follows directly from (3.75). For k odd, we have saddles:

$$DF(\pm\pi,0) = \begin{pmatrix} -\frac{g}{\ell} & 0 \\ 0 & 1 \end{pmatrix}. \tag{3.78}$$

In this case, segments of the real axis provide heteroclinic orbits, from $(0,0)$ to $(-\pi,0)$, from $(0,0)$ to $(\pi,0)$, etc.

Exercises

1. If F generates the flow Φ_F^t and $v^t(x) = v(\Phi_F^t(x))$, show that

$$D\Phi_F^t(x)F(x) = F(\Phi_F^t(x)) \tag{3.79}$$

and

$$Dv^t(x) = Dv(\Phi_F^t(x))D\Phi_F^t(x). \tag{3.80}$$

Relate these identities to the simultaneous validity of (3.10) and (3.12). *Hint.* To get (3.79), use

$$\frac{d}{dt}\Phi_F^t(x) = \frac{d}{ds}\Phi_F^t \circ \Phi_F^s(x)\Big|_{s=0} = D\Phi_F^t(x)F(x), \tag{3.81}$$

and compare (3.8).

2. Extend Proposition 3.2 as follows. Replace hypothesis (3.19) by

$$\nabla V(x) \cdot F(x) \le K, \quad \forall\, x \in \mathbb{R}^n,$$

for some $K < \infty$. Show that the flow Φ_F^t is forward complete.

3. Let $\Omega = \mathbb{R}^n$ and assume F is a C^1 vector field on Ω. Show that if

$$\|F(x)\| \le C(1 + \|x\|),$$

then the flow generated by F is complete.
(*Hint.* Recall Exercise 12 of §1.)
Show that the flow is forward complete if

$$F(x) \cdot x \le C(1 + \|x\|^2).$$

4. Let $\Omega \subset \mathbb{R}^n$ be open and F be a C^1 vector field on Ω. Let $U \subset \Omega$ be an open set whose closure $\overline{U}$ is a compact subset of Ω, and whose boundary ∂U is smooth. Let $n : \partial U \to \mathbb{R}^n$ denote the outward pointing unit normal to ∂U. Assume

$$n(x) \cdot F(x) < 0, \quad \forall\, x \in \partial U. \tag{3.82}$$

Show that $\Phi^t_F(x) \in U$ if $x \in U$ and $t \geq 0$, and deduce that Φ^t_F is forward complete on U, and also on $\overline{U}$.

5. In the setting of Exercise 4, relax the hypothesis (3.82) to

$$n(x) \cdot F(x) \leq 0, \quad \forall\, x \in \partial U. \tag{3.83}$$

Show that $\Phi^t_F(x) \in \overline{U}$ if $x \in \overline{U}$ and $t \geq 0$, and deduce that Φ^t_F is forward complete on $\overline{U}$.
Hint. Find a smooth family F_τ of C^1 vector fields on Ω such that $F_0 = F$ and, for $0 < \tau < 1$, F_τ has the property given in (3.82). Then make use of Exercise 4 and of results of §2.

6. In the setting of Exercise 5, replace the hypothesis (3.83) by

$$n(x) \cdot F(x) = 0, \quad \forall\, x \in \partial U. \tag{3.84}$$

Show that Φ^t_F is complete on $\overline{U}$, and that $\Phi^t_F(x) \in \partial U$ whenever $x \in \partial U$ and $t \in \mathbb{R}$.

7. Show that if F is given by (3.24), then

$$\operatorname{div} F = 0,$$

while if F is given by (3.50), then

$$\operatorname{div} F = -\frac{\alpha}{m},$$

and if F is given by (3.72), then

$$\operatorname{div} F = \frac{g}{\ell}\cos\theta + 1.$$

8. Let $A \in M(2, \mathbb{R})$, and take J as in (3.29). Show that if A is positive definite then $A = P^2$ with P positive definite. Show that

$$-JA \text{ and } -PJP \text{ are similar,}$$

and deduce that

$$A \in M(2, \mathbb{R}) \text{ positive definite} \Longrightarrow B = -JA \text{ has 2 purely imaginary eigenvalues.}$$

Relate this to Proposition 3.3.

9. Give an alternative proof of Proposition 3.4, avoiding use of Lemma 3.5, starting with the representation of $y(t) = \Phi_F^t(x_0 + y_0) - x_0$ as

$$y(t) = e^{tL} y_0 + \int_0^t e^{(t-s)L} R(y(s))\, ds,$$

and using the fact that (3.55) implies

$$\|e^{tL}\| \le Ce^{-Kt},$$

for some $C, K \in (0, \infty)$.

10. Consider the system

$$\frac{dx}{dt} = y, \quad \frac{dy}{dt} = 1 - x^2. \tag{3.85}$$

Take $E(x, y) = y^2/2 + x^3/3 - x$. Show that if $(x(t), y(t))$ solves (3.85), then $dE(x(t), y(t)) = 0$. Show that the associated vector field has two critical points, one a center and the other a saddle. Sketch level curves of E and put in arrows to show the phase space portrait of F. Show that there is a homoclinic orbit connecting the saddle to itself.

11. Returning to the context of Exercise 1, show that (2.2) gives

$$\frac{d}{dt} D\Phi_F^t(x) = DF(\Phi_F^t(x)) D\Phi_F^t(x), \quad D\Phi_F^0(x) = I. \tag{3.86}$$

Recall from (8.6)–(8.10) of Chapter 3 that, for an $n \times n$ matrix function $M(t)$,

$$\frac{d}{dt} M(t) = A(t)M(t) \Longrightarrow \frac{d}{dt} \det M(t) = (\operatorname{Tr} A(t)) \det M(t).$$

Deduce that

$$\begin{aligned} \frac{d}{dt} \det D\Phi_F^t(x) &= \operatorname{Tr} DF(\Phi_F^t(x)) \det D\Phi_F^t(x) \\ &= \operatorname{div} F(\Phi_F^t(x)) \det D\Phi_F^t(x)). \end{aligned} \tag{3.87}$$

Relate this to (3.13), using the change of variable formula

$$\int u(x)\, dx = \int u(\Phi_F^t(x)) \det D\Phi_F^t(x)\, dx. \tag{3.88}$$

12. Use (8.10) of Chapter 3 to conclude from (3.87) that

$$\det D\Phi_F^t(x) = \exp\Bigl\{\int_0^t \operatorname{div} F(\Phi_F^s(x))\,ds\Bigr\}. \tag{3.89}$$

13. Let $\overline{U} \subset \Omega \subset \mathbb{R}^n$ be a smoothly bounded domain. The divergence theorem says that if F is a C^1 vector field on Ω,

$$\int\limits_U \operatorname{div} F(x)\,dx = \int\limits_{\partial U} n(x)\cdot F(x)\,dS(x), \tag{3.90}$$

where $n(x)$ is the outward pointing unit normal to ∂U and $dS(x)$ is $(n-1)$-dimensional surface area on ∂U (arc length if $n=2$). Given this identity, we see that, in the setting of Proposition 3.1, (3.17) is equivalent to

$$\frac{d}{dt}\operatorname{Vol}\Phi_F^t(B) = \int\limits_{\partial\Phi_F^t(B)} n(x)\cdot F(x)\,dS(x). \tag{3.91}$$

Show that this holds if and only if for each smoothly bounded $\overline{U}\subset\Omega$,

$$\frac{d}{dt}\operatorname{Vol}\Phi_F^t(U)\big|_{t=0} = \int\limits_{\partial U} n(x)\cdot F(x)\,dS(x). \tag{3.92}$$

Try to provide a direct demonstration of (3.92) (at least for $n=2$).

4. Gradient vector fields

As mentioned in §3, a vector field F on an open subset $\Omega\subset\mathbb{R}^n$ is a gradient vector field provided there exists $u\in C^1(\Omega)$ such that

$$F = \nabla u, \tag{4.1}$$

i.e., $F=(F_1,\dots,F_n)^t$ with $F_k = \partial u/\partial x_k$. It is of interest to characterize which vector fields are gradient fields. Here is one necessary condition. Suppose $u\in C^2(\Omega)$ and (4.1) holds. Then

$$\frac{\partial F_k}{\partial x_j} = \frac{\partial}{\partial x_j}\frac{\partial u}{\partial x_k}, \tag{4.2}$$

and

$$\frac{\partial}{\partial x_j}\frac{\partial u}{\partial x_k} = \frac{\partial}{\partial x_k}\frac{\partial u}{\partial x_j}, \tag{4.3}$$

so if (4.1) holds then

$$\frac{\partial F_k}{\partial x_j} = \frac{\partial F_j}{\partial x_k}, \quad \forall\, j,k \in \{1,\dots,n\}. \tag{4.4}$$

We will establish the following converse.

Proposition 4.1. *Assume $\Omega \subset \mathbb{R}^n$ is a connected open set satisfying the condition (4.13) given below. Let F be a C^1 vector field on Ω. If (4.4) holds on Ω, then there exists $u \in C^2(\Omega)$ such that (4.1) holds.*

We will construct u as a line integral. Namely, fix $p \in \Omega$, and for each $x \in \Omega$ let γ be a smooth path from p to x:

$$\gamma : [0,1] \longrightarrow \Omega, \quad \gamma(0) = p,\ \gamma(1) = x. \tag{4.5}$$

We propose that, under the hypotheses of Proposition 4.1, we can take

$$u(x) = \int\limits_{\gamma} F(y) \cdot dy. \tag{4.6}$$

Here the line integral is defined by

$$\int\limits_{\gamma} F(y) \cdot dy = \int_0^t F(\gamma(t)) \cdot \gamma'(t)\, dt. \tag{4.7}$$

For this to work, we need to know that (4.6) is independent of the choice of such a path. A key step to getting this is to consider a smooth 1-parameter family of paths γ_s from p to x:

$$\begin{gathered} \gamma_s(t) = \gamma(s,t), \quad \gamma : [0,1] \times [0,1] \longrightarrow \Omega, \\ \gamma(s,0) = p, \quad \gamma(s,1) = x. \end{gathered} \tag{4.8}$$

Lemma 4.2. *If F is a C^1 vector field satisfying (4.4) and γ_s is a smooth family satisfying (4.8), then*

$$\int\limits_{\gamma_s} F(y) \cdot dy \ \ \textit{is independent of} \ \ s \in [0,1]. \tag{4.9}$$

Proof. We compute the s-derivative of this family of line integrals, i.e., of

$$\begin{aligned}\int_0^1 F(\gamma(s,t)) \cdot \frac{\partial \gamma}{\partial t}(s,t)\,dt \\ = \int_0^1 \sum_j F_j(\gamma(s,t))\frac{\partial \gamma_j}{\partial t}(s,t)\,dt.\end{aligned} \tag{4.10}$$

The s-derivative of the integrand is obtained via the product rule and the chain rule. We obtain

$$\begin{aligned}\frac{d}{ds}\int_{\gamma_s} F(y)\cdot dy = \int_0^1 \sum_{j,k} \frac{\partial F_j}{\partial x_k}(\gamma(s,t))\frac{\partial}{\partial s}\gamma_k(s,t)\,\frac{\partial}{\partial t}\gamma_j(s,t)\,dt \\ + \int_0^1 \sum_j F_j(\gamma(s,t))\,\frac{\partial}{\partial s}\frac{\partial}{\partial t}\gamma_j(s,t)\,dt.\end{aligned} \tag{4.11}$$

We can apply the identity

$$\frac{\partial}{\partial s}\frac{\partial}{\partial t}\gamma_j(s,t) = \frac{\partial}{\partial t}\frac{\partial}{\partial s}\gamma_j(s,t)$$

to the second integrand on the right side of (4.11) and then integrate by parts. This involves applying $\partial/\partial t$ to $F_j(\gamma(s,t))$, and hence another application of the chain rule. When this is done, the second integral on the right side of (4.11) becomes

$$-\int_0^1 \sum_{j,k} \frac{\partial F_j}{\partial x_k}(\gamma(s,t))\frac{\partial}{\partial t}\gamma_k(s,t)\,\frac{\partial}{\partial s}\gamma_j(s,t)\,dt. \tag{4.12}$$

Now if we interchange the roles of j and k in (4.12), we cancel the first integral on the right side of (4.11), provided (4.4) holds. This proves the lemma.

Given $\Omega \subset \mathbb{R}^n$ open and connected, we say Ω is *simply connected* provided it has the following property:

(4.13) Given $p, x \in \Omega$, if γ_0 and γ_1 are smooth paths from p to x,
they are connected by a smooth family γ_s of paths from p to x.

Here is a class of such domains.

Lemma 4.3. *If $\Omega \subset \mathbb{R}^n$ is an open convex domain, then Ω is simply connected.*

Proof. If Ω is convex, two paths γ_0 and γ_1 from $p \in \Omega$ to $x \in \Omega$ are connected by

$$\gamma_s(t) = (1-s)\gamma_0(t) + s\gamma_1(t), \quad 0 \le s \le 1. \tag{4.14}$$

Of course there are many other simply connected domains, as the reader is invited to explore.

Now that we have Lemma 4.2, under the hypotheses of Proposition 4.1 we simply write

$$u(x) = \int_p^x F(y) \cdot dy. \tag{4.15}$$

Note that if q is another point in Ω, we can take a smooth path from p to x, passing through q, and write

$$u(x) = \int_p^q F(y) \cdot dy + \int_q^x F(y) \cdot dy. \tag{4.16}$$

Again using the path independence, we see we can independently choose paths from p to q and from q to x in (4.16); these paths need not match up smoothly at q.

We are now in a position to complete the proof of Proposition 4.1. Take $\delta > 0$ so that $\{y \in \mathbb{R}^n : \|x-y\| \le \delta\} \subset \Omega$. Take $k \in \{1, \dots, n\}$, fix q_k such that $|q_k - x_k| < \delta$, and write

$$u(x) = \int_p^{(x_1,\dots,q_k,\dots,x_n)} F(y) \cdot dy + \int_{(x_1,\dots,q_k,\dots,x_n)}^x F(y) \cdot dy. \tag{4.17}$$

Here the intermediate point is obtained by replacing x_k in $x = (x_1, \dots, x_n)$ by q_k. The first term on the right side of (4.17) is independent of x_k, so

$$\begin{aligned} \frac{\partial u}{\partial x_k}(x) &= \frac{\partial}{\partial x_k} \int_{(x_1,\dots,q_k,\dots,x_n)}^x F(y) \cdot dy \\ &= \frac{\partial}{\partial x_k} \int_{q_k}^{x_k} F_k(x_1, \dots, x_{k-1}, s, x_{k+1}, \dots, x_n)\, ds \\ &= F_k(x), \end{aligned} \tag{4.18}$$

the last identity by the fundamental theorem of calculus. This proves Proposition 4.1.

An example of a domain that is not simply connected is the punctured plane $\mathbb{R}^2 \setminus 0$. Consider on this domain the vector field

$$F(x) = \frac{Jx}{\|x\|^2}, \quad J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \tag{4.19}$$

with components

$$F_1(x) = \frac{-x_2}{x_1^2 + x_2^2}, \quad F_2(x) = \frac{x_1}{x_1^2 + x_2^2}. \tag{4.20}$$

We have

$$\frac{\partial F_1}{\partial x_2} = \frac{x_2^2 - x_1^2}{\|x\|^4} = \frac{\partial F_2}{\partial x_1}, \quad \text{on} \;\; \mathbb{R}^2 \setminus 0. \tag{4.21}$$

However, F is not a gradient vector field on $\mathbb{R}^2 \setminus 0$. Up to an additive constant, the only candidate for u in (4.1) is the angular coordinate θ:

$$F(x) = \nabla \theta(x), \tag{4.22}$$

and this identity is true on any region Ω formed by removing from $\mathbb{R}^2$ a ray starting from the origin. However, θ cannot be defined as a smooth, single valued function on $\mathbb{R}^2 \setminus 0$.

Let us linger on the case $n = 2$ and make contact with the concept of “exact equations.” Consider a 2×2 system

$$\frac{dx}{dt} = f_1(x, y), \quad \frac{dy}{dt} = f_2(x, y). \tag{4.23}$$

We take $(x, y) \in \Omega \subset \mathbb{R}^2$ and assume $f_j \in C^1(\Omega)$. This system turns into a single differential equation for y as a function of x:

$$\frac{dy}{dx} = \frac{f_2(x, y)}{f_1(x, y)}, \tag{4.24}$$

which we rewrite as

$$\begin{aligned} &g_1(x, y)\, dx + g_2(x, y)\, dy = 0, \\ &g_1(x, y) = f_2(x, y), \quad g_2(x, y) = -f_1(x, y). \end{aligned} \tag{4.25}$$

The equation (4.25) is called exact if there exists $u \in C^2(\Omega)$ such that

$$g_1 = \frac{\partial u}{\partial x}, \quad g_2 = \frac{\partial u}{\partial y}. \tag{4.26}$$

If there is such a u, solutions to (4.24) or (4.25) are given by

$$u(x,y)=C. \tag{4.27}$$

Now (4.26) is the condition that $G=(g_1,g_2)^t$ be a gradient vector field on Ω. Note that the relation between $F=(f_1,f_2)^t$ and $G=(g_1,g_2)^t$, with components given by (4.25), is

$$G=-JF, \tag{4.28}$$

where

$$J=\begin{pmatrix}0 & -1\\ 1 & 0\end{pmatrix}. \tag{4.29}$$

As we have seen, when Ω is simply connected, (4.26) holds for some u if and only if

$$\frac{\partial g_1}{\partial y}=\frac{\partial g_2}{\partial x}. \tag{4.30}$$

Note that this is equivalent to

$$\operatorname{div} F=0. \tag{4.31}$$

Remark. If $F=(F_1,F_2,F_3)^t$ is a vector field on $\Omega\subset\mathbb{R}^3$, its curl is defined as

$$\begin{aligned}\operatorname{curl} F&=\nabla\times F\\ &=\det\begin{pmatrix} i & j & k\\ \partial/\partial x & \partial/\partial y & \partial/\partial z\\ F_1 & F_2 & F_3\end{pmatrix}\\ &=\Bigl(\frac{\partial F_3}{\partial y}-\frac{\partial F_2}{\partial z}\Bigr)i+\Bigl(\frac{\partial F_1}{\partial z}-\frac{\partial F_3}{\partial x}\Bigr)j+\Bigl(\frac{\partial F_2}{\partial x}-\frac{\partial F_1}{\partial y}\Bigr)k.\end{aligned} \tag{4.32}$$

We see that

$$\text{(4.4) holds} \Longleftrightarrow \operatorname{curl} F=0. \tag{4.33}$$

We conclude with some remarks on how to construct $u(x)$, satisfying

$$\frac{\partial u}{\partial x_j}(x)=F_j(x),\quad 1\le j\le n, \tag{4.34}$$

given the compatibility conditions (4.4), without evaluating line integrals. We start with

$$u_n(x) = \int F_n(x)\,dx_n, \quad \text{so} \quad \frac{\partial u_n}{\partial x_n} = F_n(x). \tag{4.35}$$

Then $\partial(u - u_n)/\partial x_n = 0$, so

$$u(x) = u_n(x) + v(x'), \quad x' = (x_1, \dots, x_{n-1}). \tag{4.36}$$

It remains to find v, a function of fewer variables. It must solve

$$\frac{\partial v}{\partial x_j} = F_j(x) - \frac{\partial u_n}{\partial x_j}, \quad 1 \le j \le n-1. \tag{4.37}$$

Note that the left side is independent of x_n, which requires that the right side have this property. To check this, we calculate

$$\begin{aligned} \frac{\partial}{\partial x_n}\Big(F_j(x) - \frac{\partial u_n}{\partial x_j}\Big) &= \frac{\partial F_j}{\partial x_n} - \frac{\partial}{\partial x_n}\frac{\partial u_n}{\partial x_j} \\ &= \frac{\partial F_n}{\partial x_j} - \frac{\partial}{\partial x_j}\frac{\partial u_n}{\partial x_n} \\ &= 0, \end{aligned} \tag{4.38}$$

the second identity by (4.4) (and (4.3)). Thus (4.37) takes the form

$$\frac{\partial v}{\partial x_j} = G_j(x'), \quad 1 \le j \le n-1, \tag{4.39}$$

with $G_j(x') = F_j(x) - \partial u_n/\partial x_j$. Note that, for $1 \le j, k \le n-1$,

$$\begin{aligned} \frac{\partial G_j}{\partial x_k} &= \frac{\partial F_j}{\partial x_k} - \frac{\partial}{\partial x_k}\frac{\partial u_n}{\partial x_j} \\ &= \frac{\partial F_k}{\partial x_j} - \frac{\partial}{\partial x_j}\frac{\partial u_n}{\partial x_k} \\ &= \frac{\partial G_k}{\partial x_j}, \end{aligned} \tag{4.40}$$

so the task of solving (4.39) is just like that in (4.34), but with one fewer variable. An iteration yields the solution to (4.34).

Example. Take

$$F(x, y, z) = (y, x + z^2, 2yz)^t. \tag{4.41}$$

One readily verfies (4.4), or equivalently that curl $F = 0$. Here (4.35) gives

$$u_3(x,y,z) = \int 2yz\,dz = yz^2, \tag{4.42}$$

so

$$u(x,y,z) = yz^2 + v(x,y). \tag{4.43}$$

Next, requiring $\partial u/\partial y = x + z^2$ means

$$\frac{\partial v}{\partial y} = x, \tag{4.44}$$

so

$$v(x,y) = xy + w(x). \tag{4.45}$$

Then, requiring $\partial u/\partial x = y$ means $\partial w/\partial x = 0$, so we get

$$u(x,y,z) = yz^2 + xy, \tag{4.46}$$

as the unique function on $\mathbb{R}^3$ such that $\nabla u = F$, up to an additive constant.

One can turn the method given by (4.35)–(4.40) into an alternative proof of Proposition 4.1, at least if Ω is an n-dimensional box. The reader is invited to look into what happens when this method is applied to F given on $\mathbb{R}^2 \setminus 0$ by (4.19).

Exercises

For (1)–(4), identify which vector fields are gradient fields. If the field is a gradient field ∇u, find u.

(1) $(yz, xz, xy),$

(2) $(xy, yz, xz),$

(3) $(2x, z, y),$

(4) $(2x, y, z).$

For (5)–(8), identify which equations are exact. If the equation is exact, write down the solution, in implicit form (4.27).

(5) $$(2x + y)\,dx + x\,dy = 0,$$

(6) $$x\,dx + (2x + y)\,dy = 0,$$

(7) $$dx + x\,dy = 0,$$

(8) $$e^y\,dx + xe^y\,dy = 0.$$

Given $f(x, y)\,dx + g(x, y)\,dy$, a function $u(x, y)$ is called an *integrating factor* if $uf\,dx + ug\,dy$ is exact. For example, e^y is an integrating factor for $dx + x\,dy$. Find integrating factors for the left sides of (9)–(12), and use them to find solutions, in implicit form.

(9) $$(x^2 + y^2 - 1)\,dx - 2xy\,dy = 0,$$

(10) $$x^2y^3\,dx + x(1 + y^2)\,dy = 0,$$

(11) $$y\,dx + (2x - ye^y)\,dy = 0,$$

(12) $$dx + 2xy\,dy = 0.$$

13. Establish the following variant of Lemma 4.2:

Lemma 4.2A. *If F is a C^1 vector field on Ω satisfying (4.4) and γ_s is a smooth family satisfying*

$$\gamma_s(t) = \gamma(s, t), \quad \gamma : [0, 1] \times [0, 1] \to \Omega, \quad \gamma(s, 0) \equiv \gamma(s, 1),$$

then

$$\int_{\gamma_s} F(y) \cdot dy \quad \text{is independent of} \quad s \in [0, 1].$$

5. Newtonian equations

In Chapter 1 we saw how Newton's law $F = ma$ leads to a second order differential equation for the motion on a line of a single particle, acted on by a force. Newton's laws also apply to a system of m interacting particles, moving in n-dimensional space, to give a second order system of the form

(5.1) $$m_k \frac{d^2x_k}{dt^2} = \sum_{\{j:j\neq k\}} F_{jk}(x_k - x_j), \quad 1 \le k \le m.$$

Each x_k takes values in $\mathbb{R}^n$, so $x = (x_1, \dots, x_m)$ takes values in $\mathbb{R}^{mn}$. Here x_k is the location of a particle of mass m_k. The law that each action produces an equal and opposite reaction translates to

$$F_{jk}(x_k - x_j) = -F_{kj}(x_j - x_k). \tag{5.2}$$

A particularly important class of forces $F_{jk}(x_k - x_j)$ are those parallel (or antiparallel) to the line from x_j to x_k:

$$F_{jk}(x_k - x_j) = f_{jk}(\|x_k - x_j\|)(x_k - x_j). \tag{5.3}$$

In such a case, (5.2) is equivalent to

$$f_{jk}(r) = f_{kj}(r). \tag{5.4}$$

A force field of the form (5.3) is a gradient vector field:

$$\begin{gathered} f_{jk}(\|u\|)u = -\nabla V_{jk}(u), \\ V_{jk}(u) = v_{jk}(\|u\|), \quad v'_{jk}(r) = -r f_{jk}(r). \end{gathered} \tag{5.5}$$

If (5.4) holds,

$$V_{jk}(u) = V_{kj}(u). \tag{5.6}$$

The total energy of this system of interacting particles is

$$E = \frac{1}{2} \sum_k m_k \Big\| \frac{dx_k}{dt} \Big\|^2 + \frac{1}{2} \sum_{j \neq k} V_{jk}(x_k - x_j). \tag{5.7}$$

The first sum is the total kinetic energy and the second sum is the total potential energy. The following calculations yield conservation of energy. First,

$$\frac{dE}{dt} = \sum_k m_k \frac{d^2 x_k}{dt^2} \cdot \frac{dx_k}{dt} + \frac{1}{2} \sum_{j \neq k} \nabla V_{jk}(x_k - x_j) \cdot \Big(\frac{dx_k}{dt} - \frac{dx_j}{dt} \Big). \tag{5.8}$$

Next, (5.1) implies that the first sum on the right side of (5.8) is equal to

$$\sum_{j \neq k} F_{jk}(x_k - x_j) \cdot \frac{dx_k}{dt}, \tag{5.9}$$

and (5.3)–(5.5) imply that the second sum on the right side of (5.8) is equal to

$$-\frac{1}{2} \sum_{j \neq k} F_{jk}(x_k - x_j) \cdot \Big(\frac{dx_k}{dt} - \frac{dx_j}{dt} \Big) = -\sum_{j \neq k} F_{jk}(x_k - x_j) \cdot \frac{dx_k}{dt}. \tag{5.10}$$

Comparing (5.9) and (5.10), we have energy conservation:

$$\frac{dE}{dt} = 0. \tag{5.11}$$

We can convert the second order system (5.1) for mn variables into a first order system for $2mn$ variables. One way would be to introduce the velocities $v_k = x_k'$, but we get a better mathematical structure by instead using the *momenta*:

$$p_k = m_k \frac{dx_k}{dt}, \quad 1 \le k \le m. \tag{5.12}$$

We can express the energy E in (5.7) as a function of position $x = (x_1, \dots, x_m)$ and momentum $p = (p_1, \dots, p_m)$:

$$E(x,p) = \sum_k \frac{1}{2m_k} \|p_k\|^2 + \frac{1}{2} \sum_{j \ne k} V_{jk}(x_k - x_j). \tag{5.13}$$

Recall that $x_k = (x_{k1}, \dots, x_{kn}) \in \mathbb{R}^n$ and $p_k = (p_{k1}, \dots, p_{kn}) \in \mathbb{R}^n$. We have

$$\frac{\partial E}{\partial p_{k\ell}} = \frac{1}{m_k} p_{k\ell}, \tag{5.14}$$

and

$$\frac{\partial E}{\partial x_{k\ell}} = \sum_{\{j : j \ne k\}} \frac{\partial V_{jk}}{\partial u_\ell}(x_k - x_j), \tag{5.15}$$

invoking (5.6). Let us write (5.14)–(5.15) in vector form,

$$\frac{\partial E}{\partial p_k} = \frac{1}{m_k} p_k, \quad \frac{\partial E}{\partial x_k} = \sum_{\{j : j \ne k\}} \nabla V_{jk}(x_k - x_j), \tag{5.16}$$

where $\partial E / \partial p_k = (\partial E / \partial p_{k1}, \dots, \partial E / \partial p_{kn})^t$, etc. Now the system (5.1) yields the first order system

$$\frac{dx_k}{dt} = \frac{1}{m_k} p_k, \quad \frac{dp_k}{dt} = \sum_{\{j : j \ne k\}} F_{jk}(x_k - x_j), \tag{5.17}$$

which in turn, given (5.3)–(5.5), gives

$$\frac{dx_k}{dt} = \frac{\partial E}{\partial p_k}, \quad \frac{dp_k}{dt} = -\frac{\partial E}{\partial x_k}. \tag{5.18}$$

The system (5.18) is said to be in *Hamiltonian form.*

We can place the study of Hamiltonian equations in a more general framework, as follows. Let $\mathbb{R}^{2K}$ have points (x,p), $x=(x_1,\dots,x_K)$, $p=(p_1,\dots,p_K)$. Let $\Omega\subset\mathbb{R}^{2K}$ be open and $E\in C^1(\Omega)$. Consider the system

$$\begin{aligned}\frac{dx_k}{dt}&=\frac{\partial E}{\partial p_k},\\ \frac{dp_k}{dt}&=-\frac{\partial E}{\partial x_k},\end{aligned}\tag{5.19}$$

for $1\le k\le K$. This is called a Hamiltonian system. It is of the form

$$\frac{d}{dt}\begin{pmatrix}x\\ p\end{pmatrix}=X_E(x,p),\tag{5.20}$$

where X_E is a vector field on Ω, called the Hamiltonian vector field associated to E. In this general setting, E is constant on each solution curve $(x(t),p(t))$ of (5.19). Indeed, in such a case,

$$\begin{aligned}\frac{d}{dt}E(x(t),p(t))&=\sum_k\frac{\partial E}{\partial x_k}\cdot\frac{dx_k}{dt}+\sum_k\frac{\partial E}{\partial p_k}\cdot\frac{dp_k}{dt}\\ &=\sum_k\frac{\partial E}{\partial x_k}\cdot\frac{\partial E}{\partial p_k}-\sum_k\frac{\partial E}{\partial p_k}\cdot\frac{\partial E}{\partial x_k}\\ &=0.\end{aligned}\tag{5.21}$$

Returning to the setting (5.1)–(5.2), we next discuss the conservation of the total momentum

$$P=\sum_k p_k=\sum_k m_k\frac{dx_k}{dt}.\tag{5.22}$$

Indeed,

$$\begin{aligned}\frac{dP}{dt}&=\sum_k m_k\frac{d^2x_k}{dt^2}\\ &=\sum_{j\ne k}F_{jk}(x_k-x_j)\\ &=0,\end{aligned}\tag{5.23}$$

the last identity by (5.2). Thus, for each solution $x(t)$ to (5.1), there exist $a,b\in\mathbb{R}^n$ such that

$$\frac{1}{M}\sum_k m_kx_k(t)=a+bt,\quad M=\sum_k m_k.\tag{5.24}$$

The left side is the *center of mass* of the system of interacting particles. The vectors $a, b \in \mathbb{R}^n$ are given by the initial data for (5.1):

$$a = \frac{1}{M} \sum_k m_k x_k(0), \quad b = \frac{1}{M} \sum_k m_k x_k'(0). \tag{5.25}$$

Given this, we can obtain a system similar to (5.1) for the variables

$$y_k(t) = x_k(t) - (a + bt). \tag{5.26}$$

We have $y_k'' = x_k''$ and $y_k - y_j = x_k - x_j$, so (5.1) gives

$$m_k \frac{d^2 y_k}{dt^2} = \sum_{\{j : j \neq k\}} F_{jk}(y_k - y_j), \quad 1 \le k \le m. \tag{5.27}$$

In this case we have the identity

$$\sum_k m_k y_k(t) \equiv 0, \tag{5.28}$$

as a consequence of (5.24). We can use this to reduce the size of (5.27), from a system of mn equations to a system of $(m-1)n$ equations, by substituting

$$y_m = -\frac{1}{m_m} \sum_{\ell=1}^{m-1} m_\ell y_\ell \tag{5.29}$$

into (5.27), for $1 \le k \le m - 1$. One calls $(y_1, \dots, y_m)$ center of mass coordinates.

In case $m = 2$, this substitution works out quite nicely. We have

$$y_2 = -\frac{m_1}{m_2} y_1, \tag{5.30}$$

and the system (5.27) reduces to

$$m_1 \frac{d^2 y_1}{dt^2} = F_{21}\Bigl(\Bigl(1 + \frac{m_1}{m_2}\Bigr) y_1\Bigr), \tag{5.31}$$

the equation of motion of a *single* particle in an external force field. Alternatively, for $x = x_1 - x_2 = y_1 - y_2 = (1 + m_1/m_2) y_1$,

$$\frac{m_1 m_2}{m_1 + m_2} \frac{d^2 x}{dt^2} = F_{21}(x). \tag{5.32}$$

For $m > 2$, the resulting equations are not so neat. For example, for $m = 3$, we have

$$y_3 = -\frac{m_1}{m_3} y_1 - \frac{m_2}{m_3} y_2, \tag{5.33}$$

and the system (5.27) reduces to

$$\begin{aligned} m_1 y_1'' &= F_{21}(y_1 - y_2) + F_{31}\Bigl(\Bigl(1 + \frac{m_1}{m_2}\Bigr) y_1 + \frac{m_2}{m_1} y_2\Bigr), \\ m_2 y_2'' &= F_{12}(y_2 - y_1) + F_{32}\Bigl(\frac{m_1}{m_3} y_1 + \Bigl(1 + \frac{m_2}{m_3}\Bigr) y_2\Bigr). \end{aligned} \tag{5.34}$$

Exercises

In (1)–(5), take $n = 1$, $m = 3$, and $m_1 = m_2 = m_3 = 1$. Set up the equations of motion in center of mass coordinates and analyze the solution.

$$F_{jk}(x) = x \tag{1}$$

$$F_{jk}(x) = -x \tag{2}$$

$$F_{12}(x) = F_{13}(x) = x, \quad F_{23}(x) = -x \tag{3}$$

$$F_{12}(x) = F_{13}(x) = -x, \quad F_{23}(x) = x. \tag{4}$$

$$F_{12}(x) = F_{23}(x) = -x, \quad F_{23}(x) = 1. \tag{5}$$

In all cases, (5.2) must be enforced.

6. Central force problems and two-body planetary motion

As seen in §5, one can transform the m-body problem (5.1) to center of mass coordinates, under the hypothesis (5.2), and obtain a smaller system, which for $m = 2$ is given by (5.32). Changing notation, we rewrite (5.32) as

$$m\frac{d^2x}{dt^2} = F(x). \tag{6.1}$$

Here $x \in \mathbb{R}^n$. We assume $F \in C^1(\mathbb{R}^n \setminus 0)$ but allow blowup at $x = 0$. Under hypotheses (5.3)–(5.4) for the two body problem, we have

$$F(x) = f(\|x\|)x. \tag{6.2}$$

In such a case, (6.1) is called a *central force problem.* Parallel to (5.5), we have

$$\begin{aligned} F(x) &= -\nabla V(x), \\ V(x) &= v(\|x\|), \quad v'(r) = -rf(r). \end{aligned} \tag{6.3}$$

The total energy is given by

$$E = \frac{1}{2}m\Big\|\frac{dx}{dt}\Big\|^2 + V(x), \tag{6.4}$$

and if $x(t)$ solves (6.1), then

$$\frac{dE}{dt} = m\frac{d^2x}{dt}\cdot\frac{dx}{dt} + \nabla V(x)\cdot\frac{dx}{dt} = 0, \tag{6.5}$$

yielding conservation of energy.

There are further conservation laws, starting with the following.

Proposition 6.1. *Assume $x(0) \neq 0$ and let $W \subset \mathbb{R}^n$ be the linear span of $x(0)$ and $x'(0)$. If $x(t)$ solves (6.1) for $t \in I$ and (6.2) holds, we have*

$$x(t) \in W, \quad \forall\, t \in I. \tag{6.6}$$

Proof. One way to see this is to note that (6.1) is a well posed system for $x(t)$ taking values in W. Then uniqueness of solutions yields (6.6). Here is another demonstration.

