

# Introduction to Differential Equations

Michael Taylor

MATH. DEPT., UNC

*E-mail address:* `met@math.unc.edu`

2010 *Mathematics Subject Classification.* 34-01, 34A12, 34A30, 34A34

*Key words and phrases.* differential equations, power series, exponential function, vector spaces, linear transformations, matrices, matrix exponential, vector field, flows, phase plane, critical points, variational problems, Newton's laws, Lagrangian, Hamiltonian, chaos



---

# Contents

Preface	xi
Some basic notation	xv
Chapter 1. Single differential equations	1
1.1. The exponential and trigonometric functions	3
Exercises	9
1.2. First order linear equations	14
Exercises	17
1.3. Separable equations	18
Exercises	22
1.4. Second order equations – reducible cases	23
Exercises	24
1.5. Newton’s equations for motion in 1D	25
Exercises	27
1.6. The pendulum	29
Exercises	36
1.7. Motion with resistance	38
Exercises	39
1.8. Linearization	40
Exercises	41
1.9. Second order constant-coefficient linear equations – homogeneous	42
Exercises	46
1.10. Nonhomogeneous equations I – undetermined coefficients	48
Exercises	52
1.11. Forced pendulum – resonance	53

---

Exercises	55
1.12. Spring motion	57
Exercises	58
1.13. RLC circuits	59
Exercises	61
1.14. Nonhomogeneous equations II – variation of parameters	62
Exercises	64
1.15. Variable coefficient second order equations	65
Exercises	67
1.16. Bessel’s equation	69
Exercises	73
1.17. Higher order linear equations	75
Exercises	76
1.18. The Laplace transform	78
Exercises	85
1.A. The genesis of Bessel’s equation: PDE in polar coordinates	86
1.B. Euler’s gamma function	88
1.C. Differentiating power series	91
Chapter 2. Linear algebra	97
2.1. Vector spaces	99
Exercises	100
2.2. Linear transformations and matrices	102
Exercises	104
2.3. Basis and dimension	106
Exercises	109
2.4. Matrix representation of a linear transformation	111
Exercises	113
2.5. Determinants and invertibility	114
Exercises	119
2.6. Eigenvalues and eigenvectors	123
Exercises	124
2.7. Generalized eigenvectors and the minimal polynomial	125
Exercises	129
2.8. Triangular matrices	131
Exercises	133
2.9. Inner products and norms	135
Exercises	138
2.10. Norm, trace, and adjoint of a linear transformation	140

---

Exercises	142
2.11. Self-adjoint and skew-adjoint transformations	144
Exercises	146
2.12. Unitary and orthogonal transformations	148
Exercises	151
2.A. The Jordan canonical form	154
2.B. Schur's upper triangular representation	156
2.C. The fundamental theorem of algebra	157
Chapter 3. Linear systems of differential equations	159
3.1. The matrix exponential	161
Exercises	167
3.2. Exponentials and trigonometric functions	171
Exercises	172
3.3. First-order systems derived from higher-order equations	174
Exercises	175
3.4. Non-homogeneous equations and Duhamel's formula	177
Exercises	178
3.5. Simple electrical circuits	180
Exercises	183
3.6. Second-order systems	185
Exercises	191
3.7. Curves in $\mathbb{R}^3$ and the Frenet-Serret equations	192
Exercises	196
3.8. Variable coefficient systems	198
Exercises	203
3.9. Variation of parameters and Duhamel's formula	206
Exercises	207
3.10. Power series expansions	209
Exercises	215
3.11. Regular singular points	217
Exercises	226
3.A. Logarithms of matrices	229
3.B. The matrix Laplace transform	231
3.C. Complex analytic functions	234
Chapter 4. Nonlinear systems of differential equations	237
4.1. Existence and uniqueness of solutions	240
Exercises	245
4.2. Dependence of solutions on initial data and other parameters	249

---

Exercises	252
4.3. Vector fields, orbits, and flows	253
Exercises	268
4.4. Gradient vector fields	272
Exercises	277
4.5. Newtonian equations	279
Exercises	282
4.6. Central force problems and two-body planetary motion	283
Exercises	289
4.7. Variational problems and the stationary action principle	292
Exercises	297
4.8. The brachistochrone problem	302
Exercises	305
4.9. The double pendulum	307
Exercises	311
4.10. Momentum-quadratic Hamiltonian systems	312
Exercises	316
4.11. Numerical study – difference schemes	317
Exercises	322
4.12. Limit sets and periodic orbits	324
Exercises	333
4.13. Predator-prey equations	336
Exercises	348
4.14. Competing species equations	352
Exercises	356
4.15. Chaos in multidimensional systems	359
Exercises	369
4.A. The derivative in several variables	372
4.B. Convergence, compactness, and continuity	375
4.C. Critical points that are saddles	379
4.D. Blown up phase portrait at a critical point	388
4.E. Periodic solutions of $x'' + x = \varepsilon\psi(x)$	396
4.F. A dram of potential theory	402
4.G. Brouwer's fixed-point theorem	405
4.H. Geodesic equations on surfaces	407
4.I. Rigid body motion in $\mathbb{R}^n$ and geodesics on $SO(n)$	411
Bibliography	419
Index	423

---

# Preface

The first time I cracked open a differential equations text, it was instant love. I don't remember the exact title of the book, or the author's name, but I do remember that the book was thinner than the ones I have seen selected for use in introductory courses in recent times. It was a book a student could read from cover to cover, while taking a course in the subject. I began to write course notes, with the aim of producing a text more to my liking. After a few years of this, the current book emerged.

This book has four chapters. I use Chapter 1 and parts of Chapters 2 and 3 for a first semester introduction to differential equations, and I use the rest of Chapters 2 and 3 together with Chapter 4 for the second semester.

Chapter 1 deals with single differential equations, first equations of order 1,

$$(0.0.1) \quad \frac{dx}{dt} = f(t, x),$$

then equations of order 2,

$$(0.0.2) \quad \frac{d^2x}{dt^2} = f(t, x, x').$$

We have a brief discussion of higher order equations. For second order equations, we concentrate on the case

$$(0.0.3) \quad \frac{d^2x}{dt^2} = f(x, x'),$$

which can be reduced to a first order equation for  $v = x'$ , as a function of  $x$ . Newton's law  $F = ma$  for motion of a particle on a line gives such equations. We also specialize (0.0.2) to the linear case,

$$(0.0.4) \quad x'' + bx' + cx = f(t),$$

and discuss techniques for solving such equations.

While the study of single equations is the place to start, the subject of differential equations is and always has been mainly about *systems* of equations. This



study requires a healthy dose of linear algebra. For a number of good reasons, it is not desirable to require a course in linear algebra as a prerequisite (or even a corequisite) for a course in differential equations, but rather the course includes some basic instruction in linear algebra. Chapter 2 provides the needed minicourse in linear algebra. We differ from most introductions to differential equations in providing complete proofs of the relevant results, including material on determinants, eigenvalues and eigenvectors of a linear transformation, and also generalized eigenvectors.

Chapter 3 deals with linear systems of differential equations. We start with the  $n \times n$  system

$$(0.0.5) \quad \frac{dx}{dt} = Ax, \quad x(0) = x_0 \in \mathbb{C}^n,$$

where  $A$  is an  $n \times n$  matrix, and define the matrix exponential  $e^{tA}$ , which produces the solution

$$(0.0.6) \quad x(t) = e^{tA}x_0.$$

Material from Chapter 2 plays a central role in analyzing this matrix exponential. We proceed from (0.0.5) to the inhomogeneous system

$$(0.0.7) \quad \frac{dx}{dt} = Ax + f(t), \quad x(0) = x_0.$$

We also study variable coefficient equations

$$(0.0.8) \quad \frac{dx}{dt} = A(t)x + f(t).$$

In particular, we study power series expansions for the solution, when  $A(t)$  and  $f(t)$  are given by convergent power series. We also consider expansions when (0.0.8) has a “regular singular point.” These power series topics are usually introduced in the context of a single second order equation, before the study of systems. Indeed, in Chapter 1, §1.15 touches on this, and §1.16 goes into some detail in the important special case of Bessel’s equation. We have saved the general study for Chapter 3, both to speed the introduction to systems and because the presentation in the system context is both more compact and more general than in the context of a single, second order equation.

Chapter 4 crowns the text, with a study of nonlinear systems of differential equations. These can have the form

$$(0.0.9) \quad \frac{dx}{dt} = F(t, x), \quad x(t_0) = y,$$

which resembles (0.0.1), except that now  $x(t)$  and  $F(t, x)$  take values in  $\mathbb{R}^n$ . We begin with general existence and uniqueness results. For this, we convert (0.0.9) to the integral equation

$$(0.0.10) \quad x(t) = y + \int_{t_0}^t F(s, x(s)) ds.$$

and use the Picard iteration to produce the solution, for  $|t - t_0|$  subject to certain limitations, as a limit of a certain sequence of approximate solutions. This is followed by results on how large the interval of existence can be taken. Next, we

look into results on the smooth dependence of solutions  $x = x(t, y)$  to (0.0.9) on the initial data  $y$ . An important role is played by the linearization of (0.0.9).

From here we proceed to some qualitative studies of solutions, particularly in the autonomous case,  $F(t, x) = F(x)$ , in which we interpret  $F$  as a vector field, and the solution as the flow  $\Phi^t$  generated by this vector field. One useful tool is the phase portrait, which depicts the behavior of solution curves (also called orbits) for nonlinear  $n \times n$  systems. From the point of view of visualization, the portraits work particularly well for  $n = 2$ , and are also quite useful for  $n = 3$ .

We study a variety of problems from mathematical physics, including the planetary motion problem, for two bodies interacting by the gravitational force, whose solution by Newton was a seminal inspiration to the field of ODE. We also bring in further advances in the study of equations of physics, due to Euler and Lagrange, involving the variational method. This theory impacts both physical and geometrical applications of ODE, the latter including equations for geodesics on surfaces in  $\mathbb{R}^n$ .

By this point we are looking at nonlinear systems whose solutions are not necessarily amenable to formulas. In addition to qualitative studies of the nature of these solutions, numerical studies arise as an important tool. This is taken up in §4.11. We introduce difference schemes, with emphasis on the Runge-Kutta scheme, as a very useful computational tool.

In Sections 4.13–4.14 we turn to some problems arising in mathematical biology. This is followed with some results on systems with chaotic dynamics, which arise in dimension  $\geq 3$ . This chapter closes with a number of appendices, some providing useful background in calculus, and others taking up further topics in nonlinear systems of ODE.

We follow this introduction with a record of some standard notation that will be in use throughout the text.



---

## Some basic notation

$\mathbb{R}$  is the set of real numbers.

$\mathbb{C}$  is the set of complex numbers.

$\mathbb{Z}$  is the set of integers.

$\mathbb{Z}^+$  is the set of integers  $\geq 0$ .

$\mathbb{N}$  is the set of integers  $\geq 1$  (the “natural numbers”).

$\mathbb{Q}$  is the set of rational numbers.

$x \in \mathbb{R}$  means  $x$  is an element of  $\mathbb{R}$ , i.e.,  $x$  is a real number.

$(a, b)$  denotes the set of  $x \in \mathbb{R}$  such that  $a < x < b$ .

$[a, b]$  denotes the set of  $x \in \mathbb{R}$  such that  $a \leq x \leq b$ .

$\{x \in \mathbb{R} : a \leq x \leq b\}$  denotes the set of  $x$  in  $\mathbb{R}$  such that  $a \leq x \leq b$ .

$[a, b) = \{x \in \mathbb{R} : a \leq x < b\}$  and  $(a, b] = \{x \in \mathbb{R} : a < x \leq b\}$ .

$\bar{z} = x - iy$  if  $z = x + iy \in \mathbb{C}$ ,  $x, y \in \mathbb{R}$ .

$f : A \rightarrow B$  denotes that the function  $f$  takes points in the set  $A$  to points in  $B$ . One also says  $f$  maps  $A$  to  $B$ .

$x \rightarrow x_0$  means the variable  $x$  tends to the limit  $x_0$ .

## Single differential equations

This first chapter is devoted to differential equations for a single unknown function, with emphasis on equations of the first and second order, i.e.,

$$(1.0.1) \quad \frac{dx}{dt} = f(t, x),$$

and

$$(1.0.2) \quad \frac{d^2x}{dt^2} = f\left(t, x, \frac{dx}{dt}\right).$$

Section 1.1 looks at the simplest case of (1.0.1), namely

$$(1.0.3) \quad \frac{dx}{dt} = cx.$$

We construct the solution  $x(t)$  to (1.0.3) such that  $x(0) = 1$  as a power series, defining the exponential function

$$(1.0.4) \quad x(t) = e^{ct}.$$

More generally,  $x(t) = e^{ct}$  solves  $dx/dt = cx$ , with  $x(0) = 1$ . This holds for all real  $c$  and also for *complex*  $c$ . Taking  $c = i$  and investigating basic properties of  $x(t) = e^{it}$ , we establish Euler's formula,

$$(1.0.5) \quad e^{it} = \cos t + i \sin t,$$

which in turn leads to a self-contained exposition of basic results on the trigonometric functions.

Section 1.2 treats first order linear equations, of the form

$$(1.0.6) \quad \frac{dx}{dt} + a(t)x = b(t), \quad x(t_0) = x_0,$$

and produces solutions in terms of the exponential function and integrals. Section 1.3 considers some nonlinear first order equations, particularly equations for which "separation of variables" allows one to produce a solution, in terms of various integrals.

We differ from many introductions in not lingering on the topic of first order equations. For example, we do not treat exact equations and integrating factors in this chapter. We consider it more important to get on to the study of second order equations. In any case, exact equations do get their due, in §4.4 of Chapter 4.

In §1.4 we take up second order differential equations. We concentrate there on two special classes, each allowing for a reduction to first order equations. In §1.5 we consider differential equations arising from some physical problems for motion in one space dimension, making use of Newton's law  $F = ma$ . The equations that arise in this context are amenable to methods of §1.4. In §1.5 we restate these methods in terms that celebrate the physical quantities of kinetic and potential energy, and the conservation of total energy. Section 1.6 deals with the classical pendulum, a close relative of motion on a line. In §1.7 we discuss motion in the presence of resistance, including the pendulum with resistance.

Formulas from §1.6 give rise to complicated integrals, and problems of §1.7 have additional complications. These complications arise because of nonlinearities in the equations. In §1.8 we discuss "linearization" of these equations. The associated linear differential equations are amenable to explicit analysis.

Sections 1.9–1.15 are devoted to linear second order differential equations, starting with constant coefficient equations

$$(1.0.7) \quad a \frac{d^2x}{dt^2} + b \frac{dx}{dt} + cx = f(t),$$

first with  $f \equiv 0$  in §1.9, then allowing  $f$  to be nonzero. In §1.10 we consider certain special forms of  $f(t)$ , including

$$(1.0.8) \quad e^{\kappa t}, \quad \sin \sigma t, \quad \cos \sigma t, \quad t^k,$$

treating these cases by the "method of undetermined coefficients." We discuss implications of results here, when  $f(t) = A \sin \sigma t$ , for the forced, linearized pendulum, in §1.11. Sections 1.12–1.13 treat other physical problems leading to equations of the form (1.0.7), namely spring motion problems and models of certain simple electrical circuits, called RLC circuits. In §1.14 we bring up another method, "variation of parameters," which applies to general functions  $f$  in (1.0.7).

Section 1.15 gives some results on variable coefficient second order linear differential equations. Techniques brought to bear on these equations include power series representations, extending the power series attack used on (1.0.3), and the Wronskian, first introduced in the constant-coefficient context in §1.12. In §1.16 we concentrate on a particularly important second-order ODE with variable coefficients, Bessel's equation, further pushing power series techniques and the use of the Wronskian. In §1.17 we discuss differential equations of order  $\geq 3$ . In §1.18 we introduce the Laplace transform as a tool to treat nonhomogeneous differential equations, such as (1.0.7) and higher order variants. Material introduced in §§1.15–1.18 will be covered, on a much more general level, in Chapter 3.

We end this chapter with three appendices. Appendix 1.A explains how Bessel functions arise in the search for solutions to some basic partial differential equations. Appendix 1.B has some basic material on Euler's gamma function, of use in §1.16. Appendix 1.C establishes that convergent power series can be differentiated term by term. We also derive the power series of  $f(t) = (1 - t)^{-r}$ .

### 1.1. The exponential and trigonometric functions

We construct the exponential function to solve the differential equation

$$(1.1.1) \quad \frac{dx}{dt} = x, \quad x(0) = 1.$$

We seek a solution as a power series

$$(1.1.2) \quad x(t) = \sum_{k=0}^{\infty} a_k t^k.$$

If such a power series converges for  $t$  in an interval in  $\mathbb{R}$ , it can be differentiated term-by-term. (See (1.1.45)–(1.1.50) below, and also §1.C, for more on this.) In such a case,

$$(1.1.3) \quad \begin{aligned} x'(t) &= \sum_{k=1}^{\infty} k a_k t^{k-1} \\ &= \sum_{\ell=0}^{\infty} (\ell + 1) a_{\ell+1} t^{\ell}, \end{aligned}$$

so for (1.1.1) to hold we need

$$(1.1.4) \quad a_0 = 1, \quad a_{k+1} = \frac{a_k}{k+1},$$

i.e.,  $a_k = 1/k!$ , where  $k! = k(k-1) \cdots 2 \cdot 1$ . Thus (1.1.1) is solved by

$$(1.1.5) \quad x(t) = e^t = \sum_{k=0}^{\infty} \frac{1}{k!} t^k, \quad t \in \mathbb{R}.$$

This defines the exponential function  $e^t$ .

More generally, we can define

$$(1.1.6) \quad e^z = \sum_{k=0}^{\infty} \frac{1}{k!} z^k, \quad z \in \mathbb{C}.$$

The issue of convergence for complex power series is essentially the same as for real power series. Given  $z = x + iy$ ,  $x, y \in \mathbb{R}$ , we have  $|z| = \sqrt{x^2 + y^2}$ . If also  $w \in \mathbb{C}$ , then  $|z + w| \leq |z| + |w|$  and  $|zw| = |z| \cdot |w|$ . Hence

$$\left| \sum_{k=m}^{m+n} \frac{1}{k!} z^k \right| \leq \sum_{k=m}^{m+n} \frac{1}{k!} |z|^k.$$

The ratio test then shows that the series (1.1.6) is absolutely convergent for all  $z \in \mathbb{C}$ , and uniformly convergent for  $|z| \leq R$ , for each  $R < \infty$ . Note that

$$(1.1.7) \quad e^{at} = \sum_{k=0}^{\infty} \frac{a^k}{k!} t^k$$

solves

$$(1.1.8) \quad \frac{d}{dt} e^{at} = a e^{at},$$

and this works for each  $a \in \mathbb{C}$ .



We claim that  $e^{at}$  is the only solution to

$$(1.1.9) \quad \frac{dy}{dt} = ay, \quad y(0) = 1.$$

To see this, compute the derivative of  $e^{-at}y(t)$ :

$$(1.1.10) \quad \frac{d}{dt}(e^{-at}y(t)) = -ae^{-at}y(t) + e^{-at}ay(t) = 0,$$

where we use the product rule, (1.1.8) (with  $a$  replaced by  $-a$ ) and (1.1.9). Thus  $e^{-at}y(t)$  is independent of  $t$ . Evaluating at  $t = 0$  gives

$$(1.1.11) \quad e^{-at}y(t) = 1, \quad \forall t \in \mathbb{R},$$

whenever  $y(t)$  solves (1.1.9). Since  $e^{at}$  solves (1.1.9), we have  $e^{-at}e^{at} = 1$ , hence

$$(1.1.12) \quad e^{-at} = \frac{1}{e^{at}}, \quad \forall t \in \mathbb{R}, \quad a \in \mathbb{C}.$$

Thus multiplying both sides of (1.1.11) by  $e^{at}$  gives the asserted uniqueness:

$$(1.1.13) \quad y(t) = e^{at}, \quad \forall t \in \mathbb{R}.$$

We can draw further useful conclusions from applying  $d/dt$  to products of exponential functions. In fact, let  $a, b \in \mathbb{C}$ ; then

$$(1.1.14) \quad \begin{aligned} & \frac{d}{dt} \left( e^{-at} e^{-bt} e^{(a+b)t} \right) \\ &= -ae^{-at} e^{-bt} e^{(a+b)t} - be^{-at} e^{-bt} e^{(a+b)t} + (a+b)e^{-at} e^{-bt} e^{(a+b)t} \\ &= 0, \end{aligned}$$

so again we are differentiating a function that is independent of  $t$ . Evaluation at  $t = 0$  gives

$$(1.1.15) \quad e^{-at} e^{-bt} e^{(a+b)t} = 1, \quad \forall t \in \mathbb{R}.$$

Using (1.1.12), we get

$$(1.1.16) \quad e^{(a+b)t} = e^{at} e^{bt}, \quad \forall t \in \mathbb{R}, \quad a, b \in \mathbb{C},$$

or, setting  $t = 1$ ,

$$(1.1.17) \quad e^{a+b} = e^a e^b, \quad \forall a, b \in \mathbb{C}.$$

We next record some properties of  $\exp(t) = e^t$  for real  $t$ . The power series (1.1.5) clearly gives  $e^t > 0$  for  $t \geq 0$ . Since  $e^{-t} = 1/e^t$ , we see that  $e^t > 0$  for all  $t \in \mathbb{R}$ . Since  $de^t/dt = e^t > 0$ , the function is monotone increasing in  $t$ , and since  $d^2e^t/dt^2 = e^t > 0$ , this function is convex. Note that

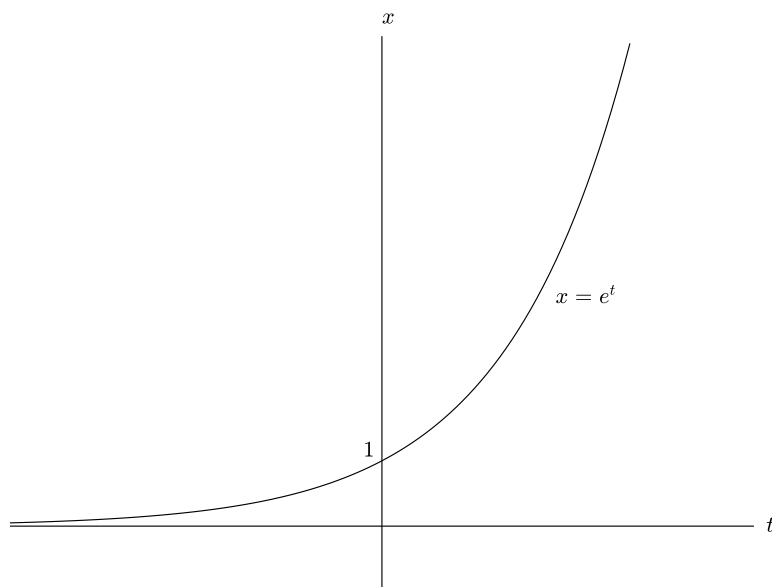
$$(1.1.18) \quad e^1 = 1 + 1 + \frac{1}{2} + \cdots > 2,$$

so  $e^k > 2^k \nearrow +\infty$  as  $k \rightarrow +\infty$ . Hence

$$(1.1.19) \quad \lim_{t \rightarrow +\infty} e^t = +\infty.$$

Since  $e^{-t} = 1/e^t$ ,

$$(1.1.20) \quad \lim_{t \rightarrow -\infty} e^t = 0.$$



**Figure 1.1.1.** Exponential function

As a consequence,

$$(1.1.21) \quad \exp : \mathbb{R} \longrightarrow (0, \infty)$$

is smooth and one-to-one and onto, with positive derivative, so the inverse function theorem of one-variable calculus applies. There is a smooth inverse

$$(1.1.22) \quad L : (0, \infty) \longrightarrow \mathbb{R}.$$

We call this inverse the natural logarithm:

$$(1.1.23) \quad \log x = L(x).$$

See Figures 1.1.1 and 1.1.2 for graphs of  $x = e^t$  and  $t = \log x$ .

Applying  $d/dt$  to

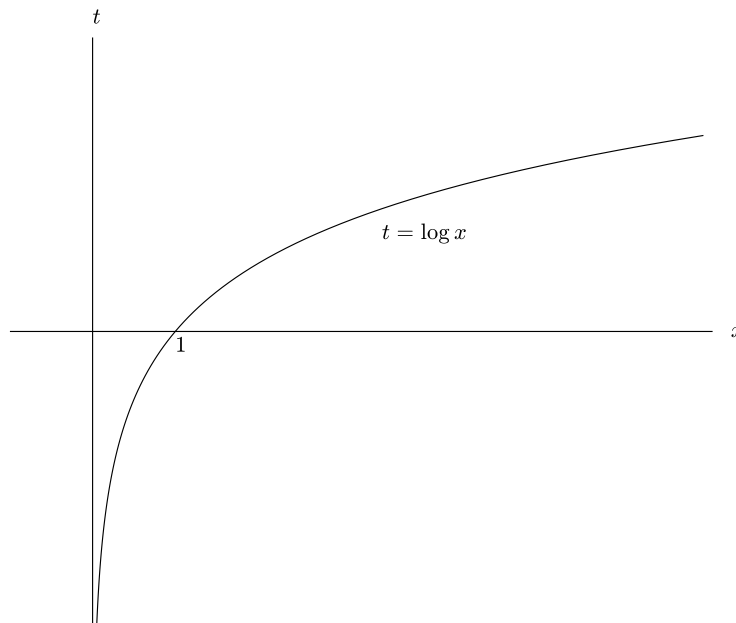
$$(1.1.24) \quad L(e^t) = t$$

gives

$$(1.1.25) \quad L'(e^t)e^t = 1, \quad \text{hence } L'(e^t) = \frac{1}{e^t},$$

i.e.,

$$(1.1.26) \quad \frac{d}{dx} \log x = \frac{1}{x}.$$



**Figure 1.1.2.** The logarithm

Since  $\log 1 = 0$ , we get

$$(1.1.27) \quad \log x = \int_1^x \frac{dy}{y}.$$

An immediate consequence of (1.1.17) (for  $a, b \in \mathbb{R}$ ) is the identity

$$(1.1.28) \quad \log xy = \log x + \log y, \quad x, y \in (0, \infty).$$

We move on to a study of  $e^z$  for purely imaginary  $z$ , i.e., of

$$(1.1.29) \quad \gamma(t) = e^{it}, \quad t \in \mathbb{R}.$$

This traces out a curve in the complex plane, and we want to understand which curve it is. Let us set

$$(1.1.30) \quad e^{it} = c(t) + is(t),$$

with  $c(t)$  and  $s(t)$  real valued. First we calculate  $|e^{it}|^2 = c(t)^2 + s(t)^2$ . For  $x, y \in \mathbb{R}$ ,

$$(1.1.31) \quad z = x + iy \implies \bar{z} = x - iy \implies z\bar{z} = x^2 + y^2 = |z|^2.$$

It is elementary that

$$(1.1.32) \quad \begin{aligned} z, w \in \mathbb{C} \implies \overline{z\bar{w}} = \bar{z}\bar{\bar{w}} \implies \overline{z^n} = \bar{z}^n, \\ \text{and } \overline{z + w} = \bar{z} + \bar{w}. \end{aligned}$$

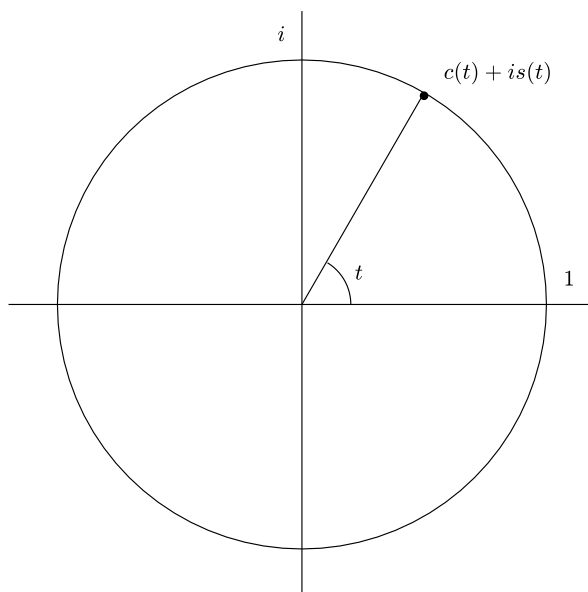


Figure 1.1.3. Behind Euler's formula

Hence

$$(1.1.33) \quad \overline{e^z} = \sum_{k=0}^{\infty} \frac{\overline{z^k}}{k!} = e^{\overline{z}}.$$

In particular,

$$(1.1.34) \quad t \in \mathbb{R} \implies |e^{it}|^2 = e^{it}e^{-it} = 1.$$

Hence  $t \mapsto \gamma(t) = e^{it}$  has image in the unit circle centered at the origin in  $\mathbb{C}$ . Also

$$(1.1.35) \quad \gamma'(t) = ie^{it} \implies |\gamma'(t)| \equiv 1,$$

so  $\gamma(t)$  moves at unit speed on the unit circle. We have

$$(1.1.36) \quad \gamma(0) = 1, \quad \gamma'(0) = i.$$

Thus, for  $t$  between 0 and the circumference of the unit circle, the arc from  $\gamma(0)$  to  $\gamma(t)$  is an arc on the unit circle, pictured in Figure 1.1.3, of length

$$(1.1.37) \quad \ell(t) = \int_0^t |\gamma'(s)| ds = t.$$

Standard definitions from trigonometry say that the line segments from 0 to 1 and from 0 to  $\gamma(t)$  meet at angle whose measurement in radians is equal to the length of the arc of the unit circle from 1 to  $\gamma(t)$ , i.e., to  $\ell(t)$ . The cosine of this

angle is defined to be the  $x$ -coordinate of  $\gamma(t)$  and the sine of the angle is defined to be the  $y$ -coordinate of  $\gamma(t)$ . Hence the computation (1.1.37) gives

$$(1.1.38) \quad c(t) = \cos t, \quad s(t) = \sin t.$$

Thus (1.1.30) becomes

$$(1.1.39) \quad e^{it} = \cos t + i \sin t,$$

which is Euler's formula. The identity

$$(1.1.40) \quad \frac{d}{dt} e^{it} = i e^{it},$$

applied to (1.1.39), yields

$$(1.1.41) \quad \frac{d}{dt} \cos t = -\sin t, \quad \frac{d}{dt} \sin t = \cos t.$$

We can use (1.1.17) to derive formulas for sin and cos of the sum of two angles. Indeed, comparing

$$(1.1.42) \quad e^{i(s+t)} = \cos(s+t) + i \sin(s+t)$$

with

$$(1.1.43) \quad e^{is} e^{it} = (\cos s + i \sin s)(\cos t + i \sin t)$$

gives

$$(1.1.44) \quad \begin{aligned} \cos(s+t) &= (\cos s)(\cos t) - (\sin s)(\sin t), \\ \sin(s+t) &= (\sin s)(\cos t) + (\cos s)(\sin t). \end{aligned}$$

Returning to basics, we recall that the calculations done so far in this section were all predicated on the fact that the power series (1.1.7) can be differentiated term by term. This is a special case of a general result about convergent power series, established in §1.C. However, making use of the special structure of (1.1.7), we include a direct demonstration here. To begin, look at

$$(1.1.45) \quad E_n^a(t) = \sum_{k=0}^n \frac{a^k}{k!} t^k,$$

which satisfies

$$(1.1.46) \quad \begin{aligned} \frac{d}{dt} E_n^a(t) &= \sum_{k=1}^n \frac{a^k}{(k-1)!} t^{k-1} \\ &= \sum_{\ell=0}^{n-1} \frac{a^{\ell+1}}{\ell!} t^\ell \\ &= a E_{n-1}^a(t). \end{aligned}$$

Integration gives

$$(1.1.47) \quad a \int_0^t E_{n-1}^a(s) ds = E_n^a(t) - 1.$$

Now we have

$$(1.1.48) \quad E_{n-1}^a(s) \longrightarrow e^{as}, \quad E_n^a(t) \longrightarrow e^{at},$$

uniformly on finite intervals, as  $n \rightarrow \infty$ , and then the integral estimate

$$\left| \int_0^t (E(s) - F(s)) ds \right| \leq |t| \max_{0 \leq s \leq t} |E(s) - F(s)|$$

implies

$$(1.1.49) \quad \int_0^t E_{n-1}^a(s) ds \longrightarrow \int_0^t e^{as} ds,$$

as  $n \rightarrow \infty$ . Consequently, we can pass to the limit  $n \rightarrow \infty$  in (1.1.47) and get

$$(1.1.50) \quad a \int_0^t e^{as} ds = e^{at} - 1.$$

Applying  $d/dt$  to the left side of (1.1.50) gives  $ae^{at}$ , by the fundamental theorem of calculus. Hence this must be the derivative of the right side of (1.1.50), and this gives (1.1.8).