Define $A \in \mathcal{L}(\mathbb{R}^n)$ by

$$\begin{aligned} Av &= v, & \forall\, v \in W, \\ Av &= -v, & \forall v \in W^{\perp}. \end{aligned} \tag{6.7}$$

Note that A is an orthogonal transformation. Let $y(t) = Ax(t)$. The hypothesis on the initial data gives

$$y(0) = x(0), \quad y'(0) = x'(0). \tag{6.8}$$

Also, given $F(x)$ of the form (6.2), we have $AF(x) = F(y)$, so $y(t)$ solves (6.1). The basic uniqueness result proven in §1 implies $y \equiv x$ on I, which in turn gives (6.6).

A third proof of Proposition 6.1, valid for $n = 3$, can be obtained from conservation of angular momentum, established in (6.11) below.

Proposition 6.1 guarantees that each path $x(t)$ solving (6.1) lies in a plane, and we can take $n = 2$. For the next step, it is actually convenient to take $n = 3$. Thus $x(t)$ solves (6.1) and $x(t)$ is a path in $\mathbb{R}^3$. We define the *angular momentum*

$$\alpha(t) = mx(t) \times x'(t). \tag{6.9}$$

We then have, under hypothesis (6.2),

$$\begin{aligned} \alpha'(t) &= mx(t) \times x''(t) \\ &= x(t) \times F(x) \\ &= f(\|x\|)\, x(t) \times x(t) \\ &= 0. \end{aligned} \tag{6.10}$$

This yields conservation of angular momentum:

$$x(t) \times x'(t) \equiv L, \tag{6.11}$$

where $L = x(0) \times x'(0) \in \mathbb{R}^3$. In case $x(t) = (x_1(t), x_2(t), 0)$, we have

$$x(t) \times x'(t) = (0, 0, x_1(t)x_2'(t) - x_1'(t)x_2(t)), \tag{6.12}$$

so the conservation law (6.11) gives

$$x_1(t)x_2'(t) - x_1'(t)x_2(t) \equiv L_3. \tag{6.13}$$

Let's return to the planar setting, and also use complex notation:

$$x(t) = x_1(t) + ix_2(t) = r(t)e^{i\theta(t)}. \tag{6.14}$$

A computation gives

$$\begin{aligned} x' &= (r' + ir\theta')e^{i\theta}, \\ x'' &= [r'' - r(\theta')^2 + i(2r'\theta' + r\theta'')]e^{i\theta}, \end{aligned} \tag{6.15}$$

so (6.1)–(6.2) becomes

$$m\big[r'' - r(\theta')^2 + i(2r'\theta' + r\theta'')\big] = f(r)r. \tag{6.16}$$

Equating real and imaginary parts separately, we get

$$\begin{aligned} &r'' - r(\theta')^2 = \frac{f(r)r}{m}, \\ &2r'\theta' + r\theta'' = 0. \end{aligned} \tag{6.17}$$

Note that

$$\frac{d}{dt}(r^2\theta') = r(2r'\theta' + r\theta''), \tag{6.18}$$

so the second equation in (6.17) says $r^2\theta'$ is independent of t. This is actually equivalent to the conservation of angular momentum, (6.13). In fact, we have $x_1 = r\cos\theta,\ x_2 = r\sin\theta$, hence

$$x_1' = r'\cos\theta - r\theta'\sin\theta, \quad x_2' = r'\sin\theta + r\theta'\cos\theta, \tag{6.19}$$

and hence

$$x_1x_2' - x_1'x_2 = r^2\theta'. \tag{6.20}$$

Thus we have in two ways derived the identity

$$r^2\theta' = L. \tag{6.21}$$

(For notational simplicity, we drop the subscript 3 from (6.13).)

There is the following geometrical interpretation of (6.21). The (signed) area $A(t)$ swept out by the ray from 0 to $x(s)$, as s runs from t_0 to t, is given by

$$A(t) = \frac{1}{2}\int_{\theta(t_0)}^{\theta(t)} r^2\,d\theta = \frac{1}{2}\int_{t_0}^{t} r(s)^2\theta'(s)\,ds, \tag{6.22}$$

so

$$A'(t) = \frac{1}{2}r^2\theta' = \frac{L}{2}. \tag{6.23}$$

This says

(6.24) Equal areas are swept out in equal times,

which, as we will discuss below, is Kepler's second law.

Next, we can plug $\theta' = L/r^2$ into the first equation of (6.17), obtaining

$$\frac{d^2r}{dt^2} = \frac{f(r)r}{m} + \frac{L^2}{r^3}. \tag{6.25}$$

This has the form

$$\frac{d^2r}{dt^2} = g(r), \tag{6.26}$$

treated in Chapter 1, §5. We recall that treatment. Take $w(r)$ such that $g(r) = -w'(r)$, so (6.26) becomes

$$\frac{d^2r}{dt^2} = -w'(r). \tag{6.27}$$

Then form the "energy"

$$E = \frac{1}{2}\Bigl(\frac{dr}{dt}\Bigr)^2 + w(r), \tag{6.28}$$

and compute that if $r(t)$ solves (6.27) then

$$\frac{dE}{dt} = \frac{d^2r}{dt^2}\frac{dr}{dt} + w'(r)\frac{dr}{dt} = 0, \tag{6.29}$$

so for each solution to (6.27), there is a constant E such that

$$\frac{dr}{dt} = \pm\sqrt{2E - 2w(r)}. \tag{6.30}$$

Separation of variables gives

$$\int \frac{dr}{\sqrt{2E - 2w(r)}} = \pm t + C. \tag{6.31}$$

This integral can be quite messy.

Note that dividing (6.30) by (6.21) yields a differential equation for r as a function of θ:

$$\frac{dr}{d\theta} = \pm \frac{r^2}{L} \sqrt{2E - 2w(r)}, \tag{6.32}$$

which separates to

$$L \int \frac{dr}{r^2 \sqrt{2E - 2w(r)}} = \pm \theta + C. \tag{6.33}$$

Let us recall that

$$w'(r) = -\frac{f(r)r}{m} - \frac{L^2}{r^3}. \tag{6.34}$$

Typically the integral in (6.33) is as messy as the one in (6.31). These integrals do turn out to be tractable in one very important case, the Kepler problem, to which we now turn.

This problem is named after the astronomer Johannes Kepler, who from observations formulated the following three laws for planetary motion.

1. The planets move on ellipses with the sun at one focus.

2. The line segment from the sun to a planet sweeps out equal areas in equal time intervals.

3. The period of revolution of a planet is proportional to $a^{3/2}$, where a is the semi-major axis of its ellipse.

The Kepler problem is to provide a theoretical framework in which to derive these three laws. This was solved by Isaac Newton, who formulated his universal law of gravitation, used it to derive a differential equation for the position of a planet, and solved the differential equation.

Newton's law of gravitation specifies the force between two objects, of mass m_1 and m_2, located at points x_1 and x_2 in $\mathbb{R}^3$. Let us say the center of the planet is at x_1 and the center of the sun is at x_2. In the framework of (5.1), this means specifying the vector field F_{21} on $\mathbb{R}^3$. The formula is

$$F_{21}(x) = -Gm_1m_2\frac{x}{\|x\|^3}. \tag{6.35}$$

Here G is the universal gravitational constant. If we go to center of mass coordinates, the motion of the planet is governed by (5.32), yielding (6.1) with

$$F(x) = -Km\frac{x}{\|x\|^3}, \quad K = G(m+m_2). \tag{6.36}$$

Here $m = m_1$ is the mass of the planet and m_2 is the mass of the sun. Consequently we have (6.17) with

$$\frac{f(r)r}{m} = -\frac{K}{r^2}, \tag{6.37}$$

and (6.25) becomes

$$\frac{d^2r}{dt^2} = -\frac{K}{r^2} + \frac{L^2}{r^3}. \tag{6.38}$$

Thus $w(r)$ in (6.27)–(6.34) is given by

$$w(r) = -\frac{K}{r} + \frac{L^2}{2r^2}. \tag{6.39}$$

Thus the integral in (6.31) is

$$\int \frac{r\,dr}{\sqrt{2Er^2+2Kr-L^2}}, \tag{6.40}$$

and the integral in (6.33) is

$$\int \frac{dr}{r\sqrt{2Er^2+2Kr-L^2}}. \tag{6.41}$$

The integral (6.40) can be evaluated by completing the square for $2Er^2 + 2Kr - L^2$. The integral (6.41) can also be evaluated, but rather than tackling this directly, we instead produce a differential equation for u, defined by

$$u = \frac{1}{r}. \tag{6.42}$$

By the chain rule,

$$\frac{dr}{dt} = -r^2\frac{du}{dt} = -r^2\frac{du}{d\theta}\frac{d\theta}{dt} = -L\frac{du}{d\theta}, \tag{6.43}$$

the last identity by (6.21). Taking another t-derivative gives

$$\frac{d^2r}{dt^2} = -L\frac{d}{dt}\frac{du}{d\theta} = -L\frac{d^2u}{d\theta^2}\frac{d\theta}{dt} = -L^2u^2\frac{d^2u}{d\theta^2}, \tag{6.44}$$

again using (6.21). Comparing this with (6.38), we get

$$-L^2u^2\frac{d^2u}{d\theta^2} = L^2u^3 - Ku^2, \tag{6.45}$$

or equivalently

$$\frac{d^2u}{d\theta^2} + u = \frac{K}{L^2}. \tag{6.46}$$

Miraculously, we have obtained a linear equation! The general solution to (6.46) is

$$u(\theta) = A\cos(\theta - \theta_0) + \frac{K}{L^2}, \tag{6.47}$$

which by (6.42) gives

$$r\Bigl[A\cos(\theta - \theta_0) + \frac{K}{L^2}\Bigr] = 1. \tag{6.48}$$

This is equivalent to

$$r\bigl[1 + e\cos(\theta - \theta_0)\bigr] = p, \quad p = \frac{L^2}{K}, \quad e = A\frac{L^2}{K}. \tag{6.49}$$

If $e = 0$, this is the equation of a circle. If $0 < e < 1$, it is the equation of an ellipse. If $e = 1$, it is the equation of a parabola, and if $e > 1$, it is the equation of one branch of a hyperbola. Among these curves, those that are bounded are the ellipses, and the circle, which we regard as a special case of an ellipse.

Since planets move in bounded orbits, this establishes Kepler's first law (with caveats, which we discuss below). Kepler's second law holds for general central force problems, as noted already in (6.24). To establish the third law, recall from (6.23) that $L/2$ is the rate at which such area is swept out, so the period T of the orbit satisfies

$$\begin{aligned}\frac{L}{2}T &= \text{ area enclosed by the ellipse}\\ &= \pi ab,\end{aligned} \tag{6.50}$$

where a is the semi-major axis and b the semi-minor axis. For an ellipse given by (6.49), we have

$$a = \frac{p}{1-e^2}, \quad b = \frac{p}{\sqrt{1-e^2}} = p^{1/2}a^{1/2}, \tag{6.51}$$

which yields

$$T = \frac{2\pi ab}{L} = 2\pi\frac{\sqrt{p}}{L}a^{3/2} = \frac{2\pi}{\sqrt{K}}a^{3/2}. \tag{6.52}$$

This establishes Kepler's third law.

We now discuss some caveats. Our solar system has nine planets, plus numerous other satellites. In the calculations above, all but one planet was ignored. One can expect this approximation to work best for Jupiter. Jupiter has about 10^{-3} the sun's mass, and its distance from the sun is about 400 times the sun's radius. Hence the center of mass of Jupiter and the sun is located about 0.4 times the sun's radius from the center of the sun. The sun and Jupiter engage in a close to circular elliptical orbit with a focus at this center of mass. Clearly this motion is going to influence the orbits of the other planets. In fact, each planet influences all the others, including Jupiter, in ways not captured by the calculations of this section. Realization of this situation led to a vigorous development of the subject known as celestial mechanics, from Newton's time on. Material on this can be found in [**AM**] and [**Gr**], and references given there.

Advances in celestial mechanics led to the discovery of the planet Neptune. By the early 1900s, this subject was sufficiently well developed that astronomers were certain that an observed anomaly in the motion of Mercury could not be explained by the Newtonian theory. This discrepancy was accounted for by Einstein's theory of general relativity, which provided a new foundation for the theory of gravity. This is discussed in [**ABS**] and also in Chapter 18 of [**T**]. While a derivation is well outside the scope of this book, we mention that the relativistic treatment leads to the following variant of (6.46):

$$\frac{d^2u}{d\theta^2} + u = A + \varepsilon u^2, \tag{6.53}$$

where $A \approx K/L^2$ and ε is a certain (small) positive constant, determined by the mass of the sun. This can be converted into the first order system

$$\frac{du}{d\theta} = v, \quad \frac{dv}{d\theta} = -u + A + \varepsilon u^2. \tag{6.54}$$

In analogy with (6.26)–(6.29), we can form

$$F(u,v) = \frac{1}{2}v^2 + \frac{1}{2}u^2 - Au - \frac{\varepsilon}{3}u^3, \tag{6.55}$$

and check that if $(u(\theta), v(\theta))$ solves (6.54), then

$$\frac{d}{d\theta} F(u, v) = 0, \tag{6.56}$$

so the orbits for (6.54) lie on level curves of F. As long as $A\varepsilon \in (0, 1/4)$, F has two critical points, a minimum and a saddle. Thus (6.54) has some solutions periodic in θ. However, the period is generally not equal to 2π. (See Appendix D for results related to computing this period.) This fact leads to the precession of the perihelion of the planet orbiting the sun, where the perihelion is the place where u is maximal, so r is minimal. In the non-relativistic situation covered by (6.46), all the solutions in (6.47) are periodic in θ of period 2π.

Exercises

1. Solve explicitly
$$w''(t) = -w(t),$$
for w taking values in $\mathbb{R}^2 = \mathbb{C}$. Show that
$$|w(t)|^2 + |w'(t)|^2 = 2E$$
is constant on each orbit.

2. For $w(t)$ taking values in $\mathbb{C}$, define a new curve by
$$z(s) = w(t)^2, \quad \frac{ds}{dt} = |w(t)|^2.$$
Show that if $w''(t) = -w(t)$, then
$$z''(s) = -4E\frac{z(s)}{|z(s)|^3},$$
so $z(s)$ solves the Kepler problem.

3. Take $u = 1/r$ as in (6.42), and generalize the calculations (6.43)–(6.46) to obtain a differential equation for u as a function of θ, for more general central forces. Consider particularly $f(x) = -\nabla V(x)$ in the cases
$$V(x) = -K\|x\|^2, \quad V(x) = -K\|x\|.$$

4. Take the following steps to show that if $p > 0$ and $0 < e < 1$, then

$$r(1 + e \cos \theta) = p \tag{6.57}$$

is the equation in polar coordinates of an ellipse.

(a) Show that (6.57) describes a closed, bounded curve, since $1+e \cos \theta > 0$ for all θ if $0 < \theta < 1$, and $\cos \theta$ is periodic in θ of period 2π. Denote the curve by $\gamma(\theta) = (x(\theta), y(\theta))$, in Cartesian coordinates.

(b) Show that this curve is symmetric about the x-axis and cuts the axis at two points, whose distance apart is

$$2a = r(0) + r(\pi),$$

so

$$a = \frac{p}{1 - e^2}. \tag{6.58}$$

(c) Show that the midpoint between $\gamma(0)$ and $\gamma(\pi)$ is given by

$$x_0 = -ea, \quad y_0 = 0.$$

(d) For $\gamma(\theta) = (x(\theta), y(\theta))$, as in part (a), show that

$$\frac{(x + ea)^2}{a^2} + \frac{y^2}{b^2} = 1, \tag{6.59}$$

i.e., that

$$\frac{(r \cos \theta + ea)^2}{a^2} + \frac{r^2(1 - \cos^2 \theta)}{b^2} = 1,$$

provided (6.57) holds, when a is given by (6.58) and

$$b = \frac{p}{\sqrt{1 - e^2}}. \tag{6.60}$$

5. As an approximation, assume that the earth has a circular orbit about the sun with a radius

$$a = 1.496 \times 10^{11} \text{ m}, \tag{6.61}$$

and its period is one year, i.e.,

$$T = 31.536 \times 10^6 \ \ \text{sec}. \tag{6.62}$$

The gravitational constant G has been measured as

$$G = 6.674 \times 10^{-11} \ \ \text{m}^3/(\text{kg sec}^2). \tag{6.63}$$

With this information, use (6.36) and (6.52) to calculate the mass m_2 of the sun. Assume the mass of the earth is negligible compared to m_2. You should get

$$m_2 = \alpha \times 10^{30} \ \ \text{kg}, \tag{6.64}$$

with α between 1 and 10.

Remark. Historically, T was measured by the position of the "fixed stars." Modern methods to measure a involve bouncing a radar signal off Venus to measure its distance, given that we have an accurate measurement of the speed of light. Then trigonometry is used to determine a. See [**GM**] for a discussion of how G has been measured; this is the most difficult issue.

6. The force of gravity the earth exerts on a body of mass m at the earth's surface is

$$-Gmm_e r^{-2}, \tag{6.65}$$

where G is given in Exercise 5,

$$r = 6.38 \times 10^6 \ \ \text{m} \tag{6.66}$$

is the radius of the earth, and m_e is the mass of the earth. It is observed that the earth's gravity accelerates objects at its surface downward at 9.8 m/sec^2, so we have

$$9.8 \ \ \text{m/sec}^2 = Gm_e r^{-2}. \tag{6.67}$$

Use this to compute m_e. You should get

$$m_e = \beta \times 10^{24} \ \ \text{kg}, \tag{6.68}$$

with β between 1 and 10.
Remark. See Appendix E for more on (6.65).

7. As an approximation, assume that the moon has a circular orbit about the earth, of radius

$$a = 3.8 \times 10^8 \ \text{ m},$$

and its period is 27.3 days, i.e.,

$$T = 2.359 \times 10^6 \ \text{ sec}.$$

Assume the mass of the moon is negligible compared to the mass of the earth. Use the method of Exercise 5 to calculate the mass of the earth. Compare your result with that of Exercise 6.

8. Use the data presented in Exercises 5 and 7 to calculate the ratio of the masses of the earth and the sun, irrespective of the knowledge of G.

9. Jupiter has a moon, Ganymede, which orbits the planet at a distance 1.07×10^9 m, with a period of 7.15 earth days. Using the method of Exercise 5 (or 8), compute the mass m_J of Jupiter. You should get

$$m_J \approx 318 \ m_e.$$

7. Variational problems and the stationary action principle

A rich source of second order systems of differential equations is provided by variational problems, which we will consider here. Let $\Omega \subset \mathbb{R}^n$ be open, and let $L \in C^2(\Omega \times \mathbb{R}^n)$, say $L = L(x, v)$. For a path $u : [a, b] \to \Omega$, consider

$$I(u) = \int_a^b L(u(t), u'(t)) \, dt. \tag{7.1}$$

We desire to find equations for a path that minimizes $I(u)$, among all such paths for which the endpoints $u(a) = p$ and $u(b) = q$ are fixed. More generally, we desire to specify when u is a stationary path, meaning that

$$\frac{d}{ds} I(u_s)\Big|_{s=0} = 0, \tag{7.2}$$

for all smooth families of paths u_s such that $u_0 = u$, $u_s(a) = p$, and $u_s(b) = q$. Let us write

$$\frac{\partial}{\partial s} u_s(t)\Big|_{s=0} = w(t), \tag{7.3}$$

so $w : [a, b] \to \mathbb{R}^n$ is an arbitrary smooth function such that $w(a) = w(b) = 0$. To compute $(d/ds)I(u_s)$, let us denote

$$L_{x_k} = \frac{\partial L}{\partial x_k}, \quad L_{v_k} = \frac{\partial L}{\partial v_k}. \tag{7.4}$$

Then

$$\begin{aligned}\frac{d}{ds}I(u_s)\Big|_{s=0} &= \int_a^b \sum_k L_{x_k}(u(t),u'(t))w_k(t)\,dt \\ &\quad + \int_a^b \sum_k L_{v_k}(u(t),u'(t))w_k'(t)\,dt.\end{aligned} \tag{7.5}$$

We can apply integration by parts to the last integral. The condition that $w_k(a)=w_k(b)=0$ implies that there are no endpoint contributions, so

$$\frac{d}{dt}I(u_s)\Big|_{s=0} = \int_a^b \sum_k \Big[L_{x_k}(u(t),u'(t)) - \frac{d}{dt}L_{v_k}(u(t),u'(t))\Big]w_k(t)\,dt. \tag{7.6}$$

For this to vanish for all smooth w_k that vanish at $t=a$ and b, it is necessary and sufficient that

$$\frac{d}{dt}L_{v_k}(u(t),u'(t)) - L_{x_k}(u(t),u'(t)) = 0, \quad \forall\, k. \tag{7.7}$$

This system is called the Lagrange equation for stationarity of (7.1). Applying the chain rule to the first sum, we can expand this out as

$$\begin{aligned}\sum_\ell L_{v_k v_\ell}(u(t),u'(t))u_\ell''(t) + \sum_\ell L_{v_k x_\ell}(u(t),u'(t))u_\ell'(t) \\ -L_{x_k}(u(t),u'(t)) \ = 0, \quad \forall\, k.\end{aligned} \tag{7.8}$$

This can be converted to a first order system for $(u(t),u'(t))$, to which the results of §1 apply, provided the $n \times n$ matrix

$$\Big(L_{v_k v_\ell}(x,v)\Big) \tag{7.9}$$

of second order partial derivatives of $L(x,v)$ with respect to v is invertible.

The Newtonian equations of motion can be put into this Lagrangian framework, as follows. A particle of mass m, position x, and velocity v, moving in a force field $F(x) = -\nabla V(x)$, has kinetic energy and potential energy

$$T = \frac{1}{2}m\|v\|^2, \quad \text{and} \quad V = V(x), \tag{7.10}$$

respectively. The Lagrangian $L(x,v)$ is given by the *difference*:

$$L(x,v) = T - V = \frac{1}{2}m\|v\|^2 - V(x). \tag{7.11}$$

Figure 7.1

In such a case,

$$L_{v_k}(x,v) = mv_k, \quad L_{x_k}(x,v) = -\frac{\partial V}{\partial x_k}, \tag{7.12}$$

and the Lagrange system (7.7) becomes the standard Newtonian system

$$m\frac{d^2u}{dt^2} = -\nabla V(u). \tag{7.13}$$

In this setting, the integral (7.1) is called the *action.* The assertion that the laws of motion are given by the stationary condition for (7.1) where L is the Lagrangian (7.11) is the stationary action principle.

The Lagrangian approach can be particularly convenient in situations where coordinates other than Cartesian coordinates are used. As an example, we consider the simple pendulum problem, and give a treatment that can be compared and contrasted with that given in §6 of Chapter 1. As there, we have a rigid rod, of length ℓ, suspended at one end. We assume the rod has negligible mass, except for an object of mass m at the other end. See Fig. 7.1. The rod makes an angle θ with the downward vertical. We seek a differential equation for θ as a function of t.

The end with the mass m traces out a path in a plane, which, as in Chapter 1, we identify with the complex plane, with the origin at the point where the pendulum is suspended and the real axis pointing vertically down. We can write the path as

$$z(t) = \ell e^{i\theta(t)}. \tag{7.14}$$

The velocity is

$$v(t) = z'(t) = i\ell\theta'(t)e^{i\theta(t)}, \tag{7.15}$$

so the kinetic energy is

$$T = \frac{1}{2}m\|v(t)\|^2 = \frac{m\ell^2}{2}\theta'(t)^2. \tag{7.16}$$

Meanwhile the potential energy, due to the force of gravity, is

$$V = -mg\ell\cos\theta. \tag{7.17}$$

Taking $\psi = \theta'$, we have the Lagrangian

$$\begin{gathered} L(\theta,\psi) = \frac{m\ell^2}{2}\psi^2 + mg\ell\cos\theta, \\ L_\psi(\theta,\psi) = m\ell^2\psi, \quad L_\theta(\theta,\psi) = -mg\ell\sin\theta, \end{gathered} \tag{7.18}$$

and Lagrange's equation

$$\frac{d}{dt}L_\psi(\theta(t),\theta'(t)) - L_\theta(\theta(t),\theta'(t)) = 0 \tag{7.19}$$

yields the pendulum equation

$$\frac{d^2\theta}{dt^2} + \frac{g}{\ell}\sin\theta = 0, \tag{7.20}$$

in agreement with (6.6) of Chapter 1.

The approach above avoided a computation of the force acting on the pendulum (cf. (6.4) of Chapter 1), and is arguably a bit simpler than the approach given in Chapter 1. The Lagrangian approach can be *very much* simpler in more complex situations, such as the double pendulum, which we will discuss in §9.

An important variant of these variational problems is the class of *constrained variational problems*, which we now discuss. For the sake of definiteness, let M be either a smooth curve in $\Omega \subset \mathbb{R}^2$ or a smooth surface in $\Omega \subset \mathbb{R}^3$, and let $n(x)$ be a smooth unit normal to M, for $x \in M$. Again, let $L \in C^2(\Omega \times \mathbb{R}^n)$, $n = 2$ or 3, and define $I(u)$ by (7.1). We look for equations for

$$u : [a,b] \longrightarrow M, \tag{7.21}$$

satisfying the stationary condition (7.2), not for all smooth families of paths u_s such that $u_0 = u$ and $u_s(0) = p$, $u_s(b) = q$, but rather for all such paths satisfying the constraint

$$u_s : [a,b] \longrightarrow M. \tag{7.22}$$

Again we take $w(t)$ as in (7.3), and this time we obtain an arbitrary smooth function $w : [a, b] \to \mathbb{R}^n$, satisfying $w(a) = w(b) = 0$, and the additional constraint

$$w(t) \cdot n(u(t)) \equiv 0. \tag{7.23}$$

The calculations (7.4)–(7.6) still apply, but from here we get a conclusion different from (7.7). Since (7.6) holds for all $w(t)$ described as just above, the conclusion is

$$\frac{d}{dt} L_v(u(t), u'(t)) - L_x(u(t), u'(t)) \text{ is parallel to } n(u(t)), \tag{7.24}$$

where $L_v = (L_{v_1}, \dots, L_{v_n})^t$ and $L_x = (L_{x_1}, \dots, L_{x_n})^t$. In case $n = 3$, an equivalent formulation of (7.24) is

$$\Big[\frac{d}{dt} L_v(u(t), u'(t)) - L_x(u(t), u'(t))\Big] \times n(u(t)) = 0. \tag{7.25}$$

Let's specialize this constrained variational problem to the case

$$L(x, v) = \frac{1}{2} \|v\|^2. \tag{7.26}$$

The associated integral

$$E(u) = \frac{1}{2} \int_a^b \|u'(t)\|^2 \, dt \tag{7.27}$$

is called the *energy* of $u : [a, b] \to M$. In this case, $L_v = v$ and $L_x = 0$, so (7.24) becomes

$$u''(t) \text{ is parallel to } n(u(t)). \tag{7.28}$$

That is, $u''(t) = a(t)n(u(t))$. Taking the inner product with $n(t)$ gives $a(t) = n(u(t)) \cdot u''(t)$, so (7.28) yields

$$u''(t) = n(u(t)) \cdot u''(t) n(u(t)). \tag{7.29}$$

An equation with a better form can be obtained by differentiating

$$u'(t) \cdot n(u(t)) \equiv 0, \tag{7.30}$$

to get

$$u'' \cdot n(u(t)) = -u'(t) \cdot \frac{d}{dt} n(u(t)). \tag{7.31}$$

Plugging this into the right side of (7.29) gives the differential equation

$$u''(t)+u'(t)\cdot\Bigl(\frac{d}{dt}n(u(t))\Bigr)n(u(t))=0. \tag{7.32}$$

Note by (7.28) that u'' is orthogonal to $u'(t)$, so

$$\frac{d}{dt}\|u'(t)\|^2=2u'(t)\cdot u''(t)\equiv 0. \tag{7.33}$$

Thus stationary paths $u:[a,b]\rightarrow M$ for the energy have constant speed.

Such curves on M are *geodesics*. These curves are also constant speed curves on M that are stationary curves for the arclength:

$$\ell(u)=\int_a^b \|u'(t)\|\,dt. \tag{7.34}$$

We will not go further into this here. The reader can consult texts on elementary differential geometry, such as [**DoC**], [**Hen**]], or [**Op**], or see [**T**], Chapter 1, §11.

We next present another approach to finding equations for stationary paths of (7.27). Suppose $\Omega=\mathcal{O}\times\mathbb{R}$ and M is the graph of a function $z=\varphi(x_1,x_2)$, for $x=(x_1,x_2)\in\mathcal{O}$. Then a curve $u:[a,b]\rightarrow M$ has the form

$$u(t)=\bigl(x(t),\varphi(x(t))\bigr), \tag{7.35}$$

and

$$u'(t)=(x'(y),\nabla\varphi(x(t))\cdot x'(t)), \tag{7.36}$$

so

$$\begin{aligned}\|u'(t)\|^2&=\|x'(t)\|^2+(\nabla\varphi(x(t))\cdot x'(t))^2\\&=x'(t)\cdot G(x(t))x'(t),\end{aligned} \tag{7.37}$$

where

$$G(x)=\begin{pmatrix}1+\varphi_1(x)^2 & \varphi_1(x)\varphi_2(x)\\ \varphi_1(x)\varphi_2(x) & 1+\varphi_2(x)^2\end{pmatrix},\quad \varphi_j(x)=\frac{\partial\varphi}{\partial x_j}. \tag{7.38}$$

Thus the problem of finding a constrained stationary path $u(t)$ for the energy (7.27) is equivalent to the problem of finding an unconstrained stationary path $x(t)$ for

$$\mathcal{E}(x)=\frac{1}{2}\int_a^b x'(t)\cdot G(x(t))x(t)\,dt. \tag{7.39}$$

In this case,

$$\begin{aligned} L(x,v) &= \frac{1}{2} v \cdot G(x)v, \\ L_v(x,v) &= G(x)v, \quad \text{and}, \\ L_x(x,v) &= \frac{1}{2} v \cdot \nabla G(x)v, \end{aligned} \tag{7.40}$$

where the last identity means

$$L_{x_k}(x,v) = \frac{1}{2} v \cdot \frac{\partial G}{\partial x_k} v. \tag{7.41}$$

In this setting, the Lagrange equation (7.7) becomes

$$\frac{d}{dt}\Big[G(x(t))x'(t)\Big] - \frac{1}{2} x'(t) \cdot \nabla G(x(t))x'(t) = 0, \tag{7.42}$$

i.e.,

$$\frac{d}{dt} \sum_j G_{kj}(x(t))x_j'(t) - \frac{1}{2} \sum_{i,j} x_i'(t) \frac{\partial G_{ij}}{\partial x_k} x_j'(t) = 0, \quad \forall\, k. \tag{7.43}$$

Exercises

1. Given a Lagrangian $L(x,v)$, we define the "energy"

$$\begin{aligned} E(x,v) &= L_v(x,v) \cdot v - L(x,v) \\ &= \sum_k L_{v_k}(x,v)v_k - L(x,v). \end{aligned} \tag{7.44}$$

Show that if $u(t)$ solves the Lagrange equation (7.7), then

$$\frac{d}{dt} E(u(t),u'(t)) \equiv 0. \tag{7.45}$$

This is energy conservation,onservation of energy in this setting.

2. Suppose

$$L(x,v) = \frac{m}{2} v \cdot G(x)v - V(x), \tag{7.46}$$

Assume $G(x) \in M(n,\mathbb{R})$ is symmetric and invertible, and define $E(x,v)$ as in (7.44). Show that

$$E(x,v) = \frac{m}{2} v \cdot G(x)v + V(x). \tag{7.47}$$

3. Let $L(x,v)$ be given by (7.46). Show that the Lagrange equation (7.7) is

$$m\frac{d}{dt}\Big[G(u(t))u'(t)\Big] - \frac{m}{2}u'(t)\cdot\nabla G(t)u'(t) = -\nabla V(u(t)), \tag{7.48}$$

where the second term is evaluated as in (7.42)–(7.43). Show in turn that this yields the first order system

$$\frac{du_k}{dt} = v_k$$

$$m\sum_j G_{kj}(u(t))\frac{dv_j}{dt} + m\sum_{i,j} v_i(t)\Big[\frac{\partial G_{kj}}{\partial x_i} - \frac{1}{2}\frac{\partial G_{ij}}{\partial x_k}\Big]v_j(t) = -\frac{\partial V}{\partial x_k}(u(t)).$$

Produce a variant by symmetrizing the term in brackets in the second sum, with respect to i and j.

4. Consider the setting of constrained motion on $M \subset \Omega$, as in (7.21)–(7.24), and consider the following generalization of (7.26):

$$L(x,v) = \frac{m}{2}\|v\|^2 - V(x). \tag{7.49}$$

Establish the following replacement for (7.32):

$$mu''(t) + mu'(t)\cdot\Big(\frac{d}{dt}n(u(t))\Big)n(u(t)) = -P_M(u(t))\nabla V(u(t)), \tag{7.50}$$

where, for $x \in M,\ w \in \mathbb{R}^n$,

$$P_M(x)w = w - \Big(n(x)\cdot w\Big)n(x). \tag{7.51}$$

This describes motion of a particle in a force field $F(x) = -\nabla V(x)$, constrained to move on M.

5. Motion of a spherical pendulum in $\mathbb{R}^3$, in the presence of Earth's gravitational field, is described as in Exercise 4 with

$$M = \{x \in \mathbb{R}^3 : \|x\| = \ell\}, \tag{7.52}$$

and $L(x, v)$ as in (7.49), with $V(x) = mg(x \cdot k)$, where $k = (0, 0, 1)^t$. Show that in this case, (7.50) produces, for

$$u(t) = \ell\omega(t), \tag{7.53}$$

the system

$$\omega''(t) + \|\omega'(t)\|^2\omega(t) = -\frac{g}{\ell}k + \frac{g}{\ell}(\omega(t) \cdot k)\omega(t). \tag{7.54}$$

6. Results of Exercise 5 are also valid in the setting where $\mathbb{R}^3$ is replaced by $\mathbb{R}^2$. Show that, in this setting, with

$$\omega(t) = (\sin\theta(t), -\cos\theta(t))^t, \quad k = (0, 1)^t, \tag{7.55}$$

the equation (7.54) leads to the (planar) pendulum equation

$$\theta''(t) + \frac{g}{\ell}\sin\theta(t) = 0. \tag{7.56}$$

7. Let us return to the setting of Exercise 2, and set

$$p = L_v(x, v) = mG(x)v. \tag{7.57}$$

Also set

$$\mathcal{E}(x, p) = E(x, v) = E(x, G(x)^{-1}p/m). \tag{7.58}$$

Show that

$$\mathcal{E}(x, p) = \frac{1}{2m}p \cdot G(x)^{-1}p + V(x). \tag{7.59}$$

Show that the Lagrange equation (7.48) for $u(t) = x(t)$ is equivalent to the following Hamiltonian system:

$$\frac{dx_k}{dt} = \frac{\partial\mathcal{E}}{\partial p_k}, \quad \frac{dp_k}{dt} = -\frac{\partial\mathcal{E}}{\partial x_k}. \tag{7.60}$$

Hint. To get started on (7.60), note that if (7.59) holds, then

$$\frac{\partial\mathcal{E}}{\partial p} = \frac{1}{m}G(x)^{-1}p = v, \tag{7.61}$$

and that the Lagrange equation implies

$$\frac{dp_k}{dt} = L_{x_k}(x,v) = \frac{m}{2} v \cdot \frac{\partial G}{\partial x_k}(x) v - \frac{\partial V}{\partial x_k}(x). \tag{7.62}$$

Furthermore, as in (8.13) of Chapter 3,

$$\frac{\partial}{\partial x_k} G(x)^{-1} = -G(x)^{-1} \frac{\partial G}{\partial x_k}(x) G(x)^{-1}. \tag{7.63}$$

Remark. More general cases in which the change of variable $p = L_v(x,v)$ converts Lagrange's equation to Hamiltonian form are discussed in [**AM**], [**Ar**], and Chapter 1 of [**T**].

Exercises 8–11 study sufaces of revolution that are surfaces of "least area." To set this up, let $u : [0,1] \to (0,\infty)$ be smooth, and rotate the graph of $y = u(x)$ about the x-axis in (x,y,z)-space. Elementary calculus gives the formula

$$A(u) = 2\pi \int_0^1 u(t)\sqrt{1 + u'(t)^2}\, dt \tag{7.64}$$

for the area of the resulting surface of revolution. The problem is to find u for which the area is minimal, given constraints

$$u(0) = \alpha, \quad u(1) = \beta, \quad \alpha, \beta > 0. \tag{7.65}$$

8. In (7.64), $L(x,v) = x\sqrt{1+v^2}$. Show that the "energy" $E(x,v)$ in (7.44) is given by

$$E(x,v) = -\frac{x}{\sqrt{1+v^2}}. \tag{7.66}$$

9. Using (7.45), show that if $u(t)$ solves the Lagrange equation (7.7) in this setting, then there is a constant a such that

$$\frac{u(t)}{\sqrt{1+u'(t)^2}} = a, \tag{7.67}$$

hence

$$\frac{du}{dt} = \pm\sqrt{b^2u^2 - 1}, \quad b = \frac{1}{a}. \tag{7.68}$$

10. Separate variables in (7.68) and use the substitution $bu = \cosh v$ to evaluate the u-integral and conclude that

$$u(t) = \frac{1}{b}\cosh(bt + c), \tag{7.69}$$

for some constant c. Equation (7.69) is the equation of a catenary, seen before in (3.24) of Chapter 1, for the hanging cable.

11. Consider the problem of finding b and c in (7.69) such that the constraints (7.65) are satisfied. Show that sometimes no solutions exist, and sometimes two solutions exist, but one gives a smaller area than the other.

Exercises 12–15 take another look at the hanging cable problem mentioned in Exercise 10. Here we state it as the problem of minimizing the potential energy, which is mg times

$$V(u) = \int_{-A}^{A} u(t)\sqrt{1 + u'(t)^2}\, dt, \tag{7.70}$$

subject to the boundary conditions

$$u(-A) = u(A) = 0, \tag{7.71}$$

and the constraint that the curve $y = u(x)$, $-A \le x \le A$, have length L,

$$\ell(u) = \int_{-A}^{A} \sqrt{1 + u'(t)^2}\, dt = L. \tag{7.72}$$

Such a curve describes a cable, of length L, hanging from the two points $(-A, 0)$ and $(A, 0)$, under the force of gravity. To deal with the constraint (7.72), we bring in the Lagrange multiplier method. That is, we set

$$I_\lambda(u) = V(u) + \lambda\ell(u), \tag{7.73}$$

find the stationary path for (7.73) (subject to (7.71)) as a function of λ, and then find for which λ the constraint (7.72) holds. Note that $I_\lambda(u)$ has the form (7.1) with

$$L_\lambda(x, v) = (x + \lambda)\sqrt{1 + v^2}. \tag{7.74}$$

12. Show that the "energy" $E_\lambda(x,v)$ in (7.44) is given by

$$E_\lambda(x,v)=\frac{x+\lambda}{\sqrt{1+v^2}}. \tag{7.75}$$

13. Using (7.45), show that if $u(t)$ solves the Lagrange equation (7.7) in this setting, then there exists a constant a (maybe depending on λ) such that

$$\frac{u(t)+\lambda}{\sqrt{1+u'(t)^2}}=a, \tag{7.76}$$

hence

$$\frac{du}{dt}=\pm\sqrt{b^2(u+\lambda)^2-1},\quad b=\frac{1}{a}. \tag{7.77}$$

14. Separate variables in (7.77) and use the substitution $b(u+\lambda)=\cosh v$ to evaluate the u-integral and obtain

$$u(t)=-\lambda+\frac{1}{b}\cosh(bt+c),$$

for some constant c. Show that (7.71) forces $c=0$, so

$$u(t)=-\lambda+\frac{1}{b}\cosh bt. \tag{7.78}$$

15. Calculate the length of the curve $y=u(x)$, $-A\le x\le A$, when u is given by (7.78), and show that the constraints (7.71)–(7.72) yield the equations

$$\sinh bA=\frac{bL}{2},\quad \lambda=\frac{1}{b}\cosh bA. \tag{7.79}$$

Note that the first equation has a unique solution $b\in(0,\infty)$ if and only if $L>2A$.

16. Recall the planar pendulum problem illustrated in Fig. 7.1. Instead of assuming all the mass is at the end of the rod, assume the rod has a mass

Figure 8.1

distribution $m(s)\,ds$, $0 \le s \le \ell$, so the total mass is $m = \int_0^\ell m(s)\,ds$. Show that for the potential energy V you replace (7.17) by

$$V = -m_a g\ell \cos\theta, \quad m_a = \int_0^\ell m(s)\frac{s}{\ell}\,ds, \tag{7.80}$$

and for the kinetic energy T, you replace (7.16) by

$$T = \frac{m_b \ell^2}{2}\theta'(t)^2, \quad m_b = \int_0^\ell m(s)\Bigl(\frac{s}{\ell}\Bigr)^2\,ds. \tag{7.81}$$

Write down the replacement for the pendulum equation (7.20) in this setting. Specialize the calculation to the case

$$m(s) = \frac{m}{\ell}, \quad 0 \le s \le \ell, \tag{7.82}$$

which represents a rod with uniform mass distribution.

8. The brachistochrone problem

The early masters of calculus enjoyed posing challenging problems to each other. The most famous of these is called the *brachistrochrone problem.* It was posed by Johann Bernoulli in 1696, and solved by him, by his brother Jakob, and also by Newton and by Leibniz. The problem is to find the curve along which a particle will slide without friction in the minimum time, from one given point p in the (x, y)-plane to another, q, starting at rest at p. Say $p = (0, 0)$ and $q = (a, b)$. We assume $a > 0$ and $b < 0$; see Fig. 8.1. The force of gravity acts in the direction of the negative y-axis, with acceleration g.

Our approach to this problem will involve two applications of the variational method developed in §7. (In fact, this problem helped spark the *creation* of the variational method.) First, let $\varphi : [0,a] \to \mathbb{R}$ with $\varphi(0) = 0,\ \varphi(a) = b$, and consider the constrained motion of a particle,

$$u : [0,t_0] \longrightarrow M = \{(x,\varphi(x)) : 0 \le x \le a\}, \tag{8.1}$$

under the force of gravity. Thus, in place of (7.27), we look for stationary paths for

$$I(u) = \int_0^a \Bigl[\frac{m}{2}\|u'(t)\|^2 - V(u(t))\Bigr]\,dt, \tag{8.2}$$

subject to the constraint (8.1), and with

$$V(x,y) = mgy. \tag{8.3}$$

We can convert this to an unconstrained variational problem as was done in (7.35)–(7.42), now with a nonzero V, and with lower dimension. We have

$$u(t) = \bigl(x(t), \varphi(x(t))\bigr), \tag{8.4}$$

and

$$\|u'(t)\|^2 = \bigl(1 + \varphi'(x(t))^2\bigr)x'(t)^2, \tag{8.5}$$

so the problem of finding a constrained stationary path $u(t)$ for (8.2) is equivalent to the problem of finding an unconstrained stationary path $x(t)$ for

$$J(x) = \int_0^a L(x(t), x'(t))\,dt, \tag{8.6}$$

with

$$L(x,v) = \frac{m}{2}\bigl(1 + \varphi'(x)^2\bigr)v^2 - mg\varphi(x). \tag{8.7}$$

The path $x(t)$ is governed by the differential equation

$$\frac{d}{dt}L_v(x(t), x'(t)) - L_x(x(t), x'(t)) = 0. \tag{8.8}$$

We need not write this more explicitly, since by now our experience tells us that to describe solutions to such a single equation, all we need is conservation of energy:

$$E(x,v) = \frac{m}{2}\bigl(1 + \varphi'(x)^2\bigr)v^2 + mg\varphi(x), \tag{8.9}$$

that is, for a solution to (8.8),

$$\frac{m}{2}\big(1+\varphi'(x(t))^2\big)x'(t)^2+mg\varphi(x(t))=E \tag{8.10}$$

is constant. In the current set-up, $x(0)=0$ and $x'(0)=0$, so $E=0$. We get

$$\frac{dx}{dt}=\pm\sqrt{\frac{-2g\varphi(x)}{1+\varphi'(x)^2}}, \tag{8.11}$$

which separates to

$$\frac{1}{\sqrt{2g}}\int_0^a\sqrt{\frac{1+\varphi'(x)^2}{-\varphi(x)}}\,dx=\int_0^{t_0}dt. \tag{8.12}$$

In other words, the elapsed time for the particle to move from $p=(0,0)$ to $q=(a,b)$ along the path $y=\varphi(x)$ is given by the left side of (8.12).

Hence the brachistochrone problem is reduced to the problem of finding $\varphi:[0,a]\longrightarrow\mathbb{R}$, minimizing

$$K(\varphi)=\int_0^a\mathcal{L}(\varphi(x),\varphi'(x))\,dx, \tag{8.13}$$

subject to the condition

$$\varphi(0)=0,\quad \varphi(a)=b, \tag{8.14}$$

where

$$\mathcal{L}(\varphi,\psi)=\sqrt{\frac{1+\psi^2}{-\varphi}}. \tag{8.15}$$

Stationary paths for (8.13) satisfy the Lagrange equation

$$\frac{d}{dt}\mathcal{L}_\psi(\varphi(x),\varphi'(x))-\mathcal{L}_\varphi(\varphi(x),\varphi'(x))=0. \tag{8.16}$$

Note that

$$\mathcal{L}_\psi(\varphi,\psi)=\frac{\psi}{\sqrt{-\varphi(1+\psi^2)}},\quad L_\varphi(\varphi,\psi)=-\frac{1}{2}\frac{\sqrt{-\varphi(1+\psi^2)}}{\varphi^2}. \tag{8.17}$$

Solutions to (8.16) have the property that

$$\mathcal{E}(\varphi(t),\varphi'(t))=E \tag{8.18}$$

is constant, where (parallel to (7.44))

$$\mathcal{E}(\varphi,\psi)=\mathcal{L}_\psi(\varphi,\psi)-\mathcal{L}(\varphi,\psi). \tag{8.19}$$

Using (8.15) and (8.17), we have

$$\begin{aligned}\mathcal{E}(\varphi,\psi)&=\frac{\psi^2}{\sqrt{-\varphi(1+\psi^2)}}-\sqrt{\frac{1+\psi^2}{-\varphi}}\\ &=-\frac{1}{\sqrt{-\varphi(1+\psi^2)}}.\end{aligned} \tag{8.20}$$

Thus, if $\varphi(x)$ satisfies (8.16), then

$$\varphi(x)\bigl(1+\varphi'(x)^2\bigr)=-k^2,\quad \text{const.}, \tag{8.21}$$

where we have written the constant as $-k^2$ to enforce the condition that $\varphi(x)<0$ for $0<x\le a$. For notational convenience, we make the change of variable

$$y(x)=-\varphi(x), \tag{8.22}$$

so (8.21) becomes

$$y(x)\bigl(1+y'(x)^2\bigr)=k^2, \tag{8.23}$$

giving

$$\frac{dy}{dx}=\sqrt{\frac{k^2}{y}-1}. \tag{8.24}$$

The equation (8.24) separates to

$$\int\frac{dy}{\sqrt{\frac{k^2}{y}-1}}=\int dx. \tag{8.25}$$

The left integral has the form of (5.15) in Chapter 1, with $E_0=-1,\ Km=k^2$. Rather then recall the formulas (5.16)–(5.22) of Chapter 1, we implement the method previewed in Exercise 3 of that section. We use the change of variable

$$y=k^2\sin^2\tau,\quad 2\tau=\theta. \tag{8.26}$$

Then

$$dy=2k^2\sin\tau\cos\tau\,d\tau,\qquad \sqrt{\frac{k^2}{y}-1}=\frac{\cos\tau}{\sin\tau}, \tag{8.27}$$

Figure 8.2

so

$$\begin{aligned}\int \frac{dy}{\sqrt{\frac{k^2}{y}-1}} &= 2k^2 \int \sin^2 \tau \, d\tau \\ &= \frac{k^2}{2} \int (1-\cos\theta)\, d\theta \\ &= \frac{k^2}{2}(\theta - \sin\theta),\end{aligned} \tag{8.28}$$

the second identity because $\sin^2\tau = (1-\cos 2\tau)/2$. Thus the curve $(x, y(x))$, $x \in [0, a]$, is parametrized by

$$\begin{aligned} x = x(\theta) &= \frac{k^2}{2}(\theta - \sin\theta), \\ y = y(\theta) &= \frac{k^2}{2}(1-\cos\theta). \end{aligned} \tag{8.29}$$

The choice of $k^2 > 0$ is dictated by the implication

$$0 < \theta < \pi k^2, \quad \frac{k^2}{2}(\theta - \sin\theta) = a \Longrightarrow \frac{k^2}{2}(1-\cos\theta) = |b|. \tag{8.30}$$

This solves the brachistochrone problem. The curve defined by (8.29) is known as a *cycloid*. See Fig. 8.2. Here $\rho = k^2/2$.

Remark. Note that $y'(0) = +\infty$, so the optimal path starts directly down.

Exercises

1. Show that for each $a, |b| \in (0, \infty)$, there is a unique $k^2 > 0$ such that $(a, |b|) \in \mathbb{R}^2_+$ lies on the curve (8.29), for some $\theta \in (0, \pi k^2)$.
Hint. Consult Fig. 8.2.

2. In the setting of Exercise 1, show that if $|b|/a < 2/\pi$, then $\theta > \pi k^2/2$, and the optimal path dips below b before reaching the endpoint $q = (a, b)$.

3. With $x(\theta)$ and $y(\theta)$ as in (8.29), set $\varphi(\theta) = -y(\theta)$. Let

$$\theta_1 = \frac{k^2}{2}\pi, \quad \theta_0 \in [0, \theta_1). \tag{8.31}$$

Show that the time it takes a particle starting at rest at $(x(\theta_0), \varphi(\theta_0))$ to slide down the curve $(x(\theta), \varphi(\theta))$, $\theta_0 \le \theta \le \theta_1$, to the point $(x(\theta_1), \varphi(\theta_1))$ (the bottom of the cycloid) is independent of θ_0. One says the cycloid also solves the *tautochrone problem.*

9. The double pendulum

Here we study the motion of a double pendulum, such as illustrated in Fig. 9.1. We have a pair of rigid rods, of lengths ℓ_1 and ℓ_2, of negligible mass except for objects of mass m_1 and m_2 attached to one end of each rod. The other end of rod 1 is attached to a fixed point, and the end of rod 2 not containing mass 2 is attached to rod 1 at mass 1. The rods are assumed free to swing back and forth in a plane. Thus the configuration at time t is described by the angles $\theta_1(t)$ and $\theta_2(t)$, that the rods make with the vertical. Gravity acts on the masses m_j, with a downward force of $m_j g$.

We identify the plane mentioned above with the complex plane, with rod 1 attached to the origin and the real axis pointing down. Thus the position of mass 1 is

$$z_1(t) = \ell_1 e^{i\theta_1(t)}, \tag{9.1}$$

and the position of mass 2 is

$$z_2(t) = z_1(t) + \ell_2 e^{i\theta_2(t)}. \tag{9.2}$$

Their velocities are

$$\begin{aligned} z_1' &= i\ell_1\theta_1' e^{i\theta_1}, \\ z_2' &= i\ell_1\theta_1' e^{i\theta_1} + i\ell_2\theta_2' e^{i\theta_2}, \end{aligned} \tag{9.3}$$

Figure 9.1

with square norms

$$
\begin{aligned}
|z_1'|^2 &= \ell_1^2(\theta_1')^2, \\
|z_2'|^2 &= (\ell_1\theta_1' e^{i\theta_1} + \ell_2\theta_2' e^{i\theta_2})(\ell_1\theta_1' e^{-i\theta_1} + \ell_2\theta_2' e^{-i\theta_2}) \\
&= \ell_1^2(\theta_1')^2 + \ell_2^2(\theta_2')^2 + 2\ell_1\ell_2\theta_1'\theta_2'\cos(\theta_1 - \theta_2).
\end{aligned} \tag{9.4}
$$

The potential energy of this system is given by

$$
\begin{aligned}
V &= -m_1 g \operatorname{Re} z_1(t) - m_2 g \operatorname{Re} z_2(t) \\
&= -m_1 g \ell_1 \cos\theta_1 - m_2 g(\ell_1\cos\theta_1 + \ell_2\cos\theta_2),
\end{aligned} \tag{9.5}
$$

and the kinetic energy by

$$
T = \frac{m_1}{2}|z_1'(t)|^2 + \frac{m_2}{2}|z_2'(t)|^2. \tag{9.6}
$$

If we write

$$
\theta = \begin{pmatrix}\theta_1 \\ \theta_2\end{pmatrix}, \quad \psi = \begin{pmatrix}\psi_1 \\ \psi_2\end{pmatrix} = \begin{pmatrix}\theta_1' \\ \theta_2'\end{pmatrix}, \tag{9.7}
$$

then (9.4) gives

$$
T = \frac{1}{2}\psi \cdot G(\theta)\psi, \tag{9.8}
$$

with

$$G(\theta)=\begin{pmatrix}(m_1+m_2)\ell_1^2 & m_2\ell_1\ell_2\cos(\theta_1-\theta_2)\\ m_2\ell_1\ell_2\cos(\theta_1-\theta_2) & m_2\ell_2^2\end{pmatrix}. \tag{9.9}$$

Thus the Lagrangian $L=T-V$ is given by

$$L(\theta,\psi)=\frac{1}{2}\psi\cdot G(\theta)\psi-V(\theta), \tag{9.10}$$

with $V(\theta)$ as in (9.5), and the equation of motion for the double pendulum is

$$\frac{d}{dt}L_\psi(\theta,\theta')-L_\theta(\theta,\theta')=0. \tag{9.11}$$

As in (7.48), this expands out to the 2 by 2 system

$$\frac{d}{dt}\sum_j G_{kj}(\theta(t))\theta_j'(t)-\frac{1}{2}\sum_{i,j}\theta_i'(t)\frac{\partial G_{ij}}{\partial\theta_k}\theta_j'(t)=-\frac{\partial V}{\partial\theta_k}(\theta(t)), \tag{9.12}$$

for $k=1,2$. Making explicit use of (9.5) and (9.9), we have

$$\begin{aligned}L_{\psi_1}(\theta,\psi)&=(m_1+m_2)\ell_1^2\psi_1+m_2\ell_1\ell_2\psi_2\cos(\theta_1-\theta_2),\\ L_{\psi_2}(\theta,\psi)&=m_2\ell_2^2\psi_2+m_2\ell_1\ell_2\psi_1\cos(\theta_1-\theta_2),\end{aligned} \tag{9.13}$$

and

$$\begin{aligned}L_{\theta_1}(\theta,\psi)&=-m_2\ell_1\ell_2\psi_1\psi_2\sin(\theta_1-\theta_2)-(m_1+m_2)g\ell_1\sin\theta_1,\\ L_{\theta_2}(\theta,\psi)&=m_2\ell_1\ell_2\psi_1\psi_2\sin(\theta_1-\theta_2)-m_2g\ell_2\sin\theta_2.\end{aligned} \tag{9.14}$$

Thus the explicit version of (9.11)–(9.12) is the pair of equations

$$\begin{aligned}&(m_1+m_2)\ell_1^2\theta_1''+m_2\ell_1\ell_2\frac{d}{dt}\Bigl[\theta_2'\cos(\theta_1-\theta_2)\Bigr]\\ &\qquad=-m_2\ell_1\ell_2\theta_1'\theta_2'\sin(\theta_1-\theta_2)-(m_1+m_2)g\ell_1\sin\theta_1,\end{aligned} \tag{9.15}$$

and

$$\begin{aligned}&\ell_2^2\theta_2''+\ \ell_1\ell_2\frac{d}{dt}\Bigl[\theta_1'\cos(\theta_1-\theta_2)\Bigr]\\ &\qquad=\ell_1\ell_2\theta_1'\theta_2'\sin(\theta_1-\theta_2)-g\ell_2\sin\theta_2.\end{aligned} \tag{9.16}$$

Note that the masses m_1 and m_2 do not appear in (9.16); m_1 does not appear in either term of $(d/dt)L_{\psi_2}-L_{\theta_2}$, and m_2 factors out.

As in (7.44)–(7.47), we have the energy

$$E(\theta,\psi)=\frac{1}{2}\psi\cdot G(\theta)\psi+V(\theta), \tag{9.17}$$

and if $\theta(t)$ solves (9.11), or equivalently (9.15)–(9.16), then

$$\frac{d}{dt}E(\theta(t),\theta'(t))=0. \tag{9.18}$$

By (9.5) and (9.9), the explicit form of the energy is

$$\begin{aligned}E(\theta,\psi)=&\frac{1}{2}(m_1+m_2)\ell_1^2\psi_1^2+m_2\ell_1\ell_2\psi_1\psi_2\cos(\theta_1-\theta_2)\\&+\frac{1}{2}m_2\ell_2^2\psi_2^2-m_1g\ell_1\cos\theta_1-m_2g(\ell_1\cos\theta_1+\ell_2\cos\theta_2).\end{aligned} \tag{9.19}$$

As in (7.57)–(7.60), we can convert the equations of motion to Hamiltonian form, by setting

$$p=G(\theta)\psi. \tag{9.20}$$

The energy (9.17) becomes

$$\begin{aligned}\mathcal{E}(\theta,p)&=E(\theta,G(\theta)^{-1}p)\\&=\frac{1}{2}p\cdot G(\theta)^{-1}p+V(\theta),\end{aligned} \tag{9.21}$$

and (9.11) is equivalent to

$$\frac{d\theta_k}{dt}=\frac{\partial\mathcal{E}}{\partial p_k},\quad \frac{dp_k}{dt}=-\frac{\partial\mathcal{E}}{\partial\theta_k}. \tag{9.22}$$

Note that, for $G(\theta)$ given by (9.9),

$$G(\theta)^{-1}=\frac{1}{\det G(\theta)}\begin{pmatrix}m_2\ell_2^2 & -m_2\ell_1\ell_2\cos(\theta_1-\theta_2)\\ -m_2\ell_1\ell_2\cos(\theta_1-\theta_2) & (m_1+m_2)\ell_1^2\end{pmatrix}, \tag{9.23}$$

and

$$\det G(\theta)=m_1m_2\ell_1^2\ell_2^2+m_2^2\ell_1^2\ell_2^2\sin^2(\theta_1-\theta_2). \tag{9.24}$$

For notational simplicity we write

$$\mathcal{E}(\theta,p)=\frac{1}{2}p\cdot H(\theta)p+V(\theta),\quad H(\theta)=G(\theta)^{-1}. \tag{9.25}$$

Solutions to (9.22) are orbits of the flow generated by the Hamiltonian vector field

$$\begin{aligned} X_{\mathcal{E}}(\theta,p) &= -J\nabla_{\theta,p}\mathcal{E}(\theta,p) \\ &= \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \begin{pmatrix} \nabla_\theta \mathcal{E} \\ \nabla_p \mathcal{E} \end{pmatrix} \\ &= \begin{pmatrix} \nabla_p \mathcal{E} \\ -\nabla_\theta \mathcal{E} \end{pmatrix}. \end{aligned} \tag{9.26}$$

Here $I \in M(2,\mathbb{R})$ is the identity matrix and $J \in M(4,\mathbb{R})$ is defined by the second identity in (9.26). From this formula we see that the critical points of $X_{\mathcal{E}}$ coincide with the critical points of $\mathcal{E}$. Note that

$$\nabla_p \mathcal{E}(\theta,p) = H(\theta)p, \tag{9.27}$$

and $H(\theta)$ is invertible for all θ, so if $\mathcal{E}$ has a critical point at (θ,p), $p=0$. Now

$$\nabla_\theta \mathcal{E}(\theta,0) = \nabla V(\theta), \tag{9.28}$$

so we deduce that (θ,p) is a critical point of $X_{\mathcal{E}}$ if and only if $p=0$ and $\nabla V(\theta)=0$. Rewriting (9.5) as

$$V(\theta) = -(m_1+m_2)g\ell_1\cos\theta_1 - m_2 g\ell_2\cos\theta_2, \tag{9.29}$$

we see that

$$\nabla V(\theta) = \begin{pmatrix} (m_1+m_2)g\ell_1\sin\theta_1 \\ m_2 g\ell_2 \sin\theta_2 \end{pmatrix}, \tag{9.30}$$

so the critical points of V consist of $\theta_1 = j\pi,\ \theta_2 = k\pi,\ j,k\in\mathbb{Z}$. In summary, the critical points of $X_{\mathcal{E}}$ consist of

$$(\theta_1,\theta_2,p_1,p_2) = (j\pi,k\pi,0,0), \quad j,k\in\mathbb{Z}. \tag{9.31}$$

Towards the goal of understanding the behavior of $X_{\mathcal{E}}$ near these critical points, we examine its derivative. We have

$$DX_{\mathcal{E}}(\theta,0) = \begin{pmatrix} 0 & H(\theta) \\ -D^2V(\theta) & 0 \end{pmatrix}. \tag{9.32}$$

The matrix $H(\theta)$ is positive definite for all θ, and in particular, since $\sin j\pi = 0$ and $\cos j\pi = (-1)^j$,

$$H(j\pi,k\pi) = \frac{1}{m_1m_2\ell_1^2\ell_2^2} \begin{pmatrix} m_2\ell_2^2 & (-1)^{j-k+1}m_2\ell_1\ell_2 \\ (-1)^{j-k+1}m_2\ell_1\ell_2 & (m_1+m_2)\ell_1^2 \end{pmatrix}. \tag{9.33}$$

Also,

$$D^2V(j\pi,k\pi) = \begin{pmatrix} (-1)^j(m_1+m_2)g\ell_1 & 0 \\ 0 & (-1)^k m_2 g\ell_2 \end{pmatrix}. \tag{9.34}$$

We are set up to examine the linearization of the flow generated by $X_{\mathcal{E}}$ at the critical points. This will be pursued, in a more general setting, in the next section.

Exercises

1. Pass to the limit $m_2 \to 0$ in the double pendulum system (9.15)–(9.16) and derive the limiting system

(9.35)
$$\theta_1'' + \frac{g}{\ell_1} \sin \theta_1 = 0,$$

(9.36)
$$\theta_2'' + \frac{\ell_1}{\ell_2} \frac{d}{dt} \Big[\theta_1' \cos(\theta_1 - \theta_2) \Big] = \frac{\ell_1}{\ell_2} \theta_1' \theta_2' \sin(\theta_1 - \theta_2) - \frac{g}{\ell_2} \sin \theta_2.$$

2. Recall the spherical pendulum, introduced in Exercise 5 of §7. Derive equations of motion for a double spherical pendulum.

3. Instead of assuming all the mass of rods 1 and 2 is concentrated at an end, assume that rod j has mass distribution $m_j(s)\, ds,\ 0 \le s \le \ell_j$, so the total mass of rod j is $m_j = \int_0^{\ell_j} m_j(s)\, ds,\ j = 1, 2$. Obtain formulas for the potential and kinetic energy, replacing (9.5) and (9.6), and then obtain equations of motion, replacing (9.15)–(9.16).
Note. See Exercise 16 in §7 to get started.

10. Momentum-quadratic Hamiltonian systems

Most of the Lagrangians arising in the last three sections have been of the form

$$L(x, v) = \frac{1}{2} v \cdot G(x) v - V(x), \tag{10.1}$$

for $x \in \Omega \subset \mathbb{R}^n,\ v \in \mathbb{R}^n$, where $G(x) \in M(n, \mathbb{R})$ is symmetric and invertible, in fact positive definite, but for awhile we will work in this more general setting. As exercises in §7 have revealed, making the change of variables $(x, v) \mapsto (x, p)$ with $p = G(x)v$, one can convert the Lagrange system of differential equations to Hamiltonian form,

$$\frac{dx_k}{dt} = \frac{\partial \mathcal{E}}{\partial p_k}, \quad \frac{dp_k}{dt} = -\frac{\partial \mathcal{E}}{\partial x_k}, \tag{10.2}$$

where

$$\mathcal{E}(x, p) = \frac{1}{2} p \cdot H(x) p + V(x), \quad H(x) = G(x)^{-1}. \tag{10.3}$$

We call such systems momentum-quadratic Hamiltonian systems. Note that $H(x)$ is also symmetric and invertible, and furthermore positive definite if $G(x)$ is. Solutions of (10.2) are orbits of the flow generated by the Hamiltonian vector field

$$\begin{aligned} X_{\mathcal{E}}(x,p) &= -J\nabla_{x,p}\mathcal{E}(x,p) \\ &= \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \begin{pmatrix} \nabla_x \mathcal{E} \\ \nabla_p \mathcal{E} \end{pmatrix} \\ &= \begin{pmatrix} \nabla_p \mathcal{E} \\ -\nabla_x \mathcal{E} \end{pmatrix}. \end{aligned} \tag{10.4}$$

Here, $I \in M(n,\mathbb{R})$ is the identity matrix, and $J \in M(2n,\mathbb{R})$ is defined by the second identity in (10.4).

We record some general results about the critical points of such fields, and their linearizations. To begin, the critical points of $X_{\mathcal{E}}$ coincide with the critical points of $\mathcal{E}$. Note that

$$\nabla_p \mathcal{E}(x,p) = H(x)p, \tag{10.5}$$

so, since $H(x)$ is invertible, we see that if $\mathcal{E}$ has a critical point at (x,p), then $p = 0$. Now

$$\nabla_x \mathcal{E}(x,0) = \nabla V(x), \tag{10.6}$$

so we deduce that the critical points of $X_{\mathcal{E}}$ consist of

$$\{(x,0) : \nabla V(x) = 0\}. \tag{10.7}$$

We next look at the linearization (cf. (3.32)) of $X_{\mathcal{E}}$ at a critical point $(x_0, 0)$, given by

$$DX_{\mathcal{E}}(x_0,0) = \begin{pmatrix} 0 & H(x_0) \\ -D^2V(x_0) & 0 \end{pmatrix}. \tag{10.8}$$

From here on, we assume $H(x_0)$ is positive definite. For notational simplicity, we set

$$H = H(x_0), \quad W = D^2V(x_0), \quad L = \begin{pmatrix} 0 & H \\ -W & 0 \end{pmatrix}. \tag{10.9}$$

Then the linearization of (10.2) at $(x_0, 0)$ is

$$\frac{dx}{dt} = Hp, \quad \frac{dp}{dt} = -Wx. \tag{10.10}$$

To analyze the structure of solutions to (10.10), it is convenient to directly tackle the second order system

$$\frac{d^2x}{dt^2} = -HWx, \tag{10.11}$$

and to do this we bring in the following.

Lemma 10.1. *Given that $H \in M(n,\mathbb{R})$ is positive definite, there exists a positive definite $A \in M(n,\mathbb{R})$ such that*

$$H = A^2. \tag{10.12}$$

Proof. From Chapter 2 we know that $\mathbb{R}^n$ has an orthonormal basis $\{v_j\}$ of eigenvectors of H, so $Hv_j = \lambda_j v_j,\ 1 \le j \le n$. Each λ_j is positive, so we can define A by $Av_j = \sqrt{\lambda_j} v_j,\ 1 \le j \le n$.

If we make the change of variable

$$x = Ay, \tag{10.13}$$

then (10.11) is converted to

$$y'' + AWAy = 0. \tag{10.14}$$

Note that $W \in M(n,\mathbb{R})$ is symmetric and so is AWA. Also AWA is invertible if and only if W is. This invertibility is equivalent to the assertion that $(x_0, 0)$ is a nondegenerate critical point of $X_{\mathcal{E}}$. We restrict attention to such cases. The following result will be useful.

Lemma 10.2. *Let $W \in M(n,\mathbb{R})$ be a symmetric matrix, and assume*

$$W \text{ has } k \text{ positive and } n-k \text{ negative eigenvalues.} \tag{10.15}$$

Then so does AWA, when $A \in M(n,\mathbb{R})$ is positive definite.

Proof. Write $\mathbb{R}^n = \mathcal{W}_+ \oplus \mathcal{W}_-$, where $\mathcal{W}_+$ is the linear span of the eigenvectors of W with positive eigenvalue, $\mathcal{W}_-$ the linear span of the eigenvectors of W with negative eigenvalue. Similarly, write $\mathbb{R}^n = \widetilde{\mathcal{W}}_+ \oplus \widetilde{\mathcal{W}}_-$, with W replaced by AWA. The image $A\widetilde{\mathcal{W}}_+$ of $\widetilde{\mathcal{W}}_+$ under A is a linear subspace of $\mathbb{R}^n$, and

$$v = Aw \in A\widetilde{\mathcal{W}}_+ \Longrightarrow v \cdot Wv = w \cdot AWAw \ge 0 \Longrightarrow v \in \mathcal{W}_+. \tag{10.16}$$

Thus

$$A : \widetilde{W}_+ \longrightarrow \mathcal{W}_+, \quad \text{injectively}, \tag{10.17}$$

so

$$\dim \widetilde{\mathcal{W}}_+ \le \dim \mathcal{W}_+. \tag{10.18}$$

A similar argument gives

$$\dim \widetilde{\mathcal{W}}_- \le \dim \mathcal{W}_-, \tag{10.19}$$

and finishes the proof.

To continue, under the hypotheses of Lemma 10.2, we have an orthonormal basis $\{u_1, \dots, u_n\}$ of $\mathbb{R}^n$ such that, with $\mu_j \in (0, \infty)$,

$$\begin{aligned} AWAu_j &= \mu_j^2 u_j, & j \le k, \\ AWAu_j &= -\mu_j^2 u_j, & j > k. \end{aligned} \tag{10.20}$$

in such a case, the general solution to (10.14) is

$$\begin{aligned} y(t) = &\sum_{j \le k} (a_j \sin \mu_j t + b_j \cos \mu_j t) u_j \\ &+ \sum_{j > k} (a_j e^{\mu_j t} + b_j e^{-\mu_j t}) u_j. \end{aligned} \tag{10.21}$$

Such $y(t)$ leads to

$$\begin{pmatrix} Ay(t) \\ A^{-1} y'(t) \end{pmatrix} = \begin{pmatrix} x(t) \\ p(t) \end{pmatrix} = e^{tL} \begin{pmatrix} v_0 \\ v_1 \end{pmatrix}, \tag{10.22}$$

for general $v_0, v_1 \in \mathbb{R}^n$. As a result, we have the following.

Proposition 10.3. *Under the hypotheses of Lemma 10.2, L, given by (10.9), is diagonalizable, and its eigenvalues are*

$$\begin{aligned} \pm i\mu_j \quad &\textit{for} \quad j \le k, \\ \pm \mu_j \quad &\textit{for} \quad j > k. \end{aligned} \tag{10.23}$$

Proof. The eigenvalues of L are what appear in the exponents in the matrix coefficients of e^{tL}. If L were not diagonalizable, some matrix coefficients would also contain terms of the form $t^\ell e^{t\lambda}$, $\ell \ge 1$, where $\mu = \pm i\mu_j$ or $\pm \mu_j$ in (10.23), depending on j.

A critical point of $X_{\mathcal{E}}$ is said to be *hyperbolic* if all of the eigenvalues of $DX_{\mathcal{E}}$ have nonzero real part. From the analysis above, we have the following.

Proposition 10.4. *A critical point $(x_0, 0)$ of $X_{\mathcal{E}}$ is hyperbolic if and only if*

$$D^2V(x_0) \quad \text{is negative definite.} \tag{10.24}$$

If (10.24) holds, $DX_{\mathcal{E}}(x_0, 0)$ has n positive eigenvalues and n negative eigenvalues.

Whenever a vector field X (Hamiltonian or not) has a hyperbolic critical point, say at z_0, the phase portrait near z_0 for the flow generated by X has a similar appearance to that for the flow generated by its linearization at z_0. This is a generalization of the two dimensional result mentioned below (3.69). See Appendix C for further discussion.

The opposite extreme can also be read off from (10.23).

Proposition 10.5. *At a critical point $(x_0, 0)$ of $X_{\mathcal{E}}$, all the eigenvalues of $DX_{\mathcal{E}}$ are purely imaginary if and only if*

$$D^2V(x_0) \quad \text{is positive definite.} \tag{10.25}$$

Recalling that $\mathcal{E}(x, p)$ is given by (10.3), we see that (10.25) is equivalent to

$$D^2\mathcal{E}(x_0, 0) \in M(2n, \mathbb{R}) \quad \text{is positive definite,} \tag{10.26}$$

in which case $\mathcal{E}$ has a local minimum at $(x_0, 0)$.