Having the integral formula (1.1.50), we proceed to obtain formulas for  $\int t^n e^{at} dt$ . In fact, from (1.1.46), (1.1.8), and the product rule, we obtain

$$(1.1.51) \quad \begin{aligned} \frac{d}{dt}(e^{-at} E_n^a(t)) &= -ae^{-at} E_n^a(t) + ae^{-at} E_{n-1}^a(t) \\ &= -\frac{a^{n+1}}{n!} t^n e^{-at}. \end{aligned}$$

Then the fundamental theorem of calculus gives

$$(1.1.52) \quad \begin{aligned} \int t^n e^{-at} dt &= -\frac{n!}{a^{n+1}} E_n^a(t) e^{-at} + C \\ &= -\frac{n!}{a^{n+1}} \left( 1 + at + \frac{a^2 t^2}{2!} + \cdots + \frac{a^n t^n}{n!} \right) e^{-at} + C. \end{aligned}$$

We have an analogous formula for  $\int t^n e^{at} dt$ , by replacing  $a$  by  $-a$ .

## Exercises

1. As noted, if  $z = x + iy$ ,  $x, y \in \mathbb{R}$ , then  $|z| = \sqrt{x^2 + y^2}$  is equivalent to  $|z|^2 = z\bar{z}$ . Use this to show that if also  $w \in \mathbb{C}$ ,

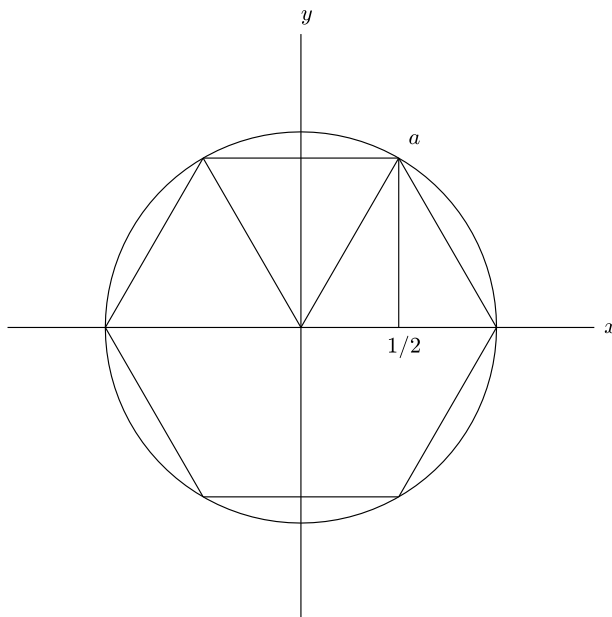
$$|zw| = |z| \cdot |w|.$$

Note that

$$\begin{aligned} |z + w|^2 &= (z + w)(\bar{z} + \bar{w}) \\ &= |z|^2 + |w|^2 + w\bar{z} + z\bar{w} \\ &= |z|^2 + |w|^2 + 2 \operatorname{Re} z\bar{w}. \end{aligned}$$

Show that  $\operatorname{Re}(z\bar{w}) \leq |zw|$  and use this in concert with an expansion of  $(|z| + |w|)^2$  and the first identity above to deduce that

$$|z + w| \leq |z| + |w|.$$



**Figure 1.1.4.** Hexagon

2. Define  $\pi$  to be the smallest positive number such that  $e^{\pi i} = -1$ . Show that

$$e^{\pi i/2} = i, \quad e^{\pi i/3} = \frac{1}{2} + \frac{\sqrt{3}}{2}i.$$

*Hint.* See Figure 1.1.4, showing  $a = e^{\pi i/3}$ .

3. Show that

$$\cos^2 t + \sin^2 t = 1,$$

and

$$1 + \tan^2 t = \sec^2 t,$$

where

$$\tan t = \frac{\sin t}{\cos t}, \quad \sec t = \frac{1}{\cos t}.$$

4. Show that

$$\frac{d}{dt} \tan t = \sec^2 t = 1 + \tan^2 t,$$

$$\frac{d}{dt} \sec t = \sec t \tan t.$$

5. Evaluate

$$\int_0^y \frac{dx}{1+x^2}.$$

*Hint.* Set  $x = \tan t$ .

6. Evaluate

$$\int_0^y \frac{dx}{\sqrt{1-x^2}}.$$

*Hint.* Set  $x = \sin t$ .

7. Show that

$$\frac{\pi}{6} = \int_0^{1/2} \frac{dx}{\sqrt{1-x^2}}.$$

*Hint.* Show that  $\sin \pi/6 = 1/2$ . Use Exercise 2 and the identity  $e^{\pi i/6} = e^{\pi i/2} e^{-\pi i/3}$ .

8. Set

$$\cosh t = \frac{1}{2}(e^t + e^{-t}), \quad \sinh t = \frac{1}{2}(e^t - e^{-t}).$$

Show that

$$\frac{d}{dt} \cosh t = \sinh t, \quad \frac{d}{dt} \sinh t = \cosh t,$$

and

$$\cosh^2 t - \sinh^2 t = 1.$$

9. Evaluate

$$\int_0^y \frac{dx}{\sqrt{1+x^2}}.$$

*Hint.* Set  $x = \sinh t$ .

10. Evaluate

$$\int_0^y \sqrt{1+x^2} dx.$$

11. Using Exercise 4, verify that

$$\begin{aligned} \frac{d}{dt}(\sec t + \tan t) &= \sec t(\sec t + \tan t), \\ \frac{d}{dt}(\sec t \tan t) &= \sec^3 t + \sec t \tan^2 t, \\ &= 2 \sec^3 t - \sec t. \end{aligned}$$

12. Next verify that

$$\begin{aligned} \frac{d}{dt} \log |\sec t| &= \tan t, \\ \frac{d}{dt} \log |\sec t + \tan t| &= \sec t. \end{aligned}$$



13. Now verify that

$$\begin{aligned}\int \tan t \, dt &= \log |\sec t|, \\ \int \sec t \, dt &= \log |\sec t + \tan t|, \\ 2 \int \sec^3 t \, dt &= \sec t \tan t + \int \sec t \, dt.\end{aligned}$$

(Here, we omit the arbitrary additive constants.)

14. Here is another approach to the evaluation of  $\int \sec t \, dt$ . We evaluate

$$I(u) = \int_0^u \frac{dv}{\sqrt{1+v^2}}$$

in two ways.

(a) Using  $v = \sinh y$ , show that

$$I(u) = \int_0^{\sinh^{-1} u} dy = \sinh^{-1} u.$$

(b) Using  $v = \tan t$ , show that

$$I(u) = \int_0^{\tan^{-1} u} \sec t \, dt.$$

Deduce that

$$\int_0^x \sec t \, dt = \sinh^{-1}(\tan x), \quad \text{for } |x| < \frac{\pi}{2}.$$

Deduce from the formula above that also

$$\cosh\left(\int_0^x \sec t \, dt\right) = \sec x,$$

and hence that

$$\exp\left(\int_0^x \sec t \, dt\right) = \sec x + \tan x.$$

Compare these formulas with the analogue in Exercise 13.

15. For  $E_n^a(t)$  as in (1.1.45),  $k \geq 1$ ,  $0 < T < \infty$ , show that

$$(1.1.53) \quad \max_{|t| \leq T} |E_{n+k}^a(t) - E_n^a(t)| \leq \frac{|aT|^{n+1}}{(n+1)!} \left(1 + \frac{|aT|}{n+2} + \frac{|aT|^2}{(n+2)(n+3)} + \cdots\right),$$

and that this is

$$(1.1.54) \quad \leq 2 \frac{|aT|^{n+1}}{(n+1)!}, \quad \text{for } n+2 > 2|aT|.$$

Deduce that

$$(1.1.55) \quad \max_{|t| \leq T} |e^{at} - E_n^a(t)|$$

satisfies (1.1.54). Show that, for each  $a, T$ , (1.1.54) tends to 0 as  $n \rightarrow \infty$ , yielding the assertion made about convergence in (1.1.48).

16. Show that

$$\left| \int_0^t e^{as} ds - \int_0^t E_n^a(s) ds \right| \leq |t| \max_{|s| \leq |t|} |e^{as} - E_n^a(s)|,$$

and observe how this, together with Exercise 15, yields (1.1.49).

17. Show that

$$(1.1.56) \quad |t| < 1 \Rightarrow \log(1+t) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} t^k = t - \frac{t^2}{2} + \frac{t^3}{3} - \dots.$$

*Hint.* Rewrite (1.1.27) as

$$\log(1+t) = \int_0^t \frac{ds}{1+s},$$

expand

$$\frac{1}{1+s} = 1 - s + s^2 - s^3 + \dots, \quad |s| < 1,$$

and integrate term by term.

18. Use (1.1.52) with  $a = -i$  to produce formulas for

$$\int t^n \cos t dt \quad \text{and} \quad \int t^n \sin t dt.$$

19. Figure 1.1.5 (a)–(b) shows graphs of the image of

$$\gamma(t) = e^{\alpha t}, \quad 0 \leq t \leq 6\pi,$$

for

$$\alpha = -\frac{1}{4} + i,$$

$$\alpha = -\frac{1}{8} - i.$$

Match each value of  $\alpha$  to (a) or (b).

20. Given  $t > 0$  and  $a \in \mathbb{C}$ , we define  $t^a$  by

$$t^a = e^{a \log t}.$$

Show that, for  $t > 0$ ,

$$\frac{d}{dt} t^a = a t^{a-1}.$$

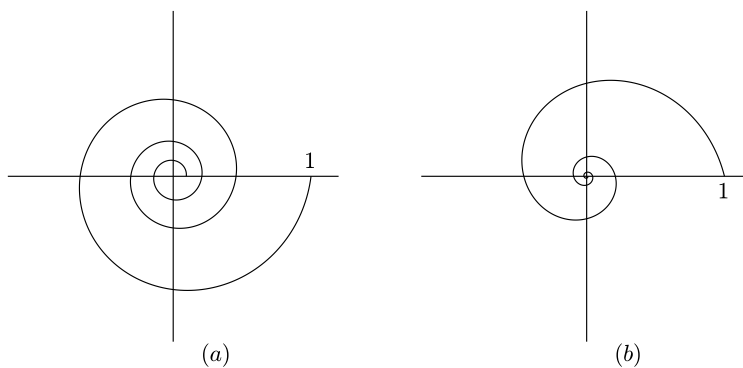


Figure 1.1.5. Spirals

## 1.2. First order linear equations

Here we tackle first order linear equations. These are equations of the form

$$(1.2.1) \quad \frac{dx}{dt} + a(t)x = b(t), \quad x(t_0) = x_0,$$

given functions  $a(t)$  and  $b(t)$ , continuous on some interval containing  $t_0$ . As a warm-up, we first treat

$$(1.2.2) \quad \frac{dx}{dt} + ax = b, \quad x(0) = x_0,$$

with  $a$  and  $b$  constants. One key to solving (1.2.2) is the identity

$$(1.2.3) \quad \frac{d}{dt}(e^{at}x) = e^{at}\left(\frac{dx}{dt} + ax\right),$$

which follows by applying the product formula and (1.1.8). Thus, multiplying both sides of (1.2.2) by  $e^{at}$  gives

$$(1.2.4) \quad \frac{d}{dt}(e^{at}x) = e^{at}b,$$

and then integrating both sides from 0 to  $t$  gives

$$(1.2.5) \quad e^{at}x(t) = x_0 + \int_0^t e^{as}b \, ds.$$

We can carry out the integral, using (1.1.45), and get

$$(1.2.6) \quad e^{at}x(t) = x_0 + \frac{e^{at} - 1}{a}b,$$

and finally division by  $e^{at}$  yields

$$(1.2.7) \quad \begin{aligned} x(t) &= e^{-at}x_0 + \frac{b}{a}(1 - e^{-at}) \\ &= \frac{b}{a} + e^{-at}\left(x_0 - \frac{b}{a}\right). \end{aligned}$$

In order to tackle (1.2.1), we need a replacement for (1.2.3). To get it, note that if  $A(t)$  is differentiable, the chain rule plus (1.1.8) gives

$$(1.2.8) \quad \frac{d}{dt}e^{A(t)} = e^{A(t)}A'(t).$$

Hence

$$(1.2.9) \quad \frac{d}{dt}(e^{A(t)}x) = e^{A(t)}\left(\frac{dx}{dt} + A'(t)x\right).$$

Thus we can multiply (1.2.1) by  $e^{A(t)}$  and get

$$(1.2.10) \quad \frac{d}{dt}(e^{A(t)}x) = e^{A(t)}b(t),$$

provided

$$(1.2.11) \quad A'(t) = a(t).$$

To arrange this, we can set

$$(1.2.12) \quad A(t) = \int_{t_0}^t a(s) ds.$$

Then we can integrate (1.2.10) from  $t_0$  to  $t$ , to get

$$(1.2.13) \quad e^{A(t)}x(t) = x_0 + \int_{t_0}^t e^{A(s)}b(s) ds,$$

and hence

$$(1.2.14) \quad x(t) = e^{-A(t)}x_0 + e^{-A(t)} \int_{t_0}^t e^{A(s)}b(s) ds.$$

For example, consider

$$(1.2.15) \quad \frac{dx}{dt} - tx = b(t), \quad x(0) = x_0.$$

From (1.2.12) we get

$$(1.2.16) \quad A(t) = -\frac{t^2}{2},$$

and (1.2.10) becomes

$$(1.2.17) \quad \frac{d}{dt}(e^{-t^2/2}x) = e^{-t^2/2}b(t),$$

hence

$$(1.2.18) \quad e^{-t^2/2}x(t) = x_0 + \int_0^t e^{-s^2/2}b(s) ds.$$

Let us look at two special cases. First,

$$(1.2.19) \quad b(t) = t.$$

Then the integral in (1.2.18) is

$$(1.2.20) \quad \int_0^t e^{-s^2/2}s ds = \int_0^{t^2/2} e^{-\sigma} d\sigma = 1 - e^{-t^2/2}.$$

The second case is

$$(1.2.21) \quad b(t) = 1.$$

Then the integral in (1.2.18) is

$$(1.2.22) \quad \int_0^t e^{-s^2/2} ds.$$

This is not an elementary function, but it can be related to the special function

$$(1.2.23) \quad \operatorname{Erf}(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-s^2/2} ds.$$

Namely,

$$(1.2.24) \quad \frac{1}{\sqrt{2\pi}} \int_0^t e^{-s^2/2} ds = \operatorname{Erf}(t) - \operatorname{Erf}(0).$$

Note that

$$(1.2.25) \quad \operatorname{Erf}(0) = \frac{1}{2} \operatorname{Erf}(\infty) = \frac{1}{2} \frac{1}{\sqrt{2\pi}} I,$$

where

$$(1.2.26) \quad \begin{aligned} I = \int_{-\infty}^{\infty} e^{-s^2/2} ds &\Rightarrow I^2 = \int_{\mathbb{R}^2} e^{-|x|^2/2} dx \\ &= \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2} r dr d\theta \\ &= 2\pi \int_0^{\infty} e^{-s} ds \\ &= 2\pi. \end{aligned}$$

Hence we have

$$(1.2.27) \quad \operatorname{Erf}(\infty) = 1, \quad \operatorname{Erf}(0) = \frac{1}{2}.$$

## Bernoulli equations

Equations of the form

$$(1.2.28) \quad \frac{dx}{dt} + a(t)x = b(t)x^n$$

are called Bernoulli equations. Such an equation is not linear if  $n \neq 1$  or  $0$ , but in these cases one gets a linear equation by the substitution

$$(1.2.29) \quad y = x^{1-n}.$$

In fact, (1.2.29) gives  $y' = (1 - n)x^{-n}x'$ , and plugging in (1.2.28) gives

$$(1.2.30) \quad \frac{dy}{dt} = (1 - n)[b(t) - a(t)y],$$

which is linear.

---

### Exercises

Solve the following initial value problems. Do the integrals if you can.

1.

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

2.

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

3.

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

4.

$$\frac{dx}{dt} + tx = t^3, \quad x(0) = 1.$$

5.

$$\frac{dx}{dt} + tx = x^3, \quad x(0) = 1.$$

6.

$$\frac{dx}{dt} + (\tan t)x = \cos t, \quad x(0) = 1.$$

7.

$$\frac{dx}{dt} + (\sec t)x = \cos t, \quad x(0) = 1.$$

### 1.3. Separable equations

A separable differential equation is one for which the method of separation of variables, which we introduce in this section, is applicable. We illustrate this with another approach to the equation (1.2.2), which we rewrite as

$$(1.3.1) \quad \frac{dx}{dt} = b - ax, \quad x(0) = x_0.$$

Separating variables involves moving the  $x$ -dependent objects to the left and the  $t$ -dependent objects to the right, when possible. In case (1.3.1), this is possible; we have

$$(1.3.2) \quad \frac{dx}{b - ax} = dt.$$

We next integrate both sides. A change of variable allows us to use (1.1.27), to obtain

$$(1.3.3) \quad \int \frac{dx}{b - ax} = -\frac{1}{a} \int \frac{dx}{x - b/a} = -\frac{1}{a} \log \left| x - \frac{b}{a} \right| + C.$$

Hence (1.3.2) yields

$$(1.3.4) \quad -\frac{1}{a} \log \left| x - \frac{b}{a} \right| = t - C,$$

hence

$$(1.3.5) \quad x(t) - \frac{b}{a} = \pm e^{-at+aC} = Ke^{-at}.$$

Here  $K$  is a constant, which can be found by using the initial condition  $x(0) = x_0$ . We get  $x_0 - b/a = K$ , so (1.3.5) yields

$$(1.3.6) \quad x(t) = \frac{b}{a} + e^{-at} \left( x_0 - \frac{b}{a} \right),$$

consistent with (1.2.7).

Generally, a separable differential equation is one that can be put in the form

$$(1.3.7) \quad \frac{dx}{dt} = f(x)g(t),$$

and then separation of variables gives

$$(1.3.8) \quad \frac{dx}{f(x)} = g(t) dt,$$

integrating to

$$(1.3.9) \quad \int \frac{dx}{f(x)} = \int g(t) dt.$$

Here is another basic example:

$$(1.3.10) \quad \frac{dx}{dt} = x^2, \quad x(0) = 1.$$

We get

$$(1.3.11) \quad \frac{dx}{x^2} = dt,$$

which integrates to

$$(1.3.12) \quad -\frac{1}{x} = t + C,$$

hence  $x = -1/(t + C)$ . The initial condition in (1.3.10) gives  $C = -1$ , so the solution to (1.3.10) is

$$(1.3.13) \quad x(t) = \frac{1}{1-t}.$$

Note that this solution blows up as  $t \nearrow 1$ .

### The hanging cable

Suppose a length of cable, lying in the  $(x, y)$ -plane, is fastened at  $(-a, 0)$  and at  $(a, 0)$ , and hangs down freely, in equilibrium, as pictured in Figure 1.3.1. The force of gravity acts in the direction of the negative  $y$ -axis. We want the equation of the curve traced out by the cable, which we assume to have length  $2L$  (not stretchable) and uniform mass density.

To tackle this problem, we introduce  $\theta(x)$ , the angle the tangent to the curve at  $(x, y(x))$  makes with the  $x$ -axis, which is given by

$$(1.3.14) \quad \tan \theta(x) = y'(x).$$

We will derive a differential equation for  $\theta(x)$ , as follows.

At each point  $(x, y(x))$ , there is a tension on the cable, of magnitude  $T(x)$ , and the physical laws governing the behavior of the cable are the following. First, the horizontal component of the tension, given by  $T(x) \cos \theta(x)$ , is constant. Second, the vertical component of the tension, given by  $T(x) \sin \theta(x)$ , is proportional to the weight of the cable lying below  $y = y(x)$ , hence to the length  $L(x)$  of the cable, from  $(0, y(0))$  to  $(x, y(x))$ . In other words, we have

$$(1.3.15) \quad \begin{aligned} T(x) \cos \theta(x) &= T_0, \\ T(x) \sin \theta(x) &= \kappa L(x), \end{aligned}$$

where  $T_0$  and  $\kappa$  are certain constants (whose quotient will be specified below). As for  $L(x)$ , we have

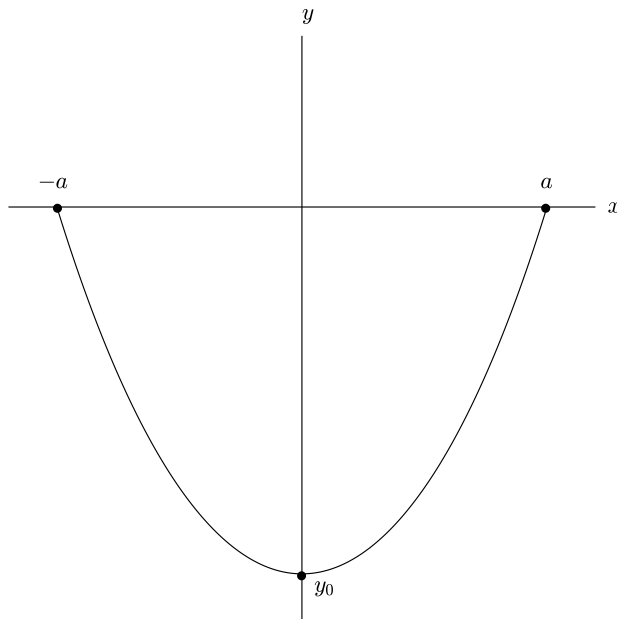
$$(1.3.16) \quad \begin{aligned} L(x) &= \int_0^x \sqrt{1 + y'(t)^2} dt \\ &= \int_0^x \sec \theta(t) dt, \end{aligned}$$

by (1.3.14) and Exercise 3 of §1.1.

Taking the quotient of the two identities in (1.3.15) yields

$$(1.3.17) \quad \tan \theta(x) = \beta \int_0^x \sec \theta(t) dt, \quad \beta = \frac{\kappa}{T_0}.$$





**Figure 1.3.1.** Catenary

Differentiating (1.3.17) with respect to  $x$  and using Exercise 4 of §1.1, we get

$$(1.3.18) \quad \sec^2 \theta(x) \frac{d\theta}{dx} = \beta \sec \theta(x),$$

i.e.,

$$(1.3.19) \quad \frac{d\theta}{dx} = \beta \cos \theta.$$

We can separate variables here, to obtain

$$(1.3.20) \quad \int \sec \theta \, d\theta = \int \beta \, dx.$$

Exercise 14 of §1.1 applies to the integral on the left, and we get

$$(1.3.21) \quad \sec \theta(x) = \cosh(\beta x + \alpha).$$

To yield the expected result  $\theta(0) = 0$  (see Figure 1.3.1 again), we set  $\alpha = 0$ .

To get a formula for  $y(x)$ , use (1.3.14) to write

$$(1.3.22) \quad y(x) = y_0 + \int_0^x \tan \theta(t) \, dt, \quad y_0 = y(0).$$

Now, by Exercises 3 and 8 of §1.1, together with (1.3.21), we have

$$(1.3.23) \quad \tan^2 \theta(x) = \sec^2 \theta(x) - 1 = \cosh^2 \beta x - 1 = \sinh^2 \beta x,$$

so (3.22) gives

$$(1.3.24) \quad \begin{aligned} y(x) &= y_0 + \int_0^x \sinh \beta t \, dt \\ &= y_0 - \frac{1}{\beta} + \frac{1}{\beta} \cosh \beta x. \end{aligned}$$

The graph of such a curve is called a *catenary*.

If we are given that the endpoints of the cable are at  $(\pm a, 0)$  and that the total length is  $2L$  (necessarily  $L > a$ ), we can recover  $\beta$  and  $y_0$  in (1.3.24), as follows. From (1.3.16) and (1.3.21),

$$(1.3.25) \quad L = \int_0^a \cosh \beta t \, dt = \frac{1}{\beta} \sinh \beta a,$$

so  $\beta$  is uniquely determined by the property that

$$(1.3.26) \quad \frac{\sinh \tau}{\tau} = \frac{L}{a}, \quad \beta = \frac{\tau}{a} > 0.$$

Note that  $h(\tau) = (\sinh \tau)/\tau$  is smooth,  $h(0) = 1$ ,  $h'(\tau) > 0$  for  $\tau > 0$ , and  $h(\tau) \nearrow +\infty$  as  $\tau \nearrow +\infty$ . Once one has  $\beta$ , then the identity  $y(a) = 0$  gives

$$(1.3.27) \quad y_0 = \frac{1}{\beta} - \frac{1}{\beta} \cosh \beta a.$$

### Homogeneous equations, separable in new variables

One can make a change of variable to convert a differential equation of the form

$$(1.3.28) \quad \frac{dx}{dt} = f(t, x)$$

to a separable equation when  $f(t, x)$  has the following homogeneity property:

$$(1.3.29) \quad f(rt, rx) = f(t, x), \quad \forall r \in \mathbb{R} \setminus 0.$$

In such a case,  $f$  has the form

$$(1.3.30) \quad f(t, x) = g\left(\frac{x}{t}\right).$$

We can set

$$(1.3.31) \quad y = \frac{x}{t},$$

so  $x = ty$ ,  $x' = ty' + y$ , and (1.3.28) turns into

$$(1.3.32) \quad \frac{dy}{dt} = \frac{g(y) - y}{t},$$

which is separable.

For example, consider

$$(1.3.33) \quad \frac{dx}{dt} = \frac{x^2 - t^2}{x^2 + t^2} + \frac{x}{t}.$$

In this case, (1.3.29) applies, and we can take  $g(y) = (y^2 - 1)/(y^2 + 1) + y$  in (1.3.30), so with  $y$  as in (1.3.31) we have

$$(1.3.34) \quad \frac{dy}{dt} = \frac{1}{t} \frac{y^2 - 1}{y^2 + 1},$$

which separates to

$$(1.3.35) \quad \left(1 + \frac{2}{y^2 - 1}\right) dy = \frac{dt}{t}.$$

To integrate the left side of (1.3.35), write

$$(1.3.36) \quad \frac{2}{y^2 - 1} = \frac{1}{y + 1} - \frac{1}{y - 1},$$

to get

$$(1.3.37) \quad \int \frac{2}{y^2 - 1} dy = \log |y + 1| - \log |y - 1| \\ = \log \left| \frac{y + 1}{y - 1} \right|,$$

the latter identity by (1.1.28). Thus the solution to (1.3.33) is given implicitly by

$$(1.3.38) \quad \frac{x}{t} + \log \left| \frac{x + t}{x - t} \right| = \log |t| + C.$$

## Exercises

Solve the following initial value problems. Do the integrals, if you can.

1.

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

2.

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

3.

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

4.

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

5.

$$\frac{dx}{dt} = e^{x-t}, \quad x(0) = 0.$$

6.