In case (10.25) holds, we can deduce from (10.21)–(10.22), with $k = n$, that the orbits of e^{tL} all lie in n-dimensional tori. As for the flow generated by $X_{\mathcal{E}}$ itself, we know that its orbits all lie on level surfaces of $\mathcal{E}$. Near $(x, p) = (x_0, 0)$, these level sets look like $(2n - 1)$-dimensional spheres in $\mathbb{R}^n$. In case $n = 1$, these are closed curves in $\mathbb{R}^2$, and indeed the phase portrait for the flow generated by $X_{\mathcal{E}}$ near $(x_0, 0)$ looks like that for the flow generated by its linearization. In such a case, $(x_0, 0)$ is a *center*, discussed in §3. In case $n > 1$, the orbits of the flow generated by $X_{\mathcal{E}}$ near $(x_0, 0)$ do not necessarily lie on n-dimensional tori. The analysis of this behavior is much more subtle than in the case of hyperbolic critical points. There will be n-dimensional invariant tori that are invariant under the flow, arising rather densely near $(x_0, 0)$, but the flow generated by $X_{\mathcal{E}}$ often has chaotic behavior on the complement of these tori. Study of this situation is part of the deep Kolmogorov-Arnold-Moser (KAM) theory. Discussion of this, and references to further work, can be found in [**AM**], Chapter 8, and [**Ar**], Appendices 7–8.

For $n \geq 2$, there can be cases intermediate between those covered by Proposition 10.4 and those covered by Proposition 10.5.

Proposition 10.6. *If $(x_0, 0)$ is a critical point for $X_{\mathcal{E}}$ and*

$$D^2V(x_0) \ \text{has } k \text{ positive eigenvalues and } n-k \text{ negative eigenvalues,} \tag{10.27}$$

then

$$\begin{aligned} DX_{\mathcal{E}}(x_0, 0) \ &\text{has } 2k \text{ imaginary eigenvalues, and} \\ &n-k \text{ positive, and } n-k \text{ negative eigenvalues.} \end{aligned} \tag{10.28}$$

In such cases, with $k \geq 1$ and $n \geq 2$, the phase portrait for the flow generated by $X_{\mathcal{E}}$ near $(x_0, 0)$ will generally differ from that of its linearization in important details, with some exceptions, arising when $X_{\mathcal{E}}$ is "integrable." We refer to the sources cited above for more on this.

Let us specialize these results to the case of the double pendulum, discussed in §9. There V was given by (9.29), and the critical points by (9.31), i.e., $(j\pi, k\pi, 0, 0)$, and $D^2V(j\pi, k\pi)$ by (9.34). We have

$$\begin{aligned} &j \text{ and } k \text{ even } \Longrightarrow D^2V(j\pi, k\pi) \text{ positive definite,} \\ &j \text{ and } k \text{ odd } \Longrightarrow D^2V(j\pi, k\pi) \text{ negative definite,} \\ &j \text{ and } k \text{ of opposite parity } \Longrightarrow D^2V(j\pi, k\pi) \text{ indefinite.} \end{aligned} \tag{10.29}$$

In the first case Proposition 10.5 applies, in the second case Proposition 10.4 applies, and in the third case Proposition 10.6 applies, with $k = 1$ and $n - k = 1$.

Exercises

1. Establish analogues of Propositions 10.3, 10.5, and 10.6 in case H is allowed to be indefinite (nondegenerate), and we assume

$$D^2V(x_0) \ \text{ is either positive definite or negative definite.} \tag{10.30}$$

Exercises 2–6 deal with the 2×2 system

$$\frac{d^2}{dt^2}\begin{pmatrix} x \\ y \end{pmatrix} = -\nabla_{x,y} V(x, y), \tag{10.31}$$

for various functions V. The associated energy function, as in (10.3), is

$$\mathcal{E}(x, y, p, q) = \frac{1}{2}(p^2 + q^2) + V(x, y). \tag{10.32}$$

In each case, do the following.
(a) Find all the critical points of $\mathcal{E}$.
(b) Determine the type of each critical point of $\mathcal{E}$.
(c) Determine the behavior of the eigenvalues of $DX_{\mathcal{E}}$ at each such critical point (via Proposition 10.6).

2. Take
$$V(x,y) = (\cos x)(\cos y).$$

3. Take
$$V(x,y) = x^2 + xy + y^4.$$

4. Take
$$V(x,y) = x^4 + xy + y^4.$$

5. Take
$$V(x,y) = x^4 - xy + y^4.$$

6. Take
$$V(x,y) = x^4 - x^2y + y^4.$$

7. Do analogues of Exercises 2–6 with (10.32) replaced by

$$\mathcal{E}(x,y,p,q) = \frac{1}{2}(p^2 - q^2) + V(x,y). \tag{10.33}$$

Now Proposition 10.6 will not apply, but Exercise 1 might (or might not).

11. Numerical study – difference schemes

We describe some ways of numerically approximating the solution to a system of differential equations

$$\frac{dx}{dt} = F(x), \quad x(t_0) = x_0. \tag{11.1}$$

Higher order systems can be transformed to first order systems and treated by these methods, which are known as difference schemes.

To start, we pick a time step h and attempt an approximation to the solution to (11.1) at times $t_0 + nh$:

$$x_n \approx x(t_0 + nh). \tag{11.2}$$

Noting that a smooth solution to (11.1) satisfies

$$\begin{aligned} x(t+h) &= x(t) + hx'(t) + O(h^2) \\ &= x(t) + hF(x(t)) + O(h^2), \end{aligned} \tag{11.3}$$

we have the following crude difference scheme:

$$x_{n+1} = x_n + hF(x_n). \tag{11.4}$$

This is said to be first order accurate, meaning that over an interval of unit length one carries out $1/h$ such operations, each with error $O(h^2)$, giving an accumulated error $O(h)$, i.e., on the order of h to the first power. This method of approximating the solution $x(t)$ is often called the Euler method, though considering what a great master of computation Euler was, it is hard to believe he actually took it seriously. Shortly we will present a fourth order accurate method, which is generally satisfactory, after describing some second order accurate methods.

These better difference schemes will be suggested by higher order accurate methods of numerical integration. The connection between the two comes from rewriting (11.1) as

$$x(t+h) = x(t) + \int_0^h F(x(t+s))\, ds. \tag{11.5}$$

Consider methods of approximating

$$\int_0^h g(s)\, ds \tag{11.6}$$

better than $hg(0) + O(h^2)$, for smooth g. Two simple improvements are

$$\frac{h}{2}\Bigl[g(0) + g(h)\Bigr] + O(h^3), \tag{11.7}$$

the trapezoidal method, and

$$hg\Bigl(\frac{h}{2}\Bigr) + O(h^3), \tag{11.8}$$

the midpoint method. These lead respectively to

$$x(t+h) = x(t) + \frac{h}{2}\Big[F(x(t)) + F(x(t+h))\Big] + O(h^3) \tag{11.9}$$

and

$$x(t+h) = x(t) + hF\Big(x\Big(t+\frac{h}{2}\Big)\Big) + O(h^3). \tag{11.10}$$

Neither of them immediately converts to an explicit difference scheme, but in (11.9) we can substitute $F(X(t+h)) = F\big(X(t)+hF(X(t))\big)+O(h^2)$ and in (11.10) we can substitute $F\big(X(t+h/2)\big) = F\big(X(t)+(h/2)F(X(t))\big)+O(h^2)$, to obtain the second order accurate difference schemes

$$x_{n+1} = x_n + \frac{h}{2}\Big[F(x_n) + F\big(x_n + hF(x_n)\big)\Big] \tag{11.11}$$

and

$$x_{n+1} = x_n + hF\Big(x_n + \frac{h}{2}F(x_n)\Big). \tag{11.12}$$

Often (11.11) is called Heun's method and (11.12) a modified Euler method.

We now come to the heart of the matter for this section. The Runge-Kutta scheme for (11.1) is specified as follows. The approximation x_n to $x(t_0+nh)$ is given recursively by

$$x_{n+1} = x_n + \frac{h}{6}\Big(K_{n1} + 2K_{n2} + 2K_{n3} + K_{n4}\Big), \tag{11.13}$$

where

$$\begin{aligned} K_{n1} &= F(x_n), \\ K_{n2} &= F\Big(x_n + \frac{1}{2}hK_{n1}\Big), \\ K_{n3} &= F\Big(x_n + \frac{1}{2}hK_{n2}\Big), \\ K_{n4} &= F(x_n + hK_{n3}). \end{aligned} \tag{11.14}$$

This scheme is 4th order accurate. It is one of the most popular and important difference schemes used for numerical studies of systems of differential equations. We make some comments about its derivation.

We will consider a method of deriving 4th order accurate difference schemes, based on Simpson's formula

$$\int_0^h g(s)\,ds = \frac{h}{6}\Big(g(0) + 4g\big(\frac{h}{2}\big) + g(h)\Big) + O(h^5). \tag{11.15}$$

This formula is derived by producing a quadratic polynomial $p(s)$ such that $p(s) = g(s)$ at $s = 0,\ h/2$, and h, and then exactly integrating $p(s)$. The formula can be verified by rewriting it as

$$\int_{-h}^{h} G(s)\,ds = \frac{h}{3}\Big[G(-h) + 4G(0) + G(h)\Big] + O(h^5). \tag{11.16}$$

The main part on the right is exact for all odd $G(s)$, and it is also exact for $G(s) = 1$ and $G(s) = s^2$, so it is exact when $G(s)$ is a polynomial of degree ≤ 3. Making a power series expansion $G(s) = \sum_{j=0}^{3} a_j s^j + O(s^4)$ then yields (11.16).

Now, write the equation (11.1) as the integral equation (11.5). By (11.15),
(11.17)

$$\int_0^h F(X(t+s))\,ds = \frac{h}{6}\Big[F(X(t)) + 4F\Big(X\Big(t+\frac{h}{2}\Big)\Big) + F(X(t+h))\Big] + O(h^5).$$

We then have as an immediate consequence the following result on producing accurate difference schemes.

Proposition 11.1. *Suppose the approximation*

$$x(t+h) \approx x(t) + \Phi(x(t),h) = \mathcal{X}(x(t),h) \tag{11.18}$$

produces a jth order accurate difference scheme for the solution to (11.1). If $j \le 3$, then a difference scheme accurate of order $j+1$ is given by

$$x_{n+1} = x_n + \frac{h}{6}\Big[F(x_n) + 4F\Big(\mathcal{X}\Big(x_n, \frac{h}{2}\Big)\Big) + F(\mathcal{X}(x_n,h))\Big]. \tag{11.19}$$

Furthermore, if $x(t+h) \approx \mathcal{X}_\ell(x(t),h)$ both work in (11.18), $\ell = 0, 1$, then you can use

$$x_{n+1} = x_n + \frac{h}{6}\Big[F(x_n) + 4F\Big(\mathcal{X}_0\Big(x_n, \frac{h}{2}\Big)\Big) + F(\mathcal{X}_1(x_n,h))\Big]. \tag{11.20}$$

We apply this to two second order methods derived before:

$$\mathcal{X}_0(x_n,h) = x_n + \frac{h}{2}\Big[F(x_n) + F(x_n + hF(x_n))\Big], \ \text{Heun}, \tag{11.21}$$

and

$$\mathcal{X}_1(x_n,h) = x_n + hF\Big(x_n + \frac{h}{2}F(x_n)\Big), \ \text{modified Euler}. \tag{11.22}$$

Thus a third order accurate scheme is produced. The last term in (11.19) becomes

(11.23)
$$\frac{h}{6}\Big[F(x_n)+4F\Big(x_n+\frac{h}{4}\Big[F(x_n)+F\Big(x_n+\frac{h}{2}F\Big)\Big]\Big)+F\Big(x_n+hF\Big(x_n+\frac{h}{2}F\Big)\Big)\Big],$$

where $F = F(x_n)$. In terms of $K_{n1},\ K_{n2}$ as defined in (11.14), we have

$$\frac{h}{6}\Big[K_{n1}+4F\Big(x_n+\frac{h}{4}[K_{n1}+K_{n2}]\Big)+F(x_n+hK_{n2})\Big]. \tag{11.24}$$

This could be used in a 3rd order accurate scheme, but some simplification of the middle term is desirable. Note that, for smooth H,

$$H\Big(x+\frac{1}{2}\eta\Big)=\frac{1}{2}H(x)+\frac{1}{2}H(x+\eta)+O(|\eta|^2). \tag{11.25}$$

Consequently, as $|K_{n1}-K_{n2}| = O(h)$, by (11.14),

(11.26)
$$F\Big(x_n+\frac{h}{4}[K_{n1}+K_{n2}]\Big)=\frac{1}{2}F\Big(x_n+\frac{h}{2}K_{n1}\Big)+\frac{1}{2}F\Big(x_n+\frac{h}{2}K_{n2}\Big)+O(h^4).$$

Therefore we have the following.

Proposition 11.2. *A third order accurate difference scheme for (11.1) is given by*

$$x_{n+1}=x_n+\frac{h}{6}[K_{n1}+2K_{n2}+2K_{n3}+L_{n4}] \tag{11.27}$$

where $K_{n1},\ K_{n2},\ K_{n3}$ are given by (11.14) and

$$L_{n4}=F(x_n+hK_{n2}). \tag{11.28}$$

We can now produce a 4th order accurate difference scheme by applying Proposition 11.1 with $\mathscr{X}(x_n,h)$ defined by (11.27). Thus we obtain the difference scheme.

$$\begin{aligned} x_{n+1}=x_n+\frac{h}{6}\Big\{K_{n1}&+4F\Big(x_n+\frac{h}{12}[K_{n1}+2k_{n2}+2k_{n3}+\ell_{n4}]\Big)\\ &+F\Big(x_n+\frac{h}{6}[K_{n1}+2K_{n2}+2K_{n3}+L_{n4}]\Big)\Big\},\end{aligned} \tag{11.29}$$

where $K_{nj},\ L_{n4}$ are as above and

$$\begin{aligned} k_{n2}&=F\Big(x_n+\frac{h}{4}K_{n1}\Big),\\ k_{n3}&=F\Big(x_n+\frac{h}{4}k_{n2}\Big),\\ \ell_{n4}&=F\Big(x_n+\frac{h}{2}k_{n2}\Big).\end{aligned} \tag{11.30}$$

This formula is more complicated than the Runge-Kutta formula (11.13). We say no more about how to obtain (11.13), which represents a masterpiece of insight.

We have dealt specifically with autonomous systems in (11.1), but a non-autonomous system

$$\frac{dx}{dt} = G(t,x), \quad x(t_0) = x_0, \tag{11.31}$$

can be treated similarly, as one can see by writing its autonomous analogue

$$\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} G(y,x) \\ 1 \end{pmatrix}, \quad \begin{pmatrix} x(t_0) \\ y(t_0) \end{pmatrix} = \begin{pmatrix} x_0 \\ t_0 \end{pmatrix}, \tag{11.32}$$

and applying the formulas just derived to (11.32).

We move briefly to another class of difference schemes, based on power series. It derives from the expansion

$$x(t+h) = x(t) + hx'(t) + \frac{h^2}{2}x''(t) + \cdots + \frac{h^k}{k!}x^{(k)}(t) + O(h^{k+1}). \tag{11.33}$$

To begin, differentiate (11.1), producing

$$x''(t) = F_2(x,x'), \quad F_2(x,x') = DF(x)x'. \tag{11.38}$$

Continue differentiating, getting

$$x^{(j)}(t) = F_j(x,x',\dots,x^{(j-1)}), \quad j \le k. \tag{11.35}$$

Then one obtains a difference scheme for an approximation x_n to $x(t_0+nh)$, of the form

$$x_{n+1} = x_n + hx_n' + \frac{h^2}{2}x_n'' + \cdots + \frac{h^k}{k!}x_n^{(k)}, \tag{11.36}$$

where

$$x_n' = F(x_n), \quad x_n'' = F_2(x_n,x_n'), \tag{11.37}$$

and, inductively,

$$x_n^{(j)} = F_j(x_n,x_n',\dots,x_n^{(j-1)}). \tag{11.38}$$

This difference scheme is kth order accurate. In practice, this is not usually a good method, because the formulas for F_j tend to become rapidly more complex. However, in some cases the functions F_j happen not to become very complex, and then this is a good method.

To mention a couple of examples, first consider the central force problem

$$\begin{aligned} x' &= v, \\ y' &= w, \\ v' &= -x(x^2+y^2)^{-3/2}, \\ w' &= -y(x^2+y^2)^{-3/2}. \end{aligned} \tag{11.39}$$

Here, the power series method is not nearly as convenient as the Runge-Kutta method. On the other hand, for the pendulum problem, which for $g/\ell = 1$ we can write as

$$\theta' = \psi, \quad \psi' = -\sin\theta, \tag{11.40}$$

we have

$$\begin{aligned} \theta'' &= \psi', \quad \psi'' = -\psi\cos\theta, \\ \theta^{(3)} &= \psi'', \quad \psi^{(3)} = -\psi'\cos\theta + \psi^2\sin\theta, \\ \theta^{(4)} &= \psi^{(3)}, \quad \psi^{(4)} = -\psi''\cos\theta + 3\psi'\psi\sin\theta + \psi^3\cos\theta, \end{aligned} \tag{11.41}$$

from which one can get a workable fourth order difference scheme of the form (11.36)–(11.38).

There are other classes of difference schemes, such as "predictor-corrector" methods, which we will not discuss here. More about this can be found in numerical analysis texts, such as [**At**] and [**Sh**].

Readers with a working knowledge of a general purpose computer programming language, such as FORTRAN or C, will find it interesting to implement the Runge-Kutta method on a variety of systems of differential equations, including (11.39) and (11.40). Be sure to use double precision arithmetic, which makes computations to 16 digits of accuracy. Alternatively, specialized programming tools such as MATLAB and Mathematica can be used. These tools have built-in graphics capability, with which one can produce phase portraits, and they also have built-in differential equation solvers, whose output one can compare with the output from one's own program. Useful literature on these latter tools for the study of differential equations can be found in [**PA**] and [**GMP**].

When running such programs, pay attention to the way solutions behave when the step size h is changed. As a rule of thumb, if the solution does not change appreciably when the step size is halved, the solution is accurate. To be sure, there is frequently more to obtaining accurate solutions than just choosing a small step size. For more on this, we recommend numerical analysis texts, such as cited above, and of course we also recommend lots of practice on various systems of differential equations.

Exercises

The following exercises are for readers who can use a programming language.

1. Write a program to apply the Runge-Kutta method to the pendulum problem (11.40).

2. Write a program to apply the power series method described in (11.33)–(11.38) to (11.40). Produce a fourth order accurate method.

3. Consider applying the Runge-Kutta scheme to the problem of motion in a planar force field,

$$x'' = f(x,y), \quad y'' = g(x,y), \tag{11.42}$$

which can be written as the first order system

$$\begin{aligned} x' = v, &\quad v' = f(x,y), \\ y' = w, &\quad w' = g(x,y). \end{aligned} \tag{11.43}$$

Show that (11.13)–(11.14) in this context become

$$\begin{aligned} x &\mapsto x + \frac{h}{6}(v + 2v_2 + 2v_3 + v_4), \\ y &\mapsto y + \frac{h}{6}(w + 2w_2 + 2w_3 + w_4), \\ v &\mapsto v + \frac{h}{6}(a_1 + 2a_2 + 2a_3 + a_4), \\ w &\mapsto w + \frac{h}{6}(b_1 + 2b_2 + 2b_3 + b_4), \end{aligned} \tag{11.44}$$

where a_j, b_j, v_j, and w_j are computed as follows. First,

$$a_1 = f(x,y), \quad b_1 = g(x,y); \tag{11.45}$$

then

$$\begin{aligned} x_2 = x + \frac{h}{2}v, &\quad y_2 = y + \frac{h}{2}w, \\ v_2 = v + \frac{h}{2}a_1, &\quad w_2 = w + \frac{h}{2}b_1, \end{aligned} \tag{11.46}$$

and

$$a_2 = f(x_2, y_2), \quad b_2 = g(x_2, y_2); \tag{11.47}$$

then

$$\begin{aligned} x_3 &= x + \frac{h}{2} v_2, \quad y_3 = y + \frac{h}{2} w_2, \\ v_3 &= v + \frac{h}{2} a_2, \quad w_3 = w + \frac{h}{2} b_2, \end{aligned} \tag{11.48}$$

and

$$a_3 = f(x_3, y_3), \quad b_3 = g(x_3, y_3); \tag{11.49}$$

then

$$\begin{aligned} x_4 &= x + h v_3, \quad y_4 = y + h w_3, \\ v_4 &= v + h a_3, \quad w_4 = w + h b_3, \end{aligned} \tag{11.50}$$

and finally,

$$a_4 = f(x_4, y_4), \quad b_4 = g(x_4, y_4). \tag{11.51}$$

Write a program to implement this difference scheme. Test it for various functions $f(x, y)$ and $g(x, y)$. Consider particularly

$$f(x, y) = -\frac{x}{(x^2 + y^2)^{3/2}}, \quad g(x, y) = -\frac{y}{(x^2 + y^2)^{3/2}}, \tag{11.52}$$

arising in the Kepler problem, (11.39).

4. Extend the scope of Exercise 3 to treat

$$x'' = f(x, y, x', y'), \quad y'' = g(x, y, x', y').$$

5. Write a program to apply the Runge-Kutta method to the double pendulum problem (9.15)–(9.16).

12. Limit sets and periodic orbits

Let F be a C^1 vector field on an open set $\mathcal{O} \subset \mathbb{R}^n$, generating the flow Φ^t. Take $x \in \mathcal{O}$. If $\Phi^t(x)$ is well defined for all $t \geq 0$, we define the ω-limit set $L_\omega(x)$ to consist of all points $y \in \mathcal{O}$ such that there exist $t_k \nearrow +\infty$

Figure 12.1

Figure 12.2

with $\Phi^{t_k}(x) \to y$. Similarly, if $\Phi^t(x)$ is well defined for all $t \le 0$, we define the α-limit set $L_\alpha(x)$ to consist of all points $y \in \mathcal{O}$ such that there exist $t_k \searrow -\infty$ with $\Phi^{t_k}(x) \to y$. Sinks are ω-limit sets for all nearby points. Other examples of ω-limit sets are pictured in Figs. 12.1–12.3. In Fig. 12.1, $L_\omega(x)$ is a periodic orbit, i.e., for some $T \in (0, \infty)$, $\Phi^T(y) = y$. In Fig. 12.2, $L_\omega(x)$ is a figure eight, containing a hyperbolic critical point of the vector field. In Fig. 12.3, $L_\omega(x)$ contains several critical points.

The following result, characterizing ω-limit sets in the plane without critical points (under a few additional hypotheses), is called the Poincaré-Bendixson theorem.

Theorem 12.1. *Let $\mathcal{O}$ be a planar domain, and let F generate a flow Φ^t on $\mathcal{O}$. Assume there is a set $K \subset \mathcal{O}$ that is a closed, bounded subset of $\mathbb{R}^2$ and satisfies $\Phi^t(K) \subset K$ for all $t > 0$. Take $x \in K$. If $L_\omega(x)$ contains no critical point of F, then it is a periodic orbit of Φ.*

Figure 12.3

An important ingredient in the proof of the Poincaré-Bendixson theorem is the following classical result about closed curves in the plane.

Jordan Curve Theorem. *Let C be a simple closed curve in $\mathbb{R}^2$, i.e., a continuous, one-to-one image of the unit circle. Then $\mathbb{R}^2 \setminus C$ consists of two connected pieces. Any curve from a point in one of these pieces to a point in the other must cross C.*

We will not present a proof of the Jordan curve theorem. Proofs can be found in [**GrH**], §18, and in [**Mun**]. We do mention that actually we will need this result only for piecewise smooth simple closed curves, where a simpler proof exists; see [**Sto**], pp. 34–40, or [**T**], Chapter 1, §19. The ability of a simple closed curve to separate $\mathbb{R}^n$ fails for $n \geq 3$, which makes the Poincaré-Bendixson theorem an essentially two-dimensional result. Examples discussed in §15 illustrate how much more complex matters can be in higher dimension.

To tackle Theorem 12.1, first note that the hypotheses imply $L_\omega(x)$ is a nonempty subset of K. Let $y \in L_\omega(x)$, and say

$$y_k = \Phi^{t_k}(x), \quad t_k \nearrow +\infty, \quad y_k \to y. \tag{12.1}$$

We have $F(y) \neq 0$. Let Γ be a smooth curve segment in $\mathcal{O}$, containing y, such that the tangent to Γ at y is linearly independent of $F(y)$. Shrinking Γ if necessary, we can assume that for each $z \in \Gamma$, the tangent to Γ at z is linearly independent of $F(z)$. We say F is transverse to Γ; cf. Fig. 12.4.

With y_k as in (12.1), we can assume all y_k are sufficiently close to y to lie in orbits through Γ, and adjusting each t_k as needed, we can take

$$y_k \in \Gamma, \quad \forall\, k. \tag{12.2}$$

Figure 12.4

At this point, is is useful to revise the list $\{t_k\}$ slightly. Let $t_1 \in \mathbb{R}^+$, $y_1 = \Phi^{t_1}(x)$ be as above. Now let $t_k \nearrow +\infty$ denote all the successive times when $\Phi^t(x)$ intersects Γ, so we may be adding times to the set denoted t_k in (12.1). Shortly we will show that (12.1) continues to hold for this expanded set of points $y_k = \Phi^{t_k}(x)$. First, we make the following useful observation.

Lemma 12.2. *With $t_j < t_{j+1} < t_{j+2}$ as above,*

$$y_{j+1} \ \ \textit{lies between} \ \ y_j \ \ \textit{and} \ \ y_{j+2} \ \ \textit{on} \ \ \Gamma. \tag{12.3}$$

Proof. Consider the curve C_j starting at y_j, running to y_{j+1} along $\Phi^t(x)$, $t_j \le t \le t_{j+1}$, and returning to y_j along Γ. Cf. Fig. 12.5. This is a simple closed curve, and the Jordan curve theorem applies.

Now for s and σ small and positive, and $z \in \Gamma$, not on the opposite side of y_{j+1} from y_j, we have $\Phi^s(y_{j+1}) = \Phi^{t_j+s}(x)$ and $\Phi^{-\sigma}(z)$ in the two different connected components of $\mathbb{R}^2 \setminus C_j$. Since $\{\Phi^s(y_{j+1}) : s \ge 0\}$ cannot cross C_j at any point but a point in Γ, we must have

$$\Phi^{-\sigma}(y_{j+2}) = \Phi^{t_{j+2}-\sigma}(x)$$

in the opposite component of $\mathbb{R}^2 \setminus C_j$ from that containing such $\Phi^{-\sigma}(z)$, so y_{j+2} must be on the opposite side of y_{j+1} from y_j in Γ.

Having Lemma 12.2, we see that the expanded set of points $\{y_k\} \subset \Gamma$ interlaces the original set, so (12.1) continues to hold. We see that the convergence of y_k to y is monotone on Γ. If by chance some $y_j = y$, then all $y_k = y$. Otherwise, all the points y_k lie on the same side of y, i.e., on the same connected component of $\Gamma \setminus \{y\}$.

The main thing we need to establish to prove Theorem 12.1 is that the orbit through y is periodic. The next result takes us closer to that goal.

Lemma 12.3. *Suppose $s > 0$ and $\Phi^s(y) \in \Gamma$. Then $\Phi^s(y) = y$.*

Figure 12.5

Proof. We have

$$\sup_{0\le t\le s+1} \|\Phi^t(y_k)-\Phi^t(y)\| = \varepsilon_k \to 0, \quad \text{as} \ \ k\to\infty. \tag{12.4}$$

It follows that there exist $\delta_k \to 0$ such that $\Phi^{s+\delta_k}(y_k)\in\Gamma$, and hence

$$\Phi^{s+\delta_k}(y_k) = y_{k+\ell(k)}, \quad \text{for some} \ \ \ell(k)\in\{1,2,3,\dots\}. \tag{12.5}$$

Thus

$$\Phi^s(y) = \lim_{k\to\infty} \Phi^{s+\delta_k}(y_k) = \lim_{k\to\infty} y_{k+\ell(k)} = y, \tag{12.6}$$

as asserted.

We are ready for the endgame in the proof of Theorem 12.1. Let $s_j \nearrow +\infty$ and consider $z_j = \Phi^{s_j}(y)$. We have each $z_j \in K$, and passing to a subsequence, we can assume

$$z_j = \Phi^{s_j}(y) \longrightarrow z \in K. \tag{12.7}$$

We have $F(z)\neq 0$, so there is a curve segment $\widetilde{\Gamma}$ through z, transverse to F. Adjusting s_j, we can arrange

$$z_j \in \widetilde{\Gamma}_j. \tag{12.8}$$

We need only two such points in such a curve $\widetilde{\Gamma}$; say, upon relabeling,

$$z_1 = \Phi^{s_1}(y), \ z_2 = \Phi^{s_2}(y) = \Phi^{s_2-s_1}(z_1) \in \widetilde{\Gamma}. \tag{12.9}$$

Figure 12.6

See Fig. 12.6.

Note that

$$\Phi^{t_k+s_1}(x) \longrightarrow z_1, \tag{12.10}$$

so we can use the previous results, with t_k replaced by $t_k + s_1$ and y by z_1, and Γ by $\widetilde{\Gamma}$. In this case, the analogue of the hypothesis in Lemma 12.3 applies:

$$s_2 - s_1 > 0, \quad \Phi^{s_2-s_1}(z_1) \in \widetilde{\Gamma}. \tag{12.11}$$

The conclusion of Lemma 12.3 is

$$\Phi^{s_2-s_1}(z_1) = z_1, \tag{12.12}$$

i.e., actually $z_2 = z_1$.

Thus the orbit of Φ through y is periodic, of period $s_2 - s_1$. Since $y \in L_\omega(x)$, it follows that this periodic orbit is contained in $L_\omega(x)$. It is also readily seen that no other point in $\mathcal{O}$ can belong to $L_\omega(x)$, so Theorem 12.1 is proved.

The following equation, known as the van der Pol equation, illustrates the workings of Theorem 12.1. The equation is

$$x'' - \mu(1 - x^2)x' + x = 0. \tag{12.13}$$

Here μ is a positive parameter. This models the current in a nonlinear circuit that amplifies a weak current ($|x| < 1$) and damps a strong current ($|x| > 1$). See the exercises for more on this. The equation (12.13) converts to the first order system

$$x' = y, \quad y' = -x + \mu(1 - x^2)y. \tag{12.14}$$

Figure 12.7

Fig. 12.7 is a phase portrait for the case $\mu = 1$. The vector field F associated with (12.14) has one critical point, at the origin. The linearization of (12.14) at the origin is

$$\frac{d}{dt}\begin{pmatrix} \xi \\ \eta \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & \mu \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix}, \tag{12.15}$$

and the eigenvalues of this matrix are

$$\frac{\mu}{2} \pm \frac{1}{2}\sqrt{\mu^2 - 4}. \tag{12.16}$$

Thus the origin is a source whenever $\mu > 0$. It is a spiral source provided also $\mu < 2$. Note that when $(x(t), y(t))$ solves (12.14),

$$\frac{d}{dt}(x^2 + y^2) = 2\mu(1 - x^2)y^2, \tag{12.17}$$

which is ≥ 0 for $|x| \leq 1$, and in particular is ≥ 0 near the origin.

An examination of Fig. 12.7 indicates the presence of a periodic orbit, attracting all the other orbits. Let us see how this fits into the set-up of Theorem 12.1. To do this, we need to describe a closed bounded set $K \subset \mathbb{R}^2$ such that $\Phi^t(K) \subset K$ for all $t > 0$, where Φ^t is the flow generated by F, and such that F has no critical points in K. We construct K as follows. Look at the orbit of F starting at the point A on the positive y-axis, shown in Fig. 12.7 and again in Fig. 12.8.

A numerical integration of (12.14) (using the Runge-Kutta scheme) shows

Figure 12.8

that

$$\begin{array}{ll} & \Phi^t(A) \ \text{ winds clockwise about the origin,} \\ (12.18) & \text{and again hits the positive } y\text{-axis} \\ & \text{at the point } B, \text{ lying below } A. \end{array}$$

To this path from A to B, one adds the line segment (on the y-axis) from B to A, producing a simple closed curve $\mathcal{C}$. It follows readily from (12.14) that on this line segment the vector field F points to the right. Thus the closed region $\widetilde{K}$ bounded by this curve has the invariance property

$$\Phi^t(\widetilde{K}) \subset \widetilde{K}, \quad \forall\, t \geq 0. \tag{12.19}$$

We then pick $\varepsilon > 0$ small enough (in particular < 1), and set

$$K = \widetilde{K} \setminus \{(x,y) : x^2 + y^2 < \varepsilon^2\}. \tag{12.20}$$

The fact that

$$\Phi^t(K) \subset K, \quad \forall\, t \geq 0 \tag{12.21}$$

follows from (12.19) and (12.17). We have removed the only critical point of F, so K contains no critical points, and Theorem 12.1 applies.

It must be said that the validity of the argument just given relies on the accuracy of the statement (12.18) about the orbit through A. Here we have relied on a numerical approximation to that orbit. We applied the Runge-Kutta scheme, described in §11, with step sizes $h = 10^{-2}$, 10^{-3}, and 10^{-4}, using double precision (16 digit) variables, and got consistent results in all three cases. The last case involves quite a small step size, and if one were to use 8 digit arithmetic, there could be a danger of accumulating truncation errors. In any case, with today's computers there is no point in using 8 digit arithmetic.

Theorem 12.1 is a special case of the following result.

Bendixson's Theorem. *Let F be a C^1 vector field on $\mathcal{O} \subset \mathbb{R}^2$, generating a flow Φ^t. Assume there is a set $K \subset \mathcal{O}$ that is a closed, bounded subset of $\mathbb{R}^2$ and satisfies $\Phi^t(K) \subset K$ for all $t > 0$. Assume F has at most finitely many critical points in K. Then if $x \in K$, $L_\omega(x)$ is one of the following:*
(a) a critical point,
(b) a periodic orbit,
(c) a cyclic graph consisting of critical points joined by orbits.

A proof can be found in [**CL**], Chapter 16, or in [**Lef**], Chapter 10. Note that alternative (c) is illustrated in Figs. 12.2 and 12.3. We emphasize that both this result and Theorem 12.1 are results for planar vetor fields. In higher dimension, matters are completely different, as we will discuss in §15.

We recall a device already used to deal with alternative (a), and develop it a little further. Suppose F is a C^1 vector field on $\mathcal{O} \subset \mathbb{R}^n$, and there is a function $V \in C^1(\mathcal{O})$. Assume V has a unique minimum, at $p \in K$. If $x(t) = \Phi^t(x_0)$, then, by the chain rule,

$$\frac{d}{dt} V(x(t)) = \nabla V(x(t)) \cdot F(x(t)). \tag{12.22}$$

If also V has the property

$$\nabla V(y) \cdot F(y) < 0, \quad \forall\, y \in \mathcal{O} \setminus p, \tag{12.23}$$

we say V is a strong Lyapunov function for F. In such a case,

$$\frac{d}{dt} V(x(t)) < 0, \quad \text{whenever} \ \ x(t) \neq p. \tag{12.24}$$

If we replace (12.23) by the weaker property

$$\nabla V(y) \cdot F(y) \le 0, \quad \forall\, y \in \mathcal{O}, \tag{12.25}$$

we say V is a Lyapunov function for F. In such a case,

$$\frac{d}{dt} V(x(t)) \le 0, \quad \forall\, t \ge 0. \tag{12.26}$$

Thus, as $t \nearrow +\infty$, $V(x(t))$ monotonically approaches a limit, V_0, which must be $\ge V(p)$, and furthermore,

$$\lim_{t \to +\infty} \frac{d}{dt} V(x(t)) = 0. \tag{12.27}$$

This has the following immediate consequence.

Proposition 12.4. *Let F be a C^1 vector field on $\mathcal{O} \subset \mathbb{R}^n$, generating a flow Φ^t. Assume there is a set $K \subset \mathcal{O}$ that is a closed, bounded subset of $\mathbb{R}^n$ and satisfies $\Phi^t(K) \subset K$ for all $t > 0$. Take $x_0 \in K$. Assume $V \in C^1(\mathcal{O})$ is a Lyapunov function for F. Then*

$$L_\omega(x_0) \subset \{y \in \mathcal{O} : \nabla V(y) \cdot F(y) = 0\}. \tag{12.28}$$

If V is a strong Lyapunov function, then

$$L_\omega(x_0) = \{p\}. \tag{12.29}$$

Exercises

1. Let $\mathcal{O} \subset \mathbb{R}^n$ be open and $\overline{\Omega} \subset \mathcal{O}$ a closed bounded set with smooth boundary $\partial\Omega$, with outward pointing normal n. Let F be a C^1 vector field on $\mathcal{O}$, generating the flow Φ^t. Assume

$$F \cdot n \le 0 \quad \text{on} \quad \partial\Omega. \tag{12.30}$$

Show that

$$\Phi^t(\overline{\Omega}) \subset \overline{\Omega}, \quad \forall\, t \ge 0. \tag{12.31}$$

2. In the setting of Exercise 1, show that

$$\Phi^t(\overline{\Omega}) \subset \Phi^s(\overline{\Omega}) \quad \text{for} \quad 0 < s < t. \tag{12.32}$$

Set

$$\mathcal{B} = \bigcap_{t \in \mathbb{R}^+} \Phi^t(\overline{\Omega}) = \bigcap_{k \in \mathbb{Z}^+} \Phi^k(\overline{\Omega}). \tag{12.33}$$

Show that

$$\Phi^t(\mathcal{B}) = \mathcal{B}, \quad \forall\, t \ge 0. \tag{12.34}$$

Remark. It can be shown from material in Appendix B that $\mathcal{B}$ is nonempty, closed, and bounded.

3. In the setting of Exercise 2, show that

$$\forall\, x \in \overline{\Omega}, \quad L_\omega(x) \subset \mathcal{B}. \tag{12.35}$$

4. In the setting of Exercise 2, show that

$$\operatorname{div} F < 0 \;\text{ on }\; \overline{\Omega} \Longrightarrow \mathrm{Vol}(\mathcal{B}) = 0. \tag{12.36}$$

5. In the setting of Exercise 4, assume that $n = 2$, and that F has no critical points in $\overline{\Omega}$, so by Theorem 12.1 there is a periodic orbit of Φ in $\overline{\Omega}$. Show that, due to (12.36), there can be only *one* periodic orbit of Φ in $\overline{\Omega}$.
   Hint. Feel free to use the Jordan Curve Theorem.

   Exercises 6–8 deal with a nonlinear RLC circuit, as pictured in Fig. 12.9. The setup is as in §13 of Chapter 1 (see also Chapter 3, §5), except that Ohm's law is modified. The voltage drop across the "resistor" is given by

$$V = f(I), \tag{12.37}$$

   where f can be nonlinear, and not necessarily monotonic. As an example, one could have

$$f(I) = \mu\Bigl(\frac{1}{3}I^3 - I\Bigr). \tag{12.38}$$

   Vacuum tubes and transistors can behave as such circuit elements. The voltage drop across the capacitor and the inductor are, as before, given respectively by

$$V = L\frac{dI}{dt}, \quad V = \frac{Q}{C}. \tag{12.39}$$

   Units of current, etc., are as in §13 of Chapter 1.

6. Modify the computations done in (14.1)–(14.7) of Chapter 1 and show that the current $I(t)$ satisfies the differential equation

$$\frac{d^2I}{dt^2} + \frac{f'(I)}{L}\frac{dI}{dt} + \frac{1}{LC}I = \frac{E'(t)}{L}. \tag{12.40}$$

Figure 12.9

Show that rescaling I and t leads to (12.14), when $f(I)$ is given by (12.38) and $E \equiv 0$. More generally, rescale (12.40) to

$$\frac{d^2x}{dt^2} + f'(x)\frac{dx}{dt} + x = g(t). \tag{12.41}$$

7. Assume $g \equiv 0$ in (12.41). Parallel to (12.14), one can convert this equation to the first order system

$$x' = y, \quad y' = -x - f'(x)y.$$

Show that you can also convert it to the first order system

$$\begin{aligned} \frac{dx}{dt} &= y - f(x), \\ \frac{dy}{dt} &= -x. \end{aligned} \tag{12.43}$$

This is called a Lienard equation.

8. Show that if $(x(t), y(t))$ solves (12.43), then

$$\frac{d}{dt}(x^2 + y^2) = -2xf(x). \tag{12.44}$$

Figure 13.1

13. Predator-prey equations

Here and in the following section we consider differential equations that model population densities. We start with one species. The simplest model is the exponential growth model:

$$\frac{dx}{dt} = ax. \tag{13.1}$$

Here $x(t)$ denotes the population of the species (or rather, an approximation to what would be an integer valued function). The model simply states that the rate of growth of the population is proportional to the population itself. The solution to (13.1) is our old friend $x(t) = e^{at}x(0)$. This unbounded increase in population is predicated on the existence of limitless resources to nourish the species. An alternative to (13.1) posits that the resources can support a population no greater than K. The following is called the logistic equation:

$$\frac{dx}{dt} = ax(1 - bx), \tag{13.2}$$

where $b = 1/K$. In this model, (13.1) is a good approximation for small x, but the rate of growth slows down to 0 as x approaches its upper limit K. The equation (13.2) can be solved by separation of variables:

$$\frac{dx}{x(1 - bx)} = a\,dt. \tag{13.3}$$

The reader can perform the integration as an exercise.

The function $F(x) = ax(1 - bx)$ on the right side of (13.2) is a one-dimensional vector field, with critical points at $x = 0$ and $x = 1/b$. The intervals $(-\infty, 0)$, $(0, 1/b)$, and $(1/b, \infty)$ are all invariant under the flow generated by F, although only the interval $(0, 1/b)$ has biological relevance. See Fig. 13.1 for the "phase portrait."

We turn to a class of 2×2 systems called "predator-prey" equations. For this, we set

$$\begin{aligned} x(t) &= \text{ population of predators,} \\ y(t) &= \text{ population of prey,} \\ z(t) &= \text{ rate at which each predator eats prey.} \end{aligned} \tag{13.4}$$

Figure 13.2

Depending on the choice of the exponential growth model or the logistic model for the species of prey in the absence of predators, the following systems arise to model these populations:

$$\begin{aligned} \frac{dx}{dt} &= -ax + bzx, \\ \frac{dy}{dt} &= ry - zx, \end{aligned} \tag{13.5}$$

or

$$\begin{aligned} \frac{dx}{dt} &= -ax + bzx, \\ \frac{dy}{dt} &= ry(1 - cy) - zx. \end{aligned} \tag{13.6}$$

Here, a, b, c, and r are positive constants. As for the rate of feeding z, we assume

$$z = \zeta(y). \tag{13.7}$$

Clearly if $y = 0$ then $z = 0$. One possibility that is used is

$$\zeta(y) = \kappa y, \tag{13.8}$$

for some positive constant κ. This posits that the rate of feeding of a predator is proportional to the rate of close encounters of that predator with members of the other species, which in turn is proportional to the population y. This seems intuitively reasonable if y is not large, but most creatures stop eating once they are full, so a more reasonable candidate for $\zeta(y)$ might be as pictured in Fig. 13.2, representing a feeding rate bounded by β.

A class of functions of this sort is given by

$$\zeta(y) = \frac{\kappa y}{1 + \gamma y}, \quad \frac{\kappa}{\gamma} = \beta. \tag{13.9}$$

Another class is

$$\zeta(y) = \beta(1 - e^{-\gamma y}), \quad \beta\gamma = \kappa. \tag{13.10}$$

Let us examine various cases in more detail.

Volterra-Lotka equations

The case (13.5) with z given by (13.8) produces systems called Volterra-Lotka equations:

$$\begin{aligned} \frac{dx}{dt} &= -ax + \sigma xy, \qquad \sigma = b\kappa, \\ \frac{dy}{dt} &= ry - \kappa xy. \end{aligned} \tag{13.11}$$

Note that the x-axis and y-axis are invariant under the flow defined by this system. We have $x' = -ax$ on the x-axis and $y' = ry$ on the y-axis. It follows that the first quadrant, where $x \geq 0$ and $y \geq 0$, is invariant under the flow. This is the region in the (x, y)-plane of biological significance. The vector field $V(x, y) = (-ax + \sigma xy, ry - \kappa xy)^t$ has two critical points. One is the origin. Note that

$$DV(0,0) = \begin{pmatrix} -a & 0 \\ 0 & r \end{pmatrix}, \tag{13.12}$$

so the origin is a saddle. The other critical point is

$$(x_0, y_0) = \Bigl(\frac{r}{\kappa}, \frac{a}{\sigma}\Bigr). \tag{13.13}$$

Note that

$$DV(x_0, y_0) = \begin{pmatrix} 0 & \sigma x_0 \\ -\kappa y_0 & 0 \end{pmatrix}, \tag{13.14}$$

with purely imaginary eigenvalues, so we have a center for the linearization of V at (x_0, y_0). In fact, (x_0, y_0) is a center for V, as we now show.

From (13.11) we get

$$\frac{dy}{dx} = \frac{y(r - \kappa x)}{x(\sigma y - a)}, \tag{13.15}$$

which separates to

$$\Bigl(\sigma - \frac{a}{y}\Bigr)\, dy = \Bigl(\frac{r}{x} - \kappa\Bigr)\, dx. \tag{13.16}$$

Figure 13.3

Integrating yields

$$\sigma y - a \log y = r \log x - \kappa x + C. \tag{13.17}$$

We deduce that the following smooth function on the region $x, y > 0$,

$$H(x,y) = \sigma y - a \log y + \kappa x - r \log x, \tag{13.18}$$

is constant on orbits of (13.11), i.e., these orbits lie on level curves of H. Note that

$$\nabla H(x,y) = \begin{pmatrix} \kappa - \frac{r}{x} \\ \sigma - \frac{a}{y} \end{pmatrix}, \quad D^2 H(x,y) = \begin{pmatrix} \frac{r}{x^2} & 0 \\ 0 & \frac{a}{y^2} \end{pmatrix}, \tag{13.19}$$

hence, with (x_0, y_0) as in (13.13),

$$\nabla H(x_0,y_0) = 0, \quad D^2 H(x_0,y_0) = \begin{pmatrix} \frac{r}{x_0^2} & 0 \\ 0 & \frac{a}{y_0^2} \end{pmatrix}, \tag{13.20}$$

the latter matrix being positive definite, so H has a minimum at (x_0, y_0), which implies that (x_0, y_0) is a center for V. The phase portrait for orbits of (13.11) is pictured in Fig. 13.3.

The system (13.11) was studied independently by Lotka and Volterra around 1925, by Lotka as a model of some chemical reactions and by Volterra as a predator-prey model, specifically for sharks preying on another species of fish. Volterra made the following further observation. Bring in another type of predator, fishermen. Assume the fishermen keep everything they catch and that the probability of getting caught in their nets is the same for sharks and their prey. Then the system (13.11) gets revised to

$$\begin{aligned} \frac{dx}{dt} &= -ax + \sigma xy - ex, \\ \frac{dy}{dt} &= ry - \kappa xy - ey. \end{aligned} \tag{13.21}$$

Now (13.21) has the same form as (13.11), with a replaced by $a+e$ and with r replaced by $r-e$, all these constants remaining positive as long as

$$0 < e < r. \tag{13.22}$$

Then the previous analysis applies. The system (13.21) has a stable critical point at

$$(x_1, y_1) = \Big(\frac{r-e}{\kappa}, \frac{a+e}{\sigma}\Big). \tag{13.23}$$

Note that at this critical point there are fewer sharks and more prey, compared to (13.13). Of course, this depends on the hypothesis (13.22). If $e > r$, things are catastrophically different.

First modification

We turn from Volterra-Lotka equations to predator-prey models given by (13.6), still keeping (13.8). Then we have the following system:

$$\begin{aligned} \frac{dx}{dt} &= -ax + \sigma xy, \qquad \sigma = b\kappa, \\ \frac{dy}{dt} &= ry(1-cy) - \kappa xy. \end{aligned} \tag{13.24}$$

As with (13.11), the x-axis and y-axis are invariant under the flow defined by this system. We have $x' = -ax$ on the x-axis and $y' = ry(1-cy)$ on the y-axis. Again, the first quadrant ($x \ge 0, y \ge 0$) is invariant under the flow. Note furthermore that, for

$$V(x,y) = (-ax + \sigma xy, ry(1-cy) - \kappa xy)^t, \tag{13.25}$$

we have

$$V\Big(x, \frac{1}{c}\Big) = \Big(\Big(\frac{\sigma}{c} - a\Big)x, -\frac{\kappa}{c}x\Big)^t, \tag{13.26}$$

which points downward for $x > 0$. It follows that

$$\mathcal{R} = \Big\{(x,y) : x \ge 0,\, 0 \le y \le \frac{1}{c}\Big\} \tag{13.27}$$

is invariant under this flow. It is this region in the (x, y)-plane that is of biological significance.

To proceed, we find the critical points of $V(x, y)$, given by (13.25). Two of these are

$$(0,0) \text{ and } \Bigl(0, \frac{1}{c}\Bigr). \tag{13.28}$$

$DV(0,0)$ is again given by (13.12), so $(0,0)$ is a saddle. Also,

$$DV\Bigl(0, \frac{1}{c}\Bigr) = \begin{pmatrix} -a + \frac{\sigma}{c} & 0 \\ -\frac{\kappa}{c} & -r \end{pmatrix}. \tag{13.29}$$

V has a third critical point, at

$$y_0 = \frac{a}{\sigma}, \quad x_0 = \frac{r}{\kappa}\Bigl(1 - \frac{ca}{\sigma}\Bigr) = \frac{rc}{\kappa\sigma}\Bigl(\frac{\sigma}{c} - a\Bigr). \tag{13.30}$$

Note how this point is shifted to the left from the point (13.13). There are three cases to consider.

Case I. $\sigma/c - a < 0$.
In this case, the critical point (13.30) is not in the first quadrant, so V has only the critical points (13.28) in $\mathcal{R}$. In this case (13.29) has two negative eigenvalues, so the critical point $(0, 1/c)$ is a sink. Note that the x-component of $V(x, y)$ is

$$x(\sigma y - a) \le x\Bigl(\frac{\sigma}{c} - a\Bigr), \quad \text{for } x \ge 0,\ y \le \frac{1}{c}, \tag{13.31}$$

so V points to the left everywhere in $\mathcal{R}$ except the left edge. Consequently, the population of predators is driven to extinction as $t \to +\infty$, whatever the initial condition.

Case II. $\sigma/c - a > 0$.
In this case the third critical point (x_0, y_0) is in the first quadrant. In fact, $y_0 = a/\sigma < 1/c$, so $(x_0, y_0) \in \mathcal{R}$. Now (13.29) has one positive and one negative eigenvalue, so the critical point $(0, 1/c)$ is a saddle. As for the nature of (x_0, y_0), we have

$$\begin{aligned} DV(x_0, y_0) &= \begin{pmatrix} -a + \sigma y_0 & \sigma x_0 \\ -\kappa y_0 & r(1 - 2cy_0) - \kappa x_0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & \frac{rc}{\kappa}(\frac{\sigma}{c} - a) \\ -\frac{\kappa a}{\sigma} & -\frac{rca}{\sigma} \end{pmatrix}. \end{aligned} \tag{13.32}$$

Figure 13.4

Note that

$$\begin{aligned} \det DV(x_0,y_0) &= \frac{rca}{\sigma}\Bigl(\frac{\sigma}{c}-a\Bigr) > 0, \\ \operatorname{Tr} DV(x_0,y_0) &= -\frac{rca}{\sigma} < 0. \end{aligned} \tag{13.33}$$

It follows that the eigenvalues of $DV(x_0,y_0)$ are either both negative or have negative real part. Hence (x_0,y_0) is a sink.

We claim that the orbit through each point in $\mathcal{R}$ not on the x or y-axis approaches (x_0,y_0) as $t \to +\infty$. To see this, we construct a Liapunov function. We do this by modifying $H(x,y)$ in (13.18), which has a minimum at the point (13.13), to one that has a minimum at the point (13.30). We take

$$\widetilde{H}(x,y) = \sigma y - a\log y + \kappa x - r\Bigl(1-\frac{ca}{\sigma}\Bigr)\log x. \tag{13.34}$$

If $(x(t),y(t))$ solves (13.24), a computation gives

$$\frac{d}{dt}\widetilde{H}(x,y) = -\frac{rc}{\sigma}(\sigma y - a)^2. \tag{13.35}$$

By Proposition 12.4, if we take any point $p \in \mathcal{R}$, with positive x and y-coordinates (so it is in the domain of $\widetilde{H}$), the ω-limit set of p satisfies

$$L_\omega(p) \subset \Bigl\{(x,y) \in \mathcal{R} : y = \frac{a}{\sigma}\Bigr\}. \tag{13.36}$$

The right side is a horizontal line to which V is clearly transverse except at the critical point (x_0,y_0), so indeed $L_\omega(p) = (x_0,y_0)$.

See Fig. 13.4 for a phase portrait treating Case II.

Case III. $\sigma/c - a = 0$.
In this case $(x_0, y_0) = (0, 1/c)$. In (13.29) the eigenvalues are 0 and $-r$, so $(0, 1/c)$ is a degenerate critical point. In place of (13.31) we have that the x-component of $V(x, y)$ is

$$x(\sigma y - a) \le 0, \quad \text{for } x \ge 0,\ y \le \frac{1}{c}, \tag{13.37}$$

and it is strictly negative for $x > 0,\ y < 1/c$. Hence, as in Case I, the population of predators is driven to extinction as $t \to +\infty$.

Second modification

We now move to the next level of sophistication, using the system (13.6) with $z = \zeta(y)$, described as in Fig. 13.2. Thus, we look at systems of the form

$$\begin{aligned} \frac{dx}{dt} &= -ax + bx\zeta(y), \\ \frac{dy}{dt} &= ry(1 - cy) - x\zeta(y). \end{aligned} \tag{13.37}$$

As before, a, b, c, and r are all positive constants. To be precise about what we mean when we say $\zeta(y)$ behaves as in Fig. 13.2, we make the following hypotheses:

$$\begin{array}{ll} \text{(a)} & \zeta : [0, \infty) \to [0, \infty) \ \text{ is smooth,} \\ \text{(b)} & \zeta(0) = 0, \\ \text{(c)} & \zeta'(y) > 0, \quad \forall\, y \ge 0, \\ \text{(d)} & \sup \zeta(y) = \beta < \infty, \\ \text{(e)} & \zeta''(y) \le 0. \end{array} \tag{13.38}$$

All these conditions are satisfied by the examples (13.9) and (13.10). Hypothesis (c) implies ζ is strictly monotone increasing, and hypothesis (e) implies ζ is concave.

In this case, the vector field is

$$V(x, y) = (x(b\zeta(y) - a), ry(1 - cy) - x\zeta(y))^t. \tag{13.39}$$

Parallel to (13.26),

$$V\Bigl(x, \frac{1}{c}\Bigr) = \Bigl(\Bigl(b\zeta\Bigl(\frac{1}{c}\Bigr) - a\Bigr)x, -\zeta\Bigl(\frac{1}{c}\Bigr)x\Bigr)^t, \tag{13.40}$$

which points downward for $x > 0$, and again it follows that the region $\mathcal{R}$, given by (13.27), is invariant under the flow Φ^t generated by V, for $t \geq 0$, and this is the region in the (x, y)-plane that is of biological significance.

Next, we find the critical points of $V(x, y)$. Again, two of them are

$$(0,0) \quad \text{and} \quad \Bigl(0, \frac{1}{c}\Bigr),$$

and again $DV(0,0)$ is given by (13.12), so $(0,0)$ is a saddle. This time,

$$DV\Bigl(0, \frac{1}{c}\Bigr) = \begin{pmatrix} b\zeta(\frac{1}{c}) - a & 0 \\ -\zeta(\frac{1}{c}) & -r \end{pmatrix}. \tag{13.41}$$

Also, a critical point would occur at (x_0, y_0) if these coordinates satisfy

$$\zeta(y_0) = \frac{a}{b}, \quad x_0 = \frac{b}{a} r y_0 (1 - c y_0). \tag{13.42}$$

Under the hypotheses (13.38), the first equation in (13.42) has a (unique) solution if and only if

$$\frac{a}{b} < \beta. \tag{13.43}$$

From here on we will assume (13.43) holds, and leave it to the reader to consider the behavior of the flow when (13.43) fails. Given (13.43), x_0 and y_0 are well defined by (13.42). Parallel to the study of (13.30), again we have three cases.

Case I. $1 - c y_0 < 0$,
Case II. $1 - c y_0 > 0$,
Case III. $1 - c y_0 = 0$.

In Case I, (x_0, y_0) is not in the first quadrant, and in Case III, $(x_0, y_0) = (0, 1/c)$. Again we leave these cases to the reader to think about. We concentrate on Case II.

In Case II, $x_0 > 0$ and $0 < y_0 < 1/c$, so

$$(x_0, y_0) \in \mathcal{R}. \tag{13.44}$$

Given $\zeta(y_0) = a/b$ and the hypotheses (13.38) on ζ, we have

$$\zeta\Bigl(\frac{1}{c}\Bigr) > \frac{a}{b} \Longleftrightarrow \frac{1}{c} > y_0 \Longleftrightarrow 1 - c y_0 > 0, \tag{13.45}$$

and hence in Case II, $DV(0,1/c)$ has one positive eigenvalue and one negative eigenvalue, so

$$\Big(0,\frac{1}{c}\Big) \ \text{ is a saddle.} \tag{13.46}$$

(In Case I, the eigenvalues of $DV(0,1/c)$ are both negative, so $(0,1/c)$ is a sink, and in Case III these eigenvalues are 0 and $-r$.) Next, a computation gives the following analogue of (13.32):

$$\begin{aligned} DV(x_0,y_0) &= \begin{pmatrix} b\zeta(y_0)-a & b\zeta'(y_0)x_0 \\ -\zeta(y_0) & r(1-2cy_0)-x_0\zeta'(y_0) \end{pmatrix} \\ &= \begin{pmatrix} 0 & b\zeta'(y_0)x_0 \\ -\frac{a}{b} & r(1-2cy_0)-x_0\zeta'(y_0) \end{pmatrix}, \end{aligned} \tag{13.47}$$

and parallel to (13.33) we have

$$\begin{aligned} \det DV(x_0,y_0) &= ax_0\zeta'(y_0) > 0, \\ \operatorname{Tr} DV(x_0,y_0) &= r(1-2cy_0) - x_0\zeta'(y_0) \\ &= r\Big[-cy_0 + (1-cy_0)\Big\{1-\frac{\zeta'(y_0)y_0}{\zeta(y_0)}\Big\}\Big]. \end{aligned} \tag{13.48}$$

Let us set

$$Z_0 = 1 - \frac{\zeta'(y_0)y_0}{\zeta(y_0)}. \tag{13.49}$$

Given ζ, this is a function of a/b, but it is independent of c and r. Note that, since $\zeta(0)=0$,

$$\frac{\zeta(y_0)}{y_0} = \zeta'(\tilde{y}), \quad \text{for some } \ \tilde{y}\in(0,y_0), \tag{13.50}$$

by the mean value theorem, so the hypotheses on ζ in (13.38) imply

$$0 < Z_0 < 1. \tag{13.51}$$

(Note that in the context of the previous model, with $\zeta(y)$ given by (13.8), $Z_0 = 0$.) We have

$$\operatorname{Tr} DV(x_0,y_0) = r\big[Z_0(1-cy_0) - cy_0\big]. \tag{13.52}$$

This gives rise to three cases.

Case IIA. $Z_0 < cy_0/(1-cy_0)$.
Then $\operatorname{Tr} DV(x_0,y_0) < 0$, so, by (13.48),

$$(x_0,y_0) \ \text{ is a sink.} \tag{13.53}$$

Figure 13.5

Case IIB. $Z_0 > cy_0/(1 - cy_0)$.
Then $\operatorname{Tr} DV(x_0, y_0) > 0$, so, by (13.48),

$$(x_0, y_0) \text{ is a source.} \tag{13.54}$$

Case IIC. $Z_0 = cy_0/(1 - cy_0)$.
Then $\operatorname{Tr} DV(x_0, y_0) = 0$, so, by (13.48), the eigenvalues of $DV(x_0, y_0)$ are (nonzero) purely imaginary numbers. In this case, (x_0, y_0) is a center for the linearization of V.

We will concentrate on Cases IIA and IIB. Before pursuing these cases further, we want to describe a family of bounded domains in $\mathcal{R}$ that are invariant under the flow Φ^t for $t \ge 0$. Namely, consider the triangle $\mathcal{T}_\mu$ with vertices at $(0, 1/c)$, $(0, 0)$, and $(\mu, 0)$, as pictured in Fig. 13.5.

Claim. If $\mu > 0$ is large enough, the triangle $\mathcal{T}_\mu$ is invariant under Φ^t, for $t \ge 0$.

Proof. Note that V is vertical on the left edge of $\mathcal{T}_\mu$, with critical points at the endpoints of this line segment. Also V points horizontally to the left on the bottom edge of $\mathcal{T}_\mu$. It remains to show that V points into $\mathcal{T}_\mu$ along the line segment from $(0, 1/c)$ to $(\mu, 0)$, provided μ is sufficiently large. This line segment is given by

$$x = \mu(1 - cy), \quad 0 \le y \le \frac{1}{c}, \tag{13.55}$$

and the vector

$$N_\mu = \begin{pmatrix} 1 \\ \mu c \end{pmatrix} \tag{13.56}$$

is normal to this segment, and points away from $\mathcal{T}_\mu$. We want to show that $V \cdot N_\mu \le 0$ along this line segment, for μ large. Indeed, from (13.39),

$$\begin{aligned} V(\mu(1-cy),y)\cdot N_\mu &= (1-cy)\big[\mu(b\zeta(y)-a)+\mu cry-\mu^2 c\zeta(y)\big] \\ &= \mu(1-cy)\big[-a+cry-(\mu c-b)\zeta(y)\big], \end{aligned} \tag{13.57}$$

and under the hypotheses (13.38) on ζ, this is

$$\le 0, \quad \forall\, y \in \Big[0,\frac{1}{c}\Big], \tag{13.58}$$

if μ is sufficiently large, say $\mu \ge \mu_0$.

A similar computation shows that, if $\mu_1 > \mu_0$, then, for each $p \in \mathcal{R}$, $\Phi^t(p) \in \mathcal{T}_{\mu_1}$ for all sufficiently large t.

Back to Cases IIA and IIB, as we have seen, in Case IIA (x_0, y_0) is a sink. It is possible to show that

$$\text{in Case IIA,} \quad \Phi^t(p) \longrightarrow (x_0,y_0), \quad \text{as} \quad t \to +\infty, \tag{13.59}$$

for all p in the interior of $\mathcal{R}$, so the phase portrait has qualitative features similar to Fig. 13.4. On the other hand, in Case IIB, (x_0, y_0) is a source. Hence there is an open set U containing (x_0, y_0) such that

$$\mathcal{T}_{\mu_0} \setminus U \ \text{ is invariant under } \Phi^t, \text{ for } t \ge 0. \tag{13.60}$$

This region does contain the two critical points $(0,0)$ and $(0,1/c)$, on its boundary, but since they are saddles, the argument used to establish the Poincaré-Bendixson theorem, Theorem 12.1, shows that

$$\text{in Case IIB,} \quad L_\omega(p) \ \text{ is a periodic orbit,} \tag{13.61}$$

for all $p \ne (x_0, y_0)$ in the interior of $\mathcal{R}$. The phase portrait is depicted in Fig. 13.6.

Exercises

Exercises 1–5 deal with the system (13.37), i.e.,

$$\begin{aligned} x' &= -ax + bx\zeta(y), \\ y' &= ry(1-cy) - x\zeta(y), \end{aligned} \tag{13.62}$$

Figure 13.6

where $\zeta(y)$ is given by (13.9), i.e.,

$$\zeta(y) = \frac{\kappa y}{1+\gamma y}, \quad \frac{\kappa}{\gamma} = \beta. \tag{13.63}$$

As usual, $a, b, c, \kappa, \gamma, r \in (0,\infty)$. The exercises deal with when Cases I–III, specified below (13.43), hold. Recall these cases apply if and only if there is a critical point (x_0, y_0) given by (13.42), i.e., of and only if

$$\frac{a}{b} < \beta = \frac{\kappa}{\gamma}. \tag{13.64}$$

We will assume this holds.

1. Show that the critical point (x_0, y_0) is given by

$$y_0 = \frac{a}{b\kappa - a\gamma}, \quad x_0 = \frac{b}{a} r y_0 (1 - c y_0). \tag{13.65}$$

2. Show that

$$\begin{aligned} \text{Case I} &\Longleftrightarrow ac > b\kappa - a\gamma, \\ \text{Case II} &\Longleftrightarrow ac < b\kappa - a\gamma, \\ \text{Case III} &\Longleftrightarrow ac = b\kappa - a\gamma. \end{aligned}$$

3. Let Z_0 be given by (13.49), i.e.,

$$Z_0 = 1 - \frac{\zeta'(y_0) y_0}{\zeta(y_0)}. \tag{13.66}$$

Show that

$$Z_0 = \frac{a\gamma}{b\kappa}. \tag{13.67}$$

4. In Case II, recall Cases IIA–IIC, specified below (13.52). Show that

$$\begin{aligned}\text{Case IIA} &\Longleftrightarrow \frac{\gamma}{\kappa} < \frac{c}{b\kappa - a\gamma - ac}, \\ \text{Case IIB} &\Longleftrightarrow \frac{\gamma}{\kappa} > \frac{c}{b\kappa - a\gamma - ac}, \\ \text{Case IIC} &\Longleftrightarrow \frac{\gamma}{\kappa} = \frac{c}{b\kappa - a\gamma - ac},\end{aligned}$$

5. Let us take

$$a = 1, \quad b = 2, \quad \kappa = 1, \quad \gamma = 1. \tag{13.68}$$

Note that (13.64) holds. Show that

$$\begin{aligned}\text{Case I} &\Longleftrightarrow c > 1, \\ \text{Case II} &\Longleftrightarrow c < 1, \\ \text{Case III} &\Longleftrightarrow c = 1.\end{aligned}$$

In Case II, show that

$$\begin{aligned}\text{Case IIA} &\Longleftrightarrow c > \frac{1}{3}, \\ \text{Case IIB} &\Longleftrightarrow c < \frac{1}{3}, \\ \text{Case IIC} &\Longleftrightarrow c = \frac{1}{3}.\end{aligned}$$

Exercises 6–10 deal with the system (13.62), where $\zeta(y)$ is given by (13.10), i.e.,

$$\zeta(y) = \beta(1 - e^{-\gamma y}), \quad \beta\gamma = \kappa. \tag{13.69}$$

Again there is a critical point (x_0, y_0), given by (13.42), if and only if (13.64) holds. We assume this holds, so $b\beta > a$.

6. Show that the critical point (x_0, y_0) is given by

$$y_0 = \frac{1}{\gamma} \log \frac{b\beta}{b\beta - a}, \quad x_0 = \frac{b}{a} r y_0 (1 - c y_0). \tag{13.70}$$

7. For Z_0, defined by (13.66), show that

$$Z_0 = 1 - \frac{b\beta - a}{a} \log \frac{b\beta}{b\beta - a}. \tag{13.71}$$

8. Parallel to Exercise 2, study when Cases I–III hold.

9. Parallel to Exercise 4, study when Cases IIA–IIC hold.

10. Take $a, b, \kappa,$, and γ as in (13.68). Work out a parallel to Exercise 5.

For Exercises 11–12, consider the following system, for x predators and y prey, presented in [**Tau**], p. 376:

$$\begin{aligned} x' &= ax\Bigl(b - \frac{x}{y}\Bigr), \\ y' &= ry(1 - cy) - x\zeta(y). \end{aligned} \tag{13.72}$$

Here the equation for y is as in (13.62), modeling the population of prey in terms of the logistic equation, modified by how fast the prey is eaten. The equation for x has a different basis, a sort of logistic equation in which the population y determines the population limit of x, at any given time.

11. Work out an analysis of the system (13.72) as parallel as possible to the analysis done in this section for (13.62).

12. Take $\zeta(y)$ as in (13.63) and work out results parallel to those of Exercises 1–5.

Exercises 13–15 are for readers who can use a programming language, with graphics capabilities.

13. The following system is known as the basic model of virus dynamics (cf. [**NM**], p. 100, [**W**], p. 26):

$$\begin{aligned}\frac{dx}{dt} &= \lambda - dx - \beta xv,\\ \frac{dy}{dt} &= \beta xv - ay,\\ \frac{dv}{dt} &= ky - uv.\end{aligned} \tag{13.73}$$

Here, x represents the uninfected cell population, y the infected cell population, and v the virus population. The positive parameters λ, d, β, a, k, and u are taken to be constant. The ratio

$$R_0 = \frac{\lambda\beta k}{adu} \tag{13.74}$$

is called the basic reproducive ratio. Graph solution curves for (13.73), with various choices of parameters. Account for the assertion that if $R_0 < 1$ the virus cannot maintain an infection, but if $R_0 > 1$ the system converges to an equilibrium, in which $v > 0$.

14. The simplifying assumption that the virus population is proportional to the infected cell population (say $\beta v = by$) leads to the system

$$\begin{aligned}\frac{dx}{dt} &= \lambda - dx - bxy,\\ \frac{dy}{dt} &= -ay + bxy.\end{aligned} \tag{13.75}$$

Study this system, with an eye to comparison with the Volterra-Lotka system (13.11). Here, replace (13.74) by

$$R_0 = \frac{b\lambda}{ad}. \tag{13.76}$$

15. The following system modifies (13.75) by introducing $z(t)$, the population of "killer T cells," which kill off infected cells, thereby negatively affecting y:

$$\begin{aligned}\frac{dx}{dt} &= \lambda - dx - bxy,\\ \frac{dy}{dt} &= bxy - ay - pyz,\\ \frac{dz}{dt} &= cyz - bz,\end{aligned} \tag{13.77}$$

now with positive parameters λ, d, b, a, p, and c. Continue to define R_0 by (13.76). Consider particularly cases where

$$R_0 > 1, \quad c\Big(\frac{\lambda}{a} - \frac{d}{b}\Big) > b. \tag{13.78}$$

Account for the assertion that in this case the virus population first grows, stimulating the production of killer T cells, which in turn fight the infection and lead to an equilibrium.

For more on these models, see [**NM**] and [**W**], and references therein.

14. Competing species equations

The following system models the populations $x(t)$ and $y(t)$ of two competing species:

$$\begin{aligned} \frac{dx}{dt} &= ax(1-bx) - cxy, \\ \frac{dy}{dt} &= \alpha y(1-\beta y) - \gamma xy. \end{aligned} \tag{14.1}$$

In this model, each population is governed by a logistic equation in the absence of the other species. The presence of the other species reduces the population of its opponent, at a rate proportional to xy. Setting $X = bx$ and $Y = \beta y$ produces an equation like (14.1), but with $X(1-X)$ and $Y(1-Y)$ in place of $x(1-bx)$ and $y(1-\beta y)$, and with different factors. A change of notation gives the system

$$\begin{aligned} \frac{dx}{dt} &= ax(1-x) - cxy, \\ \frac{dy}{dt} &= \alpha y(1-y) - \gamma xy. \end{aligned} \tag{14.2}$$

which we will consider henceforth. We take $a, c, \alpha, \gamma \in (0, \infty)$. Associated to this system is the vector field

$$V = \begin{pmatrix} ax(1-x) - cxy \\ \alpha y(1-y) - \gamma xy \end{pmatrix}. \tag{14.3}$$

Note that $V(x, 0) = (ax(1-x), 0)^t$ and $V(0, y) = (0, \alpha y(1-y))^t$, so the x-axis and y-axis are invariant under the flow Φ^t generated by V. Hence the quadrant $\{x \ge 0, y \ge 0\}$, which is the region of biological significance, is invariant under Φ^t. Note also that

$$V(x, 1) = \begin{pmatrix} ax(1-x) - cx \\ -\gamma x \end{pmatrix}, \quad V(1, y) = \begin{pmatrix} -cy \\ \alpha y(1-y) - \gamma y \end{pmatrix}, \tag{14.4}$$

so Φ^t leaves invariant the region

$$\mathcal{B} = \{(x, y) : 0 \le x, y \le 1\}, \tag{14.5}$$

for $t \ge 0$.

The vector field V has the following critical points,

$$(0,0), \quad (0,1), \quad (1,0), \tag{14.6}$$

and a fourth critical point (x_0, y_0), satisfying

$$cy_0 = a(1 - x_0), \quad \gamma x_0 = \alpha(1 - y_0). \tag{14.7}$$

A calculation gives

$$x_0 = \alpha \frac{a - c}{a\alpha - c\gamma}, \quad y_0 = a \frac{\alpha - \gamma}{a\alpha - c\gamma}. \tag{14.8}$$

The point (x_0, y_0) may or may not lie in the first quadrant. We investigate this further below.