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

### 1.4. Second order equations – reducible cases

Second order differential equations have the form

$$(1.4.1) \quad x'' = f(t, x, x'), \quad x(t_0) = x_0, \quad x'(t_0) = v_0.$$

There are some important cases, with special structure, which reduce to first order equations for

$$(1.4.2) \quad v(t) = \frac{dx}{dt}.$$

One such case is

$$(1.4.3) \quad x'' = f(t, x'),$$

which for  $v$  given by (1.4.2) yields

$$(1.4.4) \quad \frac{dv}{dt} = f(t, v), \quad v(t_0) = v_0.$$

Depending on the nature of  $f(t, v)$ , methods discussed in §§1.2–1.3 might apply to (1.4.4). Once one has  $v(t)$ , then

$$(1.4.5) \quad x(t) = x_0 + \int_{t_0}^t v(s) ds.$$

The following is a more significant special case:

$$(1.4.6) \quad x'' = f(x, x').$$

Direct substitution of  $v$ , given by (1.4.2), yields

$$(1.4.7) \quad \frac{dv}{dt} = f(x, v),$$

which is not satisfactory, since (1.4.7) contains too many variables. One route to success is to rewrite the equation as one for  $v$  as a function of  $x$ , using

$$(1.4.8) \quad \frac{dv}{dt} = \frac{dv}{dx} \frac{dx}{dt} = v \frac{dv}{dx}.$$

Substitution into (1.4.7) gives the first order equation

$$(1.4.9) \quad \frac{dv}{dx} = \frac{f(x, v)}{v}, \quad v(x_0) = v_0.$$

Again, depending on the nature of  $f(x, v)/v$ , methods developed in §§2.2–2.3 might apply to (1.4.9).

An important special case of (1.4.6) is

$$(1.4.10) \quad x'' = f(x),$$

in which case (1.4.9) becomes

$$(1.4.11) \quad \frac{dv}{dx} = \frac{f(x)}{v},$$

which is separable:

$$(1.4.12) \quad v dv = f(x) dx,$$

hence

$$(1.4.13) \quad \frac{1}{2}v^2 = g(x) + C, \quad \int f(x) dx = g(x) + C.$$

Thus

$$(1.4.14) \quad \frac{dx}{dt} = v = \pm\sqrt{2g(x) + 2C},$$

which in turn is separable:

$$(1.4.15) \quad \pm \int \frac{dx}{\sqrt{2g(x) + 2C}} = t + C_2.$$

The constants  $C$  and  $C_2$  are determined by the initial conditions.

---

### Exercises

Use  $v = dx/dt$  to transform each of the following equations to first order equations, either for  $v = v(t)$  or for  $v = v(x)$ , as appropriate. Solve these first order equations, if you can.

1.

$$\frac{d^2x}{dt^2} = t \frac{dx}{dt}.$$

2.

$$\frac{d^2x}{dt^2} = \frac{dx}{dt} + t.$$

3.

$$\frac{d^2x}{dt^2} = x \frac{dx}{dt}.$$

4.

$$\frac{d^2x}{dt^2} = \frac{dx}{dt} + x.$$

5.

$$\frac{d^2x}{dt^2} = x^2.$$

### 1.5. Newton's equations for motion in 1D

Newton's law for motion in 1D of a particle of mass  $m$ , subject to a force  $F$ , is

$$(1.5.1) \quad F = ma,$$

where  $a$  is acceleration:

$$(1.5.2) \quad a(t) = \frac{dv}{dt} = \frac{d^2x}{dt^2},$$

the rate of change of the velocity  $v(t) = dx/dt$ . In general one might have  $F = F(t, x, x')$ . If  $F$  is  $t$ -independent,  $F = F(x, x')$ , which puts us in the setting of (1.4.6).

Frequently one has  $F = F(x)$ , which puts us in the setting of (1.4.10). We revisit this setting, bringing in some more concepts from physics. We set

$$(1.5.3) \quad F(x) = -V'(x).$$

$V(x)$ , defined up to an additive constant, is called the potential energy. The total energy is the sum of the potential energy and the kinetic energy,  $mv^2/2$ :

$$(1.5.4) \quad E = \frac{1}{2}mv(t)^2 + V(x(t)).$$

Note that

$$(1.5.5) \quad \begin{aligned} \frac{dE}{dt} &= mv(t)v'(t) + V'(x(t))x'(t) \\ &= ma(t)v(t) - F(x(t))v(t) \\ &= 0, \end{aligned}$$

the last identity by (1.5.1). This identity celebrates energy conservation. Given that  $x$  solves

$$(1.5.6) \quad m \frac{d^2x}{dt^2} = -V'(x), \quad x(t_0) = x_0, \quad x'(t_0) = v_0,$$

one has from (1.5.5) that for all  $t$ ,

$$(1.5.7) \quad \frac{1}{2}mx'(t)^2 + V(x(t)) = E_0,$$

where

$$(1.5.8) \quad E_0 = \frac{1}{2}mv_0^2 + V(x_0).$$

The equation (1.5.7) is equivalent to

$$(1.5.9) \quad \frac{dx}{dt} = \pm \sqrt{\frac{2}{m}(E_0 - V(x))},$$

which separates to

$$(1.5.10) \quad \int \frac{dx}{\sqrt{E_0 - V(x)}} = \pm \sqrt{\frac{2}{m}}t + C,$$

or, alternatively,

$$(1.5.11) \quad \int_{x_0}^x \frac{dy}{\sqrt{E_0 - V(y)}} = \pm \sqrt{\frac{2}{m}}(t - t_0).$$

Note that (1.5.7) and (1.5.10) recover (1.4.13) and (1.4.15).

### Projectile problem

Let's look in more detail at a special case, modeling the motion of a projectile of mass  $m$  traveling directly away from (or toward) the earth. In such a case, Newton's law of gravity gives

$$(1.5.12) \quad F(x) = -\frac{Km}{x^2}, \quad \text{hence } V(x) = -\frac{Km}{x}, \quad x \in (0, \infty).$$

In such a case, the conserved energy is

$$(1.5.13) \quad E_0 = \frac{m}{2} \left( v^2 - \frac{2K}{x} \right) = \frac{m}{2} \mathcal{E}(x, v).$$

See Figure 1.5.1 for a sketch of level curves of the function  $\mathcal{E}(x, v)$ . There are three cases to consider:

$$(1.5.14) \quad \begin{aligned} &\mathcal{E} = -a^2 < 0, \quad \mathcal{E} = 0, \quad \mathcal{E} = a^2 > 0, \quad \text{i.e.,} \\ &E_0 = -\frac{m}{2}a^2 < 0, \quad E_0 = 0, \quad E_0 = \frac{m}{2}a^2 > 0. \end{aligned}$$

In the first case,  $x(t)$  has a maximum at  $x_{\max} = 2K/a^2$ . In the other two cases,  $x(t) \rightarrow +\infty$  as  $t \rightarrow +\infty$  (if  $v_0 > 0$ ) or as  $t \rightarrow -\infty$  (if  $v_0 < 0$ ). Given  $x_0 \in (0, \infty)$ , the velocity  $v_0 \in (0, \infty)$  for which  $\mathcal{E}(x_0, v_0) = 0$  is called the "escape velocity."

We investigate the integral on the left side of (1.5.10), i.e.,

$$(1.5.15) \quad \int \frac{dx}{\sqrt{E_0 + Km/x}},$$

which in the three cases in (1.5.14) is  $\sqrt{2/m}$  times

$$(1.5.16) \quad \int \frac{x dx}{\sqrt{2Kx - a^2x^2}}, \quad \int \sqrt{\frac{x}{2K}} dx, \quad \int \frac{x dx}{\sqrt{2Kx + a^2x^2}},$$

respectively. The second integral in (1.5.16) is easy; we investigate how to compute the other two, which we rewrite as

$$(1.5.17) \quad \frac{1}{a} \int \frac{x dx}{\sqrt{2kx - x^2}}, \quad \frac{1}{a} \int \frac{x dx}{\sqrt{2kx + x^2}}, \quad k = \frac{K}{a^2}.$$

We can compute these integrals by completing the square:

$$(1.5.18) \quad x^2 - 2kx = (x - k)^2 - k^2, \quad x^2 + 2kx = (x + k)^2 - k^2.$$

The respective change of variables  $y = x - k$  and  $y = x + k$  turn the integrals in (1.5.17) into the respective integrals

$$(1.5.19) \quad \int \frac{(y + k) dy}{\sqrt{k^2 - y^2}}, \quad \int \frac{(y - k) dy}{\sqrt{y^2 - k^2}}.$$

By inspection,

$$(1.5.20) \quad \int \frac{y dy}{\sqrt{k^2 - y^2}} = -\sqrt{k^2 - y^2} + C, \quad \int \frac{y dy}{\sqrt{y^2 - k^2}} = \sqrt{y^2 - k^2} + C.$$

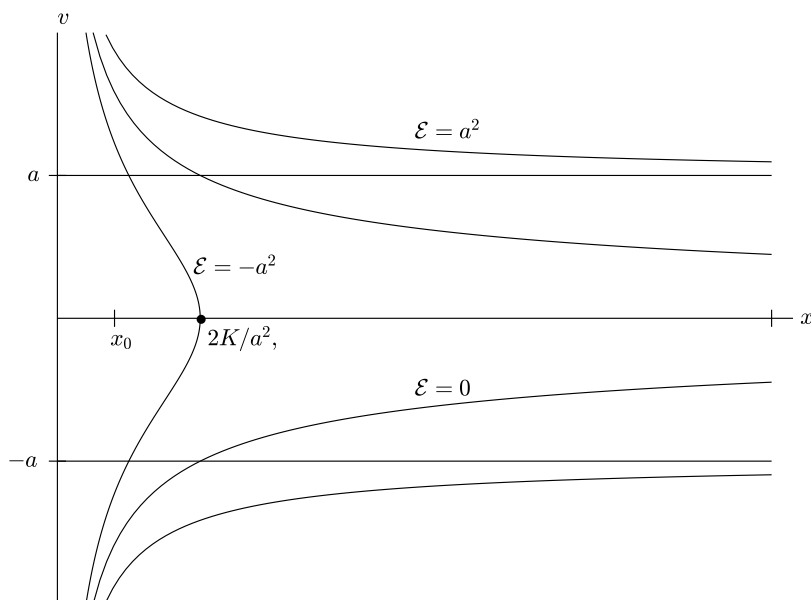


Figure 1.5.1. Projectile paths

The remaining parts of (1.5.19), after a change of variable  $y = kz$ , become

$$(1.5.21) \quad k \int \frac{dz}{\sqrt{1-z^2}}, \quad k \int \frac{dz}{\sqrt{z^2-1}}.$$

To do these integrals, use

$$(1.5.22) \quad \begin{aligned} z = \sin s &\implies \int \frac{dz}{\sqrt{1-z^2}} = \int \frac{\cos s}{\cos s} ds = s + C, \\ z = \cosh s &\implies \int \frac{dz}{\sqrt{z^2-1}} = \int \frac{\sinh s}{\sinh s} ds = s + C. \end{aligned}$$

---

## Exercises

1. Make calculations analogous to (1.5.12)–(1.5.15) for each of the following forces. Examine whether you can do the resulting integrals.

(a)

$$F(x) = -Kx.$$



(b)

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

(c)

$$F(x) = -\frac{K}{x}.$$

(d)

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

2. For such forces as given above, in each case find a potential energy  $V(x)$  and sketch the level curves in the  $(x, v)$ -plane of the energy function

$$E(x, v) = \frac{m}{2}v^2 + V(x).$$

3. Use the substitution

$$x = k^2 \sin^2 \theta$$

to evaluate

$$\int \frac{dx}{\sqrt{\frac{k^2}{x} - 1}},$$

and use

$$x = k^2 \sinh^2 u$$

to evaluate

$$\int \frac{dx}{\sqrt{\frac{k^2}{x} + 1}}.$$

Use these calculations as alternatives for evaluating (1.5.15), for  $E_0 < 0$  and  $E_0 > 0$ , respectively.

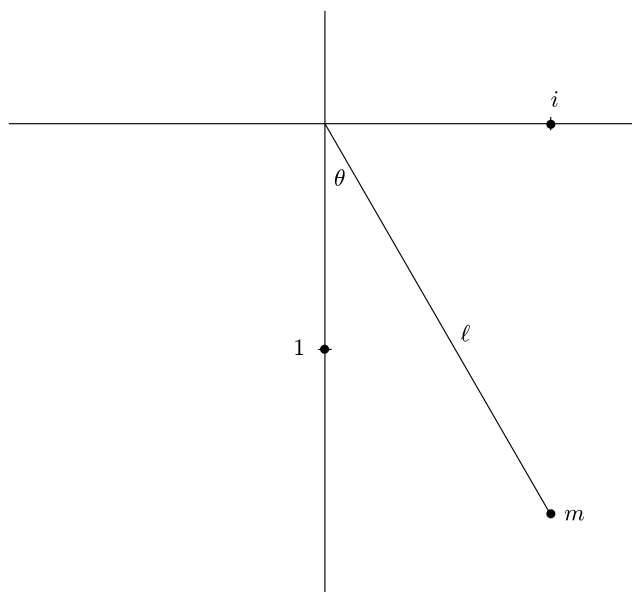


Figure 1.6.1. Pendulum

## 1.6. The pendulum

We produce a differential equation to describe the motion of a pendulum, which will be modeled by a rigid rod, of length  $\ell$ , suspended at one end. We assume the rod has negligible mass, except for an object of mass  $m$  at the other end. See Figure 1.6.1. The rod is held at an angle  $\theta = \theta_0$  from the downward pointing vertical, and released at time  $t = 0$ , after which it moves because of the force of gravity. We seek a differential equation for  $\theta$  as a function of  $t$ .

The end with the mass  $m$  traces out a path in a plane, which 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

$$(1.6.1) \quad z(t) = \ell e^{i\theta(t)}.$$

The velocity is

$$(1.6.2) \quad v(t) = z'(t) = i\ell\theta'(t)e^{i\theta(t)},$$

and the acceleration is

$$(1.6.3) \quad a(t) = v'(t) = \ell[i\theta''(t) - \theta'(t)^2]e^{i\theta(t)}.$$

The force of gravity on the mass is  $mg$ , where  $g = 32 \text{ ft/sec}^2$ , provided the pendulum is located on the surface of the Earth. The total force  $F$  on the mass is the sum of

the gravitational force and the force the rod exerts on the mass to keep it always at a distance  $\ell$  from the origin. The force the rod exerts is parallel to  $e^{i\theta(t)}$ , so

$$(1.6.4) \quad F(t) = mg + \Phi(t)e^{i\theta(t)},$$

for some real valued  $\Phi(t)$  (to be determined). We can rewrite  $mg$  as

$$(1.6.5) \quad mg = mge^{-i\theta(t)}e^{i\theta(t)} = mg[\cos\theta(t) - i\sin\theta(t)]e^{i\theta(t)},$$

and hence

$$(1.6.6) \quad F(t) = [-img\sin\theta(t) + mg\cos\theta(t) + \Phi(t)]e^{i\theta(t)}.$$

Newton's law  $F = ma$  applied to (1.6.3)–(1.6.6) gives

$$(1.6.7) \quad m\ell[i\theta''(t) - \theta'(t)^2] = -img\sin\theta(t) + (mg\cos\theta(t) + \Phi(t)).$$

Comparing imaginary parts gives

$$(1.6.8) \quad m\ell\theta''(t) = -mg\sin\theta(t),$$

or

$$(1.6.9) \quad \frac{d^2\theta}{dt^2} + \frac{g}{\ell}\sin\theta = 0.$$

This is the pendulum equation.

The kinetic energy of this pendulum is

$$(1.6.10) \quad \frac{1}{2}m|v(t)|^2 = \frac{m\ell^2}{2}\theta'(t)^2,$$

and its potential energy (up to an additive constant) is given by  $-mg$  times the real part of  $z(t)$ , i.e.,

$$(1.6.11) \quad V(\theta) = -mg\ell\cos\theta.$$

The total energy is hence

$$(1.6.12) \quad E = \frac{m\ell^2}{2}\theta'(t)^2 - mg\ell\cos\theta(t).$$

Note that

$$(1.6.13) \quad \begin{aligned} \frac{dE}{dt} &= m\ell^2\theta'(t)\theta''(t) + mg\ell(\sin\theta(t))\theta'(t) \\ &= m\ell^2\theta'(t)\left(\theta''(t) + \frac{g}{\ell}\sin\theta(t)\right), \end{aligned}$$

so the pendulum equation (1.6.9) implies  $dE/dt = 0$ , i.e., we have conservation of energy. Under the initial condition formulated at the beginning of this section,

$$(1.6.14) \quad \theta(0) = \theta_0, \quad \theta'(0) = 0,$$

we have initial energy

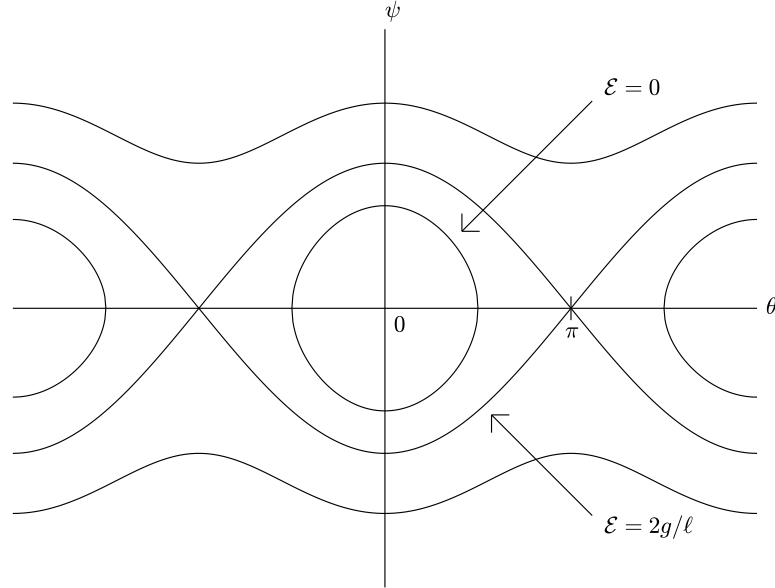
$$(1.6.15) \quad E_0 = -mg\ell\cos\theta_0,$$

and the energy conservation gives

$$(1.6.16) \quad \mathcal{E}(\theta, \theta') = \frac{2E_0}{m\ell^2} = A_0,$$

where

$$(1.6.17) \quad \mathcal{E}(\theta, \psi) = \psi^2 - \frac{2g}{\ell}\cos\theta.$$



**Figure 1.6.2.** Level curves of  $\mathcal{E}(\theta, \psi) = \psi^2 - (2g/\ell) \cos \theta$

Level curves of this function are depicted in Figure 1.6.2. If  $\theta(t)$  solves (1.6.9) and  $\psi(t) = \theta'(t)$ , then  $(\theta(t), \psi(t))$  traces out a path on one of these level curves.

Note that

$$(1.6.18) \quad \nabla \mathcal{E}(\theta, \psi) = \left( \frac{2g}{\ell} \sin \theta, 2\psi \right),$$

so  $\mathcal{E}$  has critical points at  $\theta = k\pi$ ,  $\psi = 0$ . The matrix of second order partial derivatives of  $\mathcal{E}$  is

$$(1.6.19) \quad D^2 \mathcal{E}(\theta, \psi) = \begin{pmatrix} \frac{2g}{\ell} \cos \theta & 0 \\ 0 & 2 \end{pmatrix},$$

so

$$(1.6.20) \quad D^2 \mathcal{E}(k\pi, 0) = \begin{pmatrix} (-1)^k \frac{2g}{\ell} & 0 \\ 0 & 2 \end{pmatrix}.$$

We see that at the critical point  $(k\pi, 0)$ ,  $\mathcal{E}$  has a local minimum if  $k$  is even and a saddle-type behavior if  $k$  is odd, as illustrated in Figure 1.6.2.

Note that if the initial condition (1.6.14) holds, then  $A_0 = -(2g/\ell) \cos \theta_0$ , and hence  $A_0 < 2g/\ell$ , so the curve traced by  $(\theta(t), \psi(t))$  is a closed curve. One might instead have initial data of the form

$$(1.6.21) \quad \theta(0) = \theta_0, \quad \theta'(0) = \psi_0,$$

and one could pick  $\psi_0$  so that  $\mathcal{E}(\theta_0, \psi_0) > 2g/\ell$ .

We proceed to formulas parallel to (1.5.7)–(1.5.11). Starting from the energy conservation (1.6.16), which we rewrite as

$$(1.6.22) \quad \theta'(t)^2 - \frac{2g}{\ell} \cos \theta(t) = A_0,$$

we have

$$(1.6.23) \quad \theta'(t) = \pm \sqrt{\frac{2g}{\ell} \sqrt{A_1 + \cos \theta}}, \quad A_1 = \frac{\ell}{2g} A_0 = \frac{E_0}{mg\ell},$$

which separates and integrates to

$$(1.6.24) \quad \int \frac{d\theta}{\sqrt{A_1 + \cos \theta}} = \pm \sqrt{\frac{2g}{\ell}} t + C.$$

In the current set-up, where, by (1.6.12),  $E_0 \geq -mg\ell$ , we have

$$(1.6.25) \quad A_1 \geq -1.$$

Note that to achieve  $A_1 = -1$  requires  $\theta(0) = 0$  and  $\theta'(0) = 0$ , in which case (1.6.23) yields the initial value problem

$$(1.6.26) \quad \theta'(t) = \pm \sqrt{\frac{2g}{\ell} \sqrt{-1 + \cos \theta}}, \quad \theta(0) = 0,$$

with solution

$$(1.6.27) \quad \theta(t) \equiv 0.$$

In this case (1.6.24) has no meaning. Indeed, if  $\theta > 0$  and one considers

$$(1.6.28) \quad \int_0^\theta \frac{d\varphi}{\sqrt{-1 + \cos \varphi}},$$

the integrand is imaginary and furthermore it is not integrable. Nevertheless,  $\theta(t) \equiv 0$  is a solution to the original problem.

Let us now assume  $A_1 > -1$ . Write

$$(1.6.29) \quad B_1 = A_1 + 1 > 0,$$

so

$$(1.6.30) \quad \begin{aligned} A_1 + \cos \theta &= B_1 - (1 - \cos \theta) \\ &= B_1 - 2 \sin^2 \frac{\theta}{2}, \end{aligned}$$

thanks to the identity  $\cos 2\varphi = \cos^2 \varphi - \sin^2 \varphi = 1 - 2 \sin^2 \varphi$ . We can rewrite the left side of (1.6.24) as

$$(1.6.31) \quad \begin{aligned} \int \frac{d\theta}{\sqrt{A_1 + \cos \theta}} &= \int \frac{d\theta}{\sqrt{B_1 - 2 \sin^2 \theta/2}} \\ &= \frac{\beta}{\sqrt{2}} \int \frac{d\theta}{\sqrt{1 - \beta^2 \sin^2 \theta/2}}, \end{aligned}$$

with

$$(1.6.32) \quad \beta = \sqrt{\frac{2}{B_1}} > 0.$$

The last integral in (1.6.31) is known as an elliptic integral when  $\beta^2 \neq 1$ , i.e., when  $A_1 \neq 1$ . Material on such integrals can be found in books that treat elliptic function theory, including [47].

The case  $\beta = 1$  (i.e.,  $A_1 = 1$ , or  $E_0 = mg\ell$ ) does give rise to an elementary integral, namely

$$(1.6.33) \quad \int \frac{d\theta}{\sqrt{1 + \cos \theta}} = \frac{1}{\sqrt{2}} \int \sec \frac{\theta}{2} d\theta \\ = \sqrt{2} \sinh^{-1} \left( \tan \frac{\theta}{2} \right) + C,$$

for  $|\theta| < \pi$ , the latter identity by Exercise 14 of §1.1.

### Further study of the elliptic integral in (1.6.24) – separatrices, periodic solutions, and their periods

Let us pursue the computations arising from (1.6.24) in more detail, taking the initial condition

$$(1.6.34) \quad \theta(0) = 0, \quad \theta'(0) = \psi_0, \quad \psi_0 \in (0, \infty).$$

Then (1.6.24) yields, for the solution  $\theta(t)$ ,

$$(1.6.35) \quad \int_0^{\theta(t)} \frac{d\vartheta}{\sqrt{A_1 + \cos \vartheta}} = \sqrt{\frac{2g}{\ell}} t.$$

In such a case,

$$(1.6.36) \quad A_1 = \frac{E_0}{mg\ell} = \frac{\ell}{2g} \psi_0^2 - 1, \quad \text{hence } B_1 = \frac{\ell}{2g} \psi_0^2.$$

Then (1.6.31) yields

$$(1.6.37) \quad \int_0^{\theta(t)} \frac{d\vartheta}{\sqrt{1 - \beta^2 \sin^2 \vartheta/2}} = \frac{\sqrt{2}}{\beta} \sqrt{\frac{2g}{\ell}} t \\ = \psi_0 t,$$

with

$$(1.6.38) \quad \beta = \sqrt{\frac{2}{B_1}} = \frac{2}{\psi_0} \sqrt{\frac{g}{\ell}}.$$

Let us specialize (1.6.37) to

$$(1.6.39) \quad \beta = 1, \quad \text{hence } \psi_0 = 2\sqrt{\frac{g}{\ell}}, \quad \text{so } E_0 = mg\ell.$$

By (1.6.33), we get

$$(1.6.40) \quad \tan \frac{\theta(t)}{2} = \sinh \frac{\psi_0}{2} t = \sinh \sqrt{\frac{g}{\ell}} t,$$

or

$$(1.6.41) \quad \theta(t) = 2 \tan^{-1} \sinh \sqrt{\frac{g}{\ell}} t.$$

Applying  $d/dt$  yields

$$(1.6.42) \quad \theta'(t) = \psi(t) = 2\sqrt{\frac{g}{\ell}} \frac{1}{\cosh \sqrt{\frac{g}{\ell}} t}.$$

In this case,

$$(1.6.43) \quad \theta(t) \rightarrow \pm\pi \quad \text{and} \quad \psi(t) \rightarrow 0, \quad \text{as } t \rightarrow \pm\infty.$$

The curve  $(\theta(t), \psi(t))$  and its mirror image are called *separatrices*. They separate bounded periodic solutions from unbounded solution curves.

We turn to the case

$$(1.6.44) \quad \beta > 1, \quad \text{hence } 0 < \psi_0 < 2\sqrt{\frac{g}{\ell}}, \quad \text{so } E < mgl.$$

In this case,

$$(1.6.45) \quad \theta(t) \text{ is periodic, say of period } \Pi(\psi_0),$$

and we want to find a formula for  $\Pi(\psi_0)$ . Looking at Figure 1.6.2, we see that

$$(1.6.46) \quad \psi(t) = 0 \quad \text{at} \quad t = \frac{1}{4}\Pi(\psi_0).$$

Comparison with the formula

$$(1.6.47) \quad \begin{aligned} \psi(t) = \frac{d\theta}{dt} &= \sqrt{\frac{2g}{\ell}} \sqrt{B_1 - 2\sin^2 \frac{\theta}{2}} \\ &= \beta \sqrt{\frac{g}{\ell}} \sqrt{1 - \beta^2 \sin^2 \frac{\theta}{2}} \end{aligned}$$

gives

$$(1.6.48) \quad \psi = 0 \quad \text{when} \quad \sin^2 \frac{\theta}{2} = \frac{B_1}{2},$$

and hence

$$(1.6.49) \quad \begin{aligned} \frac{1}{4}\Pi(\psi_0) &= \sqrt{\frac{\ell}{2g}} \int_0^{\theta_1} \frac{d\vartheta}{\sqrt{B_1 - 2\sin^2 \vartheta/2}}, \\ \sin^2 \frac{\theta_1}{2} &= \frac{B_1}{2} = \frac{1}{\beta^2} = \frac{\ell}{4g} \psi_0^2. \end{aligned}$$

Equivalently,

$$(1.6.50) \quad \begin{aligned} \frac{1}{4}\Pi(\psi_0) &= \frac{1}{\psi_0} \int_0^{\theta_1} \frac{d\vartheta}{\sqrt{1 - \beta^2 \sin^2 \vartheta/2}} \\ &= \frac{2}{\psi_0} \int_0^{\theta_1/2} \frac{d\varphi}{\sqrt{1 - \beta^2 \sin^2 \varphi}}, \end{aligned}$$

with  $\theta_1$  as in (1.6.49). Making the change of variable  $x = \sin \varphi$ , we get

$$(1.6.51) \quad \frac{1}{4}\Pi(\psi_0) = \frac{2}{\psi_0} \int_0^{1/\beta} \frac{dx}{\sqrt{(1-x^2)(1-\beta^2 x^2)}},$$

and finally, setting  $y = \beta x$  yields

$$(1.6.52) \quad \begin{aligned} \frac{1}{4}\Pi(\psi_0) &= \frac{2\alpha}{\psi_0} \int_0^1 \frac{dy}{\sqrt{(1-y^2)(1-\alpha^2 y^2)}} \\ &= \sqrt{\frac{\ell}{g}} \int_0^1 \frac{dy}{\sqrt{(1-y^2)(1-\alpha^2 y^2)}}, \end{aligned}$$

with

$$(1.6.53) \quad \alpha = \frac{1}{\beta} = \sqrt{\frac{B_1}{2}} = \frac{1}{2} \sqrt{\frac{\ell}{g}} \psi_0,$$

so  $0 < \alpha < 1$ . Clearly  $\alpha \rightarrow 0$  when  $\psi_0 \rightarrow 0$ , so we have

$$(1.6.54) \quad \begin{aligned} \lim_{\psi_0 \rightarrow 0} \Pi(\psi_0) &= 4 \sqrt{\frac{\ell}{g}} \int_0^1 \frac{dy}{\sqrt{1-y^2}} \\ &= 2\pi \sqrt{\frac{\ell}{g}}. \end{aligned}$$

This coincides with the period of solutions to

$$(1.6.55) \quad \frac{d^2\theta}{dt^2} + \frac{g}{\ell}\theta = 0,$$

which we will identify in §1.8 with the linearization of the pendulum equation about the zero solution.

Finally, we examine the case

$$(1.6.56) \quad 0 < \beta < 1, \quad \text{hence } \psi_0 = \frac{2}{\beta} \sqrt{\frac{g}{\ell}} > 2\sqrt{\frac{g}{\ell}}, \quad \text{so } E > mg\ell.$$

In such a case, we see from (1.6.47) that  $\theta(t)$  is monotone in  $t$ . However, it does possess the “periodicity”

$$(1.6.57) \quad \theta(t+s) = \theta(t) + 2\pi, \quad \text{with } s = \Pi(\psi_0),$$

where, when (1.6.56) holds,

$$(1.6.58) \quad \begin{aligned} \Pi(\psi_0) &= \sqrt{\frac{\ell}{2g}} \int_0^{2\pi} \frac{d\vartheta}{\sqrt{A_1 + \cos \vartheta}} \\ &= \frac{1}{\psi_0} \int_0^{2\pi} \frac{d\vartheta}{\sqrt{1 - \beta^2 \sin^2 \vartheta/2}} \\ &= \frac{2}{\psi_0} \int_0^\pi \frac{d\varphi}{\sqrt{1 - \beta^2 \sin^2 \varphi}}. \end{aligned}$$

Making the change of variable  $x = \sin \varphi$ , we get

$$(1.6.59) \quad \Pi(\psi_0) = \frac{2}{\psi_0} \int_0^1 \frac{dx}{\sqrt{(1-x^2)(1-\beta^2 x^2)}}.$$



REMARK. The integrals in (1.6.52) and (1.6.59) are called *complete elliptic integrals*. One can expand these integrals in convergent power series in  $\alpha^2$  and  $\beta^2$ , respectively, using the formula

$$(1.6.60) \quad \frac{1}{\sqrt{1-u}} = \sum_{k=0}^{\infty} a_k u^k, \quad \text{for } |u| < 1, \quad \text{with}$$

$$a_0 = 1, \quad a_k = \left(1 - \frac{1}{2}\right) \left(2 - \frac{1}{2}\right) \cdots \left(k - \frac{1}{2}\right)$$

(see Appendix 1.C), with  $u = \alpha^2 x^2$  in (1.6.52) and  $u = \beta^2 x^2$  in (1.6.59), and then integrating term by term. The coefficients in the resulting power series involve

$$(1.6.61) \quad \int_0^1 \frac{x^{2k}}{\sqrt{1-x^2}} dx = \int_0^{\pi} \sin^{2k} \varphi d\varphi$$

$$= \frac{1}{2} \left(\frac{1}{2i}\right)^{2k} \int_0^{2\pi} (e^{i\varphi} - e^{-i\varphi})^{2k} d\varphi$$

$$= \pi 2^{-2k} \binom{2k}{k}.$$

One can also express these complete elliptic integrals in terms of a function known as the Gauss *arithmetic-geometric mean* (cf. [47], Chapter 6, §4).

---

## Exercises

1. Let  $E$  be given by (6.8). Show that if  $\theta(t)$  solves (6.6) and  $|\theta(t)| < \pi/2$  for all  $t$ , then  $E < 0$ .

2. Show that the level set in Figure 1.6.2 where  $\mathcal{E} = 2g/\ell$  (i.e.,  $E = mg\ell$ ) is given by

$$\psi = \pm 2\sqrt{\frac{g}{\ell}} \cos \frac{\theta}{2}.$$

3. By (1.6.3), the component of acceleration parallel to  $e^{i\theta}$  is  $-\ell\theta'(t)^2 e^{i\theta(t)}$ . Compute the component of the gravitational force parallel to  $e^{i\theta(t)}$ , and deduce that the force the rod exerts on the mass to keep it always at a distance  $\ell$  from the origin is  $\Phi e^{i\theta(t)}$ , with

$$\Phi = -m\ell\theta'(t)^2 - mg \cos \theta.$$

Deduce that, with  $E$  as in (1.6.12),

$$\Phi(t) = \frac{E}{\ell} - \frac{3m\ell}{2} \theta'(t)^2.$$

4. Apply the change of variable  $s = \sin \varphi$  to the last integral in (1.6.31), i.e., to

$$\int \frac{d\varphi}{\sqrt{1 - \beta^2 \sin^2 \varphi}}.$$

Show that the integral becomes

$$\int \frac{ds}{\sqrt{(1-s^2)(1-\beta^2 s^2)}}.$$

Specialize to  $\beta = 1$  and obtain an alternative derivation of the formula for  $\int \sec \varphi d\varphi$  given in Exercise 13 of §1.1.

5. Suppose the mass at the end of the pendulum has a charge  $q_1$  and there is a charge  $q_2$  fixed at  $(x, y) = (2\ell, 0)$ . Then the force  $F(t)$  is modified to

$$F(t) = mg - Kq_1q_2 \frac{2\ell - \ell e^{i\theta(t)}}{|2\ell - \ell e^{i\theta(t)}|^3} + \Phi(t)e^{i\theta(t)},$$

where  $K$  is a positive constant. Use this to produce a modification of the pendulum equation.

### 1.7. Motion with resistance

In many real cases, the force acting on a moving object is the sum of a force associated with a potential and a resistance, typically depending on the velocity and acting to slow the motion down. For example, the motion of a ball of mass  $m$  falling through the air near the surface of the earth can be modeled by the differential equation

$$(1.7.1) \quad m \frac{d^2x}{dt^2} = mg - \alpha \frac{dx}{dt},$$

where the  $x$ -axis points down toward the earth. Here  $g = 32 \text{ ft/sec}^2$  and  $\alpha$  is an experimentally determined constant, depending on the size of the ball, and measures air resistance. We can rewrite (1.7.1) as an equation for  $v = dx/dt$ :

$$(1.7.2) \quad \frac{dv}{dt} = g - \frac{\alpha}{m}v,$$

an equation that is both linear and separable. Unless  $v$  is small, the formula  $-\alpha v$  for the force of air resistance is not so accurate, and a more accurate equation might be

$$(1.7.3) \quad \frac{dv}{dt} = g - \frac{\alpha}{m}v - \frac{\beta}{m}v^3.$$

This is not linear, but it is separable. For  $v$  close to the speed of sound in air, even this model loses validity.

If the ball is falling from the stratosphere toward the surface of the earth, the variation in air density, hence in air resistance, must be taken into account. One might replace the model (1.7.1) by

$$(1.7.4) \quad m \frac{d^2x}{dt^2} = mg - \alpha(x) \frac{dx}{dt}.$$

The method of (1.4.6)–(1.4.9) is applicable here, yielding for  $v = dx/dt$  the equation

$$(1.7.5) \quad \frac{dv}{dx} = \frac{mg}{v} - \alpha(x).$$

This, however, is not typically amenable to a solution in terms of elementary functions.

Another example of motion with resistance arises in the pendulum. Between air resistance and friction where the rod is attached, the pendulum equation (1.6.9) might be modified to the following damped pendulum equation:

$$(1.7.6) \quad \frac{d^2\theta}{dt^2} + \frac{\alpha}{m} \frac{d\theta}{dt} + \frac{g}{\ell} \sin \theta = 0,$$

for some positive constant  $\alpha$ . Again the method of (1.4.6)–(1.4.9) is applicable, to yield for  $\psi = d\theta/dt$  the equation

$$(1.7.7) \quad \frac{d\psi}{d\theta} = -\frac{\alpha}{m} - \frac{g \sin \theta}{\ell \psi}.$$

However, this equation is not particularly tractable, and does not yield much insight into the behavior of solutions to (1.7.6).

---

**Exercises**

1. Suppose  $v(t)$  solves (1.7.2) and  $v(0) = 0$ . Show that

$$\lim_{t \rightarrow +\infty} v(t) = \frac{mg}{\alpha},$$

and

$$v(t) < \frac{mg}{\alpha}, \quad \forall t \in [0, \infty).$$

What does it mean to call  $mg/\alpha$  the *terminal velocity*?

2. Do the analogue of Exercise 1 when  $v(t)$  solves (1.7.3) and  $v(0) = 0$ .

3. In the setting of Exercise 1, what happens if, instead of  $v(0) = 0$ , we have

$$v(0) = v_0 > \frac{mg}{\alpha}?$$

4. Apply the method of separation of variables to (1.7.3). Note that

$$g - \frac{\alpha}{m}v - \frac{\beta}{m}v^3 = p(v)$$

has three complex roots (at least one of which must be real). For what values of  $\alpha$ ,  $\beta$ , and  $m$  does  $p(v)$  have one real root and for what values does it have three real roots? How does this bear on the behavior of

$$\int \frac{dv}{p(v)}?$$

5. More general models for motion with resistance involve the following modification of (1.5.6):

$$m \frac{d^2x}{dt^2} = -V'(x) - \alpha \frac{dx}{dt}.$$

Parallel to (1.5.4), set

$$E(t) = \frac{1}{2}m \left( \frac{dx}{dt} \right)^2 + V(x(t)).$$

Show that

$$\frac{dE}{dt} \leq 0.$$

One says energy is *dissipated*, due to the resistance.

### 1.8. Linearization

As we have seen, some equations, such as the pendulum equation (1.6.9), which we rewrite here as

$$(1.8.1) \quad \frac{d^2x}{dt^2} + \frac{g}{\ell} \sin x = 0,$$

can be “solved” in terms of an integral, in this case (1.6.24), i.e.,

$$(1.8.2) \quad \int \frac{dx}{\sqrt{A_1 + \cos x}} = \pm \sqrt{\frac{2g}{\ell}} t + C.$$

However, the integral is a complicated special function. Meanwhile other equations, such as the damped pendulum equation (1.7.6), which we rewrite

$$(1.8.3) \quad \frac{d^2x}{dt^2} + \frac{\alpha}{m} \frac{dx}{dt} + \frac{g}{\ell} \sin x = 0,$$

are not even amenable to solutions as “explicit” as (1.8.2). In such cases one might nevertheless gain valuable insight into solutions that are small perturbations of some known particular solution to (1.8.1) or (1.8.3), or more generally

$$(1.8.4) \quad x''(t) = f(t, x(t), x'(t)).$$

In case (1.8.1) and (1.8.3),  $x(t) \equiv 0$  is a solution. More generally, one might have a known solution  $y(t)$  of (1.8.4); i.e.,  $y(t)$  is known and satisfies

$$(1.8.5) \quad y''(t) = f(t, y(t), y'(t)).$$

Now take  $x(t) = y(t) + \varepsilon u(t)$ . We derive an equation for  $u(t)$  so that  $x(t)$  satisfies (1.8.4), at least up to  $O(\varepsilon^2)$ , i.e.,

$$(1.8.6) \quad y''(t) + \varepsilon u''(t) = f(t, y(t) + \varepsilon u(t), y'(t) + \varepsilon u'(t)) + O(\varepsilon^2).$$

To get this equation, write, with  $f = f(t, x, v)$ ,

$$(1.8.7) \quad f(t, y + \varepsilon u, y' + \varepsilon u') = f(t, y, y') + \varepsilon \left( \frac{\partial f}{\partial x}(t, y, y')u + \frac{\partial f}{\partial v}(t, y, y')u' \right) + O(\varepsilon^2),$$

the first order Taylor polynomial approximation. Plugging this into (1.8.6) and using (1.8.5), we see that (1.8.6) holds provided  $u(t)$  satisfies the equation

$$(1.8.8) \quad u''(t) = A(t)u(t) + B(t)u'(t),$$

where

$$(1.8.9) \quad A(t) = \frac{\partial^2 f}{\partial x^2}(t, y(t), y'(t)), \quad B(t) = \frac{\partial^2 f}{\partial x \partial v}(t, y(t), y'(t)).$$

The equation (1.8.8) is a linear equation, called the *linearization* of (1.8.4) about the solution  $y(t)$ .

In case (1.8.1),  $f(t, x, v) = -(g/\ell) \sin x$ , and the linearization about  $y(t) = 0$  of this equation is

$$(1.8.10) \quad \frac{d^2u}{dt^2} + \frac{g}{\ell}u = 0.$$

In case (1.8.3),  $f(t, x, v) = (\alpha/m)v + (g/\ell) \sin x$ , and the linearization about  $y(t) = 0$  of this equation is

$$(1.8.11) \quad \frac{d^2u}{dt^2} + \frac{\alpha}{m} \frac{du}{dt} + \frac{g}{\ell}u = 0.$$

To take another example, consider

$$(1.8.12) \quad x''(t) = tx(t) - x(t)^2.$$

One solution is

$$(1.8.13) \quad y(t) = t.$$

In this case we have (1.8.4) with  $f(t, x, v) = tx - x^2$ , hence  $f_x(t, x, v) = t - 2x$  and  $f_v(t, x, v) = 0$ . Then  $f_x(t, y, y') = f_x(t, t, 1) = -t$ , and the linearization of (1.8.12) about  $y(t) = t$  is

$$(1.8.14) \quad u''(t) + tu(t) = 0.$$

---

### Exercises

Compute the linearizations of the following equations, about the given solution  $y(t)$ .

1.

$$x'' + \cosh x - \cosh 1 = 0, \quad y(t) = 1.$$

2.

$$x'' + \cosh x - \cosh t = 0, \quad y(t) = t.$$

3.

$$x'' + x' \sin x = 0, \quad y(t) = 0.$$

4.

$$x'' + x' \sin x = 0, \quad y(t) = \frac{\pi}{2}.$$

5.

$$x'' + \sin x = 0, \quad y(t) = \pi.$$

### 1.9. Second order constant-coefficient linear equations – homogeneous

Here we look into solving differential equations of the form

$$(1.9.1) \quad a \frac{d^2x}{dt^2} + b \frac{dx}{dt} + cx = 0,$$

with constants  $a$ ,  $b$ , and  $c$ . We assume  $a \neq 0$ . We impose an initial condition, such as

$$(1.9.2) \quad x(0) = \alpha, \quad x'(0) = \beta.$$

We look for solutions in the form

$$(1.9.3) \quad x(t) = e^{rt},$$

for some constant  $r$ , which worked so well for first order equations in §1.1. By results derived there, if  $x(t)$  has the form (1.9.3), then  $x'(t) = re^{rt}$  and  $x''(t) = r^2e^{rt}$ , so substitution into the left side of (1.9.1) gives

$$(1.9.4) \quad (ar^2 + br + c)e^{rt},$$

which vanishes if and only if  $r$  satisfies the equation

$$(1.9.5) \quad ar^2 + br + c = 0.$$

The polynomial  $p(r) = ar^2 + br + c$  is called the characteristic polynomial associated with the differential equation (1.9.1). Its roots are given by

$$(1.9.6) \quad r_{\pm} = -\frac{b}{2a} \pm \frac{1}{2a} \sqrt{b^2 - 4ac}.$$

There are two cases to consider:

$$(I) \quad b^2 - 4ac \neq 0,$$

$$(II) \quad b^2 - 4ac = 0.$$

In Case I, the equation (1.9.5) has two distinct roots, and we get two distinct solutions to (1.9.1),  $e^{r_+t}$  and  $e^{r_-t}$ . It is easy to see that whenever  $x_1(t)$  and  $x_2(t)$  solve (9.1), so does  $C_1x_1(t) + C_2x_2(t)$ , for arbitrary constants  $C_1$  and  $C_2$ . Hence

$$(1.9.7) \quad x(t) = C_+e^{r_+t} + C_-e^{r_-t}$$

solves (1.9.1), for all constants  $C_+$  and  $C_-$ .

Having this, we can find a solution to (1.9.1) with initial data (1.9.2) as follows. Taking  $x(t)$  as in (9.7), so  $x'(t) = C_+r_+e^{r_+t} + C_-r_-e^{r_-t}$ , we set  $t = 0$  to obtain

$$(1.9.8) \quad x(0) = C_+ + C_-, \quad x'(0) = r_+C_+ + r_-C_-,$$

so (1.9.2) holds if and only if  $C_+$  and  $C_-$  satisfy

$$(1.9.9) \quad \begin{aligned} C_+ + C_- &= \alpha, \\ r_+C_+ + r_-C_- &= \beta. \end{aligned}$$

This set of two linear equations for  $C_+$  and  $C_-$  has a unique solution if and only if  $r_+ \neq r_-$ . In fact, the first equation in (1.9.9) gives

$$(1.9.10) \quad r_-C_+ + r_-C_- = r_- \alpha,$$

and subtracting this from the second equation in (1.9.9) yields

$$(1.9.11) \quad C_+ = \frac{\beta - \alpha r_-}{r_+ - r_-},$$

and then the first equation in (1.9.9) yields

$$(1.9.12) \quad C_- = \alpha - C_+ = \frac{\alpha r_+ - \beta}{r_+ - r_-}.$$

In Case II,  $r = -b/2a$  is a double root of the characteristic polynomial, and we have the solution  $x(t) = e^{rt}$  to (1.9.1). We claim there is another solution to (1.9.1) that is not simply a constant multiple of this one. We look for a second solution in the form

$$(1.9.13) \quad x(t) = u(t)e^{rt},$$

hoping to get a simpler differential equation for  $u(t)$ . Note that then  $x' = (u' + ru)e^{rt}$  and  $x'' = (u'' + 2ru' + r^2u)e^{rt}$ , and hence

$$(1.9.14) \quad \begin{aligned} ax'' + bx' + cx &= \left\{ a(u'' + 2ru' + r^2u) + b(u' + ru) + cu \right\} e^{rt} \\ &= \left\{ au'' + (2ar + b)u' + (ar^2 + br + c)u \right\} e^{rt} \\ &= au'' e^{rt}, \end{aligned}$$

given that (1.9.5) holds with  $r = -b/2a$ . Thus the vanishing of (1.9.14) is equivalent to  $u''(t) = 0$ , i.e., to  $u(t) = C_1 + C_2t$ . Hence another solution to (1.9.1) in this case is  $te^{rt}$ , and, in place of (1.9.7), we have solutions

$$(1.9.15) \quad x(t) = C_1e^{rt} + C_2te^{rt},$$

for all constants  $C_1$  and  $C_2$ .

We can then find a solution to (1.9.1) with initial data (1.9.2) as follows. Taking  $x(t)$  as in (1.9.15), so  $x'(t) = C_1re^{rt} + C_2rte^{rt} + C_2e^{rt}$ , we set  $t = 0$  to obtain

$$(1.9.16) \quad x(0) = C_1, \quad x'(0) = C_1 + C_2,$$

so (1.9.2) is satisfied if and only if  $C_1$  and  $C_2$  satisfy

$$(1.9.17) \quad C_1 = \alpha, \quad C_1 + C_2 = \beta,$$

i.e., if and only if

$$(1.9.18) \quad C_1 = \alpha, \quad C_2 = \beta - \alpha.$$

We claim the constructions given above provide *all* of the solutions to (1.9.1), in the two respective cases. To see this, let  $x(t)$  be any solution to (1.9.1), let  $r = r_+$  (which equals  $r_-$  in Case II), and consider  $u(t) = e^{-rt}x(t)$ , as in (1.9.13). The computation (1.9.14) holds if  $r_+ = r_-$ , and if  $r_+ \neq r_-$  we get

$$(1.9.19) \quad ax'' + bx' + cx = \left\{ au'' + (2ar + b)u' \right\} e^{rt}.$$

As we have seen, when  $r_+ = r_-$  this forces  $u''(t) \equiv 0$ , which hence forces  $u(t)$  to have the form  $C_1 + C_2t$  for some constants  $C_j$ , and hence  $x(t) = C_1e^{rt} + C_2te^{rt}$ . When  $r_+ \neq r_-$ , vanishing of (1.9.19) forces

$$(1.9.20) \quad av' + (2ar + b)v = 0, \quad \text{with } v = u',$$



which, by results of §1.1, forces

$$(1.9.21) \quad \begin{aligned} v(t) &= K_0 e^{-(2r+b/a)t}, \quad \text{hence} \\ u(t) &= K_1 + K_2 e^{-(2r+b/a)t}, \end{aligned}$$

for some constants  $K_0$ ,  $K_1$ , and  $K_2$ . This in turn implies

$$(1.9.22) \quad x(t) = K_1 e^{rt} + K_2 e^{-(r+b/a)t}.$$

But (1.9.6) gives  $r_+ + r_- = -b/a$ , hence

$$(1.9.23) \quad r = r_+ \implies -\left(r + \frac{b}{a}\right) = r_-,$$

so (1.9.22) is indeed of the form (1.9.7), with  $C_+ = K_1$  and  $C_- = K_2$ .

The arguments given above show that indeed all solutions to (1.9.1) have the form (1.9.7) or (1.9.15), in Cases I and II, respectively. We say that (1.9.7) (in Case I) and (1.9.15) (in Case II) provide the *general solution* to (1.9.1). This analysis of the general solutions together with the computations giving (1.9.12) and (1.9.18), establish the following.

**Theorem 1.9.1.** *Given  $a$ ,  $b$ , and  $c$ , with  $a \neq 0$ , and given  $\alpha$  and  $\beta$ , the initial value problem (1.9.1)–(1.9.2) has a unique solution  $x(t)$ . In Case I,  $x(t)$  has the form (1.9.7), and in Case II, it has the form (1.9.15).*

The results derived above apply whether  $a$ ,  $b$ , and  $c$  are real or not. If we assume they are real, then Case I naturally divides into two sub cases:

$$\begin{aligned} \text{(IA)} \quad & b^2 - 4ac > 0, \\ \text{(IB)} \quad & b^2 - 4ac < 0. \end{aligned}$$

In Case IA, the roots of the characteristic equation (1.9.5) given by (1.9.6) are real. In Case IB, we have complex roots, of the form

$$(1.9.24) \quad r_{\pm} = r \pm i\sigma, \quad r = -\frac{b}{2a}, \quad \sigma = \frac{1}{2a} \sqrt{4ac - b^2}.$$

Hence the solutions (1.9.7) have the form

$$(1.9.25) \quad x(t) = C_+ x_+(t) + C_- x_-(t), \quad x_{\pm}(t) = e^{(r \pm i\sigma)t}.$$

From §1 we have  $e^{(r \pm i\sigma)t} = e^{rt} e^{\pm i\sigma t}$ , and also

$$(1.9.26) \quad e^{\pm i\sigma t} = \cos \sigma t \pm i \sin \sigma t.$$

Hence

$$(1.9.27) \quad x_{\pm}(t) = e^{rt} (\cos \sigma t \pm i \sin \sigma t).$$

In particular, the following are also solutions to (1.9.1):

$$(1.9.28) \quad \begin{aligned} x_1(t) &= \frac{1}{2} (x_+(t) + x_-(t)) = e^{rt} \cos \sigma t, \\ x_2(t) &= \frac{1}{2i} (x_+(t) - x_-(t)) = e^{rt} \sin \sigma t. \end{aligned}$$

We can hence rewrite (1.9.25) as  $x(t) = C_1 x_1(t) + C_2 x_2(t)$ , or equivalently

$$(1.9.29) \quad x(t) = C_1 e^{rt} \cos \sigma t + C_2 e^{rt} \sin \sigma t,$$

for some constants  $C_1$  and  $C_2$ , related to  $C_+$  and  $C_-$  by

$$(1.9.30) \quad C_1 = C_+ + C_-, \quad C_2 = i(C_+ - C_-).$$

We can combine these relations with (1.9.11)–(1.9.12) to solve the initial value problem (1.9.1)–(1.9.2).

We now apply the methods just developed to the linearized pendulum and damped pendulum equations (1.8.10) and (1.8.11), i.e.,

$$(1.9.31) \quad \frac{d^2u}{dt^2} + \frac{g}{\ell}u = 0,$$

and

$$(1.9.32) \quad \frac{d^2u}{dt^2} + \frac{\alpha}{m} \frac{du}{dt} + \frac{g}{\ell}u = 0.$$

Here,  $g, \ell, \alpha$ , and  $m$  are all  $> 0$ . Let us set

$$(1.9.33) \quad k = \sqrt{\frac{g}{\ell}}, \quad b = \frac{\alpha}{m},$$

so  $b > 0, k > 0$ , and the equations (1.9.31)–(1.9.32) become

$$(1.9.34) \quad \frac{d^2u}{dt^2} + k^2u = 0,$$

and

$$(1.9.35) \quad \frac{d^2u}{dt^2} + b \frac{du}{dt} + k^2u = 0.$$

The characteristic equation for (1.9.34) is  $r^2 + k^2 = 0$ , with roots  $r = \pm ik$ . The general solution to (1.9.34) can hence be written either as  $u(t) = C_+e^{ikt} + C_-e^{-ikt}$  or as

$$(1.9.36) \quad u(t) = C_1 \cos kt + C_2 \sin kt.$$

The resulting motion is oscillatory motion, with period  $2\pi/k$ .

The characteristic equation for (1.9.35) is  $r^2 + br + k^2 = 0$ , with roots

$$(1.9.37) \quad r_{\pm} = -\frac{b}{2} \pm \frac{1}{2}\sqrt{b^2 - 4k^2}.$$

There are three cases to consider:

$$(IB) \quad b^2 - 4k^2 < 0,$$

$$(II) \quad b^2 - 4k^2 = 0,$$

$$(IA) \quad b^2 - 4k^2 > 0.$$

In Case IB, say  $b^2 - 4k^2 = -4\kappa^2$ . Then  $r_{\pm} = -(b/2) \pm i\kappa$ , and the general solution to (1.9.35) has the form

$$(1.9.38) \quad u(t) = C_1 e^{-bt/2} \cos \kappa t + C_2 e^{-bt/2} \sin \kappa t.$$

These decay exponentially as  $t \nearrow +\infty$ . This is damped oscillatory motion. The oscillatory factors have period

$$(1.9.39) \quad \frac{2\pi}{\kappa} = \frac{2\pi}{\sqrt{k^2 - (b/2)^2}},$$

which approaches  $\infty$  as  $b \nearrow 2k$ .

In Case IA, say  $\beta = \sqrt{b^2 - 4k^2}$ , so  $r_{\pm} = (-b \pm \beta)/2$ . Note that  $0 < \beta < b$ , so both  $r_+$  and  $r_-$  are negative. The general solution to (1.9.35) then has the form

$$(1.9.40) \quad u(t) = C_1 e^{(-b+\beta)t/2} + C_2 e^{(-b-\beta)t/2}, \quad -b \pm \beta < 0.$$

These decay without oscillation as  $t \nearrow +\infty$ . One says this motion is *overdamped*. In Case II, the characteristic equation for (1.9.35) has the double root  $-b/2$ , and the general solution to (1.9.35) has the form

$$(1.9.41) \quad u(t) = C_1 e^{-bt/2} + C_2 t e^{-bt/2}.$$

These also decay without oscillation as  $t \nearrow +\infty$ . One says this motion is *critically damped*.

The nonlinear damped pendulum equation (1.7.6) can also be shown to manifest these damped oscillatory, critically damped, and overdamped behaviors.

## Exercises

1. Find the general solution to each of the following equations for  $x = x(t)$ .

(a) 
$$x'' + 25x = 0.$$

(b) 
$$x'' - 25x = 0.$$

(c) 
$$x'' - 2x' + x = 0.$$

(d) 
$$x'' + 2x' + x = 0.$$

(e) 
$$x'' + x' + x = 0.$$

2. In each case (a)–(e) of Exercise 1, find the solution satisfying the initial condition

$$x(0) = 1, \quad x'(0) = 0.$$

3. In each case (a)–(e) of Exercise 1, find the solution satisfying the initial condition

$$x(0) = 0, \quad x'(0) = 1.$$

4. For  $\varepsilon \neq 0$ , solve the initial value problem

$$x''_{\varepsilon} - 2x'_{\varepsilon} + (1 - \varepsilon^2)x_{\varepsilon} = 0, \quad x_{\varepsilon}(0) = 0, \quad x'_{\varepsilon}(0) = 1.$$

Compute the limit

$$x(t) = \lim_{\varepsilon \rightarrow 0} x_{\varepsilon}(t),$$

and show that the limit solves

$$x'' - 2x' + x = 0, \quad x(0) = 0, \quad x'(0) = 1.$$

### 1.10. Nonhomogeneous equations I – undetermined coefficients

We study nonhomogeneous, second order, constant coefficient linear equations, that is to say, equations of the form

$$(1.10.1) \quad a \frac{d^2x}{dt^2} + b \frac{dx}{dt} + cx = f(t),$$

with constants  $a, b$ , and  $c$  ( $a \neq 0$ ) and a given function  $f(t)$ . The equation (1.10.1) is called nonhomogeneous whenever  $f(t)$  is not identically 0. We might impose initial conditions, like

$$(1.10.2) \quad x(0) = \alpha, \quad x'(0) = \beta.$$

In this section we assume  $f(t)$  is a constant multiple of one of the following functions, or perhaps a finite sum of such functions:

$$(1.10.3) \quad e^{\kappa t},$$

$$(1.10.4) \quad \sin \sigma t,$$

$$(1.10.5) \quad \cos \sigma t,$$

$$(1.10.6) \quad t^k.$$

We discuss a method, called the “method of undetermined coefficients,” to solve (1.10.1) in such cases. In §1.14 we will discuss a method that applies to a broader class of functions  $f$ .