We have

$$DV(0,0) = \begin{pmatrix} a & 0 \\ 0 & \alpha \end{pmatrix}, \tag{14.9}$$

so $(0,0)$ is a source. Also,

$$DV(0,1) = \begin{pmatrix} a - c & 0 \\ -\gamma & -\alpha \end{pmatrix}, \quad DV(1,0) = \begin{pmatrix} -a & -c \\ 0 & \alpha - \gamma \end{pmatrix}, \tag{14.10}$$

and each of these might be a saddle or a sink, depending on the signs of $a - c$ and $\alpha - \gamma$. Next,

$$\begin{aligned} DV(x_0, y_0) &= \begin{pmatrix} a(1 - 2x_0) - cy_0 & -cx_0 \\ -\gamma y_0 & \alpha(1 - 2y_0) - \gamma x_0 \end{pmatrix} \\ &= \begin{pmatrix} -ax_0 & -cx_0 \\ -\gamma y_0 & -\alpha y_0 \end{pmatrix}, \end{aligned} \tag{14.11}$$

the second identity by (14.7). Hence

$$\begin{aligned} \det DV(x_0, y_0) &= (a\alpha - c\gamma) x_0 y_0, \\ \operatorname{Tr} DV(x_0, y_0) &= -ax_0 - \alpha y_0. \end{aligned} \tag{14.12}$$

At this point, it is natural to consider the following cases:

Figure 14.1

Case I. $a > c$ and $\alpha > \gamma$.
Case II. $a < c$ and $\alpha < \gamma$.
Case III. $a > c$ and $\alpha < \gamma$.
Case IV. $a < c$ and $\alpha > \gamma$.

In Case I, we see from (14.10) that

$$(0,1) \text{ and } (1,0) \text{ are saddles.} \tag{14.13}$$

In this case, $a\alpha > c\gamma$, so, by (14.8),

$$x_0 > 0, \quad y_0 > 0, \tag{14.14}$$

and the critical point (x_0, y_0) is in the first quadrant. Then we see from (14.12) that

$$\det DV(x_0, y_0) > 0, \quad \operatorname{Tr} DV(x_0, y_0) < 0, \tag{14.15}$$

so

$$(x_0, y_0) \text{ is a sink.} \tag{14.16}$$

We have

$$\Phi^t(x, y) \longrightarrow (x_0, y_0) \text{ as } t \to +\infty, \tag{14.17}$$

whenever $x > 0$ and $y > 0$. The two competing species tend to an equilibrium of coexistence. The phase portrait for this case, with $a = 2, \alpha = 2, c = 1, \gamma = 1$, is illustrated in Fig. 14.1.

Figure 14.2

In Case II, we see from (14.10) that

$$(0,1) \ \text{ and } \ (1,0) \ \text{ are sinks.} \tag{14.18}$$

In this case, $a\alpha < c\gamma$, so, by (14.8), again (14.14) holds, and the critical point (x_0, y_0) is in the first quadrant. We see from (14.12) that

$$\det DV(x_0, y_0) < 0, \tag{14.19}$$

so

$$(x_0, y_0) \ \text{ is a saddle.} \tag{14.20}$$

The phase portrait for this case, with $a = 1, \alpha = 1, c = 2, \gamma = 2$, is illustrated in Fig. 14.2. For almost all initial data (x, y) in the first quadrant, $\Phi^t(x, y)$ tends to either $(0, 1)$ or $(1, 0)$ as $t \to +\infty$. One species or the other tends toward extinction, depending on the initial conditions.

In Case III, we see from (14.10) that

$$(0,1) \ \text{ is a saddle and } \ (1,0) \ \text{ is a sink.} \tag{14.21}$$

From here two sub-cases arise, depending on the relative size of $a\alpha$ and $c\gamma$.

Case IIIA. $a\alpha > c\gamma$.
This time, by (14.8),

$$x_0 > 0, \quad y_0 < 0, \tag{14.22}$$

Figure 14.3

so the critical point (x_0, y_0) is not in the first quadrant. We see from (14.12) that

$$\det DV(x_0, y_0) < 0, \tag{14.23}$$

so

$$(x_0, y_0) \ \text{ is a saddle.} \tag{14.24}$$

The phase portrait for this case, with $a = 2, \alpha = 1, c = 1/4, \gamma = 2$, is illustrated in Fig. 14.3. We have

$$\Phi^t(x, y) \longrightarrow (1, 0) \ \text{ as } \ t \to +\infty, \tag{14.25}$$

whenever $x > 0$ and $y > 0$. Species y tends to extinction.

Case IIIB. $a\alpha < c\gamma$.
This time, by (14.8),

$$x_0 < 0, \quad y_0 > 0, \tag{14.26}$$

and again the critical point (x_0, y_0) is not in the first quadrant. We see from (14.22) that

$$\det DV(x_0, y_0) > 0. \tag{14.27}$$

Thus

$$(x_0, y_0) \ \text{ is a source or a sink,} \tag{14.28}$$

Figure 14.4

depending on the sign of $\operatorname{Tr} DV(x_0, y_0)$. The phase portrait for this case, with $a = 2, \alpha = 1/2, c = 1, \gamma = 2$, is illustrated in Fig. 14.4. (In this example, (x_0, y_0) is a sink.) Again (14.25) holds whenever $x > 0$ and $y > 0$.

To summarize Case III, the flows in the first quadrant have the same qualitative features in the two sub-cases; (14.25) holds. The features differ outside the first quadrant.

As for Case IV, this reduces to Case III by switching the roles of x and y.

Exercises

1. Note that if x and y solve (14.2), then

$$\frac{d}{dt}(x + y) = -ax^2 - \alpha y^2 - (c + \gamma)xy + ax + \alpha y.$$

Show that there exists $R \in (0, \infty)$ such that

$$x, y \geq 0,\ x^2 + y^2 \geq R^2 \Longrightarrow \frac{d}{dt}(x + y) \leq 0.$$

Deduce global existence of solutions to (14.2), for $t \geq 0$, given $(x(0), y(0))$ in the first quadrant.

2. In the setting of Exercise 1, show that whenever $x(0) > 0$ and $y(0) > 0$, we have $(x(t), y(t)) \in \mathcal{B}$, given by (14.5), for $t > 0$ sufficiently large.

3. Consider the system

$$\begin{aligned} \frac{dx}{dt} &= x(1-x) - xy, \\ \frac{dy}{dt} &= y(1-y) - \gamma xy, \end{aligned}$$

   with $\gamma \in (0, \infty)$. Specify when Cases I–IV hold. Record the possible outcomes, as regards coexistence/extinction.

4. Consider the system

$$\begin{aligned} \frac{dx}{dt} &= \frac{1}{2}x(1-x) - cxy, \\ \frac{dy}{dt} &= y(1-y) - 2xy, \end{aligned}$$

   with $c \in (0, \infty)$. Specify when Cases I–IV hold. Record the possible outcomes, as regards coexistence/extinction.

5. Consider the system

$$\begin{aligned} \frac{dx}{dt} &= ax(1-x) - xy, \\ \frac{dy}{dt} &= 2y(1-y) - xy, \end{aligned}$$

   with $a \in (0, \infty)$. Specify when Cases I–IV hold. Record the possible outcomes, as regards coexistence/extinction.

15. Chaos in multidimensional systems

As previewed in the introduction to this chapter, two phenomena conspire to limit the complexity of flows generated by autonomous planar vector fields. One is that orbits cannot cross each other, due to uniqueness (this holds in any number of dimensions). The other is that a directed curve (with nonzero velocity) in the plane divides a neighborhood of each of its points into two parts, the left and the right. This latter fact played an important role in §12. In dimension 3 and higher, this breaks down completely, and allows for far more complex flows.

Newtonian motion in a force field in the plane is described by a second order 2×2 system of differential equations, which is converted to a 4×4 first order system. Energy conservation confines the motion to a 3-dimensional constant energy surface. If the force is a central force, there is also conservation of angular momentum. These two conservation laws make for regular motion, as seen in §§5–6. These are "integrable" systems. Such integrability is special. Most systems from physics and other sources do not possess it. For example, the double pendulum equation, derived in §9, does not have this property. (We do not prove this here.)

Flows generated by vector fields on n-dimensional domains with $n \geq 3$ are thus sometimes regular, but often they lack regularity to such a degree that they are deemed "chaotic." Signatures of chaos include the inability to predict the long time behavior of orbits. This inability arises not only from the lack of a formula for the solution in terms of elementary functions. In addition, numerical approximations to the orbits of these flows reveal a "sensitive dependence" on initial conditions and other parameters. Furthermore, phase portraits of these orbits *look* complex.

Research into these chaotic flows takes the study of differential equations to the next level, beyond this introduction. We end this chapter with a discussion of two special cases of 3×3 systems, to give a flavor of the complexities that lie beyond, and we provide pointers to literature that addresses the deep questions raised by efforts to understand such systems.

Lorenz equations

The first example is the following system, produced by E. Lorenz in 1963 to model some aspects of fluid turbulence:

$$\begin{aligned} x' &= \sigma(y - x), \\ y' &= rx - y - xz, \\ z' &= xy - bz. \end{aligned} \tag{15.1}$$

An alternative presentation is

$$\frac{d}{dt}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -\sigma & \sigma & 0 \\ r & -1 & 0 \\ 0 & 0 & -b \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -x \\ 0 & x & 0 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix}. \tag{15.2}$$

Denoting the right side of (15.2) by $V(x, y, z)$, we see that the first matrix on the right side is $DV(0, 0, 0)$. One assumes the parameters σ, b, and r are all positive. Lorenz took

$$\sigma = 10, \quad b = \frac{8}{3}, \tag{15.3}$$

and considered various values of r, with emphasis on

$$r = 28. \tag{15.4}$$

Phase portraits of some orbits for (15.1), with σ and b given by (15.3) and with various values of r are given in Fig. 15.1. Each of the six portraits depicts the forward orbits through the three points

$$x = \frac{k}{100}, \quad y = 0, \quad z = 5, \quad k = -1, 0, 1. \tag{15.5}$$

The portraits start out simple, execute a sequence of changes, as r increases, reaching substantial apparent complexity at $r = 28$. We discuss some aspects of this.

First, some global results. Global forward solvability of (15.1) can be established with the help of the remarkable function

$$f(x, y, z) = rx^2 + \sigma y^2 + \sigma(z - 2r)^2. \tag{15.6}$$

A calculation shows that if $(x(t), y(t), z(t))$ solves (15.1), then

$$\frac{d}{dt} f(x, y, z) = -2\sigma(rx^2 + y^2 + bz^2 - 2brz). \tag{15.7}$$

Clearly there exists $K \in (0, \infty)$ such that

$$B = \{(x, y, z) \in \mathbb{R}^3 : f(x, y, z) \le K\} \tag{15.8}$$

is a closed, bounded subset of $\mathbb{R}^3$ and the right side of (15.7) is < 0 on the complement of B. Hence

$$\Phi^t(B) \subset B, \quad \forall\, t > 0, \tag{15.9}$$

where Φ^t is the flow generated by $V(x, y, z)$. Moreover, for each $(x, y, z) \in \mathbb{R}^3$,

$$\Phi^t(x, y, z) \in B, \quad \text{for all sufficiently large } \ t > 0. \tag{15.10}$$

Note that (15.9) plus the identity $\Phi^t = \Phi^s \circ \Phi^{t-s}$ implies

$$\Phi^t(B) \subset \Phi^s(B) \quad \text{for } \ 0 < s < t, \tag{15.11}$$

so $B(t) = \Phi^t(B)$ is a family of closed, bounded sets that is decreasing as $t \nearrow +\infty$. Now set

$$\mathcal{B} = \bigcap_{t \in \mathbb{R}^+} B(t) = \bigcap_{k \in \mathbb{Z}^+} B(k). \tag{15.12}$$

Figure 15.1

The set $\mathcal{B}$ is called the *attractor* for (15.1). We have

$$\Phi^t(\mathcal{B}) = \mathcal{B}, \quad \forall\, t \ge 0. \tag{15.13}$$

Note that

$$\operatorname{div} V = -\sigma - 1 - b < 0, \tag{15.14}$$

so results of §3 imply

$$\operatorname{Vol} \mathcal{B} = 0. \tag{15.15}$$

This attractor has a simple description for small r, but becomes very complex for larger r.

To proceed with the analysis, consider the critical points. The origin is a critical point of V for all $\sigma, b, r \in (0,\infty)$. Since $DV(0)$ is the first matrix on the right side of (15.2), we see its eigenvalues are

$$\lambda_\pm = -\frac{\sigma+1}{2} \pm \frac{1}{2}\sqrt{(\sigma+1)^2 + 4\sigma(r-1)}, \quad \lambda_3 = -b, \tag{15.16}$$

with eigenvectors

$$v_\pm = \begin{pmatrix} \sigma \\ \lambda_\pm + \sigma \\ 0 \end{pmatrix}, \quad v_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{15.17}$$

It follows from (15.16) that

(15.18)

$$\begin{aligned} 0 < r < 1 &\Longrightarrow DV(0) \ \text{ has 3 negative eigenvalues,} \\ r > 1 &\Longrightarrow DV(0) \ \text{ has 2 negative and one positive eigenvalue.} \end{aligned}$$

For $r > 1$, the positive eigenvalue is λ_+ and its associated eigenvector is v_+. There is a parallel to the results in (3.31) describing saddles. It is shown in [**Hart**] that there is a smooth 2-dimensional surface through the origin consisting of points p such that $\Phi^t(p) \to 0$ as $t \to +\infty$ and a smooth 1-dimensional curve through the origin consisting of points p such that $\Phi^t(p) \to 0$ as $t \to -\infty$. In general, a smooth k-dimensional surface in $\mathbb{R}^n$ is called a k-dimensional manifold. The sets described above are called a "stable manifold" and an "unstable manifold," respectively. See also Appendic C for further discussion.

For $r > 1$, V has two additional critical points, satisfying

$$x = y, \quad (r - 1 - z)x = 0, \quad bz = x^2, \tag{15.19}$$

i.e.,

$$C_\pm = (\pm\sqrt{b(r-1)}, \pm\sqrt{b(r-1)}, r-1). \tag{15.20}$$

We have

$$DV(C_\pm) = \begin{pmatrix} -\sigma & \sigma & 0 \\ 1 & -1 & \pm\xi \\ \pm\xi & \pm\xi & -b \end{pmatrix}, \quad \xi = \sqrt{b(r-1)}. \tag{15.21}$$

Note that $DV(C_+)$ and $DV(C_-)$ are conjugate by the action of

$$\begin{pmatrix} -1 & & \\ & -1 & \\ & & 1 \end{pmatrix}, \tag{15.22}$$

so they have the same eigenvalues. This mirrors the fact that (15.1) is invariant under the transformation $(x, y, z) \mapsto (-x, -y, z)$. Further calculations give the following results, when σ and b are given by (15.3):
(15.23)

$DV(C_\pm)$ has

3 negative eigenvalues for $1 < r < 1.346\cdots$

1 negative and 2 with negative real part for $1.346\cdots < r < 24.74\cdots$

1 negative and 2 with positive real part for $r > 24.74\cdots$.

In the first two cases in (15.23), Proposition 3.4 applies, and for all points p sufficiently close to C_+, $\Phi^t(p) \to C_+$ as $t \to +\infty$, and similarly for C_-. The third case in (15.23) is like the second case in (15.18), except the numbers are reversed. In such a case, there are a 2-dimensional unstable manifold and a 1-dimensional stable manifold through C_+, and similarly for C_-, in the language introduced below (15.18).

With these calculations in hand, let's take a closer look at the six phase portraits depicted in Fig. 15.1, orbits with initial data given by (15.5). In all cases there is a vertical line segment from $(0, 0, 5)$ to $(0, 0, 0)$, and we see from (15.1) that the z-axis is invariant under the flow for all values of the parameters. Furthermore, on the z-axis, $z' = -bz$. Now the initial points $(\pm 0.01, 0, 5)$ are close by, but for all r-values depicted, $DV(0)$ has one positive eigenvalue, and the orbits push away from the origin, in a direction close to $\pm v_+$, where v_+ is given by (15.17). The orbit from $(+0.01, 0, 5)$ spirals into the critical point C_+, and the orbit from $(-0.01, 0, 5)$ spirals into C_-, in the first two portraits, where $r = 4.667$ and 9.333. Around $r \approx 14$, something new happens. These orbits pass close to the origin.

Fig. 15.2 shows such a transition in more detail. Here we have two orbits with slightly different initial conditions, namely

$$p_\pm = \pm \varepsilon v_+, \tag{15.24}$$

with ε chosen small, to capture the unstable manifold of the origin more accurately. At a certain critical value $r_h \approx 13.926$, the unstable manifold is actually a pair of homoclinic orbits, approaching the origin both as $t \to -\infty$ and as $t \to +\infty$. For larger values of r, the orbit from p_+ (and that from $(+0.01, 0, 5)$) crosses over and spirals into C_-, while the orbit from p_- (and that from $(-0.01, 0, 5)$) spirals into C_+, as depicted in the fourth portrait in Fig. 15.1.

This spiraling into $C_\pm$ does not endure as r increases. As stated in (15.23), there is a critical $r_c \approx 24.74$ past which $DV(C_\pm)$ has two eigenvalues with positive rather than negative real part. At $r = 23.333$, this spiraling has slowed. In fact, the fifth portrait in Fig. 15.1 does not reveal spiraling all the way in. The orbits pictured there are of the form $\Phi^t(p_j)$ for $t \in [0, 40]$. If t is taken somewhat larger, one sees asymptotic approaches to $C_\pm$, with $r = 23.333$.

In the sixth phase portrait of Fig. 15.1, we have $r = 28 > r_c$. The orbits approach the unstable manifolds of $C_\pm$ and then spiral out from these critical points. After some spiraling out from C_-, the orbit starting from $(+0.01, 0, 5)$ makes a jump to the vicinity of C_+, approaches its unstable manifold, and starts spiraling out from C_+. After a while, the orbit jumps back to the vicinity of C_-, and this spiraling and jumping is endlessly repeated. The six phase portraits in Fig. 15.3 show

$$\Phi^t(0.01, 0, 5), \quad 80j < t < 80(j + 1), \quad 0 \le j \le 5. \tag{15.25}$$

The portraits differ in fine detail from each other, but they are fairly similar, and seem to reveal what is called a strange attractor.

Figures 15.1–3 were produced by numerically integrating (15.1), using a fourth-order Runge-Kutta scheme, described in §11, with step size

$$h = 0.0005. \tag{15.26}$$

Use of the step size $h = 0.001$ produced apparently identical phase portraits in Fig. 15.2, and in all but the last portrait in Fig. 15.1. There were noticeable differences in the last phase portrait of Fig. 15.1 and in the portraits of Fig. 15.3. This phenomenon gives evidence of sensitive dependence of the orbits on initial conditions, and leads to unpredictibility of orbits, which is part of the signature of chaos.

Figure 15.2

Figure 15.3

We make one further comment about Figs. 15.1–15.3. Of course, the orbits depicted are curves $(x(t), y(t), z(t))$ in $\mathbb{R}^3$. What is shown in these figures are 2-dimensional projections, namely $(u(t), v(t))$, with $u(t) = x(t) + y(t)/2,\ v(t) = z(t) - y(t)/2$.

Periodically forced Duffing equation

Our second example arises from motion in 1 dimension, in a nonlinear background field, with a periodic forcing term added:

$$\frac{d^2x}{dt^2} = f(x) + r\cos t. \tag{15.27}$$

Here r is a parameter. When converted to a first order system and put in autonomous form, this becomes

$$\begin{aligned}\frac{dx}{dt} &= y,\\ \frac{dy}{dt} &= f(x) + r\cos z,\\ \frac{dz}{dt} &= 1.\end{aligned} \tag{15.28}$$

We take

$$f(x) = x - x^3. \tag{15.29}$$

Then (15.27) is called a periodically forced Duffing equation if $r \neq 0$. For $r = 0$ it is called Duffing's equation, and it reduces to a 2×2 system, whose phase portrait is given in Fig. 15.4. There are two homoclinic orbits, each tending to the origin as $t \to \pm\infty$. All the other orbits are closed, and lie on level curves of

$$E(x, y) = \frac{y^2}{2} - \frac{x^2}{2} + \frac{x^4}{4}. \tag{15.30}$$

Fig. 15.5 shows six individual orbits of (15.28) with $r = 0$, orbits through the six points

$$x = \sqrt{2} + \frac{3k}{10}, \quad -4 \le k \le 1, \quad y = 0. \tag{15.31}$$

The orbits for (15.28) are curves $(x(t), y(t), z(t))$, but for $r = 0$ we simply plot $(x(t), y(t))$ in Fig. 15.5.

For $r \neq 0$, matters are more complicated, since z is coupled to (x, y) in (15.28). We need a different way to portray the orbits $(x(t), y(t), z(t))$. In

Figure 15.4

this case, unlike for the Lorenz system, a linear projection of (x, y, z) space onto (u, v) space is not the best way to proceed. Taking into account the periodicity of the right side of (15.28) in z, we treat $z = t$ as an angular variable, and transfer (x, y, z) space to $(\tilde{x}, \tilde{y}, \tilde{z})$ space, with

$$\tilde{x} = (x + 2)\cos t, \quad \tilde{y} = y, \quad \tilde{z} = (x + 2)\sin t.$$

This corresponds to taking the (x, y) plane pictured in Fig. 15.5 and rotating it about the vertical axis $x = -2$. We follow this with the linear map to the (u, v) plane, $u = \tilde{z} - \tilde{x}/2$, $v = \tilde{y} - \tilde{x}/2$. Consequently, to produce Figs. 15.6–15.8, we draw curves $(u(t), v(t))$, with

$$(15.32) \qquad u(t) = (x(t) + 2)\Bigl(\sin t - \frac{\cos t}{2}\Bigr), \quad v(t) = y(t) - (x(t) + 2)\frac{\cos t}{2}.$$

For initial data, we take x and y as in (15.31) and $z = 0$. We use a fourth order Runge-Kutta scheme.

Fig. 15.6 draws such curves when (x, y, z) solve (15.28) with $r = 0$. Note that in all but the fifth portrait, the orbits lie on smooth donut-shaped surfaces (called tori). The fifth portrait depicts the homoclinic orbit, which spends most of its time near the origin in (x, y)-space. It lies on a surface that is smooth except along a curve, where it has a corner.

Figure 15.7 gives this representation of orbits of (15.28), with

$$(15.33) \qquad r = 0.1.$$

Two of the six orbits seem to lie on smooth tori (one very thin, the other somewhat deformed). The other four are all apparently a mess, and also,

Figure 15.5

Figure 15.6

Figure 15.7

Figure 15.8

Figure 15.9

apparently, about the same mess. In Fig. 15.8 we present such orbits with initial data

$$x = \sqrt{2} + \frac{k}{20}, \quad 0 \le k \le 5, \quad y = 0, \tag{15.34}$$

interpolating from the fifth orbit of Fig. 15.7 halfway to the sixth orbit. Here the first three orbits appear chaotic and the last three appear to lie on smooth surfaces.

An alternative to depicting orbits of the system (15.28)–(15.29) is to depict orbits of the associated *Poincaré map*, characterized as follows. Take an initial point $p = (x_0, y_0, 0)$. Solve (15.28) with this initial data, and then set $q = (x(2\pi), y(2\pi), 2\pi)$. The nature of the mapping on the third coordinate is trivial in this case, so we just consider

$$(x(0), y(0)) \mapsto (x(2\pi), y(2\pi)). \tag{15.35}$$

This is the Poincaré map associated to the system (15.28).

The Poincaré map is defined in a more general context. Let X be a smooth vector field on $\Omega \subset \mathbb{R}^n$ and let S be an $(n-1)$-dimensional surface transversal to X, i.e., X is nowhere tangent to S. Under certain circumstances, one has a Poincaré map

$$P : \mathcal{O} \longrightarrow S, \tag{15.36}$$

defined on an open subset $\mathcal{O} \subset S$, where $p \in \mathcal{O}$ and $P(p) = q$ is the point $\Phi_X^t(p)$ with smallest $t > 0$ such that $\Phi_X^t(p) \in S$. See Fig. 15.9.

In the setting of (15.28), (15.35), the orbit for Poincaré map $(x(0), y(0)) = (x, 0)$, with x as in (15.34), is presented in Fig. 15.10, which can be appreciated in light of Fig. 15.8. Each picture in Fig. 15.10 is made of 9000 points in the orbit of the Poincaré map (or rather an approximation via a Runge-Kutta difference scheme). The first three pictures seem to show orbits spread out in a 2-dimensional region, while the last three seem to show orbits lying on smooth curves.

Going further, each of the last three pictures in Fig. 15.10 suggest the following:

Assertion. There is a region $\overline{\Omega} \subset \mathbb{R}^2$, smoothly equivalent to the disk

$$\overline{D} = \{x \in \mathbb{R}^2 : \|x\| \le 1\}, \tag{15.37}$$

that is to say, there is a smooth one-to-one map $\varphi : \overline{\Omega} \to \overline{D}$ with smooth inverse $\varphi^{-1} : \overline{D} \to \overline{\Omega}$, and the Poincaré map takes $\overline{\Omega}$ into itself, i.e.,

$$P : \overline{\Omega} \longrightarrow \overline{\Omega}. \tag{15.38}$$

Granted this, we can make use of the following result, known as Brouwer's fixed-point theorem.

Theorem. *Each smooth map*

$$\psi : \overline{D} \longrightarrow \overline{D} \tag{15.39}$$

has a fixed point, i.e., there exists $p \in \overline{D}$ such that $\psi(p) = p$.

See Appendix F for a proof of this result. Given the assertion above, we can take $\psi = \varphi \circ P \circ \varphi^{-1}$ and conclude that

$$P(q) = q, \quad q = \varphi^{-1}(p). \tag{15.40}$$

Such fixed points of the Poincaré map give rise to periodic solutions to the associated systems of differential equations (in this case, (15.27)). Establishing the existence of periodic solutions is one of many uses for Poincaré maps. We refer to references cited in the next paragraph for discussions of other uses.

Understanding how the chaotic looking orbits for the Lorenz and Duffing systems and other systems *are* chaotic has engendered a lot of work. For more material on this, we particularly recommend the Introduction to Chaos in Chapter 2 of [**GH**], which treats four examples, including the Lorenz system and the forced Duffing system. Other material on chaotic systems can be found in [**AS**], [**AP**], [**HK**], [**HSD**], [**J**], and [**LL**]. A detailed study of the Lorenz system is given in [**Sp**].

Figure 15.10

Exercises

1. Consider the double pendulum system, in the limit $m_2 = 0$, given by (9.35)–(9.36). Substitute

$$\theta_1(t) \approx r\cos\omega t, \quad \omega = \sqrt{\frac{g}{\ell_1}} \tag{15.41}$$

into (9.36), expand in powers of r, and throw away terms containing second and higher powers of r. Show that you get

$$\theta_2''(t) + \frac{g}{\ell_2}\sin\theta_2(t) = r\omega^2\frac{\ell_1}{\ell_2}\cos\theta_2(t)\,\cos\omega t. \tag{15.42}$$

Exercises 2–9 are for readers who can use a programming language, with graphics capabilities.

2. Write a program to exhibit solution curves of (15.42), in a fashion analogous to the treatment of (15.27), involving an analogue of (15.28). Try various values of $r, g/\ell_2$, etc., and see when the behavior is more chaotic or less chaotic.

3. Write a program to exhibit solutions to the full double pendulum system (9.15)–(9.16). Take, e.g., $m_1 = m_2 = 1, \ell_1 = \ell_2 = 1$, and variants.

4. Examine orbits and Poincaré maps for the periodically forced Duffing equation for other values of r, such as $r = 0.2,\ 0.05,\ 10^{-2},\ 10^{-3}$, etc.

Exercises 5–8 deal with systems of the form

$$\frac{d^2}{dt^2}\begin{pmatrix} x \\ y \end{pmatrix} = -\nabla V(x, y). \tag{15.43}$$

These are 2×2 second order systems, which convert to 4×4 first order systems. Energy conservation leads to flows on 3-dimensional constant energy surfaces. In each case, write a program to exhibit solution curves $(x(t), y(t))$. See whether the displayed solutions seem to be regular or chaotic.

5. Take

$$V(x, y) = x^2 + axy + y^4.$$

Try various $a \in [0, 10]$.

6. Take
$$V(x,y) = x^4 + axy + y^4, \quad a \in [-2, 2].$$

7. Take
$$V(x,y) = x^4 + ax^2y + y^4, \quad a \in [-1, 1].$$

8. Take
$$V(x,y) = \frac{1}{2}(x^2 + y^2) + a(x^4 - x^2y + y^4), \quad a \in [0, 1].$$

9. Taking off from models in §§13–14, see if you can construct models of interactions of 3 species that exhibit chaotic behavior.

A. The derivative in several variables

To start this section off, we define the derivative and discuss some of its basic properties. Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$, and $F : \mathcal{O} \to \mathbb{R}^m$ a continuous function. We say F is differentiable at a point $x \in \mathcal{O}$, with derivative L, if $L : \mathbb{R}^n \to \mathbb{R}^m$ is a linear transformation such that, for $y \in \mathbb{R}^n$, small,

$$F(x+y) = F(x) + Ly + R(x,y) \tag{A.1}$$

with $\|R(x,y)\| = o(\|y\|)$, i.e.,

$$\frac{\|R(x,y)\|}{\|y\|} \to 0 \quad \text{as} \quad y \to 0. \tag{A.2}$$

We denote the derivative at x by $DF(x) = L$. With respect to the standard bases of $\mathbb{R}^n$ and $\mathbb{R}^m$, $DF(x)$ is simply the matrix of partial derivatives,

$$DF(x) = \left(\frac{\partial F_j}{\partial x_k}\right), \tag{A.3}$$

so that, if $v = (v_1, \dots, v_n)^t$, (regarded as a column vector) then

$$DF(x)v = \Bigl(\sum_k \frac{\partial F_1}{\partial x_k} v_k, \dots, \sum_k \frac{\partial F_m}{\partial x_k} v_k\Bigr)^t. \tag{A.4}$$

It will be shown below that F is differentiable whenever all the partial derivatives exist and are *continuous* on $\mathcal{O}$. In such a case we say F is a C^1 function on $\mathcal{O}$. More generally, F is said to be C^k if all its partial derivatives of order $\le k$ exist and are continuous. If F is C^k for all k, we say F is C^∞.

Sometimes one might want to differentiate an $\mathbb{R}^m$-valued function $F(x,t)$ only with respect to x. In that case, if

$$F(x+y,t) = F(x,t) + Ly + R(x,y,t),$$

with $\|R(x,y,t)\| = o(\|y\|)$, we write $D_x F(x,t) = L$.

We now derive the *chain rule* for the derivative. Let $F : \mathcal{O} \to \mathbb{R}^m$ be differentiable at $x \in \mathcal{O}$, as above, let U be a neighborhood of $z = F(x)$ in $\mathbb{R}^m$, and let $G : U \to \mathbb{R}^k$ be differentiable at z. Consider $H = G \circ F$. We have

$$\begin{aligned} H(x+y) &= G(F(x+y)) \\ &= G\big(F(x) + DF(x)y + R(x,y)\big) \\ &= G(z) + DG(z)\big(DF(x)y + R(x,y)\big) + R_1(x,y) \\ &= G(z) + DG(z)DF(x)y + R_2(x,y) \end{aligned} \tag{A.5}$$

with

$$\frac{\|R_2(x,y)\|}{\|y\|} \to 0 \quad \text{as} \quad y \to 0.$$

Thus $G \circ F$ is differentiable at x, and

$$D(G \circ F)(x) = DG(F(x)) \cdot DF(x). \tag{A.6}$$

In case $k = 1$, so $G : U \to \mathbb{R}$, we can rewrite (A.6) as

$$D(G \circ F)(x) = \nabla G(F(x))^t DF(x), \tag{A.7}$$

where $\nabla G(y)^t = (\partial G/\partial y_1, \dots, \partial G/\partial y_m)$. If in addition $n = 1$, so F is a function of one variable $x \in \mathcal{O} \subset \mathbb{R}$, with values in $\mathbb{R}^m$, this in turn leads to

$$\frac{d}{dx} G(F(x)) = \nabla G(F(x)) \cdot F'(x). \tag{A.8}$$

This leads to such formulas as (3.10).

Another useful remark is that, by the Fundamental Theorem of Calculus, applied to $\varphi(t) = F(x+ty)$,

$$F(x+y) = F(x) + \int_0^1 DF(x+ty)y\,dt, \tag{A.9}$$

provided DF is continuous. A closely related application of the Fundamental Theorem of Calculus is that, if we assume $F : \mathcal{O} \to \mathbb{R}^m$ is differentiable in each variable separately, and that each $\partial F/\partial x_j$ is continuous on $\mathcal{O}$, then (A.10)

$$F(x+y) = F(x) + \sum_{j=1}^{n} \bigl[F(x+z_j) - F(x+z_{j-1})\bigr] = F(x) + \sum_{j=1}^{n} A_j(x,y) y_j,$$

$$A_j(x,y) = \int_0^1 \frac{\partial F}{\partial x_j}\bigl(x + z_{j-1} + t y_j e_j\bigr)\, dt,$$

where $z_0 = 0,\ z_j = (y_1, \dots, y_j, 0, \dots, 0)$, and $\{e_j\}$ is the standard basis of $\mathbb{R}^n$. Now (A.10) implies F is differentiable on $\mathcal{O}$, as we stated below (A.4). Thus we have established the following.

Proposition A.1. *If $\mathcal{O}$ is an open subset of $\mathbb{R}^n$ and $F : \mathcal{O} \to \mathbb{R}^m$ is of class C^1, then F is differentiable at each point $x \in \mathcal{O}$.*

For the study of higher order derivatives of a function, the following result is fundamental.

Proposition A.2. *Assume $F : \mathcal{O} \to \mathbb{R}^m$ is of class C^2, with $\mathcal{O}$ open in $\mathbb{R}^n$. Then, for each $x \in \mathcal{O},\ 1 \le j, k \le n$,*

$$\frac{\partial}{\partial x_j}\frac{\partial F}{\partial x_k}(x) = \frac{\partial}{\partial x_k}\frac{\partial F}{\partial x_j}(x). \tag{A.11}$$

To prove Proposition A.2, it suffices to treat real valued functions, so consider $f : \mathcal{O} \to \mathbb{R}$. For $1 \le j \le n$, we set

$$\Delta_{j,h} f(x) = \frac{1}{h}\bigl(f(x+he_j) - f(x)\bigr), \tag{A.12}$$

where $\{e_1, \dots, e_n\}$ is the standard basis of $\mathbb{R}^n$. The mean value theorem (for functions of x_j alone) implies that if $\partial_j f = \partial f/\partial x_j$ exists on $\mathcal{O}$, then, for $x \in \mathcal{O},\ h > 0$ sufficiently small,

$$\Delta_{j,h} f(x) = \partial_j f(x + \alpha_j h e_j), \tag{A.13}$$

for some $\alpha_j \in (0,1)$, depending on x and h. Iterating this, if $\partial_j(\partial_k f)$ exists on $\mathcal{O}$, then, for $x \in \mathcal{O},\ h > 0$ sufficiently small,

$$\begin{aligned}\Delta_{k,h}\Delta_{j,h} f(x) &= \partial_k(\Delta_{j,h} f)(x + \alpha_k h e_k) \\ &= \Delta_{j,h}(\partial_k f)(x + \alpha_k h e_k) \\ &= \partial_j \partial_k f(x + \alpha_k h e_k + \alpha_j h e_j),\end{aligned} \tag{A.14}$$

with $\alpha_j, \alpha_k \in (0,1)$. Here we have used the elementary result

$$\partial_k \Delta_{j,h} f = \Delta_{j,h}(\partial_k f). \tag{A.15}$$

We deduce the following.

Proposition A.3. *If $\partial_k f$ and $\partial_j \partial_k f$ exist on $\mathcal{O}$ and $\partial_j \partial_k f$ is continuous at $x_0 \in \mathcal{O}$, then*

$$\partial_j \partial_k f(x_0) = \lim_{h \to 0} \Delta_{k,h} \Delta_{j,h} f(x_0). \tag{A.16}$$

Clearly

$$\Delta_{k,h} \Delta_{j,h} f = \Delta_{j,h} \Delta_{k,h} f, \tag{A.17}$$

so we have the following, which easily implies Proposition A.2.

Corollary A.4. *In the setting of Proposition A.3, if also $\partial_j f$ and $\partial_k \partial_j f$ exist on $\mathcal{O}$ and $\partial_k \partial_j f$ is continuous at x_0, then*

$$\partial_j \partial_k f(x_0) = \partial_k \partial_j f(x_0). \tag{A.18}$$

If U and V are open subsets of $\mathbb{R}^n$ and $F : U \to V$ is a C^1 map, we say F is a diffeomorphism of U onto V provided F maps U one-to-one and onto V, and its inverse $G = F^{-1}$ is a C^1 map. If F is a diffeomorphism, it follows from the chain rule that $DF(x)$ is invertible for each $x \in U$. We now state a partial converse of this, the Inverse Function Theorem, which is a fundamental result in multivariable calculus.