We begin with the case (1.10.3). The first strategy is to seek a solution in the form

$$(1.10.7) \quad x(t) = Ae^{\kappa t}.$$

Here  $A$  is the “undetermined coefficient.” The goal will be to determine it. Plugging (1.10.7) into the left side of (1.10.1) gives

$$(1.10.8) \quad ax'' + bx' + cx = A(a\kappa^2 + b\kappa + c)e^{\kappa t}.$$

As long as  $\kappa$  is not a root of the characteristic polynomial  $p(r) = ar^2 + br + c$ , we get a solution to (1.10.1) in the form (1.10.7), with

$$(1.10.9) \quad A = \frac{1}{a\kappa^2 + b\kappa + c}.$$

In such a case, the equation

$$(1.10.10) \quad a \frac{d^2x}{dt^2} + b \frac{dx}{dt} + cx = Be^{\kappa t}$$

has a solution

$$(1.10.11) \quad x_p(t) = AB e^{\kappa t},$$

with  $A$  given by (1.10.9). We say  $x_p(t)$  is a *particular* solution to (1.10.10). If  $x(t)$  is another solution, then, because the equation is linear,  $y(t) = x(t) - x_p(t)$  solves the homogeneous equation

$$(1.10.12) \quad a \frac{d^2y}{dt^2} + b \frac{dy}{dt} + cy = 0,$$

which was treated in §1.9. If, for example,  $p(r)$  has distinct roots  $r_+$  and  $r_-$ , we know the general solution of (1.10.11) is

$$(1.10.13) \quad y(t) = C_+ e^{r_+ t} + C_- e^{r_- t}.$$

Then the general solution to (1.10.10) is

$$(1.10.14) \quad x(t) = \frac{B}{a\kappa^2 + b\kappa + c} e^{\kappa t} + C_+ e^{r_+ t} + C_- e^{r_- t}.$$

In (1.10.14),  $a, b, c, B$ , and  $\kappa$  are given by (1.10.10), and  $C_+$  and  $C_-$  are arbitrary constants. If the initial conditions in (1.10.2) are imposed, they will determine  $C_+$  and  $C_-$ . If  $r_+$  and  $r_-$  are complex, we could rewrite (1.10.13)–(1.10.14), using Euler's formula, as in §1.9.

Formulas (1.10.11)–(1.10.14) hold under the hypothesis that  $r_+, r_-$ , and  $\kappa$  are all distinct. If the characteristic polynomial has a double root  $r = r_+ = r_-$ , distinct from  $\kappa$ , then we replace (1.10.13) by

$$(1.10.15) \quad y(t) = C_1 e^{rt} + C_2 t e^{rt},$$

and the general solution to (1.10.10) has the form

$$(1.10.16) \quad x(t) = \frac{B}{a\kappa^2 + b\kappa + c} e^{\kappa t} + C_1 e^{rt} + C_2 t e^{rt}.$$

Again, the initial conditions (1.10.2) would determine  $C_1$  and  $C_2$ .

We turn to the case that  $\kappa$  is a root of the characteristic polynomial  $p(r)$ . In such a case, (1.10.8) vanishes, and there is not a solution to (1.10.1) in the form (1.10.7). This study splits into two cases. First assume  $p(r)$  has distinct roots. Say  $\kappa = r_+ \neq r_-$ . Then (1.10.1) (with  $f(t) = e^{\kappa t}$ ) will have a solution of the form

$$(1.10.17) \quad x(t) = A t e^{\kappa t}.$$

Indeed, a computation parallel to (1.9.14), with  $u(t) = At$ ,  $r = \kappa$ , gives

$$(1.10.18) \quad ax'' + bx' + cx = (2a\kappa + b)Ae^{\kappa t},$$

since in this case  $u'' = 0$  and  $a\kappa^2 + b\kappa + c = 0$ . Then (1.10.1) holds with  $f(t) = e^{\kappa t}$ , provided

$$(1.10.19) \quad A = \frac{1}{2a\kappa + b},$$

and more generally a particular solution to (1.10.10) is given by

$$(1.10.20) \quad x_p(t) = ABt e^{\kappa t},$$

with  $A$  given by (1.10.19). As above, the general solution to (1.10.10) then has the form

$$(1.10.21) \quad x(t) = x_p(t) + y(t),$$

where  $y(t)$  solves (1.10.12), hence has the form (1.10.13). (Recall we are assuming  $r_+ \neq r_-$ .)

To finish the analysis of (1.10.10), it remains to consider the case  $\kappa = r_+ = r_-$ . Then functions of the form (1.10.15) (with  $r = \kappa$ ) solve (1.10.12), so there is not a solution to (1.10.1) (with  $f(t) = e^{\kappa t}$ ) of the form (1.10.17). Instead, we will find a solution of the form

$$(1.10.22) \quad x(t) = At^2 e^{\kappa t}.$$

In this case, a computation parallel to (1.9.14), with  $u(t) = At^2$ ,  $r = \kappa$ , gives

$$(1.10.23) \quad ax'' + bx' + cx = 2aAe^{\kappa t},$$

since in this case  $u'' = 2A$ ,  $2a\kappa + b = 0$ , and  $a\kappa^2 + b\kappa + c = 0$ . Then (1.10.1) holds with  $f(t) = e^{\kappa t}$  provided

$$(1.10.24) \quad A = \frac{1}{2a},$$

and more generally a particular solution to (1.10.10) is given by

$$(1.10.25) \quad x_p(t) = ABt^2e^{\kappa t},$$

with  $A$  given by (1.10.24). Then the general solution to (1.10.10) has the form (1.10.21), where  $y(t)$  solves (1.10.12), hence has the form (1.10.15), with  $r = \kappa$ . (Recall we are assuming  $r_+ = r_-$ .)

As a slight extension of (1.10.10), consider the equation

$$(1.10.26) \quad a\frac{d^2x}{dt^2} + b\frac{dx}{dt} + cx = B_1e^{\kappa_1 t} + B_2e^{\kappa_2 t}.$$

This has a solution of the form

$$(1.10.27) \quad x_p(t) = x_{p1}(t) + x_{p2}(t),$$

where  $x_{pj}(t)$  are particular solutions of (1.10.10), with  $B$  replaced by  $B_j$  and  $\kappa$  replaced by  $\kappa_j$ . Then the general solution to (1.10.26) has the form (1.10.21), with  $x_p(t)$  given by (1.10.27) and  $y(t)$  solving (1.10.12).

We move on to cases of  $f(t)$  given by (1.10.4) and (1.10.5), which we combine as follows:

$$(1.10.28) \quad a\frac{d^2x}{dt^2} + c\frac{dx}{dt} + cx = b_1 \sin \sigma t + b_2 \cos \sigma t.$$

Via Euler's formula we can write

$$(1.10.29) \quad \begin{aligned} b_1 \sin \sigma t + b_2 \cos \sigma t &= B_1 e^{i\sigma t} + B_2 e^{-i\sigma t}, \\ B_1 &= \frac{b_1}{2i} + \frac{b_2}{2}, \quad B_2 = -\frac{b_1}{2i} + \frac{b_2}{2}, \end{aligned}$$

and we are back in the setting (1.10.26), with  $\kappa_1 = i\sigma$ ,  $\kappa_2 = -i\sigma$ . Thus, for example, if  $\pm i\sigma$  are not roots of the characteristic polynomial  $p(r) = ar^2 + br + c$ , we have a particular solution of the form

$$(1.10.30) \quad x_p(t) = A_1 B_1 e^{i\sigma t} + A_2 B_2 e^{-i\sigma t},$$

where  $B_1$  and  $B_2$  are as in (1.10.29) and the undetermined coefficients  $A_1$  and  $A_2$  can be obtained by plugging into (1.10.28). As an alternative presentation, we can again use Euler's formula to rewrite (1.10.30) as

$$(1.10.31) \quad x_p(t) = a_1 \sin \sigma t + a_2 \cos \sigma t,$$

where the undetermined coefficients  $a_1$  and  $a_2$  are obtained by plugging into (1.10.28).

If  $a$ ,  $b$ , and  $c$  in (1.10.1) are all real, then  $p(r)$  will not have purely imaginary roots if  $b \neq 0$ . If  $b = 0$ , the roots will be  $r_{\pm} = \pm\sqrt{-c/a}$ , which are real if  $c/a < 0$  and purely imaginary if  $c/a > 0$ . In case  $r_{\pm} = \pm i\sigma$ , considerations parallel to

(1.10.17)–(1.10.20) apply, with  $\kappa = \pm i\sigma$ . Again a further application of Euler's formula gives

$$(1.10.32) \quad x_p(t) = a_1 t \sin \sigma t + a_2 t \cos \sigma t,$$

where the coefficients  $a_1$  and  $a_2$  are obtained by plugging into (1.10.28).

We now move to cases of  $f(t)$  given by (1.10.6). Take  $k = 1$ , so we are looking at

$$(1.10.33) \quad ax'' + bx' + cx = t.$$

We try

$$(1.10.34) \quad x(t) = At + B,$$

for which  $x' = A$ ,  $x'' = 0$ , and the left side of (1.10.33) is  $cAt + (B + bA)$ . The condition that (1.10.33) hold is

$$(1.10.35) \quad cA = 1, \quad B + bA = 0,$$

solved by

$$(1.10.36) \quad A = \frac{1}{c}, \quad B = -\frac{b}{c},$$

assuming  $c \neq 0$ . If  $c = 0$ , we want to solve (for  $v = dx/dt$ )

$$(1.10.37) \quad av' + bv = t.$$

We try

$$(1.10.38) \quad v(t) = \alpha t + \beta,$$

for which  $v' = \alpha$  and the left side of (1.10.37) is  $a\alpha + b(\alpha t + \beta)$ . The condition that (1.10.37) hold is

$$(1.10.39) \quad b\alpha = 1, \quad a\alpha + b\beta = 0,$$

solved by

$$(1.10.40) \quad \alpha = \frac{1}{b}, \quad \beta = -\frac{a}{b^2},$$

assuming  $b \neq 0$ . In such a case, we can take

$$(1.10.41) \quad x(t) = \frac{\alpha}{2} t^2 + \beta t.$$

In case  $c = b = 0$ , (1.10.32) becomes

$$(1.10.42) \quad ax'' = t,$$

with solution

$$(1.10.43) \quad x(t) = \frac{1}{6a} t^3.$$

Analogous considerations apply to (1.10.6) with  $k \geq 2$ . The method can also be extended to treat  $f(t)$  in the form

$$(1.10.44) \quad t^k e^{\kappa t}, \quad t^k \sin \sigma t, \quad t^k \cos \sigma t.$$

We omit details. In such cases, it is just as convenient to use the method developed in §1.14.



See §1.16 for further insight on why the method of undetermined coefficients works for functions  $f(t)$  of the form (1.10.3)–(1.10.6), and more generally of the form (1.10.44).

---

### Exercises

1. Find the general solution to each of the following equations for  $x = x(t)$ .

(a) 
$$x'' + 25x = e^{5t}.$$

(b) 
$$x'' - 25x = e^{5t}.$$

(c) 
$$x'' - 2x + x = \sin t.$$

(d) 
$$x'' + 2x' + x = e^t.$$

(e) 
$$x'' + x' + x = \cos t.$$

2. In each case (a)–(e) of Exercise 1, find the solution satisfying the initial conditions

$$x(0) = 1, \quad x'(0) = 0.$$

3. In each case (a)–(e) of Exercises 1, find the solution satisfying the initial conditions

$$x(0) = 0, \quad x'(0) = 1.$$

4. For  $\varepsilon \neq 0$ , solve the initial value problem

$$x''_{\varepsilon} - 25x_{\varepsilon} = e^{(5+\varepsilon)t}, \quad x_{\varepsilon}(0) = 1, \quad x'_{\varepsilon}(0) = 0.$$

Compute the limit

$$x(t) = \lim_{\varepsilon \rightarrow 0} x_{\varepsilon}(t),$$

and show that the limit solves

$$x'' - 25x = e^{5t}, \quad x(0) = 1, \quad x'(0) = 0.$$

### 1.11. Forced pendulum – resonance

Here we study the following special cases of (1.10.28), modeling the linearized pendulum and damped pendulum, respectively, subjected to an additional periodic force of the form  $F_0 \sin \sigma t$ . The equations we consider are, respectively,

$$(1.11.1) \quad \frac{d^2 u}{dt^2} + \frac{g}{\ell} u = F_0 \sin \sigma t,$$

and

$$(1.11.2) \quad \frac{d^2 u}{dt^2} + \frac{\alpha}{m} \frac{du}{dt} + \frac{g}{\ell} u = F_0 \sin \sigma t.$$

The quantities  $\alpha, m, g$ , and  $\ell$  are all positive, and we take  $F_0$  and  $\sigma$  to be real. As in (1.9.33), we set

$$(1.11.3) \quad k = \sqrt{\frac{g}{\ell}}, \quad b = \frac{\alpha}{m},$$

so  $b > 0$ ,  $k > 0$ , and the equations (1.11.1)–(1.11.2) become

$$(1.11.4) \quad \frac{d^2 u}{dt^2} + k^2 u = F_0 \sin \sigma t,$$

and

$$(1.11.5) \quad \frac{d^2 u}{dt^2} + b \frac{du}{dt} + k^2 u = F_0 \sin \sigma t.$$

As long as  $k \neq \pm \sigma$ , we can set  $u(t) = a_1 \sin \sigma t$  and the left side of (1.11.4) equals  $a_1(k^2 - \sigma^2) \sin \sigma t$ , so a solution to (1.11.4) is

$$(1.11.6) \quad u_p(t) = \frac{F_0}{k^2 - \sigma^2} \sin \sigma t,$$

in such a case. Note how the coefficient  $F_0/(k^2 - \sigma^2)$  blows up as  $\sigma \rightarrow \pm k$ . If  $\sigma = k$ , then, as in (1.10.32), we need to seek a solution to (1.11.4) of the form

$$(1.11.7) \quad u_p(t) = a_1 t \sin \sigma t + a_2 t \cos \sigma t.$$

In such a case,

$$(1.11.8) \quad u_p'' + k^2 u_p = 2a_1 \sigma \cos \sigma t - 2a_2 \sigma \sin \sigma t,$$

so (1.11.4) holds provided

$$(1.11.9) \quad -2a_2 \sigma = F_0, \quad 2a_1 \sigma = 0,$$

i.e., we have

$$(1.11.10) \quad u_p(t) = -\frac{F_0}{2\sigma} t \cos \sigma t.$$

Note that  $u_p(t)$  grows without bound as  $|t| \rightarrow \infty$  in this case, as opposed to the bounded behavior in  $t$  given by (1.11.6) when  $\sigma^2 \neq k^2$ . We say we have a *resonance* at  $\sigma^2 = k^2$ .

Moving on to (1.11.5), as in (1.9.37) the characteristic polynomial  $p(r) = r^2 + br + k^2$  has roots

$$(1.11.11) \quad r_{\pm} = -\frac{b}{2} \pm \frac{1}{2} \sqrt{b^2 - 4k^2},$$

and as long as  $b > 0$ ,  $\pm i\sigma \neq r_{\pm}$ . Hence we can seek a solution to (1.11.5) in the form

$$(1.11.12) \quad u_p(t) = a_1 \sin \sigma t + a_2 \cos \sigma t.$$

A computation gives

$$(1.11.13) \quad \begin{aligned} u_p'' + bu_p' + k^2 u_p &= (-a_1 \sigma^2 - a_2 b \sigma + a_1 k^2) \sin \sigma t \\ &\quad + (-a_2 \sigma^2 + a_1 b \sigma + a_2 k^2) \cos \sigma t, \end{aligned}$$

so  $u_p$  is a solution to (1.11.5) if and only if

$$(1.11.14) \quad \begin{aligned} (k^2 - \sigma^2)a_1 - (b\sigma)a_2 &= F_0, \\ (b\sigma)a_1 + (k^2 - \sigma^2)a_2 &= 0. \end{aligned}$$

Solving for  $a_1$  and  $a_2$  gives

$$(1.11.15) \quad \begin{aligned} a_1 &= \frac{k^2 - \sigma^2}{(k^2 - \sigma^2)^2 + (b\sigma)^2} F_0, \\ a_2 &= -\frac{b\sigma}{(k^2 - \sigma^2)^2 + (b\sigma)^2} F_0. \end{aligned}$$

We can rewrite (1.11.12) as

$$(1.11.16) \quad u_p(t) = A \sin(\sigma t + \theta),$$

for some constants  $A$  and  $\theta$ , using the identity

$$(1.11.17) \quad A \sin(\sigma t + \theta) = A(\cos \theta) \sin \sigma t + A(\sin \theta) \cos \sigma t.$$

It follows that (1.11.16) is equivalent to (1.11.12) provided

$$(1.11.18) \quad A \cos \theta = a_1, \quad A \sin \theta = a_2,$$

i.e., provided

$$(1.11.19) \quad a_1 + ia_2 = Ae^{i\theta}.$$

We take  $A > 0$  such that

$$(1.11.20) \quad A^2 = a_1^2 + a_2^2 = \frac{F_0^2}{(k^2 - \sigma^2)^2 + (b\sigma)^2}.$$

Thus

$$(1.11.21) \quad A = \frac{|F_0|}{\sqrt{(k^2 - \sigma^2)^2 + (b\sigma)^2}}$$

is the amplitude of the solution (1.11.16).

If  $b$ ,  $k$ , and  $F_0$  are fixed quantities in (1.11.5) and  $\sigma$  is allowed to vary,  $A$  in (1.11.21) is maximized at the value of  $\sigma$  for which

$$(1.11.22) \quad \beta(\sigma) = (k^2 - \sigma^2)^2 + (b\sigma)^2$$

is minimal. We have

$$(1.11.23) \quad \begin{aligned} \beta'(\sigma) &= 4\sigma^3 + 2(b^2 - 2k^2)\sigma \\ &= 4\sigma \left[ \sigma^2 - \left( k^2 - \frac{b^2}{2} \right) \right]. \end{aligned}$$

Note that  $\sigma = 0$  is a critical point, and  $\beta(0) = k^4$ . There are two cases. First,

$$(1.11.24) \quad k^2 - \frac{b^2}{2} > 0 \implies \beta_{\min} = \beta\left(\pm\sqrt{k^2 - \frac{b^2}{2}}\right) \\ = b^2\left(k^2 - \frac{b^2}{4}\right),$$

since  $k^4 \geq b^2(k^2 - b^2/4)$ . (Indeed, taking  $\xi = k^2/b^2$ , this inequality is equivalent to  $\xi^2 \geq \xi - 1/4$ ; but  $\xi^2 - \xi + 1/4 = (\xi - 1/2)^2$ .) In the second case,

$$(1.11.25) \quad k^2 - \frac{b^2}{2} \leq 0 \implies \beta_{\min} = \beta(0) = k^4.$$

In these respective cases, we get

$$(1.11.26) \quad A_{\max} = \frac{|F_0|}{b} \left(k^2 - \frac{b^2}{4}\right)^{-1/2},$$

and

$$(1.11.27) \quad A_{\max} = \frac{|F_0|}{k^2}.$$

In the first case, i.e., (1.11.24), we say resonance is achieved at  $\sigma^2 = k^2 - b^2/2$ . Recall from §1.9 that critical damping occurs for  $k^2 = b^2/4$ , for the unforced pendulum, so in case (1.11.24) the unforced pendulum has damped oscillatory motion.

## Exercises

1. Find the general solution to

$$(1.11.28) \quad \frac{d^2u}{dt^2} + \frac{du}{dt} + u = 3 \sin \sigma t.$$

2. For the equation in Exercise 1, find the value of  $\sigma$  for which there is resonance.

3. Would the answer to Exercise 2 change if the right side of (1.11.28) were changed to

$$10 \sin \sigma t?$$

Explain.

4. Do analogues of Exercises 1–2 with (1.11.28) replaced by each of the following:

$$\frac{d^2u}{dt^2} + \frac{du}{dt} + 3u = \sin \sigma t,$$

$$\frac{d^2u}{dt^2} + 2\frac{du}{dt} + 3u = 2 \sin \sigma t.$$

5. Do analogues of Exercise 1 with (1.11.28) replaced by the following:

$$\frac{d^2u}{dt^2} + 2\frac{du}{dt} + u = 3 \sin \sigma t.$$

Discuss the issue of resonance in this case.

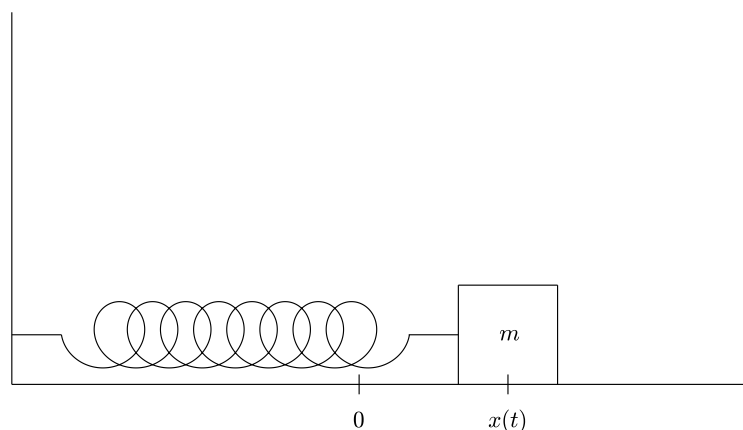


Figure 1.12.1. Mass on a spring

## 1.12. Spring motion

We consider the motion of a body of mass  $m$ , attached to one end of a spring, as depicted in Figure 1.12.1. The other end of the spring is attached to a rigid wall, and the weight slides along the floor, pushed or pulled by the spring. We assume that the force of the spring is a function of position:

$$(1.12.1) \quad F_1 = F_1(x).$$

We pick the origin to be the position where the spring is relaxed, so  $F(0) = 0$ . A good approximation, valid for small oscillations, is

$$(1.12.2) \quad F_1(x) = -Kx,$$

with a positive constant  $K$  (called the spring constant). This approximation loses accuracy if  $|x|$  is large. Sliding along the floor typically produces a frictional force that is a function of the velocity  $v = dx/dt$ . A good approximation for the frictional force is

$$(1.12.3) \quad F_2 = F_2(v) = -av,$$

where  $a$  is a positive constant, called the coefficient of friction. The total force on the mass is  $F = F_1 + F_2$ , and Newton's law  $F = ma$  yields the differential equation

$$(1.12.4) \quad m \frac{d^2x}{dt^2} + a \frac{dx}{dt} + Kx = 0.$$

This has the same form as (1.9.35), i.e.,

$$(1.12.5) \quad \frac{d^2x}{dt^2} + b\frac{dx}{dt} + k^2x = 0,$$

with

$$(1.12.6) \quad b = \frac{a}{m}, \quad k^2 = \frac{K}{m},$$

both positive, and the analysis of (1.9.35) applies here, including notions of oscillatory damped, critically damped, and overdamped motion.

One can consider systems of several masses, connected via springs. These situations lead to systems of differential equations, studied in Chapter 3.

---

### Exercises

1. Suppose one has a spring system as in Figure 1.12.1. Assume the mass  $m$  is 2 kg and the spring constant  $K$  is 6 kg/sec<sup>2</sup>. There is a frictional force of  $a$  kg/sec. Find the values of  $a$  for which the spring motion is

- (a) damped oscillatory,
- (b) critically damped,
- (c) overdamped.

2. In the context of Exercise 1, suppose there is also an external force of the form

$$10 \sin \sigma t \quad \text{kg-m/sec}^2.$$

(Assume  $x$  is given in meters.) Take

$$a = 2,$$

so (12.4) becomes

$$2\frac{d^2x}{dt^2} + 2\frac{dx}{dt} + 6x = 10 \sin \sigma t.$$

Find the value of  $\sigma$  for which there is resonance.

### 1.13. RLC circuits

Here we derive a differential equation for the current flowing through the circuit depicted in Figure 1.13.1, which consists of a resistor, with resistance  $R$  (in ohms), a capacitor, with capacitance  $C$  (in farads), and an inductor, with inductance  $L$  (in henrys). The circuit is plugged into a source of electricity, providing voltage  $E(t)$  (in volts). As stated, we want to find a differential equation for the current  $I(t)$  (in amps).

The equation is derived using two types of basic laws. The first type consists of two rules, which are special cases of Kirchhoff's laws:

- (A) The sum of the voltage drops across the three circuit elements is  $E(t)$ .
- (B) For each  $t$ , the same current  $I(t)$  flows through each circuit element.

For more complicated circuits than the one depicted in Figure 1.13.1, these rules take a more elaborate form. We return to this in Chapter 3.

The second type of law specifies the voltage drop across each circuit element:

- (a) Resistor:  $V = IR$ ,
- (b) Inductor:  $V = L \frac{dI}{dt}$ ,
- (c) Capacitor:  $V = \frac{Q}{C}$ .

As stated above,  $V$  is measured in volts,  $I$  in amps,  $R$  in ohms,  $L$  in henrys, and  $C$  in farads. In addition,  $Q$  is the charge on the capacitor, measured in coulombs. The rule (c) is supplemented by the following formula for the current across the capacitor:

$$(c2) \quad I = \frac{dQ}{dt}.$$

In (b) and (c2), time is measured in seconds.

In Figure 1.13.1, the circuit elements are numbered. We let  $V_j = V_j(t)$  denote the voltage drop across element  $j$ . Rules (A), (B), and (a) give

$$(1.13.1) \quad V_1 + V_2 + V_3 = E(t),$$

$$(1.13.2) \quad V_1 = RI.$$

Rules (B), (b), and (c)–(c2) give differential equations:

$$(1.13.3) \quad L \frac{dI}{dt} = V_3,$$

$$(1.13.4) \quad C \frac{dV_2}{dt} = I.$$

Plugging (1.13.2)–(1.13.3) into (1.13.1) gives

$$(1.13.5) \quad RI + V_2 + L \frac{dI}{dt} = E(t).$$



Applying  $d/dt$  to (1.13.5) and using (1.13.4) gives

$$(1.13.6) \quad L \frac{d^2 I}{dt^2} + R \frac{dI}{dt} + \frac{1}{C} I = E'(t).$$

This is the equation for the RLC circuit in Figure 1.13.1. If we divide by  $L$  we get

$$(1.13.7) \quad \frac{d^2 I}{dt^2} + \frac{R}{L} \frac{dI}{dt} + \frac{1}{LC} I = \frac{E'(t)}{L},$$

which has the same form as the (linearized) damped driven pendulum (1.11.5), with

$$(1.13.8) \quad b = \frac{R}{L}, \quad k^2 = \frac{1}{LC},$$

except that at this point  $E'(t)/L$  is not specified to agree with the right side of (1.11.5). However, indeed, if alternating current powers this circuit, it is reasonable to take

$$(1.13.9) \quad E(t) = E_0 \cos \sigma t,$$

so

$$(1.13.10) \quad \frac{1}{L} E'(t) = -\frac{\sigma E_0}{L} \sin \sigma t = F_0 \sin \sigma t.$$

Then analyses of solutions done in §1.11, including analyses of resonance phenomena, apply in this setting.

Actually, in this setting a different perspective on resonance is in order. The frequency  $\sigma/2\pi$  cycles/sec of the alternating current is typically fixed, while one might be able to adjust the capacitance  $C$ . Let us assume  $R$  and  $L$  are also fixed, so  $b$  in (1.13.8) is fixed but one might adjust  $k$ . Recalling the formulas (1.11.16) and (1.11.21), which in this setting take the form

$$(1.13.11) \quad I_p(t) = A \sin(\sigma t + \theta), \quad A = \frac{|F_0|}{\sqrt{(k^2 - \sigma^2)^2 + (b\sigma)^2}},$$

we see that for fixed  $b$  and  $\sigma$ , this amplitude is maximized for  $k$  satisfying

$$(1.13.12) \quad k^2 = \sigma^2,$$

i.e., for

$$(1.13.13) \quad LC = \frac{1}{\sigma^2}.$$

More elaborate circuits, containing a larger number of circuit elements, and more loops, are naturally treated in the context of systems of differential equations. See Chapter 3 for more on this.

REMARK. Consistent with formulas (a)–(c) and (c2), the units mentioned above

are related as follows:

$$\begin{aligned}
 (1.13.14) \quad & 1 \text{ amp} = 1 \frac{\text{coulomb}}{\text{sec}} \\
 & 1 \text{ farad} = 1 \frac{\text{coulomb}}{\text{volt}} \\
 & 1 \text{ henry} = 1 \frac{\text{volt-sec}}{\text{amp}} \\
 & 1 \text{ ohm} = 1 \frac{\text{volt}}{\text{amp}}.
 \end{aligned}$$

To relate these to other physical units, we mention that

$$\begin{aligned}
 (1.13.15) \quad & 1 \text{ volt} = 1 \text{ joule/coulomb} \\
 & 1 \text{ watt} = 1 \text{ volt-amp} = 1 \text{ joule/sec} \\
 & 1 \text{ joule} = 1 \text{ Newton-meter} \\
 & 1 \text{ Newton} = 1 \text{ kg-m/sec}^2.
 \end{aligned}$$

The force of gravity at the surface of the earth on a 1 kg. object is 9.8 Newtons, or, alternatively, 2.2 pounds. In other words, 1 Newton is about 0.224 pounds. Hence one joule is about 0.735 foot-pounds.

The Coulomb is a unit of charge with the following property. If two particles, of charge  $q_1$  and  $q_2$  Coulombs, are separated by  $r$  meters, the force between them is given by Coulomb's law:

$$(1.13.16) \quad F = k \frac{q_1 q_2}{r^2} \text{ Newtons, } k = 8.99 \times 10^9.$$

Investigations into the nature of electrons have shown that

$$(1.13.17) \quad -1 \text{ Coulomb} = \text{charge of } 6.24 \times 10^{18} \text{ electrons.}$$

In connection with this, we mention that one gram of water contains  $3.3 \times 10^{23}$  electrons.

## Exercises

1. Consider a circuit as in Figure 1.13.1. Assume the inductance is 4 henrys and the applied current has the form (1.13.9) with a frequency of 60 hertz, i.e., 60 cycles/sec. Find the value of the capacitance  $C$ , in farads, to achieve resonance.
2. Redo Exercise 1, this time with inductance of  $10^{-6}$  henry and applied current of the form (1.13.9) with a frequency of 120 megahertz.