Theorem A.5. *Let F be a C^k map from an open neighborhood Ω of $p_0 \in \mathbb{R}^n$ to $\mathbb{R}^n$, with $q_0 = F(p_0)$. Assume $k \geq 1$. Suppose the derivative $DF(p_0)$ is invertible. Then there is a neighborhood U of p_0 and a neighborhood V of q_0 such that $F : U \to V$ is one-to-one and onto, and $F^{-1} : V \to U$ is a C^k map. (So $F : U \to V$ is a diffeomorphism.)*

Proofs of Theorem A.5 can be found in a number of texts, including [**LS**] and Chapter 1 of [**T**].

B. Convergence, compactness, and continuity

We discuss a number of notions and results related to convergence in $\mathbb{R}^n$, of use in this chapter. First, a sequence of points (p_j) in $\mathbb{R}^n$ converges to a limit $p \in \mathbb{R}^n$ (we write $p_j \to p$) if and only if

$$\|p_j - p\| \longrightarrow 0. \tag{B.1}$$

Here $\| \cdot \|$ is the norm on $\mathbb{R}^n$ arising in §10 of Chapter 2, and the meaning of (B.1) is that for every $\varepsilon > 0$ there exists N such that

$$j \geq N \Longrightarrow \|p_j - p\| < \varepsilon. \tag{B.2}$$

A set $S \subset \mathbb{R}^n$ is said to be *closed* if and only if

$$p_j \in S,\ p_j \to p \Longrightarrow p \in S. \tag{B.3}$$

The complement $\mathbb{R}^n \setminus S$ of a closed set S is *open*. Alternatively, $\Omega \subset \mathbb{R}^n$ is open if and only if, given $q \in \Omega$, there exists $\varepsilon > 0$ such that $B_\varepsilon(q) \subset \Omega$, where

$$B_\varepsilon(q) = \{p \in \mathbb{R}^n : \|p - q\| < \varepsilon\}, \tag{B.4}$$

so q cannot be a limit of a sequence of points in $\mathbb{R}^n \setminus \Omega$.

An important property of $\mathbb{R}^n$ is *completeness*, a property defined as follows. A sequence (p_j) of points in $\mathbb{R}^n$ is called a Cauchy sequence if and only if

$$\|p_j - p_k\| \longrightarrow 0, \quad \text{as} \ \ j, k \to \infty. \tag{B.5}$$

It is easy to see that if $p_j \to p$ for some $p \in \mathbb{R}^n$, then (B.5) holds. The completeness property is the converse.

Theorem B.1. *If (p_j) is a Cauchy sequence in $\mathbb{R}^n$, then it has a limit, i.e., (B.1) holds for some $p \in \mathbb{R}^n$.*

Since convergence $p_j \to p$ in $\mathbb{R}^n$ is equivalent to convergence in $\mathbb{R}$ of each component, it is the fundamental property of completeness of $\mathbb{R}$ that is the issue. This is discussed in [**BS**], from an axiomatic viewpoint, and in [**Kr**], and also [**T2**], from a more constructive viewpoint.

Completeness provides a path to the following key notion of *compactness*. A set $K \subset \mathbb{R}^n$ is compact if and only if the following property holds.

(B.6) Each infinite sequence (p_j) in K has a subsequence that converges to a point in K.

It is clear that if K is compact, then it must be closed. It must also be bounded, i.e., there exists $R < \infty$ such that $K \subset B_R(0)$. Indeed, if K is not bounded, there exist $p_j \in K$ such that $\|p_{j+1}\| \geq \|p_j\| + 1$. In such a case, $\|p_j - p_k\| \geq 1$ whenever $j \neq k$, so (p_j) cannot have a convergent subsequence. The following converse statement is a key result.

Theorem B.2. *If $K \subset \mathbb{R}^n$ is closed and bounded, then it is compact.*

We start with a special case.

Proposition B.3. *Each closed bounded interval $I = [a, b] \subset \mathbb{R}$ is compact.*

Proof. Let (p_j) be an infinite sequence in $[a,b]$, $j \in \mathbb{Z}^+$. Divide I into two halves, $I_0 = [a,(a+b)/2]$, $I_1 = [(a+b)/2, b]$. If $p_j \in I_0$ for infinitely many j, pick some $p_{j_0} \in I_0$, and set $a_1 = 0$. Otherwise, pick some $p_{j_0} \in I_1$, and set $a_1 = 1$. Set $q_0 = p_{j_0}$.

Now divide I_{a_1} into two equal intervals, $I_{a_1 0}$ and $I_{a_1 1}$. If $p_j \in I_{a_1 0}$ for infinitely many j, pick $p_{j_1} \in I_{a_1 0}$, $j_1 > j_0$. Otherwise, pick $p_{j_1} \in I_{a_1 1}$, $j_1 > j_0$. Set $q_1 = p_{j_1}$. Continue.

One gets (q_j), a subsequence of (p_j), with the property that

$$|q_j - q_{j+k}| \le 2^{-j}|b-a|, \quad \forall\, k \ge 0. \tag{B.7}$$

Thus (q_j) is a Cauchy sequence, so by the completeness of $\mathbb{R}$, it converges, to the desired limit $p \in [a,b]$.

From Proposition B.3 it is easy enough to show that any closed, bounded box

$$\mathcal{B} = \{(x_1,\dots,x_n) \in \mathbb{R}^n : a_j \le x_j \le b_j,\ \forall\, j\} \tag{B.8}$$

is compact. If $K \subset \mathbb{R}^n$ is closed and bounded, it is a subset of such a box, and clearly every closed subset of a compact set is compact, so we have Theorem B.2.

We next discuss continuity. If $S \subset \mathbb{R}^n$, a function

$$f : S \longrightarrow \mathbb{R}^m \tag{B.9}$$

is said to be continuous at $p \in S$ provided

$$p_j \in S,\ p_j \to p \Longrightarrow f(p_j) \to f(p). \tag{B.10}$$

If f is continuous at each $p \in S$, we say f is continuous on S.

The following two results give important connections between continuity and compactness.

Proposition B.4. *If $K \subset \mathbb{R}^n$ is compact and $f : K \to \mathbb{R}^m$ is continuous, then $f(K)$ is compact.*

Proof. If (q_k) is an infinite sequence of points in $f(K)$, pick $p_k \in K$ such that $f(p_k) = q_k$. If K is compact, we have a subsequence $p_{k_\nu} \to p$ in K, and then $q_{k_\nu} \to f(p)$ in $\mathbb{R}^m$.

This leads to the second connection.

Proposition B.5. *If $K \subset \mathbb{R}^n$ is compact and $f : K \to \mathbb{R}^m$ is continuous, then there exists $p \in K$ such that*

$$\|f(p)\| = \max_{x\in K} \|f(x)\|, \tag{B.11}$$

and there exists $q \in K$ such that

$$\|f(q)\| = \min_{x\in K} \|f(x)\|. \tag{B.12}$$

The meaning of (B.11) is that $\|f(p)\| \ge \|f(x)\|$ for all $x \in K$, and the meaning of (B.12) is similar.

For the proof, consider

$$g : K \longrightarrow \mathbb{R}, \quad g(p) = \|f(p)\|. \tag{B.13}$$

This is continuous, so $g(K)$ is compact. Hence $g(K)$ is bounded; say $g(K) \subset I = [a, b]$. Repeatedly subdividing I into equal halves, as in the proof of Proposition B.3, at each stage throwing out subintervals that do not intersect $g(K)$ and keeping only the leftmost and rightmost amongst those remaining, we obtain $\alpha \in g(K)$ and $\beta \in g(K)$ such that $g(K) \subset [\alpha, \beta]$. Then $\alpha = f(q)$ and $\beta = f(p)$ for some p and $q \in K$ satisfying (B.11)–(B.12).

A variant of Proposition B.5, with a very similar proof, is that if $K \subset \mathbb{R}^n$ is compact and $f : K \to \mathbb{R}$ is continuous, then there exist $p, q \in K$ such that

$$f(p) = \max_{x\in K} f(x), \quad f(q) = \min_{x\in K} f(x). \tag{B.14}$$

We next define the *closure* $\overline{S}$ of a set $S \subset \mathbb{R}^n$, to consist of all points $p \in \mathbb{R}^n$ such that $B_\varepsilon(p) \cap S \neq \emptyset$ for all $\varepsilon > 0$. Equivalently, $p \in \overline{S}$ if and only if there exists an infinite sequence (p_j) of points in S such that $p_j \to p$.

Now we define $\sup S$ and $\inf S$. First, let $S \subset \mathbb{R}$ be nonempty and bounded from above, i.e., there exists $R < \infty$ such that $x \le R$ for all $x \in S$. Hence $x \le R$ for all $x \in \overline{S}$. In such a case, there exists an interval $[R - k, R]$ whose intersection with $\overline{S}$ is nonempty, hence compact. We set

$$\sup\, S = \max_{\overline{S}\cap[R-k,R]} x, \tag{B.15}$$

the right side well defined by (B.14), with $f(x) = x$. There is a similar definition of

$$\inf\, S, \tag{B.16}$$

when S is bounded from below.

We establish some further properties of compact sets $K \subset \mathbb{R}^n$, leading to the important result, Proposition B.9 below.

Proposition B.6. *Let $K \subset \mathbb{R}^n$ be compact. Assume $X_1 \supset X_2 \supset X_3 \supset \cdots$ form a decreasing sequence of closed subsets of K. If each $X_m \neq \emptyset$, then $\cap_m X_m \neq \emptyset$.*

Proof. Pick $x_m \in X_m$. If K is compact, (x_m) has a convergent subsequence, $x_{m_k} \to y$. Since $\{x_{m_k} : k \geq \ell\} \subset X_{m_\ell}$, which is closed, we have $y \in \cap_m X_m$.

Corolary B.7. *Let $K \subset \mathbb{R}^n$ be compact. Assume $U_1 \subset U_2 \subset U_3 \subset \cdots$ form an increasing sequence of open sets in $\mathbb{R}^n$. If $\cup_m U_m \supset K$, then $U_M \supset K$ for some M.*

Proof. Consider $X_m = K \setminus U_m$.

Before getting to Proposition B.9, we bring in the following. Let $\mathbb{Q}$ denote the set of rational numbers, and let $\mathbb{Q}^n$ denote the set of points in $\mathbb{R}^n$ all of whose components are rational. The set $\mathbb{Q}^n \subset \mathbb{R}^n$ has the following "denseness" property: given $p \in \mathbb{R}^n$ and $\varepsilon > 0$, there exists $q \in \mathbb{Q}^n$ such that $\|p - q\| < \varepsilon$. Let

$$\mathcal{R} = \{B_{r_j}(q_j) : q_j \in \mathbb{Q}^n,\ r_j \in \mathbb{Q} \cap (0, \infty)\}. \tag{B.17}$$

Note that $\mathbb{Q}$ and $\mathbb{Q}^n$ are *countable*, i.e., they can be put in one-to-one correspondence with $\mathbb{N}$. Hence $\mathcal{R}$ is a countable collection of balls. The following lemma is left as an exercise for the reader.

Lemma B.8. *Let $\Omega \subset \mathbb{R}^n$ be a nonempty open set. Then*

$$\Omega = \bigcup \{B : B \in \mathcal{R},\ B \subset \Omega\}. \tag{B.18}$$

To state the next result, we say that a collection $\{U_\alpha : \alpha \in \mathcal{A}\}$ covers K if $K \subset \cup_{\alpha \in \mathcal{A}} U_\alpha$. If each $U_\alpha \subset \mathbb{R}^n$ is open, it is called an open cover of K. If $\mathcal{B} \subset \mathcal{A}$ and $K \subset \cup_{\beta \in \mathcal{B}} U_\beta$, we say $\{U_\beta : \beta \in \mathcal{B}\}$ is a subcover.

Proposition B.9. *If $K \subset \mathbb{R}^n$ is compact, then it has the following property.*

(B.19) *Every open cover $\{U_\alpha : \alpha \in \mathcal{A}\}$ of K has a finite subcover.*

Proof. By Lemma B.8, it suffices to prove the following.

(B.20) Every countable cover $\{B_j : j \in \mathbb{N}\}$ of K by open balls has a finite subcover.

For this, we set

$$U_m = B_1 \cup \cdots \cup B_m \tag{B.21}$$

and apply Corollary B.7.

C. Critical points that are saddles

Let F be a C^3 vector field on $\Omega \subset \mathbb{R}^n$, with a critical point at $p \in \Omega$. We say p is a simple critical point if $L = DF(p)$ has no eigenvalues that are purely imaginary (or zero). From here on we assume this condition holds. As seen in Chapter 2, we can write

$$\mathbb{C}^n = W_+ \oplus W_-, \tag{C.1}$$

where W_+ is the direct sum of the generalized eigenspaces of L associated to eigenvalues with positive real part and W_- is the direct sum of the generalized eigenspaces associated to eigenvalues with negative real part. Since $L \in M(n, \mathbb{R})$, non-real eigenvalues of L must occur in complex conjugate pairs, and

$$\mathbb{R}^n = V_+ \oplus V_-, \quad V_\pm = W_\pm \cap \mathbb{R}^n. \tag{C.2}$$

We have

$$v \in W_\pm \Longrightarrow e^{tL} v \to 0 \ \text{ as } \ t \to \mp\infty, \tag{C.3}$$

and a fortiori

$$v \in V_\pm \Longrightarrow e^{tL} v \to 0 \ \text{ as } \ t \to \mp\infty. \tag{C.4}$$

We say the critical point at p is a source if $V_- = 0$, a sink if $V_+ = 0$, and a saddle if $V_- \neq 0$ and $V_+ \neq 0$. The fact that

$$V_- = \mathbb{R}^n \Longrightarrow \Phi_F^t(x) \to p \ \text{ as } \ t \to +\infty, \tag{C.5}$$

for x sufficiently close to p, where Φ_F^t is the flow generated by F, was proven in §3 (cf. Proposition 3.4), and similarly we have

$$V_+ = \mathbb{R}^n \Longrightarrow \Phi_F^t(x) \to p \ \text{ as } \ t \to -\infty, \tag{C.6}$$

for x sufficiently close to p. The purpose of this appendix is to discuss the saddle case, where $n_+ = \dim V_+ > 0$ and $n_- = \dim V_- > 0$. In such a case, as advertised in §3, there is a neighborhood U of p and there are C^1 surfaces $S_\pm$, of dimension $n_\pm$, such that

$$\{p\} = S_+ \cap S_-, \tag{C.7}$$

and

$$x \in S_\pm \Longrightarrow \Phi_F^t(x) \to p \ \text{ as } \ t \to \mp\infty. \tag{C.8}$$

The surfaces S_- and S_+ are called, respectively, the stable and unstable manifolds of F at p. They have the further property that if γ is a C^1 curve in S_+ (respectively, S_-), and $\gamma(0) = p$, then $\gamma'(0) \in V_+$ (respectively, V_-). In addition, given $\varepsilon > 0$, there exists $\delta > 0$ such that if $x \in U \setminus S_-$ but $\operatorname{dist}(x, S_-) < \delta$, then for some $t_1 > 0$, $\|\Phi_F^{t_1}(x) - p\| < \varepsilon$, and for all $t \geq t_1$, $\operatorname{dist}(\Phi_F^t(x), S_+) < \varepsilon$, at least until $\Phi_F^t(x)$ exits U. We want to demonstrate this result. For simplicity of presentation, we concentrate on the case $n = 2$ (and $n_+ = n_- = 1$). However, the argument we present can be modified to treat saddles in higher dimension.

We make some preliminary constructions. Relabeling the coordinates, we can assume $p = 0$. Altering F outside some neighborhood of $p = 0$ if necessary, we can assume F is a C^3 vector field on $\mathbb{R}^n$ and there exists $C < \infty$ such that

$$\|F(x)\| \leq C\|x\|, \quad \forall\, x \in \mathbb{R}^n. \tag{C.9}$$

Hence, as seen in §3 (Exercise 3), $\Phi_F^t(x)$ is well defined for all $x \in \mathbb{R}^n,\ t \in \mathbb{R}$. Applying the fundamental theorem of calculus twice gives

$$F(x) = Lx + \sum_{j,k} x_j x_k G_{jk}(x), \tag{C.10}$$

where

$$L = DF(0), \tag{C.11}$$

and G_{jk} are C^1 vector fields, given by

$$G_{jk}(x) = \int_0^1 \int_0^1 \frac{\partial^2}{\partial x_k \partial x_j} F(stx)\, ds\, dt. \tag{C.12}$$

We define the family of vector fields F_ε by

$$F_\varepsilon(x) = \frac{1}{\varepsilon} F(\varepsilon x), \tag{C.13}$$

for $\varepsilon > 0$. By (C.10),

$$F_\varepsilon(x) = Lx + \varepsilon G_\varepsilon(x), \tag{C.14}$$

where

$$G_\varepsilon(x) = \sum_{j,k} x_j x_k G_{jk}(\varepsilon x). \tag{C.15}$$

Passing to the limit $\varepsilon \to 0$ gives $F_0(x) = Lx$. Results of §2 yield the following.

Lemma C.1. *Given $\delta > 0$, $T < \infty$, there exists $\varepsilon_0 = \varepsilon_0(\delta, T, F) > 0$ such that for all $\varepsilon \in (0, \varepsilon_0]$,*

$$\begin{aligned} &\|x\| \le 2, \quad |t| \le T, \quad \|\Phi^s_{F_\varepsilon}(x)\| \le 2 \ \forall \, s \in [0,t] \\ &\Longrightarrow \|\Phi^t_{F_\varepsilon}(x) - e^{tL}x\| \le \delta. \end{aligned} \tag{C.16}$$

Specializing to $n = 2$, we can assume that

$$L = \begin{pmatrix} a & \\ & -b \end{pmatrix}, \quad a, b > 0. \tag{C.17}$$

We take the box

$$\mathcal{O} = \{(x_1, x_2) : |x_1|, |x_2| \le 1\}, \tag{C.18}$$

and set

$$\mathcal{O}_k = 2^{-k}\mathcal{O}. \tag{C.19}$$

We define four families of maps

$$\varphi_{\varepsilon j}, \psi_{\varepsilon j} : \Bigl[-\frac{1}{2}, \frac{1}{2}\Bigr] \longrightarrow [-1, 1], \quad j = 1, 2, \ 0 \le \varepsilon \le 1, \tag{C.20}$$

as follows. For $j = 1$, define $t_\varepsilon(s)$ as the smallest positive number such that

$$\Phi^{-t_\varepsilon(s)}_{F_\varepsilon}\Bigl(s, \frac{1}{2}\Bigr) \in \{(\sigma, 1) : -1 \le \sigma \le 1\},$$

and then set $\varphi_{\varepsilon 1}(s)$ to be the x_1-coordinate of $\Phi^{-t_\varepsilon(s)}_{F_\varepsilon}(s, 1/2)$. To give an alternative description, we are mapping the top edge of $\mathcal{O}_1$ (identified with $[-1/2, 1/2]$) to the top edge of $\mathcal{O}$ (identified with $[-1, 1]$) by the backward flow generated by F_ε. Similarly define $\varphi_{\varepsilon 2}$ via the backward flow map of the bottom edge of $\mathcal{O}_1$ to the bottom edge of $\mathcal{O}$, and define $\psi_{\varepsilon 1}$ and $\psi_{\varepsilon 2}$ via the forward flow maps of the right and left edges of $\mathcal{O}_1$ to the corresponding edges of $\mathcal{O}$. See Fig. C.1. It is readily verified that these maps are contractions for $\varepsilon = 0$, where $F_0(x) = Lx$, i.e., there exists $A = A(a, b) < 1$ such that

$$\begin{aligned} |\varphi_{\varepsilon j}(s) - \varphi_{\varepsilon j}(t)| &\le A|s - t|, \\ |\psi_{\varepsilon j}(s) - \psi_{\varepsilon j}(t)| &\le A|s - t|, \end{aligned} \tag{C.21}$$

Figure C.1

for all $s, t \in [-1/2, 1/2]$.

Results of §2 then establish the following.

Lemma C.2. *There exists $\varepsilon_1 = \varepsilon_1(F) > 0$ and $A = A(F) < 1$ such that whenever $0 \le \varepsilon \le \varepsilon_1$, the maps $\varphi_{\varepsilon j}$ and $\psi_{\varepsilon j}$ in (C.20) are well defined on $[-1/2, 1/2]$ and (C.21) holds for all $s, t \in [-1/2, 1/2]$.*

We make a further adjustment. Take $\varepsilon_2 \le \min(\varepsilon_1(F), \varepsilon_0(1/10, 10, F))$. Further shrinking ε_2 is nesessary, arrange that, whenever $\varepsilon \in (0, \varepsilon_2]$,

$$\|x\| \le 2 \Longrightarrow \|F_\varepsilon(x) - Lx\| \le \frac{1}{2}\|Lx\|, \tag{C.22}$$

so that, if $(x_1, x_2) \in \mathcal{O}$, $F_\varepsilon(x_1, x_2)$ points down if $x_2 \in [1/2, 1]$, up if $x_2 \in [-1, -1/2]$, left if $x_1 \in [1/2, 1]$, and right if $x_1 \in [-1, -1/2]$. Now replace F by F_{ε_2}, denoting this scaled vector field by F. Then (C.16) holds with $T = 10$ and $\delta = 1/10$ for all $\varepsilon \in (0, 1]$ and (C.21) holds for all $s, t \in [-1/2, 1/2]$, with $A < 1$, for all $\varepsilon \in (0, 1]$. For notational simplicity, set

$$\Phi^t_k = \Phi^t_{F_\varepsilon}, \quad \varepsilon = 2^{-k}. \tag{C.23}$$

Note that dilation by the factor 2^k takes the flow Φ^t_F on $\mathcal{O}_k$ to the flow Φ^t_k on $\mathcal{O}$.

With these preliminaries done, we start in earnest our demonstration that the flow generated by F has saddle-like behavior near the critical point $p = 0$. Denote by $\mathcal{T}, \mathcal{B}, \mathcal{L}$, and $\mathcal{R}$ the top, bottom, left, and right edges of $\mathcal{O}$, and similarly denote by $\mathcal{T}_k, \mathcal{B}_k, \mathcal{L}_k$, and $\mathcal{R}_k$ the top, bottom, left, and right

Figure C.2

sides of $\mathcal{O}_k$. Then the maps (C.20) can by slight abuse of terminology be labeled

$$\begin{aligned}\varphi_{\varepsilon 1} : \mathcal{T}_1 \to \mathcal{T}, \quad \varphi_{\varepsilon 2} : \mathcal{B}_1 \to \mathcal{B}, \\ \psi_{\varepsilon 1} : \mathcal{R}_1 \to \mathcal{R}, \quad \psi_{\varepsilon 2} : \mathcal{L}_1 \to \mathcal{L}.\end{aligned} \tag{C.24}$$

Pick two points $p_{0\ell}, p_{0r} \in \mathcal{T}$ such that for some $t_{0\ell}, t_{0r} \in (0,1)$,

$$p^*_{0\ell} = \Phi_F^{t_{0\ell}}(p_{0\ell}) \in \mathcal{L}, \quad p^*_{0r} = \Phi_F^{t_{0r}}(p_{0r}) \in \mathcal{R}, \tag{C.25}$$

and pick two points $q_{0\ell}, q_{0r} \in \mathcal{B}$ such that for some $s_{0\ell}, s_{0r} \in (0,1)$,

$$q^*_{0\ell} = \Phi_F^{s_{0\ell}}(q_{0\ell}) \in \mathcal{L}, \quad q^*_{0r} = \Phi_F^{s_{0r}}(q_{0r}) \in \mathcal{R}. \tag{C.26}$$

See Fig. C.2. The possibility to do this is guaranteed by Lemma C.1. Denote the orbits of $\Phi_0^t = \Phi_F^t$ through $p_{0\ell}, p_{0r}$ by $\gamma_{0\ell}, \gamma_{0r}$ and those through $q_{0\ell}, q_{0r}$ by $\sigma_{0\ell}, \sigma_{0r}$.

Let us call the construction just described Step 0. To continue, at Step 1, pick $\tilde{p}_{1\ell}, \tilde{p}_{1r} \in \mathcal{T}_1$ and $\tilde{q}_{1\ell}, \tilde{q}_{1r} \in \mathcal{B}_1$ such that the following holds. Note that $2\tilde{p}_{1\ell}, 2\tilde{p}_{1r} \in \mathcal{T}$ and $2\tilde{q}_{1\ell}, 2\tilde{q}_{1r} \in \mathcal{B}$. We require that, for some $t_{1\ell}, t_{1r} \in (0, T_1)$,

$$\Phi_1^{t_{1\ell}}(2\tilde{p}_{1\ell}) \in \mathcal{L}, \quad \Phi_1^{t_{1r}}(2\tilde{p}_{1r}) \in \mathcal{R}, \tag{C.27}$$

and for some $s_{1\ell}, s_{1r} \in (0, T_1)$,

$$\Phi_1^{s_{1\ell}}(2\tilde{q}_{1\ell}) \in \mathcal{L}, \quad \Phi_1^{s_{1r}}(2\tilde{q}_{1r}) \in \mathcal{R}. \tag{C.28}$$

The conditions (C.27) and (C.28) are equivalent to

$$\tilde{p}^*_{1\ell} = \Phi_F^{t_{1\ell}}(\tilde{p}_{1\ell}) \in \mathcal{L}_1, \quad \tilde{p}^*_{1r} = \Phi_F^{t_{1r}}(\tilde{p}_{1r}) \in \mathcal{R}_1, \tag{C.29}$$

Figure C.3

and

$$\tilde{q}_{1\ell}^* = \Phi_F^{s_{1\ell}}(\tilde{q}_{1\ell}) \in \mathcal{L}_1, \quad \tilde{q}_{1r}^* = \Phi_F^{s_{1r}}(\tilde{q}_{1r}) \in \mathcal{R}_1. \tag{C.30}$$

See Fig. C.3. Denote the orbits of F through $\tilde{p}_{1\ell}, \tilde{p}_{1r}$ by $\gamma_{1\ell}, \gamma_{1r}$ and those through $\tilde{q}_{1\ell}, \tilde{q}_{1r}$ by $\sigma_{1\ell}, \sigma_{1r}$. When picking $\tilde{p}_{1\ell}, \tilde{p}_{1r}, \tilde{q}_{1\ell}$, and $\tilde{q}_{1r}$, one can and should enforce the following condition. If $\gamma_{0\ell}$ intersects $\mathcal{T}_1$, $\tilde{p}_{1\ell}$ should be to the right of such an intersection, if γ_{0r} intersects $\mathcal{T}_1$, $\tilde{p}_{1r}$ should be to the left of such an intersection, and similarly for cases when $\sigma_{0\ell}$ or σ_{0r} intersect $\mathcal{B}_1$. Also, we can take $T_1 > 1$. (More on this below.)

Now we continue the orbits $\gamma_{1\ell}, \gamma_{1r}, \sigma_{1\ell}$, and σ_{1r} forward and backward, until they intersect the boundary of $\mathcal{O}$, at points $p_{1\ell}, p_{1r}, q_{1\ell}, q_{1r}$ and $p_{1\ell}^*, p_{1r}^*, q_{1\ell}^*, q_{1r}^*$, as illustrated in Fig. C.4. That such an intersection must occur is guaranteed by (C.22). This, together with the fact that orbits of F cannot intersect, guarantees that

$$p_{0\ell} < p_{1\ell} < p_{1r} < p_{0r}, \tag{C.31}$$

in the sense that $p < p'$ means p is to the left of p'. In a similar sense, made clear in Fig. C.4, we have

$$\begin{aligned} q_{0\ell} < q_{1\ell} < q_{1r} < q_{0r}, \\ p_{0r}^* < p_{1r}^* < q_{1r}^* < q_{0r}^*, \\ p_{0\ell}^* < p_{1\ell}^* < q_{1\ell}^* < q_{0\ell}^*. \end{aligned} \tag{C.32}$$

Figure C.4

Furthermore, as a consequence of (C.21), we have

$$
\begin{aligned}
|p_{1\ell} - p_{1r}| &\le A|\tilde{p}_{1\ell} - \tilde{p}_{1r}| \le A, \\
|q_{1\ell} - q_{1r}| &\le A|\tilde{q}_{1\ell} - \tilde{q}_{1r}| \le A, \\
|p^*_{1r} - q^*_{1r}| &\le A|\tilde{p}^*_{1r} - \tilde{q}^*_{1r}| \le A, \\
|p^*_{1\ell} - q^*_{1\ell}| &\le A|\tilde{p}^*_{1\ell} - \tilde{q}^*_{1r}| \le A.
\end{aligned} \tag{C.33}
$$

We proceed iteratively. At step k, pick $\tilde{p}_{k\ell}, \tilde{p}_{kr} \in \mathcal{T}_k$ and $\tilde{q}_{k\ell}, \tilde{q}_{kr} \in \mathcal{B}_k$ such that the following holds. Note that $2^k\tilde{p}_{k\ell}, 2^k\tilde{p}_{kr} \in \mathcal{T}$ and $2^k\tilde{q}_{k\ell}, 2^k\tilde{q}_{kr} \in \mathcal{B}$. We require that, for some $t_{k\ell}, t_{kr} \in (0, T_k)$,

$$
\Phi_k^{t_{k\ell}}(2^k\tilde{p}_{k\ell}) \in \mathcal{L}, \quad \Phi_k^{t_{kr}}(2^k\tilde{p}_{kr}) \in \mathcal{R}, \tag{C.34}
$$

and for some $s_{k\ell}, s_{kr} \in (0, T_k)$,

$$
\Phi_k^{s_{k\ell}}(2^k\tilde{q}_{k\ell}) \in \mathcal{L}, \quad \Phi_k^{s_{kr}}(2^k\tilde{q}_{kr}) \in \mathcal{R}. \tag{C.35}
$$

The conditions (C.34) and (C.35) are equivalent to

$$
\tilde{p}^*_{k\ell} = \Phi_F^{t_{k\ell}}(\tilde{p}_{k\ell}) \in \mathcal{L}_k, \quad \tilde{p}^*_{kr} = \Phi_F^{t_{kr}}(\tilde{p}_{kr}) \in \mathcal{R}_k, \tag{C.36}
$$

and

$$
\tilde{q}^*_{k\ell} = \Phi_F^{s_{k\ell}}(\tilde{q}_{k\ell}) \in \mathcal{L}_k, \quad \tilde{q}^*_{kr} = \Phi_F^{s_{kr}}(\tilde{q}_{kr}) \in \mathcal{R}_k. \tag{C.37}
$$

Denote the orbits of F through $\tilde{p}_{k\ell}, \tilde{p}_{kr}$ by $\gamma_{k\ell}, \gamma_{kr}$, and those through $\tilde{q}_{k\ell}, \tilde{q}_{kr}$ by $\sigma_{k\ell}, \sigma_{kr}$. When picking $\tilde{p}_{k\ell}, \tilde{p}_{kr}, \tilde{q}_{k\ell}$, and $\tilde{q}_{kr}$, one can and should enforce

Cigure C.5

the following condition. If $\gamma_{k-1,\ell}$ intersects $\mathcal{T}_k$, $\tilde{p}_{k\ell}$ should lie to the right of such a point of intersection, if $\gamma_{k-1,r}$ intersects $\mathcal{T}_k$, $\tilde{p}_{kr}$ should lie to the left of such a point of intersection, and similarly for cases where $\sigma_{k-1,\ell}$ or $\sigma_{k-1,r}$ intersect $\mathcal{B}_k$. At this point it is useful to note that, by Lemma D.1, we can take

$$T_k \to \infty \ \text{ as } \ k \to \infty, \tag{C.38}$$

and hence take (with $z = \ell$ or r)

$$\|2^k \tilde{p}_{kz} - (0,1)\| \le \eta_k, \quad \|2^k \tilde{q}_{kz} - (0,1)\| \le \eta_k, \quad \eta_k \to 0 \text{ as } k \to \infty. \tag{C.39}$$

It then follows that (again with $z = \ell$ or r)

(C.40)

$$\|\tilde{p}^*_{kz} - (2^{-k}, 0)\| \le 2^{-k}\tilde{\eta}_k, \quad \|\tilde{q}^*_{kz} - (2^{-k}, 0)\| \le 2^{-k}\tilde{\eta}_k, \quad \tilde{\eta}_k \to 0 \text{ as } k \to \infty.$$

Now we continue the orbits $\gamma_{k\ell}, \gamma_{kr}, \sigma_{k\ell}$, and σ_{kr} forward and backward, until they intersect the boundary of $\mathcal{O}$, at points $p_{k\ell}, p_{kr}, q_{k\ell}, q_{kr}$, and $p^*_{k\ell}, p^*_{kr}, q^*_{k\ell}, q^*_{kr}$, as illustrated in Fig. C.5. That such intersections must occur is guaranteed by (C.22). As before, the fact that orbits of Φ^t_F cannot intersect guarantees that

$$p_{0\ell} < \cdots < p_{k\ell} < p_{kr} < \cdots < p_{0r}, \tag{C.41}$$

in the sense specified in (C.31), and, as in (C.32),

$$\begin{aligned} q_{0\ell} < \cdots < q_{k\ell} < q_{kr} < \cdots < q_{0r}, \\ p^*_{0r} < \cdots < p^*_{kr} < q^*_{kr} < \cdots < q^*_{0r}, \\ p^*_{0\ell} < \cdots < p^*_{k\ell} < q^*_{k\ell} < \cdots < q^*_{0\ell}. \end{aligned} \tag{C.42}$$

Figure C.6

Furthermore, as a consequence of (C.21), we have

$$
\begin{aligned}
|p_{k\ell} - p_{kr}| &\le A^k |\tilde{p}_{k\ell} - \tilde{p}_{kr}| \le A^k 2^{-k} \eta_k, \\
|q_{k\ell} - q_{kr}| &\le A^k |\tilde{q}_{k\ell} - \tilde{q}_{kr}| \le A^k 2^{-k} \eta_k, \\
|p^*_{k\ell} - p^*_{kr}| &\le A^k |\tilde{p}^*_{k\ell} - \tilde{p}^*_{kr}| \le A^k 2^{-k} \tilde{\eta}_k, \\
|q^*_{k\ell} - q^*_{kr}| &\le A^k |\tilde{q}^*_{k\ell} - \tilde{q}^*_{kr}| \le A^k 2^{-k} \tilde{\eta}_k.
\end{aligned}
\tag{C.43}
$$

In particular, these distances are converging to 0 quite rapidly. We obtain limits

$$
\begin{aligned}
&p_{k\ell}, p_{kr} \to p_t \in \mathcal{T}, \quad q_{k\ell}, q_{kr} \to p_b \in \mathcal{B}, \\
&p^*_{k\ell}, q^*_{kr} \to p^*_r \in \mathcal{R}, \quad p^*_{k\ell}, q^*_{k\ell} \to p^*_\ell \in \mathcal{L}.
\end{aligned}
\tag{C.44}
$$

See Fig. C.6. We have

$$
\Phi^t_F(p_t),\ \Phi^t_F(p_b) \to 0 \ \text{ as } \ t \to +\infty,
\tag{C.45}
$$

and

$$
\Phi^t_F(p^*_\ell),\ \Phi^t_F(p^*_r) \to 0 \ \text{ as } \ t \to -\infty,
\tag{C.46}
$$

since the paths in (C.45) meet each $\mathcal{O}_k$ for large positive t and those in (C.46) meet each $\mathcal{O}_k$ for large negative t. Furthermore, by (C.39)–(C.40), plus the fact that all these paths solve $dx/dt = F(x)$, the curves in (C.45) fit together to form a C^1 curve tangent to the x_2-axis at $p = 0$, and those in (C.46) fit together to form a C^1 curve tangent to the x_1-axis at $p = 0$.