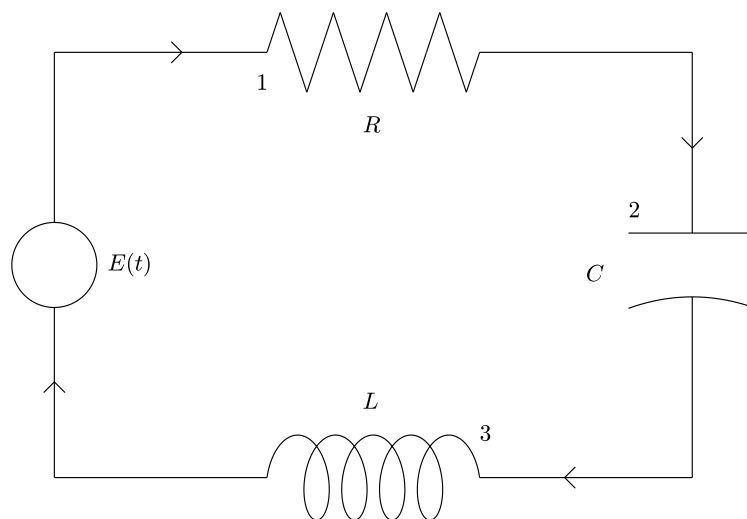


Figure 1.13.1. RLC circuit

### 1.14. Nonhomogeneous equations II – variation of parameters

Here we present another approach to solving

$$(1.14.1) \quad \frac{d^2x}{dt^2} + b\frac{dx}{dt} + cx = f(t),$$

(with constant  $b$  and  $c$ ) called the method of variation of parameters. It works as follows. Let  $y_1(t)$  and  $y_2(t)$  be a complete set of solutions of the homogeneous equation

$$(1.14.2) \quad \frac{d^2y}{dt^2} + b\frac{dy}{dt} + cy = 0.$$

The method consists of seeking a solution to (1.14.1) in the form

$$(1.14.3) \quad x(t) = u_1(t)y_1(t) + u_2(t)y_2(t),$$

and finding equations for  $u_j(t)$  that are simpler than the original equation (1.14.1). We have

$$(1.14.4) \quad x' = u_1y_1' + u_2y_2' + u_1'y_1 + u_2'y_2.$$

It will be convenient to arrange that  $x''$  not involve second order derivatives of  $u_1$  and  $u_2$ . To achieve this, we impose the condition

$$(1.14.5) \quad u_1'y_1 + u_2'y_2 = 0.$$

Then  $x'' = u_1'y_1' + u_2'y_2' + u_1y_1'' + u_2y_2''$ , and using (1.14.2) to replace  $y_j''$  by  $-by_j' - cy_j$ , we get

$$(1.14.6) \quad x'' = u_1'y_1' + u_2'y_2' - (by_1' + cy_1)u_1 - (by_2' + cy_2)u_2,$$

hence

$$(1.14.7) \quad x'' + bx' + cx = y_1'u_1' + y_2'u_2'.$$

Thus we have a solution to (1.14.1) in the form (1.14.3) provided  $u_1'$  and  $u_2'$  solve

$$(1.14.8) \quad \begin{aligned} y_1u_1' + y_2u_2' &= 0, \\ y_1'u_1' + y_2'u_2' &= f. \end{aligned}$$

This linear system for  $u_1'$  and  $u_2'$  has the explicit solution

$$(1.14.9) \quad u_1' = -\frac{y_2}{W}f, \quad u_2' = \frac{y_1}{W}f,$$

where  $W(t)$  is the following determinant, called the Wronskian determinant:

$$(1.14.10) \quad W = y_1y_2' - y_2y_1' = \det \begin{pmatrix} y_1 & y_2 \\ y_1' & y_2' \end{pmatrix}.$$

Determinants will be studied in the next chapter. The reader who has not seen them can take the first identity in (1.14.10) as a definition and ignore the second identity (for now).

Note that if the roots of the characteristic polynomial  $p(r) = r^2 + br + c$  are distinct,  $r_+ \neq r_-$ , we can take

$$(1.14.11) \quad y_1 = e^{r_+t}, \quad y_2 = e^{r_-t},$$

and then

$$(1.14.12) \quad \begin{aligned} W(t) &= r_-e^{r_+t}e^{r_-t} - r_+e^{r_-t}e^{r_+t} \\ &= (r_- - r_+)e^{(r_++r_-)t}, \end{aligned}$$

which is nowhere vanishing. If there is a double root,  $r_+ = r_- = r$ , we can take

$$(1.14.13) \quad y_1 = e^{rt}, \quad y_2 = te^{rt},$$

and then

$$(1.14.14) \quad W(t) = e^{rt}(e^{rt} + tre^{rt}) - te^{rt}re^{rt} = e^{2rt},$$

which is also nowhere vanishing.

Returning to (1.14.9), we can take

$$(1.14.15) \quad \begin{aligned} u_1(t) &= -\int_0^t \frac{y_2(s)}{W(s)} f(s) ds + C_1, \\ u_2(t) &= \int_0^t \frac{y_1(s)}{W(s)} f(s) ds + C_2, \end{aligned}$$

so

$$(1.14.16) \quad x(t) = C_1y_1(t) + C_2y_2(t) + \int_0^t [y_2(t)y_1(s) - y_1(t)y_2(s)] \frac{f(s)}{W(s)} ds.$$

Denote the last term, i.e., the integral, by  $x_p(t)$ .

Note that when the characteristic polynomial  $r^2 + br + c$  has distinct roots  $r_+ \neq r_-$  and (1.14.11)–(1.14.12) hold, we get

$$\begin{aligned} (1.14.17) \quad x_p(t) &= \frac{1}{r_- - r_+} \int_0^t [e^{r_-t} e^{r_+s} - e^{r_+t} e^{r_-s}] \frac{f(s)}{e^{(r_+ + r_-)s}} ds \\ &= \frac{1}{r_- - r_+} \int_0^t [e^{r_-(t-s)} - e^{r_+(t-s)}] f(s) ds. \end{aligned}$$

When the characteristic polynomial has double roots  $r_+ = r_- = r$  and (1.14.13)–(1.14.14) hold, we get

$$\begin{aligned} (1.14.18) \quad x_p(t) &= \int_0^t [te^{rt} e^{rs} - e^{rt} se^{rs}] \frac{f(s)}{e^{2rs}} ds \\ &= \int_0^t (t-s)e^{r(t-s)} f(s) ds. \end{aligned}$$

The next section will continue the study of the Wronskian. Further material on the Wronskian and the method of variation of parameters, in a more general context, can be found in Chapter 3.

---

### Exercises

Use the method of variation of parameters to solve each of the following for  $x = x(t)$ .

1.  $x'' + x = e^t.$
2.  $x'' + x = \sin t.$
3.  $x'' + x = t.$
4.  $x'' + x = t^2.$
5.  $x'' + x = \tan t.$

### 1.15. Variable coefficient second order equations

The general, possibly nonlinear, second order differential equation

$$(1.15.1) \quad \frac{d^2x}{dt^2} = F\left(t, x, \frac{dx}{dt}\right),$$

has already been mentioned in §1.4. If  $F(t, x, v)$  is defined and smooth on a neighborhood of  $t_0, x_0, v_0$ , and one imposes an initial condition

$$(1.15.2) \quad x(t_0) = x_0, \quad x'(t_0) = v_0,$$

it is a fundamental result that (1.15.1)–(1.15.2) has a unique solution, at least for  $t$  in some interval containing  $t_0$ . A more general result of this sort will be proven in Chapter 4.

Linear second order equations have the form

$$(1.15.3) \quad a(t)\frac{d^2x}{dt^2} + b(t)\frac{dx}{dt} + c(t)x = f(t).$$

The existence and uniqueness results stated above apply. There are many specific and much studied examples, such as Bessel's equation

$$(1.15.4) \quad \frac{d^2x}{dt^2} + \frac{1}{t}\frac{dx}{dt} + \left(1 - \frac{\nu^2}{t^2}\right)x = 0,$$

whose solutions are called Bessel functions, and Airy's equation,

$$(1.15.5) \quad \frac{d^2x}{dt^2} - tx = 0,$$

whose solutions are Airy functions, just to mention two examples. Such functions are important and show up in many contexts. We will take a closer look at Bessel's equation in the next section. Linear variable coefficient equations could arise from RLC circuits in which one has variable capacitors, resistors, and inductors, turning (1.13.6) into

$$(1.15.6) \quad L(t)\frac{d^2I}{dt^2} + R(t)\frac{dI}{dt} + \frac{1}{C(t)}I = E'(t).$$

The most frequent source of such equations as (1.15.4)–(1.15.5) comes from the theory of Partial Differential Equations. One such indication of how (1.15.4) arises is given in Appendix 1.A, at the end of this Chapter. The reader can find out much more about these equations in a text on Partial Differential Equations, such as [45]. Solutions to these equations cannot generally be given in terms of elementary functions, such as exponential functions, but are further special functions, for which many analytical techniques have been developed.

As with the exponential function, analyzed in §1.1, power series techniques are very useful. We illustrate this by producing a power series

$$(1.15.7) \quad x(t) = \sum_{k=0}^{\infty} a_k t^k$$

for the solution to the Airy equation (1.15.5), with initial data

$$(1.15.8) \quad x(0) = 1, \quad x'(0) = 0.$$

If (1.15.7) is a convergent power series, then

$$(1.15.9) \quad \begin{aligned} x''(t) &= \sum_{k=2}^{\infty} k(k-1)a_k t^{k-2} \\ &= \sum_{k=0}^{\infty} (k+2)(k+1)a_{k+2} t^k, \end{aligned}$$

while

$$(1.15.10) \quad tx(t) = \sum_{k=1}^{\infty} a_{k-1} t^k.$$

Comparison gives the recursive formula

$$(1.15.11) \quad a_{k+3} = \frac{a_k}{(k+3)(k+2)}.$$

To get started, we note that

$$(1.15.12) \quad a_0 = x(0) = 1, \quad a_1 = x'(0) = 0, \quad a_2 = \frac{1}{2}x''(0) = 0.$$

Thus  $a_{3\ell+j} = 0$  for  $j = 1, 2$ , and we get

$$(1.15.13) \quad x(t) = \sum_{\ell=0}^{\infty} \alpha_{\ell} t^{3\ell},$$

where  $\alpha_{\ell} = a_{3\ell}$  is given recursively by

$$(1.15.14) \quad \alpha_{\ell+1} = \frac{\alpha_{\ell}}{(3\ell+3)(3\ell+2)}, \quad \alpha_0 = 1.$$

The ratio test applies to show that the power series (1.15.13) converges for all  $t \in \mathbb{R}$ , yielding a solution to Airy's equation (1.15.5), with initial data (1.15.8).

A study of power series as a technique for solving ODE in a more general setting is given in §3.10 of Chapter 3.

Another useful tool is the Wronskian determinant, defined on a pair of functions  $y_1$  and  $y_2$  by

$$(1.15.15) \quad W(y_1, y_2)(t) = y_1 y_2' - y_2 y_1' = \det \begin{pmatrix} y_1 & y_2 \\ y_1' & y_2' \end{pmatrix}.$$

If  $y_1$  and  $y_2$  both solve (1.15.3) with  $f \equiv 0$ , i.e.,

$$(1.15.16) \quad a(t)y'' + b(t)y' + c(t)y = 0,$$

then substituting for  $y_j''$  in

$$(1.15.17) \quad \frac{dW}{dt} = y_1 y_2'' - y_2 y_1''$$

yields

$$(1.15.18) \quad \frac{dW}{dt} = -\frac{b(t)}{a(t)}W,$$

a useful first order linear equation for  $W$ . Note that if we have such  $y_1$  and  $y_2$ , solving (15.16) with initial condition

$$(1.15.19) \quad y(t_0) = \alpha, \quad y'(t_0) = \beta,$$

in the form  $y(t) = C_1y_1(t) + C_2y_2(t)$  involves finding  $C_1$  and  $C_2$  such that

$$(1.15.20) \quad \begin{aligned} C_1y_1(t_0) + C_2y_2(t_0) &= \alpha, \\ C_1y_1'(t_0) + C_2y_2'(t_0) &= \beta, \end{aligned}$$

which uniquely determines  $C_1$  and  $C_2$  precisely when  $W(y_1, y_2)(t_0) \neq 0$ .

In light of the existence and uniqueness statement made above (to be proved in Chapter 4), it follows that if  $y_1$  and  $y_2$  solve (1.15.16) and have nonvanishing Wronskian, on an interval on which  $a$ ,  $b$ , and  $c$  are smooth and  $a$  is nonvanishing, then the general solution to (1.15.16) has the form  $C_1y_1 + C_2y_2$ .

Recall that the Wronskian arose in the previous section, in the treatment of the method of variation of parameters. This treatment is extended to a much more general setting in Chapter 3.

---

### Exercises

Equations of the form

$$(1.15.21) \quad at^2 \frac{d^2x}{dt^2} + bt \frac{dx}{dt} + cx = 0$$

are called Euler equations.

1. Show that  $x(t) = t^r = e^{r \log t}$  solves (1.15.21) for  $t > 0$  provided  $r$  satisfies

$$(1.15.22) \quad ar(r-1) + br + c = 0.$$

2. Show that if (1.15.22) has two distinct solutions  $r_1$  and  $r_2$ , then

$$C_1t^{r_1} + C_2t^{r_2}$$

is the general solution to (1.15.21) on  $t \in (0, \infty)$ .

3. Show that if  $r$  is a double root of (1.15.22), then

$$C_1t^r + C_2(\log t)t^r$$

is the general solution to (1.15.21) for  $t \in (0, \infty)$ .

4. Find the coefficients  $a_k$  in the power series expansion

$$x(t) = \sum_{k=0}^{\infty} a_k t^k$$

for the solution to the Airy equation

$$(1.15.23) \quad \frac{d^2x}{dt^2} - tx = 0,$$

with initial data

$$x(0) = 0, \quad x'(0) = 1.$$

Show that this power series converges for all  $t$ .



5. Show that the Wronskian of two solutions to the Airy equation (1.15.23) solves the equation

$$\frac{dW}{dt} = 0.$$

### 1.16. Bessel's equation

Here we construct solutions to Bessel's equation

$$(1.16.1) \quad \frac{d^2x}{dt^2} + \frac{1}{t} \frac{dx}{dt} + \left(1 - \frac{\nu^2}{t^2}\right)x = 0.$$

This is an important equation, whose roots in partial differential equations are discussed in Appendix 1.A. Note that if the factor  $(1 - \nu^2/t^2)$  in front of  $x$  had the term 1 dropped, one would have the Euler equation

$$(1.16.2) \quad t^2x'' + tx' - \nu^2x = 0,$$

with solutions

$$(1.16.3) \quad x(t) = t^{\pm\nu},$$

as seen in (1.15.21)–(1.15.22). In light of this, we are motivated to set

$$(1.16.4) \quad x(t) = t^\nu y(t),$$

and study the resulting differential equation for  $y$ :

$$(1.16.5) \quad \frac{d^2y}{dt^2} + \frac{2\nu + 1}{t} \frac{dy}{dt} + y = 0.$$

This might seem only moderately less singular than (1.16.1) at  $t = 0$ , but in fact it has a smooth solution. To obtain it, let us note that if  $y(t)$  solves (1.16.5), so does  $y(-t)$ , hence so does  $y(t) + y(-t)$ , which is even in  $t$ . Thus, we look for a solution to (1.16.5) in the form

$$(1.16.6) \quad y(t) = \sum_{k=0}^{\infty} a_k t^{2k}.$$

Substitution into (1.16.5) yields for the left side of (1.16.5) the power series

$$(1.16.7) \quad \sum_{k=0}^{\infty} \left\{ (2k+2)(2k+2\nu+2)a_{k+1} + a_k \right\} t^{2k},$$

assuming convergence, which we will examine shortly. From this we see that, as long as

$$(1.16.8) \quad \nu \notin \{-1, -2, -3, \dots\},$$

we can fix  $a_0 = a_0(\nu)$  and solve recursively for  $a_{k+1}$ , for each  $k \geq 0$ , obtaining

$$(1.16.9) \quad a_{k+1} = -\frac{1}{4} \frac{a_k}{(k+1)(k+\nu+1)}.$$

Given (1.16.8), this recursion works, and one can readily apply the ratio test to show that the power series (1.16.6) converges for all  $t \in \mathbb{R}$ .

We will find it useful to produce an explicit solution to the recursive formula (1.16.9). For this, it is convenient to write

$$(1.16.10) \quad a_k = \alpha_k \beta_k \gamma_k,$$

with

$$(1.16.11) \quad \alpha_{k+1} = -\frac{1}{4}\alpha_k, \quad \beta_{k+1} = \frac{\beta_k}{k+1}, \quad \gamma_{k+1} = \frac{\gamma_k}{k+\nu+1}.$$

Clearly the first two equations have the explicit solutions

$$(1.16.12) \quad \alpha_k = \left(-\frac{1}{4}\right)^k \alpha_0, \quad \beta_k = \frac{\beta_0}{k!}.$$

We can solve the third if we have in hand a function  $\Gamma(z)$  satisfying

$$(1.16.13) \quad \Gamma(z+1) = z\Gamma(z).$$

Indeed, the Euler gamma function  $\Gamma(z)$ , discussed in Appendix 1.B, is a smooth function on  $\mathbb{R} \setminus \{0, -1, -2, \dots\}$  that satisfies (1.16.13). With this function in hand, we can write

$$(1.16.14) \quad \gamma_k = \frac{\tilde{\gamma}_0}{\Gamma(k+\nu+1)},$$

and putting together (1.16.10)–(1.16.14) yields

$$(1.16.15) \quad a_k = \left(-\frac{1}{4}\right)^k \frac{\tilde{a}_0}{k!\Gamma(k+\nu+1)}.$$

We initialize this with  $\tilde{a}_0 = 2^{-\nu}$ . There results the solution  $y(t) = \mathcal{J}_\nu(t)$  to (16.5), and  $x(t) = J_\nu(t) = t^\nu \mathcal{J}_\nu(t)$  to (1.16.1), given by

$$(1.16.16) \quad J_\nu(t) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\Gamma(k+\nu+1)} \left(\frac{t}{2}\right)^{2k+\nu}.$$

Supplementing the regularity of  $\Gamma(z)$  on  $\mathbb{R} \setminus \{0, -1, -2, \dots\}$ , we will see in Appendix 1.B that

$$(1.16.17) \quad \frac{1}{\Gamma(z)} \text{ is well defined and smooth in } z \in \mathbb{R} \\ \text{vanishing for } z \in \{0, -1, -2, \dots\}.$$

Consequently (1.16.16) is a valid solution to (1.16.1) for  $t \in (0, \infty)$ , for each  $\nu \in \mathbb{R}$ . In fact,

$$(1.16.18) \quad J_\nu \text{ and } J_{-\nu} \text{ solve (1.16.1), for } \nu \in \mathbb{R}.$$

The function  $J_\nu$  is called a *Bessel function*.

Let us examine the behavior of  $J_\nu(t)$  as  $t \searrow 0$ . We have

$$(1.16.19) \quad J_\nu(t) = \frac{1}{\Gamma(\nu+1)} \left(\frac{t}{2}\right)^\nu + O(t^{\nu+1}), \quad \text{as } t \searrow 0.$$

As long as  $\nu$  satisfies (1.16.8), the coefficient  $1/\Gamma(\nu+1)$  is nonzero. Furthermore,

$$(1.16.20) \quad J_{-\nu}(t) = \frac{1}{\Gamma(1-\nu)} \left(\frac{t}{2}\right)^{-\nu} + O(t^{-\nu+1}), \quad \text{as } t \searrow 0,$$

and as long as  $\nu \notin \{1, 2, 3, \dots\}$ , the coefficient  $1/\Gamma(1-\nu)$  is nonzero. In particular, we see that

$$(1.16.21) \quad \text{If } \nu \notin \mathbb{Z}, J_\nu \text{ and } J_{-\nu} \text{ are linearly independent solutions} \\ \text{to (16.1) on } (0, \infty).$$

In contrast to this, we have the following:

$$(1.16.22) \quad \text{If } n \in \mathbb{Z}, J_n(t) = (-1)^n J_{-n}(t).$$

To see this, we assume  $n \in \{1, 2, 3, \dots\}$ , and note that

$$(1.16.23) \quad \frac{1}{\Gamma(k-n+1)} = 0, \quad \text{for } 0 \leq k \leq n-1.$$

We use this, together with the restatement of (1.16.16) that

$$(1.16.24) \quad J_\nu(t) = \sum_{k=0}^{\infty} \frac{(-1)^k}{\Gamma(k+1)\Gamma(k+\nu+1)} \left(\frac{t}{2}\right)^{2k+\nu},$$

which follows from the identity  $\Gamma(k+1) = k!$ , to deduce that, for  $n \in \mathbb{N}$ ,

$$(1.16.25) \quad \begin{aligned} J_{-n}(t) &= \sum_{k=n}^{\infty} \frac{(-1)^k}{\Gamma(k+1)\Gamma(k-n+1)} \left(\frac{t}{2}\right)^{2k-n} \\ &= \sum_{\ell=0}^{\infty} \frac{(-1)^{\ell+n}}{\Gamma(\ell+1)\Gamma(\ell+n+1)} \left(\frac{t}{2}\right)^{2\ell+n} \\ &= (-1)^n J_n(t). \end{aligned}$$

Consequently  $J_\nu(t)$  and  $J_{-\nu}(t)$  are linearly independent solutions to (1.16.1) as long as  $\nu \notin \mathbb{Z}$ , but this fails for  $\nu \in \mathbb{Z}$ . We now seek a family of solutions  $Y_\nu(t)$  to (1.16.1) with the property that  $J_\nu$  and  $Y_\nu$  are linearly independent solutions, for all  $\nu \in \mathbb{R}$ . The key to this construction lies in an analysis of the Wronskian

$$(1.16.26) \quad W_\nu(t) = W(J_\nu, J_{-\nu})(t) = J_\nu(t)J'_{-\nu}(t) - J'_{\nu}(t)J_{-\nu}(t).$$

By (1.15.10), we have

$$(1.16.27) \quad \frac{dW_\nu}{dt} = -\frac{1}{t}W_\nu,$$

hence

$$(1.16.28) \quad W_\nu(t) = \frac{K(\nu)}{t}.$$

To evaluate  $K(\nu)$ , we calculate

$$(1.16.29) \quad \begin{aligned} W(J_\nu, J_{-\nu}) &= W(t^\nu \mathcal{J}_\nu, t^{-\nu} \mathcal{J}_{-\nu}) \\ &= W(\mathcal{J}_\nu, \mathcal{J}_{-\nu}) - \frac{2\nu}{t} \mathcal{J}_\nu(t) \mathcal{J}_{-\nu}(t). \end{aligned}$$

Since  $\mathcal{J}_\nu(t)$  and  $\mathcal{J}_{-\nu}(t)$  are smooth in  $t$ , so is  $W(\mathcal{J}_\nu, \mathcal{J}_{-\nu})$ , and we deduce from (1.16.28)–(1.16.29) that

$$(1.16.30) \quad W_\nu(t) = -\frac{2\nu}{t} \mathcal{J}_\nu(0) \mathcal{J}_{-\nu}(0).$$

Now, since  $\mathcal{J}_\nu(0) = 1/2^\nu \Gamma(\nu+1)$ , we have

$$(1.16.31) \quad \begin{aligned} \nu \mathcal{J}_\nu(0) \mathcal{J}_{-\nu}(0) &= \frac{\nu}{\Gamma(\nu+1)\Gamma(1-\nu)} \\ &= \frac{1}{\Gamma(\nu)\Gamma(1-\nu)}. \end{aligned}$$

An important gamma function identity, stated in Appendix 1.B, is

$$(1.16.32) \quad \Gamma(\nu)\Gamma(1-\nu) = \frac{\pi}{\sin \pi\nu}.$$

Hence (1.16.30)–(1.16.31) yields

$$(1.16.33) \quad W(J_\nu, J_{-\nu})(t) = -\frac{2}{\pi} \frac{\sin \pi \nu}{t}.$$

This motivates the following. For  $\nu \notin \mathbb{Z}$ , set

$$(1.16.34) \quad Y_\nu(t) = \frac{J_\nu(t) \cos \pi \nu - J_{-\nu}(t)}{\sin \pi \nu}.$$

Note that, by (1.16.25), numerator and denominator both vanish for  $\nu \in \mathbb{Z}$ . Now, for  $\nu \notin \mathbb{Z}$ , we have

$$(1.16.35) \quad \begin{aligned} W(J_\nu, Y_\nu)(t) &= -\frac{1}{\sin \pi \nu} W(J_\nu, J_{-\nu})(t) \\ &= \frac{2}{\pi t}. \end{aligned}$$

Consequently, for  $n \in \mathbb{Z}$ , we set

$$(1.16.36) \quad Y_n(t) = \lim_{\nu \rightarrow n} Y_\nu(t) = \frac{1}{\pi} \left[ \frac{\partial J_\nu(t)}{\partial \nu} - (-1)^n \frac{\partial J_{-\nu}(t)}{\partial \nu} \right] \Big|_{\nu=n},$$

and we also have (1.16.35) for  $\nu \in \mathbb{Z}$ . The functions  $Y_\nu$  are called Bessel functions of the second kind.

Another construction of a solution to accompany  $J_n(t)$  is given in Chapter 3, (3.11.65)–(3.11.79).

We end this section with the following integral formula for  $J_\nu(t)$ , which plays an important role in further investigations, such as the behavior of  $J_\nu(t)$  for large  $t$ .

**Proposition 1.16.1.** *If  $\nu > -1/2$ ,*

$$(1.16.37) \quad J_\nu(t) = \frac{(t/2)^\nu}{\Gamma(1/2)\Gamma(\nu+1/2)} \int_{-1}^1 (1-s^2)^{\nu-1/2} e^{ist} ds.$$

**Proof.** To verify (1.16.37), we replace  $e^{ist}$  by its power series, integrate term by term, and use some identities from Appendix 1.B. To begin, the integral on the right side of (1.16.37) is equal to

$$(1.16.38) \quad \sum_{k=0}^{\infty} \frac{1}{(2k)!} \int_{-1}^1 (ist)^{2k} (1-s^2)^{\nu-1/2} ds.$$

The identity (1.B.17) implies

$$(1.16.39) \quad \int_{-1}^1 s^{2k} (1-s^2)^{\nu-1/2} ds = \frac{\Gamma(k+1/2)\Gamma(\nu+1/2)}{\Gamma(k+\nu+1)},$$

so the right side of (1.16.37) equals

$$(1.16.40) \quad \frac{(t/2)^\nu}{\Gamma(1/2)\Gamma(\nu+1/2)} \sum_{k=0}^{\infty} \frac{1}{(2k)!} (it)^{2k} \frac{\Gamma(k+1/2)\Gamma(\nu+1/2)}{\Gamma(k+\nu+1)}.$$

As seen in (1.B.7), we have

$$(1.16.41) \quad \Gamma\left(\frac{1}{2}\right)(2k)! = 2^{2k} k! \Gamma\left(k + \frac{1}{2}\right),$$

so (1.16.40) is equal to

$$(1.16.42) \quad \left(\frac{t}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{(-1)^k}{k! \Gamma(k + \nu + 1)} \left(\frac{t}{2}\right)^{2k},$$

which agrees with our formula (1.16.16) for  $J_\nu(t)$ .  $\square$

## Exercises

1. Show that the Bessel functions  $J_\nu$  satisfy the following recursion relations:

$$\frac{d}{dt} \left( t^\nu J_\nu(t) \right) = t^\nu J_{\nu-1}(t), \quad \frac{d}{dt} \left( t^{-\nu} J_\nu(t) \right) = -t^{-\nu} J_{\nu+1}(t),$$

or equivalently

$$\begin{aligned} J_{\nu+1}(t) &= -J'_\nu(t) + \frac{\nu}{t} J_\nu(t), \\ J_{\nu-1}(t) &= J'_\nu(t) + \frac{\nu}{t} J_\nu(t). \end{aligned}$$

2. Show that  $\mathcal{J}_{-1/2}(t) = \sqrt{2/\pi} \cos t$ , and deduce that

$$J_{-1/2}(t) = \sqrt{\frac{2}{\pi t}} \cos t, \quad J_{1/2}(t) = \sqrt{\frac{2}{\pi t}} \sin t.$$

Deduce from Exercise 1 that, for  $n \in \mathbb{Z}^+$ ,

$$\begin{aligned} J_{n+1/2}(t) &= (-1)^n \left\{ \prod_{j=1}^n \left( \frac{d}{dt} - \frac{j-1/2}{t} \right) \right\} \frac{\sin t}{\sqrt{2\pi t}}, \\ J_{-n-1/2}(t) &= \left\{ \prod_{j=1}^n \left( \frac{d}{dt} - \frac{j-1/2}{t} \right) \right\} \frac{\cos t}{\sqrt{2\pi t}}. \end{aligned}$$

*Hint.* The differential equation (1.16.5) for  $\mathcal{J}_{-1/2}$  is  $y'' + y = 0$ . Since  $\mathcal{J}_{-1/2}(t)$  is even in  $t$ ,  $\mathcal{J}_{-1/2}(t) = C \cos t$ , and the evaluation of  $C$  comes from  $\mathcal{J}_{-1/2}(0) = \sqrt{2}/\Gamma(1/2) = \sqrt{2/\pi}$ , thanks to (1.B.6).

3. Show that the functions  $Y_\nu$  satisfy the same recursion relations as  $J_\nu$ , i.e.,

$$\frac{d}{dt} \left( t^\nu Y_\nu(t) \right) = t^\nu Y_{\nu-1}(t), \quad \frac{d}{dt} \left( t^{-\nu} Y_\nu(t) \right) = -t^{-\nu} Y_{\nu+1}(t).$$

4. The Hankel functions  $H_\nu^{(1)}(t)$  and  $H_\nu^{(2)}(t)$  are defined to be

$$H_\nu^{(1)}(t) = J_\nu(t) + iY_\nu(t), \quad H_\nu^{(2)}(t) = J_\nu(t) - iY_\nu(t).$$

Show that they satisfy the same recursion relations as  $J_\nu$ , i.e.,

$$\frac{d}{dt} \left( t^\nu H_\nu^{(j)}(t) \right) = t^\nu H_{\nu-1}^{(j)}(t), \quad \frac{d}{dt} \left( t^{-\nu} H_\nu^{(j)}(t) \right) = -t^{-\nu} H_{\nu+1}^{(j)}(t),$$

for  $j = 1, 2$ .

5. Show that

$$H_{-\nu}^{(1)}(t) = e^{\pi i \nu} H_{\nu}^{(1)}(t), \quad H_{-\nu}^{(2)}(t) = e^{-\pi i \nu} H_{\nu}^{(2)}(t).$$

6. Show that  $Y_{1/2}(t) = -J_{-1/2}(t)$ , and deduce that

$$H_{1/2}^{(1)}(t) = -i\sqrt{\frac{2}{\pi t}}e^{it}, \quad H_{1/2}^{(2)}(t) = i\sqrt{\frac{2}{\pi t}}e^{-it}.$$

### 1.17. Higher order linear equations

A linear differential equation of order  $n$  has the form

$$(1.17.1) \quad a_n(t) \frac{d^n x}{dt^n} + a_{n-1}(t) \frac{d^{n-1} x}{dt^{n-1}} + \cdots + a_0(t)x = f(t).$$

If  $a_j(t)$  are continuous for  $t$  in an interval  $I$  containing  $t_0$ , and  $a_n(t)$  is nonvanishing on this interval, one has a unique solution to (1.17.1) given an initial condition of the form

$$(1.17.2) \quad x(t_0) = \alpha_0, \quad x'(t_0) = \alpha_1, \dots, \quad x^{(n-1)}(t_0) = \alpha_{n-1}.$$

(As with (1.15.1)–(1.15.2), this also follows from a general result that will be established in Chapter 4.) If  $a_j(t)$  are all constant, the equation (1.17.1) has the form

$$(1.17.3) \quad a_n \frac{d^n x}{dt^n} + a_{n-1} \frac{d^{n-1} x}{dt^{n-1}} + \cdots + a_0 x = f(t).$$

It is homogeneous if  $f \equiv 0$ , in which case one has

$$(1.17.4) \quad a_n \frac{d^n x}{dt^n} + a_{n-1} \frac{d^{n-1} x}{dt^{n-1}} + \cdots + a_0 x = 0.$$

We assume  $a_n \neq 0$ .

Methods developed in §§1.9–1.10 have natural extensions to (1.17.4) and (1.17.3). The function  $x(t) = e^{rt}$  solves (1.17.4) provided  $r$  satisfies the characteristic equation

$$(1.17.5) \quad a_n r^n + a_{n-1} r^{n-1} + \cdots + a_0 = 0.$$

The fundamental theorem of algebra guarantees that (1.17.5) has  $n$  roots, i.e., there exist  $r_1, \dots, r_n \in \mathbb{C}$  such that

$$(1.17.6) \quad a_n r^n + a_{n-1} r^{n-1} + \cdots + a_0 = a_n (r - r_1) \cdots (r - r_n).$$

A proof of this theorem is given in §2.C of Chapter 2. These roots  $r_1, \dots, r_n$  may or may not be distinct. If they are distinct, the general solution to (1.17.4) has the form

$$(1.17.7) \quad x(t) = C_1 e^{r_1 t} + \cdots + C_n e^{r_n t}.$$

If  $r_j$  is a root of multiplicity  $k$ , one has solutions to (1.17.4) of the form

$$(1.17.8) \quad C_1 e^{r_j t} + C_2 t e^{r_j t} + \cdots + C_k t^{k-1} e^{r_j t}.$$

This observation can be used to yield a fresh perspective on what makes the calculations in §1.10 work. Consider for example the equation

$$(1.17.9) \quad ax'' + bx' + cx = e^{\kappa t}.$$

The right side solves the equation  $(d/dt - \kappa)e^{\kappa t} = 0$ , so any solution to (1.17.9) also solves

$$(1.17.10) \quad \left( \frac{d}{dt} - \kappa \right) \left( a \frac{d^2}{dt^2} + b \frac{d}{dt} + c \right) x = 0,$$

a homogeneous equation whose characteristic polynomial is

$$(1.17.11) \quad q(r) = (r - \kappa)(ar^2 + br + c) = (r - \kappa)p(r).$$



If  $\kappa$  is not a root of  $p(r)$ , then certainly (1.17.9) has a solution of the form  $Ae^{\kappa t}$ . If  $\kappa$  is a root of  $p(r)$ , then it is a double (or, perhaps, triple) root of  $q(r)$ , and (1.16.8) applies, leading one to (1.10.17) or (1.10.25).

One can also extend the method of variation of parameters to higher order equations (1.17.3), though the details get grim.

The equations (1.17.1)–(1.17.4) can each be recast as  $n \times n$  first order systems of differential equations, and all the results on these equations are special cases of results to be covered in Chapter 3, so we will say no more here, except to advertise that this transformation leads to a much simplified approach to the method of variation of parameters.

---

### Exercises

1. Assume the existence and uniqueness results for the solution to (1.17.1) stated in the first paragraph of this section. Show that there exist  $n$  solutions  $u_j$  to

$$a_n(t)u_j^{(n)}(t) + a_{n-1}(t)u_j^{(n-1)}(t) + \cdots + a_0(t)u_j(t) = 0$$

on  $I$  such that every solution to (1.17.1) with  $f \equiv 0$  can be written uniquely in the form

$$x(t) = C_1u_1(t) + \cdots + C_nu_n(t).$$

For general continuous  $f$ , let  $x_p$  be a particular solution to (1.17.1). Show that if  $x(t)$  is an arbitrary solution to (1.17.1), then there exist unique constants  $C_j$ ,  $1 \leq j \leq n$ , such that

$$x(t) = x_p(t) + C_1u_1(t) + \cdots + C_nu_n(t).$$

This is called the general solution to (1.17.1).

*Hint.* Require  $u_j^{(k-1)}(t_0) = \delta_{jk}$ ,  $1 \leq k \leq n$ , where  $\delta_{jk} = 1$  for  $j = k$ , 0 for  $j \neq k$ .

2. Find the general solution to each of the following equations for  $x = x(t)$ .

(a)

$$\frac{d^4x}{dt^4} - x = 0.$$

(b)

$$\frac{d^3x}{dt^3} - x = 0.$$

(c)

$$x''' - 2x'' - 4x' + 8x = 0.$$

(d)

$$x''' - 2x'' + 4x' - 8x = 0.$$

(e)

$$x''' + x = e^t.$$

3. For each of the cases (a)–(e) in Exercise 1 of §1.10, produce a third or fourth order homogeneous differential equation solved by  $x(t)$ .

Exercises 4–6 will exploit the fact that if the characteristic polynomial (1.17.6) factors as stated there, then the left side of (1.17.4) is equal to

$$a_n \left( \frac{d}{dt} - r_1 \right) \cdots \left( \frac{d}{dt} - r_n \right) x = a_n \prod_{j=1}^n \left( \frac{d}{dt} - r_j \right) x.$$

4. Show that

$$\left( \frac{d}{dt} - r_j \right) (e^{rt} u) = e^{rt} \left( \frac{d}{dt} - r_j + r \right) u,$$

and more generally

$$\prod_{j=1}^n \left( \frac{d}{dt} - r_j \right) (e^{rt} u) = e^{rt} \prod_{j=1}^n \left( \frac{d}{dt} - r_j + r \right) u.$$

5. Suppose  $r_j$  is a root of multiplicity  $k$  of (1.17.6). Show that  $x(t) = e^{r_j t} u$  solves (1.17.4) if and only if

$$\prod_{\{\ell: r_\ell \neq r_j\}} \left( \frac{d}{dt} - r_\ell + r_j \right) \left( \frac{d}{dt} \right)^k u = 0.$$

Use this to show that functions of the form (1.17.8) solve (1.17.4).

6. In light of Exercise 5, use an inductive argument to show the following. Assume the roots  $\{r_j\}$  of (1.17.6) are

$$r_\nu, \quad \text{with multiplicity } k_\nu, \quad 1 \leq \nu \leq m, \quad k_1 + \cdots + k_m = n.$$

Then the general solution to (1.17.4) is a linear combination of

$$t^{\ell_\nu} e^{r_\nu t}, \quad 0 \leq \ell_\nu \leq k_\nu - 1, \quad 1 \leq \nu \leq m.$$

### 1.18. The Laplace transform

The Laplace transform provides a tool to treat nonhomogeneous differential equations of the form

$$(1.18.1) \quad c_n \frac{d^n f}{dt^n} + c_{n-1} \frac{d^{n-1} f}{dt^{n-1}} + \cdots + c_0 f(t) = g(t),$$

for  $t \geq 0$ , with initial data

$$(1.18.2) \quad f(0) = a_0, \dots, f^{(n-1)}(0) = a_{n-1},$$

for certain classes of functions  $g$ . It is defined as follows. Assume  $f : \mathbb{R}^+ \rightarrow \mathbb{C}$  is integrable on  $[0, R]$  for all  $R < \infty$ , and satisfies

$$(1.18.3) \quad \int_0^\infty |f(t)| e^{-at} dt < \infty, \quad \forall a > A,$$

for some  $A \in \mathbb{R}$ . We define the Laplace transform of  $f$  by

$$(1.18.4) \quad \mathcal{L}f(s) = \int_0^\infty f(t) e^{-st} dt, \quad \operatorname{Re} s > A.$$

By our hypotheses, this integral is absolutely convergent for each  $s$  in the half-plane  $H_A = \{s \in \mathbb{C} : \operatorname{Re} s > A\}$ . For our current purposes, it will suffice to take  $s$  real, in  $(A, \infty)$ . Note that, for such  $s$ ,

$$(1.18.5) \quad \frac{d}{ds} \mathcal{L}f(s) = \mathcal{L}g(s), \quad g(t) = -tf(t).$$

If we assume that  $f'$  is continuous on  $[0, \infty)$  and

$$(1.18.6) \quad |f(t)| + |f'(t)| \leq C_\varepsilon e^{(A+\varepsilon)t}, \quad \text{for } t \geq 0,$$

for each  $\varepsilon > 0$ , we can integrate by parts and get

$$(1.18.7) \quad \mathcal{L}f'(s) = s\mathcal{L}f(s) - f(0),$$

and similar hypotheses for higher derivatives of  $f$  gives

$$(1.18.8) \quad \mathcal{L}f^{(k)}(s) = s^k \mathcal{L}f(s) - s^{k-1} f(0) - \cdots - f^{(k-1)}(0).$$

Hence, if  $f$  satisfies an ODE of the form (1.18.1)–(1.18.2) and if  $f, f', \dots, f^{(n-1)}$  all satisfy (1.18.6), and  $g$  satisfies (1.18.3), we have

$$(1.18.9) \quad p(s)\mathcal{L}f(s) = \mathcal{L}g(s) + q(s),$$

with

$$(1.18.10) \quad \begin{aligned} p(s) &= c_n s^n + c_{n-1} s^{n-1} + \cdots + c_0, \\ q(s) &= c_n(a_0 s^{n-1} + \cdots + a_{n-1}) + \cdots + c_1 a_0. \end{aligned}$$

If all the roots of  $p(s)$  satisfy  $\operatorname{Re} s < B$ , we have

$$(1.18.11) \quad \mathcal{L}f(s) = \frac{\mathcal{L}g(s) + q(s)}{p(s)}, \quad \operatorname{Re} s > C = \max(A, B).$$

Making use of (1.18.11) to solve (1.18.1)–(1.18.2) brings in two problems, which we now state.

I. THE RECOGNITION PROBLEM. Given the right side of (1.18.11), i.e., given

$$(1.18.12) \quad \frac{\mathcal{L}g(s) + q(s)}{p(s)} = R(s),$$

find a function  $f_1 : [0, \infty) \rightarrow \mathbb{C}$ , such that

$$(1.18.13) \quad \mathcal{L}f_1(s) = R(s), \quad \text{for } \operatorname{Re} s > C.$$

II. THE UNIQUENESS PROBLEM. Given  $f$  and  $f_1 : [0, \infty) \rightarrow \mathbb{C}$ , both satisfying (1.18.3), one wants to know that

$$(1.18.14) \quad \mathcal{L}f(s) = \mathcal{L}f_1(s), \quad \forall s > A \implies f = f_1 \text{ on } [0, \infty).$$

The uniqueness problem has a satisfactory solution. As long as  $f$  and  $f_1$  satisfy the hypotheses just stated, the result (1.18.14) is true. The proof of this can be found in §3.3 of [47]. In addition there are inversion formulas. Here is one, established in §3.3 of [47].

**Proposition 1.18.1.** *Assume  $f$  and  $f'$  are continuous on  $[0, \infty)$ , and*

$$(1.18.15) \quad |f(t)| + |f'(t)| \leq Ce^{At}, \quad t \geq 0.$$

*Then, for  $t > 0$ ,*

$$(1.18.16) \quad tf(t) = -\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{d}{ds} \mathcal{L}f(B + i\xi) e^{t(B+i\xi)} d\xi,$$

*as long as  $B > A$ , with an absolutely convergent integral on the right side.*

In light of the uniqueness, if  $f$  satisfies (1.18.3), we say

$$(1.18.17) \quad g = \mathcal{L}f \implies f = \mathcal{L}^{-1}g,$$

and call  $\mathcal{L}^{-1}$  the inverse Laplace transform.

Generally speaking, for functions  $R(s)$  that arise in (1.18.12), calculation of the integral

$$(1.18.18) \quad \int_{-\infty}^{\infty} R'(B + i\xi) e^{it\xi} d\xi$$

is not so easy, though methods of *residue calculus*, discussed in §4.1 of [47] can be effective. For the purpose of using (1.18.11) to solve (1.18.1)–(1.18.2), by finding  $f$  that satisfies

$$(1.18.19) \quad \mathcal{L}f(s) = R(s),$$

with  $R(s)$  as in (1.18.12), it is useful to have a collection of functions that are known Laplace transforms, in order to solve the recognition problem.

To start our collection, we consider the Laplace transform of  $e^{at}$ :

$$(1.18.20) \quad \int_0^{\infty} e^{at} e^{-st} dt = \int_0^{\infty} e^{-(s-a)t} dt = \frac{1}{s-a}.$$

**Table 1.** Table of Laplace transforms

	$f(t)$	$\mathcal{L}f(s)$
(a)	$\sin at$	$a/(s^2 + a^2)$
(b)	$\cos at$	$s/(s^2 + a^2)$
(c)	$\sinh at$	$a/(s^2 - a^2)$
(d)	$\cosh at$	$s/(s^2 - a^2)$

If  $a$  is real, this is valid for  $\operatorname{Re} s > a$ . However, using results from §1.1, we find it useful to note that (1.18.20) holds for *complex*  $a$ , as long as  $\operatorname{Re} s > \operatorname{Re} a$ . We can apply this to

$$(1.18.21) \quad f(t) = \cos at = \frac{1}{2}(e^{iat} + e^{-iat}),$$

for  $a \in \mathbb{R}$ , to get

$$(1.18.22) \quad \begin{aligned} \mathcal{L}f(s) &= \frac{1}{2} \left( \frac{1}{s - ia} + \frac{1}{s + ia} \right) \\ &= \frac{s}{s^2 + a^2}. \end{aligned}$$

Similar techniques yield the table of Laplace transforms presented in Table 1.

If  $a \in \mathbb{R}$ , the range of validity of (a)–(b) is  $\operatorname{Re} s > 0$ , and that of (c)–(d) is  $\operatorname{Re} s > |a|$ .

Laplace transforms of other functions, such as  $e^{-bt} \cos at$ , etc., can be identified via the identity

$$(1.18.23) \quad \mathcal{L}(e^{-bt} f)(s) = \mathcal{L}f(s + b).$$

Also, one can turn (1.18.5) around, to write

$$(1.18.24) \quad \mathcal{L}(tf)(s) = -\frac{d}{ds} \mathcal{L}f(s),$$

and, inductively,

$$(1.18.25) \quad \mathcal{L}(t^n f)(s) = (-1)^n \frac{d^n}{ds^n} \mathcal{L}f(s).$$

For example,

$$(1.18.26) \quad \begin{aligned} \mathcal{L}(t^n e^{at})(s) &= (-1)^n \frac{d^n}{ds^n} (s - a)^{-1} \\ &= \frac{n!}{(s - a)^{n+1}}, \end{aligned}$$

for  $a \in \mathbb{C}$ ,  $\operatorname{Re} s > \operatorname{Re} a$ . In particular,

$$(1.18.27) \quad f(t) = t^n \implies \mathcal{L}f(s) = n! s^{-(n+1)}.$$

Of course, by (1.18.23), the result (1.18.26) follows from its special case (1.18.27). A natural generalization of (1.18.16) arises from taking

$$(1.18.28) \quad f_z(t) = t^{z-1}, \quad z > 0.$$

We get

$$\begin{aligned}
 \mathcal{L}f_z(s) &= \int_0^\infty e^{-st}t^{z-1} dt \\
 (1.18.29) \quad &= \left( \int_0^\infty e^{-t}t^{z-1} dt \right) s^{-z} \\
 &= \Gamma(z)s^{-z},
 \end{aligned}$$

where

$$(1.18.30) \quad \Gamma(z) = \int_0^\infty e^{-t}t^{z-1} dt, \quad z > 1,$$

is the Gamma function, which plays a role in §1.16, via (1.16.13)–(1.16.16), and is treated in Appendix 1.B. Let us note that (1.18.24) implies

$$(1.18.31) \quad \mathcal{L}f_{z+1}f(s) = -\frac{d}{ds}\mathcal{L}f_z(s),$$

which in view of (1.18.29) is equivalent to the identity

$$(1.18.32) \quad \Gamma(z+1) = z\Gamma(z).$$

Also comparison of (1.18.27) and (1.18.29), with  $z = n + 1$ , yields

$$(1.18.33) \quad \Gamma(n+1) = n!$$

We can obtain another Laplace transform identity by applying  $d/dz$  to (18.28), noting that, since  $s^{-z} = e^{-z \log s}$ ,

$$(1.18.34) \quad \frac{d}{dz}s^{-z} = -(\log s)s^{-z}, \quad s > 0,$$

with an analogous formula for  $(d/dz)t^{z-1}$ :

$$(1.18.35) \quad \frac{d}{dz}t^{z-1} = (\log t)t^{z-1}.$$

Hence (18.28) yields

$$(1.18.36) \quad f(t) = (\log t)t^{z-1} \Rightarrow \mathcal{L}f(s) = (\Gamma'(z) - \Gamma(z) \log s) s^{-z}.$$

In particular,

$$\begin{aligned}
 (1.18.37) \quad f(t) = \log t \Rightarrow \mathcal{L}f(s) &= (\Gamma'(1) - \log s)s^{-1} \\
 &= -\frac{\gamma + \log s}{s},
 \end{aligned}$$

where  $\gamma = -\Gamma'(1)$  is known as Euler's constant. Taking  $s = 1$  in (1.18.37), we have the formula

$$(1.18.38) \quad \gamma = -\int_0^\infty (\log t)e^{-t} dt.$$

Collecting these results, we complement the table of Laplace transforms compiled in Table 1 with that in Table 2. Note that (h) follows from (e), via (1.18.23). One has similar variants of (f)–(g).

Another function to consider is the “impulse function”

$$(1.18.39) \quad \chi_I(t) = \begin{cases} 1, & \text{if } t \in I, \\ 0, & \text{if } t \notin I. \end{cases}$$

**Table 2.** Further Laplace transforms

	$f(t)$	$\mathcal{L}f(s)$
(e)	$t^{z-1}$	$\Gamma(z)s^{-z}$
(f)	$\log t$	$-(\gamma + \log s)/s$
(g)	$(\log t)t^{z-1}$	$(\Gamma'(z) - \Gamma(z)\log s)/s^z$
(h)	$t^{z-1}e^{at}$	$\Gamma(z)(s-a)^{-z}$

where  $I = [a, b]$  is an interval, with  $0 \leq a < b < \infty$ . We have

$$(1.18.40) \quad \mathcal{L}\chi_I(s) = \int_a^b e^{-st} dt = \frac{e^{-as} - e^{-bs}}{s}.$$

Let us apply the Laplace transform method to the following initial value problem. Take  $k, a, \alpha_0, \alpha_1 \in \mathbb{R}$ , and consider

$$(1.18.41) \quad f''(t) + k^2 f(t) = \cos at, \quad f(0) = \alpha_0, \quad f'(0) = \alpha_1.$$

From (1.18.8),

$$(1.18.42) \quad \mathcal{L}f''(s) = s^2 \mathcal{L}f(s) - \alpha_0 s - \alpha_1,$$

and since  $\mathcal{L}(\cos at)(s) = s/(s^2 + a^2)$ , (1.18.11) becomes

$$(1.18.43) \quad \mathcal{L}f(s) = \frac{s}{(s^2 + k^2)(s^2 + a^2)} + \frac{\alpha_0 s + \alpha_1}{s^2 + k^2}.$$

The last term on the right is the Laplace transform of

$$(1.18.44) \quad \alpha_0 \cos kt + \frac{\alpha_1}{k} \sin kt.$$

It remains to write the first term on the right side of (1.18.43) as a Laplace transform. For this, we apply the method of partial fractions. To start, we try

$$(1.18.45) \quad \frac{s}{(s^2 + k^2)(s^2 + a^2)} = \frac{\alpha s + \beta}{s^2 + a^2} + \frac{\gamma s + \delta}{s^2 + k^2},$$

with unknowns  $\alpha, \beta, \gamma, \delta$ . Multiplying through by  $(s^2 + k^2)(s^2 + a^2)$  and equating coefficients of various powers of  $s$  leads to four linear equations in these four unknowns. Two of them yield  $\alpha = -\gamma$  and  $\beta = -\delta$ , and then the other two become

$$(1.18.46) \quad (k^2 - a^2)\alpha = 1, \quad (k^2 - a^2)\beta = 0.$$

If  $k^2 \neq a^2$ , these are uniquely solvable, for  $\alpha = (k^2 - a^2)^{-1}$ ,  $\beta = 0$ , and (1.18.49) becomes

$$(1.18.47) \quad \frac{s}{(s^2 + k^2)(s^2 + a^2)} = \frac{1}{k^2 - a^2} \left( \frac{s}{s^2 + a^2} - \frac{s}{s^2 + k^2} \right).$$

This is the Laplace transform of

$$(1.18.48) \quad \varphi_{a,k}(t) = \frac{1}{k^2 - a^2} (\cos at - \cos kt).$$

Then the solution to the differential equation (1.18.41) is

$$(1.18.49) \quad f(t) = \varphi_{a,k}(t) + \alpha_0 \cos kt + \frac{\alpha_1}{k} \sin kt.$$

This approach fails for  $k^2 = a^2$ , paralleling the situation we encountered in examining (1.11.4). One way to treat this exceptional case is to pass to the limit in (1.18.48), obtaining

$$\begin{aligned}
 \varphi_{k,k}(t) &= \lim_{a \rightarrow k} \varphi_{a,k}(t) \\
 (1.18.50) \quad &= \lim_{a \rightarrow k} \frac{1}{k+a} \frac{\cos at - \cos kt}{k-a} \\
 &= \frac{t}{2k} \sin kt.
 \end{aligned}$$

Another approach is to refine the method of partial fractions. In lieu of (1.18.45), we have

$$\begin{aligned}
 (1.18.51) \quad \frac{s}{(s^2 + k^2)^2} &= \frac{s}{(s + ik)^2(s - ik)^2} \\
 &= \frac{i}{4k} \left( \frac{1}{(s + ik)^2} - \frac{1}{(s - ik)^2} \right).
 \end{aligned}$$

Using (1.18.26), with  $n = 1$ , we have

$$(1.18.52) \quad \mathcal{L}^{-1} \left( \frac{1}{(s \pm ik)^2} \right) (t) = te^{\mp ikt}.$$

Hence the right side of (1.18.51) is the Laplace transform of

$$(1.18.53) \quad \frac{i}{4k} (te^{-ikt} - te^{ikt}) = \frac{t}{2k} \sin kt,$$

and again we obtain the conclusion of (1.18.50), from a different perspective.

In light of this analysis, and recalling (1.18.12), we are motivated to compute the inverse Laplace transform of functions of the form  $q(s)/p(s)$ , where  $p(s)$  is a polynomial of degree  $n$ , say

$$(1.18.54) \quad p(s) = s^n + c_{n-1}s^{n-1} + \cdots + c_0,$$

and  $q(s)$  is a polynomial of degree  $\leq n-1$ . The polynomial  $p(s)$  has complex roots  $r_1, \dots, r_m$ , of multiplicity  $k_1, \dots, k_m$ , and we can write (1.18.54) as

$$(1.18.55) \quad p(s) = (s - r_1)^{k_1} \cdots (s - r_m)^{k_m}, \quad k_1 + \cdots + k_m = n.$$

This is a consequence of the fundamental theorem of algebra, which is proved in Appendix 2.C of Chapter 2. The following is an incisive result on the method of partial fractions.

**Proposition 1.18.2.** *If  $p(s)$  is a polynomial of the form (1.18.55), with  $\{r_1, \dots, r_m\}$  distinct, and if  $q(s)$  is a polynomial of degree  $\leq n-1$ , then there exist unique  $a_{j\ell} \in \mathbb{C}$ , for  $1 \leq \ell \leq m$ ,  $1 \leq j \leq k_\ell$ , such that*

$$(1.18.56) \quad \frac{q(s)}{p(s)} = \sum_{\ell=1}^m \sum_{j=1}^{k_\ell} \frac{a_{j\ell}}{(s - r_\ell)^j}.$$

**Proof.** We use some concepts developed in Chapter 2. The set of collections  $(a_{j\ell})$  of the form

$$\{a_{j\ell} \in \mathbb{C} : 1 \leq j \leq k_\ell, 1 \leq \ell \leq m\}$$

forms a vector space  $V_0$ , of dimension  $k_1 + \cdots + k_m = n$ . Meanwhile, the space  $\mathcal{P}_{n-1}$  of polynomials  $q(s)$  of degree  $\leq n-1$  is also a vector space of dimension  $n$ .



Now the correspondence in (1.18.56) yields a well defined linear map  $T$  from  $V_0$  to  $\mathcal{P}_{n-1}$ , given by  $T(a_{j\ell}) = q(s)$ , the numerator in the left side of (1.18.56), and one can verify that this map is one-to-one. Hence (cf. Corollary 2.3.7 of Chapter 2), this map is also onto, and this gives Proposition (1.18.2).  $\square$

Given the representation (1.18.56), we deduce from (1.18.26) that

$$(1.18.57) \quad \mathcal{L}^{-1}\left(\frac{q}{p}\right)(t) = \sum_{\ell=1}^m \sum_{j=1}^{k_\ell} \frac{a_{j\ell}}{(j-1)!} t^{j-1} e^{r_\ell t}.$$

Taking  $q(s) = 1$ , we obtain a function  $\varphi(t)$ , of the form (1.18.57), such that

$$(1.18.58) \quad \mathcal{L}\varphi(s) = \frac{1}{p(s)}.$$

Then the solution  $f(t)$  to (1.18.1)–(1.18.2) is equal to  $\mathcal{L}^{-1}(q/p)(t)$  plus  $f_0(t)$ , satisfying

$$(1.18.59) \quad \mathcal{L}f_0(s) = \frac{\mathcal{L}g(s)}{p(s)} = \mathcal{L}\varphi(s)\mathcal{L}g(s).$$

The following result provides a useful integral formula for  $f_0$ .

**Proposition 1.18.3.** *Let  $\varphi$  and  $g$  satisfy (1.18.3), and set*

$$(1.18.60) \quad \varphi * g(t) = \int_0^t \varphi(t-\tau)g(\tau) d\tau.$$

*Then, for  $s > A$ ,*

$$(1.18.61) \quad \mathcal{L}(\varphi * g)(s) = \mathcal{L}\varphi(s)\mathcal{L}g(s).$$

**Proof.** Given (1.18.60), we have

$$(1.18.62) \quad \begin{aligned} \mathcal{L}(\varphi * g)(s) &= \int_0^\infty e^{-st} \int_0^t \varphi(t-\tau)g(\tau) d\tau dt \\ &= \int_0^\infty \int_0^t e^{-s(t-\tau)} e^{-s\tau} \varphi(t-\tau)g(\tau) d\tau dt \\ &= \int_0^\infty \int_\tau^\infty e^{-s(t-\tau)} e^{-s\tau} \varphi(t-\tau)g(\tau) dt d\tau \\ &= \mathcal{L}\varphi(s) \int_0^\infty e^{-s\tau} g(\tau) d\tau \\ &= \mathcal{L}\varphi(s)\mathcal{L}g(s), \end{aligned}$$

as asserted.  $\square$

Recall that the method of variation of parameters, discussed in §1.14, also leads to an integral formula involving an integral over  $[0, t]$ . In fact, the method of variation of parameters and the use of the Laplace transform discussed here can both be understood as special cases of a general method, involving *Duhamel's formula*, arising when the equations are recast as first-order systems. This is explained in §3.9 of Chapter 3

---

**Exercises**

1. Compute the inverse Laplace transform of the following functions.

$$(a) \frac{1}{s^4 - 1},$$

$$(b) \frac{s + 1}{s^3 + 3s^2 + 2s}.$$

2. Use the Laplace transform to solve the following initial value problems.

$$(a) f''(t) + 3f'(t) + 2f(t) = e^{-t} \sin t, \quad f(0) = 0, \quad f'(0) = 1,$$

$$(b) f^{(4)}(t) - f(t) = \sin t, \quad f^{(j)}(0) = 0 \text{ for } 0 \leq j \leq 3.$$

3. Show that

$$f(t) = \frac{\sin t}{t} \implies \mathcal{L}f(s) = \frac{\pi}{2} - \tan^{-1} s.$$

*Hint.* By (1.18.5),

$$\frac{d}{ds} \mathcal{L}f(s) = -\mathcal{L}(tf)(s) = -\frac{1}{s^2 + 1}.$$

Integrate, and find the constant of integration using

$$\lim_{s \rightarrow \infty} \mathcal{L}f(s) = 0.$$

4. Compute the Laplace transform of

$$\frac{1 - \cos t}{t^2}.$$

### 1.A. The genesis of Bessel's equation: PDE in polar coordinates

Bessel functions, the subject of §1.16, arise in the natural generalization of the equation

$$(1.A.1) \quad \frac{d^2u}{dx^2} + k^2u = 0,$$

with solutions  $\sin kx$  and  $\cos kx$ , to partial differential equations

$$(1.A.2) \quad \Delta u + k^2u = 0,$$

where  $\Delta$  is the Laplace operator, acting on a function  $u$  on a domain  $\Omega \subset \mathbb{R}^n$  by

$$(1.A.3) \quad \Delta u = \frac{\partial^2 u}{\partial x_1^2} + \cdots + \frac{\partial^2 u}{\partial x_n^2}.$$

We can eliminate  $k^2$  from (1.A.2) by scaling. Set  $u(x) = v(kx)$ . Then equation (1.A.2) becomes

$$(1.A.4) \quad (\Delta + 1)v = 0.$$

We specialize to the case  $n = 2$  and write

$$(1.A.5) \quad \Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

For a number of special domains  $\Omega \subset \mathbb{R}^2$ , such as circular domains, annular domains, angular sectors, and pie-shaped domains, it is convenient to switch to polar coordinates  $(r, \theta)$ , related to  $(x, y)$ -coordinates by

$$(1.A.6) \quad x = r \cos \theta, \quad y = r \sin \theta.$$

In such coordinates,

$$(1.A.7) \quad \Delta v = \left( \frac{\partial^2}{\partial r^2} + \frac{1}{r} \frac{\partial}{\partial r} + \frac{1}{r^2} \frac{\partial^2}{\partial \theta^2} \right) v.$$

A special class of solutions to (1.A.4) has the form

$$(1.A.8) \quad v = w(r)e^{i\nu\theta}.$$

By (1.A.7), for such  $v$ ,

$$(1.A.9) \quad (\Delta + 1)v = \left[ \frac{d^2w}{dr^2} + \frac{1}{r} \frac{dw}{dr} + \left( 1 - \frac{\nu^2}{r^2} \right) w \right] e^{i\nu\theta},$$

so (1.A.4) holds if and only if

$$(1.A.10) \quad \frac{d^2w}{dr^2} + \frac{1}{r} \frac{dw}{dr} + \left( 1 - \frac{\nu^2}{r^2} \right) w = 0.$$

This is Bessel's equation (1.16.1) (with different variables).