We sketch how to treat the case $n = 3,\ n_+ = 2,\ n_- = 1$. In place of (C.17), we can take

$$L = \begin{pmatrix} A & \\ & -b \end{pmatrix}, \quad A \in M(2,\mathbb{R}),\ b > 0, \tag{C.47}$$

and, via Lemma 3.5, arrange that

$$Av \cdot v \ge a\|v\|^2, \quad a > 0, \quad \forall\, v \in \mathbb{R}^2. \tag{C.48}$$

In place of (C.18), we use the cylinder

$$\mathcal{O} = \{(x_1, x_2, x_3) : x_1^2 + x_2^2 \le 1, |x_3| \le 1\}. \tag{C.49}$$

with boundary

$$\partial\mathcal{O} = \mathcal{T} \cup \mathcal{B} \cup \mathcal{L}, \tag{C.50}$$

where $\mathcal{T}$ and $\mathcal{B}$ (the top and bottom) are disks and $\mathcal{L}$ (the side) is $S^1 \times [-1,1]$. We then take $\mathcal{O}_k = 2^{-k}\mathcal{O}$, with boundary $\mathcal{T}_k \cup \mathcal{B}_k \cup \mathcal{L}_k$. Parallel to (C.24), we have (at least for small ε) maps

$$\begin{gathered} \varphi_{\varepsilon 1} : \mathcal{T}_1 \to \mathcal{T}, \quad \varphi_{\varepsilon 2} : \mathcal{B}_1 \to \mathcal{B}, \\ \psi_\varepsilon : \mathcal{L}_1 \to \mathcal{L}, \end{gathered} \tag{C.51}$$

with $\varphi_{\varepsilon j}$ defined by backward flow of $\Phi^t_{F_\varepsilon}$ and ψ_ε defined by forward flow. Again the maps $\varphi_{\varepsilon j}$ are contractions for small ε. The maps ψ_ε are not contractions, but composing them on the left with the projection $S^1 \times [-1,1] \to [-1,1]$ produces a contraction, for small ε, and this is what one needs. In place of a pair of initial data on $\mathcal{T}$ and a pair on $\mathcal{B}$, one takes a circle of initial data on $\mathcal{T}$ and one on $\mathcal{B}$. Applying Φ^t_F yields a pair of flared tubes, as pictured in Fig. C.7. From here, an iteration produces nested families of such flared tubes, converging in on the one-dimensional stable manifold S_- and the two-dimensional unstable manifold S_+. The interested reader is invited to fill in the details, and work out the higher dimensional cases. See also [**CL**] and [**Hart**] for other approaches to this result.

D. Periodic solutions of $x'' + x = \varepsilon\psi(x)$

Equations of the form

$$x'' + x = \varepsilon\psi(x) \tag{D.1}$$

with "small" ε arise in a number of cases, and it is of interest to analyze various features of these solutions. For example, as mentioned in §6, the

Figure C.7

relativistic correction for planetary motion gives rise to the equation (6.53), which takes the form (D.1) for $x = u - A$, with

$$\psi(x) = (x + A)^2. \tag{D.2}$$

Another example,

$$\psi(x) = -x^3, \tag{D.3}$$

yields a special case of Duffing's equation. As we mentioned in §6, solutions to (6.53) tend not to be periodic of period 2π, and this leads to the phenomenon of precession of perihelia. It is of general interest to compute the period of a solution to (D.1), and we discuss this problem here. We assume ψ is smooth.

We rewrite (D.1) as a first order system and also explicitly record the dependence on ε:

$$\begin{aligned} x_\varepsilon'(t) &= y_\varepsilon(t), \\ y_\varepsilon'(t) &= -x_\varepsilon(t) + \varepsilon\psi(x_\varepsilon(t)). \end{aligned} \tag{D.4}$$

We pick $a \in (0, \infty)$ and impose the initial conditions

$$x_\varepsilon(0) = a, \quad y_\varepsilon(0) = 0. \tag{D.5}$$

Note that if we take

$$F_\varepsilon(x, y) = \frac{y^2}{2} + \frac{x^2}{2} - \varepsilon\Psi(x), \tag{D.6}$$

where $\Psi'(x) = \psi(x)$, then $(d/dt)F_\varepsilon(x_\varepsilon(t), y_\varepsilon(t)) = 0$ for solutions to (D.4), so orbits of (D.4) lie on level curves of F_ε. For ε sufficiently small with respect to a, the level curves of F_ε on $\{(x,y) : x^2 + y^2 \le 2a^2\}$ will be close to those of F_0, that is to say, such level curves of F_ε will be closed curves, close to circles, and the associated solutions to (D.4)–(D.5) will be periodic. The period $T(\varepsilon)$ will have the following two properties, at least for small ε:

$$T(\varepsilon) = 2\pi + O(\varepsilon), \quad y_\varepsilon(T(\varepsilon)) = 0. \tag{D.7}$$

We will calculate a more precise approximation to $T(\varepsilon)$, accurate for small ε.

The first order of business is to calculate accurate approximations to $x_\varepsilon(t)$ and $y_\varepsilon(t)$, valid uniformly for t in an interval containing $[0, 2\pi]$. It follows from §2 that $x_\varepsilon(t)$ and $y_\varepsilon(t)$ are smooth functions of ε. Hence, for each $N \in \mathbb{N}$, we can write

$$\begin{aligned} x_\varepsilon(t) &= a\cos t + \sum_{k=1}^{N} X_k(t)\varepsilon^k + R_{1N}(t,\varepsilon), \\ y_\varepsilon(t) &= -a\sin t + \sum_{k=1}^{N} Y_k(t)\varepsilon^k + R_{2N}(t,\varepsilon), \end{aligned} \tag{D.8}$$

where

$$|R_{jN}(t,\varepsilon)| \le C_{KN}\varepsilon^{N+1}, \quad \forall\, |t| \le K. \tag{D.9}$$

We write $R_{jN}(t,\varepsilon) = O(\varepsilon^{N+1})$. The coefficients $X_k(t)$ and $Y_k(t)$ satisfy differential equations, obtained as follows. We have from (D.8)

$$x''_\varepsilon(t) + x_\varepsilon(t) = \sum_{k=1}^{N} \big[X''_k(t) + X_k(t)\big]\varepsilon^k + O(\varepsilon^{N+1}), \tag{D.10}$$

while

$$\begin{aligned} \varepsilon\psi(x_\varepsilon(t)) &= \varepsilon\psi\Big(a\cos t + \sum_{k=1}^{N} X_k(t)\varepsilon^k\Big) + O(\varepsilon^{N+1}) \\ &= \varepsilon\Big[\psi(a\cos t) + \sum_{j=1}^{N} \frac{1}{j!}\psi^{(j)}(a\cos t)\Big(\sum_{k=1}^{N} X_k(t)\varepsilon^k\Big)^j\Big] + O(\varepsilon^{N+1}). \end{aligned} \tag{D.11}$$

We match up the coefficients of ε^k in (D.10) and (D.11) to obtain equations for $X_k(t)$. The case $k = 1$ gives

$$X''_1(t) + X_1(t) = \psi(a\cos t), \tag{D.12}$$

and from (D.5) the initial conditions are seen to be

$$X_1(0)=0, \quad X_1'(0)=0. \tag{D.13}$$

The solution to (D.12)–(D.13) is given by Duhamel's formula, cf. (4.10) of Chapter 3:

$$X_1(t)=\int_0^t \sin(t-s)\,\psi(a\cos s)\,ds. \tag{D.14}$$

It is convenient to expand $\sin(t-s)$ and rewrite (D.14) as

$$X_1(t)=(\sin t)\int_0^t \cos s\,\psi(a\cos s)\,ds-(\cos t)\int_0^t \sin s\,\psi(a\cos s)\,ds. \tag{D.15}$$

Regarding $Y_k(t)$, we have

$$Y_k(t)=X_k'(t), \tag{D.16}$$

for all k, and in particular

$$Y_1(t)=(\cos t)\int_0^t \cos s\,\psi(a\cos s)\,ds+(\sin t)\int_0^t \sin s\,\psi(a\cos s)\,ds. \tag{D.17}$$

In case $\psi(x)$ is given by (D.2), we have

$$\psi(a\cos s)=\frac{a^2}{2}\cos 2s+2aA\cos s+\Bigl(A^2+\frac{a^2}{2}\Bigr), \tag{D.18}$$

hence

$$\begin{aligned}&\int_0^t \cos s\,\psi(a\cos s)\,ds\\ &=\Bigl(A^2+\frac{3a^2}{4}\Bigr)\sin t+\frac{aA}{2}\sin 2t+\frac{a^2}{12}\sin 3t+aAt,\end{aligned} \tag{D.19}$$

and

$$\begin{aligned}&\int_0^t \sin s\,\psi(a\cos s)\,ds\\ &=\Bigl(A^2+\frac{aA}{2}+\frac{a^2}{2}\Bigr)-\Bigl(A^2+\frac{5a^2}{12}\Bigr)\cos t-\frac{aA}{2}\cos 2t-\frac{a^2}{12}\cos 3t.\end{aligned} \tag{D.20}$$

One can compute higher terms in (D.8). For example, matching up coefficients of ε^2 in (D.10) and (D.11) yields

$$X_2''(t)+X_2(t)=\psi'(a\cos t)X_1(t). \tag{D.21}$$

Again $X_2(0) = X_2'(0) = 0$, and, parallel to (D.14), we have

$$X_2(t) = \int_0^t \sin(t-s)\, \psi'(a\cos s) X_1(s)\, ds. \tag{D.22}$$

$Y_2(t)$ is given by (D.16). One can continue this, but we will leave off at this point.

We return to the problem of approximating the period $T(\varepsilon)$, making use of (D.7). A very effective method for solving $y_\varepsilon(T) = 0$ with $T \approx 2\pi$ is Newton's method, which gives $T(\varepsilon)$ as the limit of $T_n(\varepsilon)$, defined recursively by

$$T_0(\varepsilon) = 2\pi, \quad T_{n+1}(\varepsilon) = T_n(\varepsilon) - \frac{y_\varepsilon(T_n(\varepsilon))}{y_\varepsilon'(T_n(\varepsilon))}. \tag{D.23}$$

This sequence converges fast:

$$T(\varepsilon) = T_n(\varepsilon) + O(\varepsilon^{2^n}), \tag{D.24}$$

provided one has $y_\varepsilon(t)$ evaluated exactly. Given an approximation to $y_\varepsilon(t)$,

$$y_\varepsilon(t) = \tilde{y}_\varepsilon(t) + O(\varepsilon^N), \quad y_\varepsilon'(t) = \tilde{y}_\varepsilon'(t) + O(\varepsilon^N), \tag{D.25}$$

we can work with $\widetilde{T}_n(\varepsilon)$, given by

$$\widetilde{T}_0(\varepsilon) = 2\pi, \quad \widetilde{T}_{n+1}(\varepsilon) = \widetilde{T}_n(\varepsilon) - \frac{\tilde{y}_\varepsilon(\widetilde{T}_n(\varepsilon))}{\tilde{y}'(\widetilde{T}_n(\varepsilon))}, \tag{D.26}$$

and we get

$$T(\varepsilon) = \widetilde{T}_n(\varepsilon) + O(\varepsilon^N), \quad \text{provided } \ 2^n \ge N. \tag{D.27}$$

In particular, taking

$$y_\varepsilon(t) = a\sin t + Y_1(t)\varepsilon + O(\varepsilon^2), \tag{D.27}$$

we have

$$T(\varepsilon) = \widetilde{T}_1(\varepsilon) + O(\varepsilon^2), \tag{D.28}$$

with

$$\begin{aligned} \widetilde{T}_1(\varepsilon) &= 2\pi - \frac{\tilde{y}_\varepsilon(2\pi)}{\tilde{y}_\varepsilon'(2\pi)} \\ &= 2\pi + \frac{1}{a} Y_1(2\pi)\varepsilon, \end{aligned} \tag{D.29}$$

hence, by (D.17),

$$T(\varepsilon)=2\pi+\frac{\varepsilon}{a}\int_0^{2\pi}\cos s\,\psi(a\cos s)\,ds+O(\varepsilon^2). \tag{D.30}$$

In case $\psi(x)$ is given by (D.2), we have from (D.19) that

$$\int_0^{2\pi}\cos s\,\psi(a\cos s)\,ds=2\pi aA, \tag{D.31}$$

so in this case

$$T(\varepsilon)=2\pi(1+A\varepsilon)+O(\varepsilon^2). \tag{D.32}$$

Given an approximation $\tilde{y}_\varepsilon(t)$ satisfying (D.25) with $N=3$ or 4, we can iterate (D.26) once more, obtaining $T_2(\varepsilon)=T(\varepsilon)+O(\varepsilon^N)$, and so on. We will not pursue the details.

We now return to the problem of approximating the solution $(x_\varepsilon(t),y_\varepsilon(t))$ of (D.4), and address a limitation of the approximations of the form (D.8). As follows from (D.15)–(D.20), the first order approximation has the form

$$\begin{aligned}x_\varepsilon(t)&=a\cos t+X_1(t)\varepsilon+O(\varepsilon^2),\\ y_\varepsilon(t)&=-a\sin t+Y_1(t)\varepsilon+O(\varepsilon^2),\end{aligned} \tag{D.33}$$

and, in the case that $\psi(x)$ is given by (D.2),

$$\begin{aligned}X_1(t)&=X_1^b(t)+aAt\sin t,\\ Y_1(t)&=Y_1^b(t)-aAt\cos t,\end{aligned} \tag{D.34}$$

where $X_1^b(t)$ and $Y_1^b(t)$ are periodic in t, of period 2π, being sums of products of $\sin kt$ and $\cos kt$ $(0\le k\le 3)$. In (D.33), the notation $O(\varepsilon^2)$ means that, for any given bounded interval $[-K,K]$, the remainder is bounded by $C_K\varepsilon^2$, for $t\in[-K,K]$. However, it is apparent from (D.34) that the accuracy of this approximation breaks down severely on intervals of length $\approx 1/\varepsilon$. In fact, both $x_\varepsilon(t)$ and $y_\varepsilon(t)$ are uniformly bounded, being periodic of period $T(\varepsilon)$. As far as the terms on the right side of (D.33) are concerned,

$$\begin{aligned}&a\cos t+X_1^b(t)\varepsilon\quad\text{and}\\ &-a\sin t+Y_1^b(t)\varepsilon\end{aligned}$$

are uniformly bounded, of period 2π, but

$$aA\varepsilon t\sin t\quad\text{and}\quad-aA\varepsilon t\cos t \tag{D.35}$$

are unbounded as $|t| \to \infty$. These terms are called *secular terms*, and it is desirable to have a replacement for (D.8), in which such secular terms do not appear. To get this, we proceed as follows.

The functions

$$x_\varepsilon^\#(t) = x_\varepsilon\Big(\frac{T(\varepsilon)t}{2\pi}\Big), \quad y_\varepsilon^\#(t) = y_\varepsilon\Big(\frac{T(\varepsilon)t}{2\pi}\Big) \tag{D.36}$$

are periodic of period 2π in t and smooth in ε. Hence we have expansions

$$\begin{aligned} x_\varepsilon^\#(t) &= a\cos t + \sum_{k=1}^N X_k^\#(t)\varepsilon^k + O(\varepsilon^{N+1}), \\ y_\varepsilon^\#(t) &= -a\sin t + \sum_{k=1}^N Y_k^\#(t)\varepsilon^k + O(\varepsilon^{N+1}). \end{aligned} \tag{D.37}$$

Note that

$$\frac{d}{dt}x_\varepsilon^\#(t) = \frac{T(\varepsilon)}{2\pi}y_\varepsilon^\#(t), \tag{D.38}$$

which leads to a variant of (D.16). We have the following.

Proposition D.1. *The solution to (D.4)–(D.5) has the expansion*

$$\begin{aligned} x_\varepsilon(t) &= a\cos\frac{2\pi t}{T(\varepsilon)} + \sum_{k=1}^N X_k^\#\Big(\frac{2\pi t}{T(\varepsilon)}\Big)\varepsilon^k + O(\varepsilon^{N+1}), \\ y_\varepsilon(t) &= -a\sin\frac{2\pi t}{T(\varepsilon)} + \sum_{k=1}^N Y_k^\#\Big(\frac{2\pi t}{T(\varepsilon)}\Big)\varepsilon^k + O(\varepsilon^{N+1}). \end{aligned} \tag{D.39}$$

Each term in this series is periodic in t of period $T(\varepsilon)$, and the remainders are $O(\varepsilon^{N+1})$ uniformly for all $t \in \mathbb{R}$.

It is natural and convenient to set

$$X_0(t) = X_0^\#(t) = a\cos t, \quad Y_0(t) = Y_0^\#(t) = -a\sin t. \tag{D.40}$$

It remains to compute $X_k^\#(t)$ and $Y_k^\#(t)$ for $k \ge 1$. To this end, set

$$\frac{T(\varepsilon)}{2\pi} = 1 + \gamma(\varepsilon), \quad \gamma(\varepsilon) = \varepsilon\sum_{\ell\ge 0}\gamma_\ell\varepsilon^\ell. \tag{D.41}$$

If we compare the expressions for $x_\varepsilon(t)$ in (D.8) and (D.39) and make the substitution $s = 2\pi t/T(\varepsilon)$, we obtain

$$\begin{aligned}\sum_{k\ge 0} X_k^{\#}(s)\varepsilon^k &= \sum_{i\ge 0} X_i(s+\gamma(\varepsilon)s)\varepsilon^i \\ &= \sum_{i\ge 0}\sum_{j\ge 0}\frac{1}{j!}X_i^{(j)}(s)s^j\gamma(\varepsilon)^j\varepsilon^i \\ &= \sum_{i\ge 0}\sum_{j\ge 0}\frac{1}{j!}X_i^{(j)}(s)s^j\Bigl(\sum_{\ell\ge 0}\gamma_\ell\varepsilon^\ell\Bigr)^j\varepsilon^{i+j}.\end{aligned} \tag{D.42}$$

We conclude that $X_k^{\#}(s)$ is equal to the coefficient of ε^k in the last power series. For $k=0$, we get

$$X_0^{\#}(s) = X_0(s) = a\cos s, \tag{D.43}$$

as already noted in (D.40). For $k=1$, we get

$$\begin{aligned}X_1^{\#}(s) &= X_1(s) + \gamma_0 s X_0'(s) \\ &= X_1(s) - \gamma_0 a s\sin s.\end{aligned} \tag{D.44}$$

When $\psi(x)$ is given by (D.2), we have from (D.34) that this is

$$\begin{aligned}&= X_1^b(s) + aAs\sin s - \gamma_0 as\sin s \\ &= X_1^b(s),\end{aligned}$$

the last identity by (D.41) and (D.32), which gives $\gamma_0 = A$ in this case. Alternatively, since $X_1^{\#}(s)$ and $X_1^b(s)$ are periodic in s and the other terms are secular, these secular terms have to cancel. This holds for general $\psi(x)$; $X_1^{\#}(s)$ is obtained from $X_1(s)$ by striking out the secular terms. One can similarly characterize the higher order terms $X_k^{\#}(t)$ in (D.37). We forego the details.

We end this appendix with an indication of how to extend the scope of (D.1). We treat the pendulum equation

$$u'' + \sin u = 0, \tag{D.45}$$

and seek information on small oscillations, solving (D.45) with initial data

$$u(0) = a\sqrt{\varepsilon}, \quad u'(0) = 0. \tag{D.46}$$

Thus we set

$$x(t) = \sqrt{\varepsilon}\, u(t), \tag{D.47}$$

which solves

$$x'' + \frac{\sin \sqrt{\varepsilon} x}{\sqrt{\varepsilon}} = 0, \quad x(0) = a, \ x'(0) = 0. \tag{D.48}$$

If we set

$$\frac{\sin \tau}{\tau} = 1 - \tau^2 F(\tau), \quad F(\tau) = \frac{1}{3!} - \frac{\tau^2}{5!} + \cdots , \tag{D.49}$$

we get

$$\begin{aligned} x'' + x &= \varepsilon x^3 F(\sqrt{\varepsilon} x) \\ &= \varepsilon \frac{x^3}{3!} - \varepsilon^2 \frac{x^5}{5!} + \cdots . \end{aligned} \tag{D.50}$$

This has a form similar to (D.1), but generalized to

$$x'' + x = \varepsilon \psi(\varepsilon, x), \tag{D.51}$$

with ψ smooth in (ε, x). Treatments of the solutions to (D.1) and their periods $T(\varepsilon)$ extend to the case (D.51). The reader is invited to work out details.

E. A dram of potential theory

Newton's law of gravitation states that the force a particle of mass m_1 located at $p \in \mathbb{R}^3$ exerts on a particle of mass m_2 located at $x \in \mathbb{R}^3$ is

$$F(x) = G m_1 m_2 \frac{p - x}{\|p - x\|^3}. \tag{E.1}$$

Here G is the gravitational constant, given by (6.63). As indicated in Exercise 6 of §6, the force that a planet exerts on an external body is the same as what would be exerted if all the mass of the planet were concentrated at its center, in the Newtonian theory. In this appendix we explain why this is true, and in the course of doing so introduce an area of mathematical analysis known as potential theory. We establish this identity of force fields under the hypothesis that the mass distribution of the planet is spherically symmetric about its center. That is to say, we assume the planet, centered at p, has mass density ρ, and

$$\rho(p + Ry) = \rho(p + y), \quad \forall\, R \in O(3), \ y \in \mathbb{R}^3, \tag{E.2}$$

where we recall from Chapter 2 that $O(3)$ is the set of orthogonal transformations of $\mathbb{R}^3$. Say the planet has radius a, so

$$\|y\| > a \Longrightarrow \rho(p + y) = 0. \tag{E.3}$$

The planet's mass is

$$m_1 = \int \rho(y)\,dy. \tag{E.4}$$

If a particle of mass m_2 is located at $x \in \mathbb{R}^3$ and $\|p-x\| > a$, then the force the planet exerts on this particle is given by

$$G(x) = Gm_2 \int \frac{y-x}{\|y-x\|^3}\rho(y)\,dy. \tag{E.5}$$

We will show that if (E.2)–(E.4) hold and $\|p-x\| > a$, then $F(x) = G(x)$.

For notational simplicity, we may as well take

$$p = 0, \tag{E.6}$$

so

$$F(x) = -Gm_1m_2\frac{x}{\|x\|^3}. \tag{E.7}$$

Note that

$$F(x) = -\nabla V(x), \quad G(x) = -\nabla W(x), \tag{E.8}$$

with

$$V(x) = -\frac{Gm_1m_2}{\|x\|}, \quad W(x) = -Gm_2 \int\limits_{\|y\|\le a} \frac{1}{\|x-y\|}\rho(y)\,dy, \tag{E.9}$$

so it suffices to prove that these potential energies coincide for $\|x\| > a$, i.e.,

$$\|x\| > a \Longrightarrow V(x) = W(x). \tag{E.10}$$

As a first step toward proving (E.10), note that clearly, for all $R \in O(3)$,

$$V(Rx) = V(x), \tag{E.11}$$

and furthermore

$$\begin{aligned} W(Rx) &= -Gm_1 \int \frac{1}{\|Rx-y\|}\rho(y)\,dy \\ &= -Gm_1 \int \frac{1}{\|Rx-Rz\|}\rho(Rz)\,dz \\ &= -Gm_1 \int \frac{1}{\|x-z\|}\rho(z)\,dz \\ &= W(x), \end{aligned} \tag{E.12}$$

the second identity by change of variable and the third by (E.2). Consequently, we have

$$V(x)=v(r),\quad W(x)=w(r),\quad r=\|x\|, \tag{E.13}$$

and it remains to show that

$$r>a\Longrightarrow v(r)=w(r). \tag{E.14}$$

As another step toward showing this, we note that, given $a\in(0,\infty)$, there exists $C<\infty$ such that

$$\|y\|\le a,\ \|x\|\ge a+1\Longrightarrow\Bigl|\frac{1}{\|x\|}-\frac{1}{\|x-y\|}\Bigr|\le\frac{C}{\|x\|^2}, \tag{E.15}$$

and hence, by (E.4), (E.9) and (E.13), there exists $C_2<\infty$ such that

$$\begin{aligned} r=\|x\|\ge a+1&\Longrightarrow|V(x)-W(x)|\le\frac{C_2}{\|x\|^2}\\ &\Longrightarrow|v(r)-w(r)|\le\frac{C_2}{r^2}. \end{aligned} \tag{E.16}$$

The next step toward establishing (E.14) involves the following harmonicity,

$$\Delta V(x)=0,\quad\forall x\in\mathbb{R}^3\setminus 0, \tag{E.17}$$

where Δ is the Laplace operator,

$$\Delta f(x)=\frac{\partial^2 f}{\partial x_1^2}+\frac{\partial^2 f}{\partial x_2^2}+\frac{\partial^2 f}{\partial x_3^2}. \tag{E.18}$$

To see this, recall from (A.11) of Chapter 1 that (on $\mathbb{R}^3$)

$$f(x)=g(r)\Longrightarrow\Delta f(x)=g''(r)+\frac{2}{r}g'(r), \tag{E.19}$$

and by results on Euler equations from §15 of Chapter 1,

$$g''(r)+\frac{2}{r}g'(r)=0\Longleftrightarrow g(r)=\frac{c_1}{r}+c_2, \tag{E.20}$$

Since

$$V(x)=v(r)=-\frac{Gm_1m_2}{r}, \tag{E.21}$$

we have (E.17), and hence we also have

$$\Delta\Bigl(\frac{1}{\|x-y\|}\Bigr) = 0 \quad \text{for} \quad x \neq y, \tag{E.22}$$

so a direct consequence of the integral formula (E.9) for $W(x)$ is

$$\Delta W(x) = 0 \quad \text{for} \quad \|x\| > a. \tag{E.23}$$

Hence, by (E.13), (E.19), and (E.20),

$$\begin{aligned} r > a &\Longrightarrow w''(r) + \frac{2}{r} w'(r) = 0 \\ &\Longrightarrow w(r) = \frac{c_1}{r} + c_2, \end{aligned} \tag{E.24}$$

for some constants c_1 and c_2. This identity together with (E.21) and (E.16) proves (E.14). Hence we have (E.10), so indeed, under the hypotheses (E.2)–(E.4) (and with $p = 0$),

$$\|x\| > a \Longrightarrow F(x) = G(x). \tag{E.25}$$

We mention the following refinement of (E.23),

$$\Delta W = 4\pi G m_2 \rho. \tag{E.26}$$

This is not needed to establish (E.14), so we will not prove it here. A proof can be found in [**T**], Chapter 3, §4. Further exploration of the relation between the Laplace operator and the "potential" function W, through (E.9), leads to the subject of potential theory, addressed in Chapters 3–5 of [**T**] and in other books on partial differential equations.

The earth, the sun, and other planets and stars are approximately spherically symmetric, but not exactly so. This leads to further corrections in calculations in celestial mechanics. In addition, measurements of the strength of the earth's gravitational field give information on the inhomogeneities of the earth's composition, leading to the field of physical geodesy; cf. [**HM**].

F. Brouwer's fixed-point theorem

Here we prove the following fixed-point theorem of L. Brouwer, which arose in §15. Take

$$\overline{D} = \{x \in \mathbb{R}^2 : \|x\| \le 1\}. \tag{F.1}$$

Theorem F.1. *Each smooth map $F : \overline{D} \to \overline{D}$ has a fixed point.*

The proof proceeds by contradiction. We are claiming that $F(x) = x$ for some $x \in \overline{D}$. If not, then for each $x \in \overline{D}$ define $\varphi(x)$ to be the endpoint of the ray from $F(x)$ to x, continued until it hits

$$\partial D = \{x \in \mathbb{R}^2 : \|x\| = 1\}. \tag{F.2}$$

An explicit formula is

$$\begin{gathered}\varphi(x) = x + t(x - F(x)), \quad t = \frac{\sqrt{b^2 + 4ac} - b}{2a}, \\ a = \|x - F(x)\|^2, \quad b = 2x \cdot (x - F(x)), \quad c = 1 - \|x\|^2.\end{gathered} \tag{F.3}$$

Here t is picked to solve the equation $\|x + t(x - F(x))\|^2 = 1$. Note that $ac \geq 0$, so $t \geq 0$. It is clear that φ would have the following properties:

$$\varphi : \overline{D} \to \partial D \ \text{ smoothly}, \quad x \in \partial D \Rightarrow \varphi(x) = x. \tag{F.4}$$

Such a map is called a smooth retraction. The contradiction that proves Theorem F.1 is provided by the following result, called Brouwer's no-retraction theorem.

Theorem F.2. *There is no smooth retraction $\varphi : \overline{D} \to \partial D$ of $\overline{D}$ onto its boundary.*

Proof. This proof, also by contradiction, brings in material developed in §4. Suppose we had such a retraction φ. Consider the closed curve

$$\gamma : [0, 2\pi] \longrightarrow \partial D, \quad \gamma(t) = (\cos t, \sin t), \tag{F.5}$$

and form

$$\gamma_s(t) = \varphi(s\gamma(t)), \quad 0 \leq s \leq 1. \tag{F.6}$$

This would be a smooth family of maps

$$\gamma_s : [0, 2\pi] \longrightarrow \partial D, \quad \gamma_s(0) = \gamma_s(2\pi), \tag{F.7}$$

such that $\gamma_1 = \gamma$ and $\gamma_0(t) = \varphi(0)$ for all t. The variant of Lemma 4.2 given in Exercise 13 of §4 implies

$$\int\limits_{\gamma_s} F(y) \cdot dy \ \text{ is independent of } \ s \in [0, 1], \tag{F.7}$$

for each C^1 vector field F defined on a neighborhood of ∂D and satisfying (4.4). Clearly the line integral (F.7) is 0 for $s = 0$, so we deduce that

$$\int\limits_{\gamma} F(y) \cdot dy = 0 \tag{F.8}$$

for each such vector field. In particular, this would apply to the vector field given by (4.19)–(4.20), i.e.,

$$F(x) = \frac{1}{\|x\|^2} \begin{pmatrix} -x_2 \\ x_1 \end{pmatrix}, \tag{F.9}$$

which is smooth on $\mathbb{R}^2 \setminus 0$ and satisfies (4.4) (cf. (4.21)). On the other hand, we compute

$$\begin{aligned} \int_{\gamma} F(y) \cdot dy &= \int_0^{2\pi} (-\sin t, \cos t) \cdot (-\sin t, \cos t)\, dt \\ &= 2\pi, \end{aligned} \tag{F.10}$$

contradicting (F.8) and hence contradicting the existence of such a retraction.

The fixed-point theorem is valid for all continuous $F : \overline{D} \to \overline{D}$. In fact, an approximation argument, which we omit here, can be used to show that if such continuous F has no fixed point, there is a smooth approximation $\widetilde{F} : \overline{D} \to \overline{D}$ that would also have no fixed point.

Furthermore, Theorem F.1 holds in n dimensions, i.e., when

$$\overline{D} = \{x \in \mathbb{R}^n : \|x\| \le 1\}. \tag{F.11}$$

The reduction to Theorem F.2, in the setting of (F.11), is the same as above, but the proof of Theorem F.2 in the n-dimensional setting requires a further argument. Proofs using topology can be found in [**GrH**] and [**Mun**]. Proofs using differential forms can be found in [**Kan**], [**T**], Chapter 1, and [**T3**], Appendix G. We have no space to introduce differential forms here, but as shown in [**T**], and also in [**AM**] and [**Ar**], they give rise to many important results in the study of differential equations, at the next level.

References

[AM] R. Abraham and J. Marsden, *Foundations of Mechanics, Second Ed.*, Benjamin Cummings, Reading, Mass., 1978.

[AS] R. Abraham and C. Shaw, *Dynamics – The Geometry of Behavior, Vols. 1–3*, Aerial Press, Santa Cruz CA, 1984.

[ABS] R. Adler, M. Bazin, and M. Schiffer, *Introduction to General Relativity*, McGraw-Hill, New York, 1975.

[Ahl] L. Ahlfors, *Complex Analysis*, McGraw-Hill, New York, 1966.

[Ar] V. Arnold, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York, 1978.

[AP] D. Arrowsmith and C. Place, *An Introduction to Dynamical Systems*, Cambridge Univ. Press, Cambridge, 1990..

[At] K. Atkinson, *An introduction to Numerical Methods*, John Wiley and Sons, New York, 1978.

[BS] R. Bartle and D. Sherbert, *Introduction to Real Analysis*, J. Wiley and Sons, New York, 2000.

[BD] W. Boyce and R. DiPrima, *Elementary Differential Equations and Boundary Value Problems, Seventh Ed.*, John Wiley, New York, 2001.

[CL] E. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[DoC] M. DoCarmo, *Differential Geometry of Curves and Surfaces*, Prentice Hall, Englewood Cliffs, New Jersey, 1976.

[Gant] F. Gantmacher, *Applications of the Theory of Matrices*, Dover, New York, 2005.

[GMP] A. Gray, M. Mezzino, and M. Pinsky, *Introduction to Ordinary Differential Equations with Mathematica*, Springer-Verlag, New York, 1997.

[GrH] M. Greenberg and J. Harper, *Algebraic Topology, a First Course*, Addison-Wesley, New York, 1981.

[Gr] N. Grossman, *The Sheer Joy of Celestial Mechanics*, Birkhauser, Boston, 1996.

[GH] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.

[GM] J. Gundlach and S. Merkowitz, *Measurement of Newton's constant using a torsion balance with angular acceleration feedback*, Phys. Rev. Lett. **85** (2000,), 2869–2872.

[HK] J. Hale and H. Kocak, *Dynamics and Bifurcations*, Springer-Verlag, New York, 1991.

[Hart] P. Hartman, *Ordinary Differential Equations*, Baltimore, 1973.

[HM] W. Heiskanen and M. Moriz, *Physical Geodesy*, W. H. Freeman,, San Fransisco, 1967.

[Hen] D. Henderson, *Differential Geometry – A Geometric Introduction*, Prentice Hall, Upper Saddle River, New Jersey, 1998.

[HS] M. Hirsch and S. Smale, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, New York, 1974.

[HSD] M. Hirsch, S. Smale, and R. Devaney, *Differential Equations, Dynamical Systems, and an Introduction to Chaos*, Elsevier Academic Press, New York, 2004.

[HW] J. Hubbard and B. West, *Differential Equations: A Dynamical Systems Approach*, Springer-Verlag, New York, 1995.

[J] E. Jackson, *Perspectives of Nonlinear Dynamics, Vols. 1–2*, Cambridge Univ. Press, Cambridge, 1991.

[Kan] Y. Kannai, *An elementary proof of the no-retraction theorem*, Amer. Math. Monthly **88** (1981), 264–268.

[Kr] S. Krantz, *The Elements of Advanced Mathematics*, CRC Press, Boca Raton, 1995.

[Leb] N. Lebedev, *Special Functions and Their Applications*, Dover, New York, 1972.

[Lef] S. Lefschetz, *Differential Equations: General Theory*, Wiley-Interscience, New York, 1963.

[LL] A. Lichtenberg and M. Lieberman, *Regular and Chaotic Dynamics*, Springer-Verlag, New York, 1982.

[LS] L. Loomis and S. Sternberg, *Advanced Calculus*, Addison-Wesley, New York, 1968.

[Mun] J. Munkres, *Topology, a First Course*, Prentice Hall, Englewood Cliffs, NJ, 1974.

[Mur] J. Murray, *Mathematical Biology*, Springer-Verlag, New York, 1989.

[NM] M. Nowak and R. May, *Virus Dynamics – Mathematical Principles of Immunology and Virology*, Oxford Univ. Press, Oxford,, 2000.

[Op] J. Oprea, *Differential Geometry and its Applications*, Prentice Hall, Upper Saddle River, New Jersey, 1997.

[Per] L. Perko, *Differential Equations and Dynamical Systems*, Springer-Verlag, New York, 1991.

[PA] J. Polking and D. Arnold, *Ordinary Differential Equations Using MATLAB*, Prentice Hall, Upper Saddle River, NJ, 2003.

[PBA] J. Polking, A. Boggess, and D. Arnold, *Differential Equations with Boundary Value Problems*, Prentice-Hall, Upper Saddle River, NJ, 2006.

[Sh] L. Shampine, *Numerical Solution of Ordinary Differential Equations*, Chapman-Hall, New York, 1994.

[Sim] G. Simmons, *Differential Equations, with Applications and Historical Notes*, McGraw-Hill, New York, 1972.

[Sp] C. Sparrow, *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors*, Springer-Verlag, New York, 1982.

[Sto] J. Stoker, *Differential Geometry*, Wiley-Interscience, New York, 1969.

[Str] G. Strang, *Linear Algebra and its Applications*, Harcourt Brace Jovanovich, San Diego, 1988.

[Tau] C. Taubes, *Modeling Differential Equations in Biology*, Prentice-Hall, Upper Saddle River NJ, 2001.

[T] M. Taylor, *Partial Differential Equations, Vols. 1–3*, Springer-Verlag, New York, 1996.

[T2] M. Taylor, *Numbers*, http://www.math.unc.edu/Faculty/met/numbers.pdf.

[T3] M. Taylor, *Measure Theory and Integration, GSM #76*, American Math. Soc., Providence, RI, 2006.

[Wi] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer-Verlag, New York, 1990.

[W] D. Wodnarz, *Killer Cell Dynamics – Mathematical and Computational Approaches to Immunology*, Springer-Verlag, New York, 2007.