Note that if  $v$  solves (1.A.4) on  $\Omega \subset \mathbb{R}^2$  and if  $\Omega$  is a circular domain or an annular domain, centered at the origin, then  $\nu$  must be an integer. However, if  $\Omega$  is an angular sector or a pie-shaped domain, with vertex at the origin,  $\nu$  need not be an integer.

In  $n$  dimensions, the Laplace operator (1.A.3) can be written

$$(1.A.11) \quad \Delta v = \left( \frac{\partial^2}{\partial r^2} + \frac{n-1}{r} \frac{\partial}{\partial r} + \frac{1}{r^2} \Delta_S \right) v,$$

where  $\Delta_S$  is a second-order differential operator acting on functions on the unit sphere  $S^{n-1} \subset \mathbb{R}^n$ , called the Laplace-Beltrami operator. Generalizing (1.A.8), one looks for solutions to (1.A.4) of the form

$$(1.A.12) \quad v(x) = w(r)\psi(\omega),$$

where  $x = r\omega$ ,  $r \in (0, \infty)$ ,  $\omega \in S^{n-1}$ . Parallel to (1.A.9), for such  $v$ ,

$$(1.A.13) \quad (\Delta + 1)v = \left[ \frac{d^2 w}{dr^2} + \frac{n-1}{r} \frac{dw}{dr} + \left(1 - \frac{\nu^2}{r^2}\right)w \right] \psi(\omega),$$

provided

$$(1.A.14) \quad \Delta_S \psi = -\nu^2 \psi.$$

The equation

$$(1.A.15) \quad \frac{d^2 w}{dr^2} + \frac{n-1}{r} \frac{dw}{dr} + \left(1 - \frac{\nu^2}{r^2}\right)w = 0$$

is a variant of Bessel's equation. If we set

$$(1.A.16) \quad \varphi(r) = r^{n/2-1}w(r),$$

then (1.A.15) is converted into the Bessel equation

$$(1.A.17) \quad \frac{d^2 \varphi}{dr^2} + \frac{1}{r} \frac{d\varphi}{dr} + \left(1 - \frac{\mu^2}{r^2}\right)\varphi = 0, \quad \mu^2 = \nu^2 + \left(\frac{n-2}{2}\right)^2.$$

The study of solutions to (1.A.14) gives rise to the study of spherical harmonics, and from there to other special functions, such as Legendre functions.

The search for solutions of the form (1.A.12) is a key example of the method of separation of variables for partial differential equations. It arises in numerous other contexts. Here are a couple of other examples:

$$(1.A.18) \quad (\Delta - |x|^2 + k^2)u = 0,$$

and

$$(1.A.19) \quad \left(\Delta + \frac{K}{|x|} + k^2\right)u = 0.$$

The first describes the  $n$ -dimensional quantum harmonic oscillator. The second (for  $n = 3$ ) describes the quantum mechanical model of a hydrogen atom, according to Schrödinger. Study of these equations leads to other special functions defined by differential equations, such as Hermite functions and Whittaker functions.

Much further material on these topics can be found in books on partial differential equations, such as [45] (particularly Chapters 3 and 8).

### 1.B. Euler's gamma function

We saw in (1.16.13) the need for a function  $\Gamma(z)$  satisfying

$$(1.B.1) \quad \Gamma(z+1) = z\Gamma(z).$$

Here we produce a function that has this property, namely

$$(1.B.2) \quad \Gamma(z) = \int_0^\infty e^{-t} t^{z-1} dt, \quad \text{for } z > 0.$$

To check (1.B.1) for  $z > 0$ , we apply integration by parts.

$$(1.B.3) \quad \begin{aligned} \Gamma(z+1) &= \int_0^\infty e^{-t} t^z dt \\ &= - \int_0^\infty \left( \frac{d}{dt} e^{-t} \right) t^z dt \\ &= \int_0^\infty e^{-t} \left( \frac{d}{dt} t^z \right) dt \\ &= z\Gamma(z), \end{aligned}$$

since  $dt^z/dt = z t^{z-1}$ .

The integral (1.B.2) is readily evaluated for  $z = 1$ , yielding

$$(1.B.4) \quad \Gamma(1) = 1.$$

Then repeated use of (1.B.3) gives

$$(1.B.5) \quad \Gamma(k+1) = k!, \quad \text{for } k \in \mathbb{Z}^+.$$

There is also a useful formula for  $\Gamma(1/2)$ , given by

$$(1.B.6) \quad \begin{aligned} \Gamma\left(\frac{1}{2}\right) &= \int_0^\infty e^{-t} t^{-1/2} dt \\ &= 2 \int_0^\infty e^{-x^2} dx \\ &= \sqrt{\pi}, \end{aligned}$$

the last identity by (1.2.26). Then repeated use of (1.B.3) gives

$$(1.B.7) \quad \begin{aligned} \Gamma\left(k + \frac{1}{2}\right) &= \frac{2k-1}{2} \frac{2k-3}{2} \cdots \frac{1}{2} \Gamma\left(\frac{1}{2}\right) \\ &= 2^{-2k} \frac{(2k)!}{k!} \sqrt{\pi}. \end{aligned}$$

Having (1.B.1), we can extend  $\Gamma(z)$  to be well defined and smooth on  $\mathbb{R} \setminus \{0, -1, -2, \dots\}$ . To see this, rewrite (1.B.1) as

$$(1.B.8) \quad \Gamma(z) = \frac{1}{z} \Gamma(z+1).$$

Having  $\Gamma(z)$  defined and smooth on  $z \in (0, \infty)$ , by (1.B.2), we see that the right side of (1.B.8) is defined and smooth for  $z \in (-1, \infty)$ , except for a pole at  $z = 0$ . This extends  $\Gamma(z)$  to  $z \in (-1, \infty) \setminus \{0\}$ . Then the right side of (1.B.8) is defined

and smooth for  $z \in (-2, \infty)$ , except for poles at  $z = 0$  and  $z = -1$ . This argument can be continued. Let us further note that, by (1.B.2),

$$(1.B.9) \quad \Gamma(z) > 0 \quad \text{for } z > 0,$$

so  $1/\Gamma(z)$  is defined and smooth for  $z \in (0, \infty)$ . Rewriting (1.B.8) as

$$(1.B.10) \quad \frac{1}{\Gamma(z)} = \frac{z}{\Gamma(z+1)}$$

and arguing as above, we have  $1/\Gamma(z)$  defined and smooth for all  $z \in \mathbb{R}$ , vanishing precisely for  $z \in \{0, -1, -2, \dots\}$ .

We derive another identity that is useful for the treatment of Bessel functions in §1.16, involving the beta function  $B(x, y)$ , defined for  $x, y > 0$  by

$$(1.B.11) \quad \begin{aligned} B(x, y) &= \int_0^1 s^{x-1}(1-s)^{y-1} ds \\ &= \int_0^\infty (1+u)^{-x-y} u^{x-1} du, \end{aligned}$$

the latter identity via the change of variable  $u = s/(1-s)$ . Our asserted identity is

$$(1.B.12) \quad B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

To prove this, note that since

$$(1.B.13) \quad \Gamma(z)p^{-z} = \int_0^\infty e^{-pt} t^{z-1} dt,$$

we have

$$(1.B.14) \quad (1+u)^{-x-y} = \frac{1}{\Gamma(x+y)} \int_0^\infty e^{-(1+u)t} t^{x+y-1} dt,$$

so

$$(1.B.15) \quad \begin{aligned} B(x, y) &= \frac{1}{\Gamma(x+y)} \int_0^\infty e^{-t} t^{x+y-1} \int_0^\infty e^{-ut} u^{x-1} du dt \\ &= \frac{\Gamma(x)}{\Gamma(x+y)} \int_0^\infty e^{-t} t^{y-1} dt \\ &= \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \end{aligned}$$

as asserted.

For closer contact with (1.16.38), note that setting  $s = t^2$  in (1.B.11) gives

$$(1.B.16) \quad B(x, y) = 2 \int_0^1 t^{2x-1} (1-t^2)^{y-1} dt,$$

so, if  $k \in \mathbb{Z}^+$  and  $\nu > -1/2$ ,

$$(1.B.17) \quad B\left(k + \frac{1}{2}, \nu + \frac{1}{2}\right) = \int_{-1}^1 t^{2k} (1-t^2)^{\nu-1/2} dt.$$

There is much more that can be said about the gamma function, such as that it extends to  $\mathbb{C} \setminus \{0, -1, -2, \dots\}$ , with  $1/\Gamma(z)$  defined and smooth for all  $z \in \mathbb{C}$  (which permits one to define  $J_\nu(z)$  for complex  $\nu$ ). We refer the reader to [28],

[47], §4.3, or [45], Chapter 3, Appendix A for further material. We mention the following identity, of use in (1.16.33), whose proof can be found in these references:

$$(1.B.18) \quad \Gamma(\nu)\Gamma(1-\nu) = \frac{\pi}{\sin \pi\nu}.$$

Note that both sides are defined and smooth for  $\nu \in \mathbb{R} \setminus \mathbb{Z}$ , with singularities on  $\mathbb{Z}$ .

### 1.C. Differentiating power series

Here we establish continuity and differentiability properties for a power series

$$(1.C.1) \quad f(t) = \sum_{k=0}^{\infty} a_k t^k.$$

We allow the coefficients  $a_k$  to be complex numbers. To start, we assume this series converges for some nonzero  $t = t_0$ . This implies that the terms in this series are uniformly bounded for  $t = t_0$ :

$$(1.C.2) \quad |a_k t_0^k| \leq B < \infty, \quad \forall k.$$

The following result establishes convergence for all smaller  $|t|$ .

**Proposition 1.C.1.** *Given (1.C.2), the series (1.C.1) converges absolutely for  $|t| < T = |t_0|$ .*

**Proof.** Pick  $S \in (0, T)$ , and assume  $|t| \leq S$ . Then

$$(1.C.3) \quad |a_k t^k| \leq |a_k T^k| \left(\frac{S}{T}\right)^k \leq B r^k,$$

where  $r = S/T \in (0, 1)$ . Hence, for each  $n \in \mathbb{N}$ , if  $|t| \leq S$ ,

$$(1.C.4) \quad \sum_{k=0}^n |a_k t^k| \leq B \sum_{k=0}^n r^k.$$

Now we can evaluate the geometrical series on the right:

$$(1.C.5) \quad \begin{aligned} S_n &= \sum_{k=0}^n r^k \Rightarrow r S_n = \sum_{k=1}^{n+1} r^k \\ &\Rightarrow (1-r)S_n = 1 - r^{n+1} \\ &\Rightarrow S_n = \frac{1 - r^{n+1}}{1 - r}. \end{aligned}$$

Consequently,

$$(1.C.6) \quad \begin{aligned} 0 < r < 1 &\Rightarrow r^{n+1} \searrow 0 \text{ as } n \rightarrow \infty \\ &\Rightarrow S_n \nearrow \frac{1}{1-r} \text{ as } n \rightarrow \infty. \end{aligned}$$

This establishes the asserted absolute convergence.  $\square$

Similar arguments also lead to the following.

**Proposition 1.C.2.** *In the setting of Proposition 1.C.1, if  $0 < S < T$ , the series (1.C.1) converges uniformly on  $|t| \leq S$ .*

**Proof.** For each  $n \in \mathbb{N}$ , write

$$(1.C.7) \quad \begin{aligned} f(t) &= \sum_{k=0}^n a_k t^k + \sum_{k=n+1}^{\infty} a_k t^k \\ &= S_n(t) + R_n(t). \end{aligned}$$



The claim is that  $S_n(t) \rightarrow f(t)$ , uniformly on  $|t| \leq S$ . Indeed, for  $|t| \leq S$ ,

$$\begin{aligned}
 |R_n(t)| &\leq \sum_{k=n+1}^{\infty} |a_k t^k| \\
 &\leq B \sum_{k=n+1}^{\infty} r * k \\
 (1.C.8) \quad &= B r^{n+1} \sum_{\ell=0}^{\infty} r^{\ell} \\
 &= B \frac{r^{n+1}}{1-r},
 \end{aligned}$$

yielding  $|R_n(t)| \rightarrow 0$  uniformly for  $|t| \leq S$ .  $\square$

Before continuing our study of the power series (1.C.1), we pause to note that calculations above involving the geometric series (1.C.8) enable us to establish the following result, known as the ratio test.

**Proposition 1.C.3.** *Let  $a_k \in \mathbb{C}$  and assume there exist  $N < \infty$  and  $r < 1$  such that*

$$(1.C.9) \quad k \geq N \implies \left| \frac{a_{k+1}}{a_k} \right| \leq r.$$

*Then the series  $\sum_{k \geq 0} a_k$  is absolutely convergent.*

**Proof.** From (1.C.9) we have, by induction,

$$(1.C.10) \quad |a_{N+\ell}| \leq r^{\ell} |a_N|.$$

Hence

$$\begin{aligned}
 (1.C.11) \quad \sum_{\ell=0}^{\infty} |a_{N+\ell}| &\leq |a_N| \sum_{\ell=0}^{\infty} r^{\ell} \\
 &= \frac{|a_N|}{1-r}.
 \end{aligned}$$

This yields absolute convergence.  $\square$

We now state the main result of this appendix.

**Proposition 1.C.4.** *If the power series (1.C.1) converges for  $|t| < R$ , then  $f$  is differentiable in  $t \in (-R, R)$ , and, for such  $t$ ,*

$$(1.C.12) \quad f'(t) = \sum_{k=1}^{\infty} k a_k t^{k-1}.$$

**Proof.** It suffices to show that (1.C.12) holds for  $|t| \leq S$ , for each  $S < R$ . Pick  $T \in (S, R)$ , and note that the estimate (1.C.3) holds, when  $|t| \leq S$ , with  $r = S/T < 1$ . Hence, for  $|t| \leq S$ ,

$$\begin{aligned}
 (1.C.13) \quad |k a_k t^{k-1}| &\leq \frac{k}{T} |a_k T^k| \left( \frac{S}{T} \right)^{k-1} \\
 &\leq \frac{B}{T} k r^{k-1}.
 \end{aligned}$$

Now the ratio test applies to  $\sum_{k \geq 1} kr^{k-1}$ , given  $r < 1$ , so the series

$$(1.C.14) \quad g(t) = \sum_{k=1}^{\infty} ka_k t^{k-1}$$

is absolutely convergent, and also uniformly convergent, for  $|t| \leq S$ . It remains to show that  $g(t) = f'(t)$  for  $|t| \leq S$ , or equivalently that

$$(1.C.15) \quad \int_0^t g(s) ds = f(t) - f(0).$$

This is a consequence of the following result. □

**Proposition 1.C.5.** *Given  $b_k \in \mathbb{C}$ , assume*

$$(1.C.16) \quad g(t) = \sum_{k=0}^{\infty} b_k t^k$$

*is absolutely convergent, for  $|t| < R$ . Then, for  $|t| < R$ ,*

$$(1.C.17) \quad \int_0^t g(s) ds = \sum_{k=0}^{\infty} \frac{b_k}{k+1} t^{k+1}.$$

**Proof.** It is elementary that the series on the right side of (1.C.17) converges for  $|t| < R$ . Call the sum  $F(t)$ . As before, pick  $S < T < R$ . For  $n \in \mathbb{N}$ , write

$$(1.C.18) \quad \begin{aligned} g(t) &= \sum_{k=0}^n b_k t^k + \sum_{k=n+1}^{\infty} b_k t^k \\ &= g_n(t) + R_n(t). \end{aligned}$$

As in Proposition 1.C.2, we have  $g_n(t) \rightarrow g(t)$  and  $R_n(t) \rightarrow 0$ , uniformly for  $|t| \leq S$ , especially

$$(1.C.19) \quad \max_{|t| \leq S} |R_n(t)| \leq \varepsilon_n \rightarrow 0.$$

Clearly, for  $|t| < R$ ,

$$(1.C.20) \quad \int_0^t g_n(s) ds = \sum_{k=0}^n \frac{b_k}{k+1} t^{k+1} \rightarrow F(t),$$

as  $n \rightarrow \infty$ . Meanwhile,

$$(1.C.21) \quad \left| \int_0^t R_n(s) ds \right| \leq R\varepsilon_n.$$

Taking  $n \rightarrow \infty$  in (1.C.18)–(1.C.21) yields

$$(1.C.22) \quad \int_0^t g(s) ds = F(t),$$

as asserted. This proves Proposition 1.C.5, so we have Proposition 1.C.4. □

Having (1.C.12), we can iterate, computing the derivative of  $f'(t)$ , as

$$(1.C.23) \quad f''(t) = \sum_{k=2}^{\infty} k(k-1)a_k t^{k-2},$$

and so on,

$$(1.C.24) \quad f^{(n)}(t) = \sum_{k=n}^{\infty} k(k-1)\cdots(k-n+1)a_k t^{k-n}.$$

In particular,

$$(1.C.25) \quad f^{(n)}(0) = n!a_n, \quad \text{hence } a_n = \frac{f^{(n)}(0)}{n!}.$$

We have the following.

**Proposition 1.C.6.** *If  $f(t)$  is given by a convergent power series (1.C.1) for  $|t| < T$ ,  $T > 0$ , then*

$$(1.C.26) \quad f(t) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} t^k.$$

Frequently, one can turn this around, take a function  $f : (-T, T) \rightarrow \mathbb{R}$ , compute  $f^{(k)}(0)$ , and investigate whether (1.C.26) holds. Here is an important class of functions for which this works. Take  $r \in \mathbb{R}$ , and set

$$(1.C.27) \quad f(t) = (1-t)^{-r}.$$

We have

$$(1.C.28) \quad \begin{aligned} f'(t) &= r(1-t)^{-r-1}, \\ f''(t) &= r(r+1)(1-t)^{-r-2}, \\ &\vdots \\ f^{(n)}(t) &= r(r+1)\cdots(r+n-1)(1-t)^{-r-n}, \end{aligned}$$

hence

$$(1.C.29) \quad f^{(n)}(0) = r(r+1)\cdots(r+n-1).$$

**Claim.** For  $r \in \mathbb{R}$ , we have

$$(1.C.30) \quad (1-t)^{-r} = \sum_{k=0}^{\infty} \frac{r(r+1)\cdots(r+k-1)}{k!} t^k, \quad \text{for } |t| < 1.$$

In other words, (1.C.1) holds with

$$(1.C.31) \quad a_k = \frac{r(r+1)\cdots(r+k-1)}{k!}.$$

Note that

$$(1.C.32) \quad a_{k+1} = \frac{r+k}{k} a_k,$$

so the ratio test implies that the right side of (1.C.30) is absolutely convergent for  $|t| < 1$ , i.e., we have a well defined continuous (and differentiable) function

$$(1.C.33) \quad g(t) = \sum_{k=0}^{\infty} \frac{r(r+1) \cdots (r+k-1)}{k!} t^k.$$

Our claim is therefore that

$$(1.C.34) \quad g(t) = (1-t)^{-r}.$$

One approach to this is to estimate the remainder  $R_n(t)$  in the expansion

$$(1.C.35) \quad f(t) = \sum_{k=0}^n \frac{f^{(k)}(0)}{k!} t^k + R_n(t).$$

A discussion of this appears in §4.3 of [49]. Here is another approach. We can apply Proposition 1.C.4 to  $g(t)$  to obtain

$$(1.C.36) \quad (1-t)g'(t) = rg(t),$$

and then calculate

$$(1.C.37) \quad \begin{aligned} \frac{d}{dt}(1-t)^r g(t) &= (1-t)^r g'(t) - r(1-t)^{r-1} g(t) \\ &= (1-t)^{r-1} \left\{ (1-t)g'(t) - rg(t) \right\} \\ &= 0, \end{aligned}$$

and deduce (1.C.34), hence (1.C.30).

For an application of (1.C.30), with  $r = 1/2$ , see (1.6.60).

REMARK. Note the parallel between the use of (1.C.37) to prove (1.C.30) and the use of (1.1.10) to prove (1.1.13).



# Linear algebra

The purpose of this chapter is to provide sufficient background in linear algebra for understanding the material of Chapter 3, on linear systems of differential equations. Results here will also be useful for the development of nonlinear systems in Chapter 4.

In §2.1 we define the class of vector spaces (real and complex) and discuss some basic examples, including  $\mathbb{R}^n$  and  $\mathbb{C}^n$ , or, as we denote them,  $\mathbb{F}^n$ , with  $\mathbb{F} = \mathbb{R}$  or  $\mathbb{C}$ . In §2.2 we consider linear transformations between such vector spaces. In particular we look at an  $m \times n$  matrix  $A$  as defining a linear transformation  $A : \mathbb{F}^n \rightarrow \mathbb{F}^m$ . We define the range  $\mathcal{R}(T)$  and null space  $\mathcal{N}(T)$  of a linear transformation  $T : V \rightarrow W$ . In §2.3 we define the notion of basis of a vector space. Vector spaces with finite bases are called finite dimensional. We establish the crucial property that any two bases of such a vector space  $V$  have the same number of elements (denoted  $\dim V$ ). We apply this to other results on bases of vector spaces, culminating in the “fundamental theorem of linear algebra,” that if  $T : V \rightarrow W$  is linear and  $V$  is finite dimensional, then  $\dim \mathcal{N}(T) + \dim \mathcal{R}(T) = \dim V$ , and discuss some of its important consequences.

A linear transformation  $T : V \rightarrow V$  is said to be invertible provided it is one-to-one and onto, i.e., provided  $\mathcal{N}(T) = 0$  and  $\mathcal{R}(T) = V$ . In §2.5 we define the determinant of such  $T$ ,  $\det T$  (when  $V$  is finite dimensional), and show that  $T$  is invertible if and only if  $\det T \neq 0$ . In §2.6 we study eigenvalues  $\lambda_j$  and eigenvectors  $v_j$  of such a transformation, defined by  $Tv_j = \lambda_j v_j$ . Results of §2.5 imply  $\lambda_j$  is a root of the “characteristic polynomial”  $\det(\lambda I - T)$ . Section 2.7 extends the scope of §2.6 to a treatment of generalized eigenvectors. This topic is connected to properties of nilpotent matrices and triangular matrices, studied in §2.8.

In §2.9 we treat inner products on vector spaces, which endow them with a Euclidean geometry, in particular with a distance and a norm. In §2.10 we discuss two types of norms on linear transformations, the “operator norm” and the “Hilbert-Schmidt norm.” Then, in §§2.11–2.12, we discuss some special classes of

linear transformations on inner product spaces: self-adjoint, skew-adjoint, unitary, and orthogonal transformations.

Some appendices supplement the material of this chapter, with a treatment of the Jordan canonical form and Schur's theorem on upper triangularization. This material is not needed for Chapter 3, but for the interested reader it provides a more complete introduction to linear algebra. (A great deal more on linear algebra can be found in [48].) The third appendix gives a proof of the fundamental theorem of algebra, that every nonconstant polynomial has complex roots. This result has several applications in §§2.6–2.7.

## 2.1. Vector spaces

The reader is most likely familiar with vectors in the plane  $\mathbb{R}^2$  and 3-space  $\mathbb{R}^3$ . More generally we have  $n$ -space  $\mathbb{R}^n$ , whose elements consist of  $n$ -tuples of real numbers:

$$(2.1.1) \quad v = (v_1, \dots, v_n).$$

There is vector addition; if also  $w = (w_1, \dots, w_n) \in \mathbb{R}^n$ ,

$$(2.1.2) \quad v + w = (v_1 + w_1, \dots, v_n + w_n).$$

There is also multiplication by scalars; if  $a$  is a real number (a *scalar*),

$$(2.1.3) \quad av = (av_1, \dots, av_n).$$

We could also use complex numbers, replacing  $\mathbb{R}^n$  by  $\mathbb{C}^n$ , and allowing  $a \in \mathbb{C}$  in (2.1.3). We will use  $\mathbb{F}$  to denote  $\mathbb{R}$  or  $\mathbb{C}$ .

Many other vector spaces arise naturally. We define this general notion now. A vector space over  $\mathbb{F}$  is a set  $V$ , endowed with two operations, that of vector addition and multiplication by scalars. That is, given  $v, w \in V$  and  $a \in \mathbb{F}$ , then  $v + w$  and  $av$  are defined in  $V$ . Furthermore, the following properties are to hold, for all  $u, v, w \in V$ ,  $a, b \in \mathbb{F}$ . First there are laws for vector addition:

$$(2.1.4) \quad \text{Commutative law} \quad : \quad u + v = v + u,$$

$$(2.1.5) \quad \text{Associative law} \quad : \quad (u + v) + w = u + (v + w),$$

$$(2.1.6) \quad \text{Zero vector} \quad : \quad \exists 0 \in V, v + 0 = v,$$

$$(2.1.7) \quad \text{Negative} \quad : \quad \exists -v, v + (-v) = 0.$$

Next there are laws for multiplication by scalars:

$$(2.1.8) \quad \text{Associative law} \quad : \quad a(bv) = (ab)v,$$

$$(2.1.9) \quad \text{Unit} \quad : \quad 1 \cdot v = v.$$

Finally there are two distributive laws:

$$(2.1.10) \quad a(u + v) = au + av,$$

$$(2.1.11) \quad (a + b)u = au + bu.$$

It is easy to see that  $\mathbb{R}^n$  and  $\mathbb{C}^n$  satisfy all these rules. We will present a number of other examples below. Let us also note that a number of other simple identities are automatic consequences of the rules given above. Here are some, which the reader is invited to verify:

$$(2.1.12) \quad \begin{aligned} v + w = v &\Rightarrow w = 0, \\ v + 0 \cdot v &= (1 + 0)v = v, \\ 0 \cdot v &= 0, \\ v + w = 0 &\Rightarrow w = -v, \\ v + (-1)v &= 0 \cdot v = 0, \\ (-1)v &= -v. \end{aligned}$$



Above we represented elements of  $\mathbb{F}^n$  as *row vectors*. Often we represent elements of  $\mathbb{F}^n$  as *column vectors*. We write

$$(2.1.13) \quad v = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}, \quad av + w = \begin{pmatrix} av_1 + w_1 \\ \vdots \\ av_n + w_n \end{pmatrix}.$$

We give some other examples of vector spaces. Let  $I = [a, b]$  denote an interval in  $\mathbb{R}$ , and take a non-negative integer  $k$ . Then  $C^k(I)$  denotes the set of functions  $f : I \rightarrow \mathbb{F}$  whose derivatives up to order  $k$  are continuous. We denote by  $\mathcal{P}$  the set of polynomials in  $x$ , with coefficients in  $\mathbb{F}$ . We denote by  $\mathcal{P}_k$  the set of polynomials in  $x$  of degree  $\leq k$ . In these various cases,

$$(2.1.14) \quad (f + g)(x) = f(x) + g(x), \quad (af)(x) = af(x).$$

Such vector spaces and certain of their linear subspaces play a major role in the material developed in these notes.

Regarding the notion just mentioned, we say a subset  $W$  of a vector space  $V$  is a linear subspace provided

$$(2.1.15) \quad w_j \in W, a_j \in \mathbb{F} \implies a_1w_1 + a_2w_2 \in W.$$

Then  $W$  inherits the structure of a vector space.

## Exercises

1. Specify which of the following subsets of  $\mathbb{R}^3$  are linear subspaces:

- (a)  $\{(x, y, z) : xy = 0\}$ ,
- (b)  $\{(x, y, z) : x + y = 0\}$ ,
- (c)  $\{(x, y, z) : x \geq 0, y = 0, z = 0\}$ ,
- (d)  $\{(x, y, z) : x \text{ is an integer}\}$ ,
- (d)  $\{(x, y, z) : x = 2z, y = -z\}$ .

2. Show that the results in (2.1.12) follow from the basic rules (2.1.4)–(2.1.11).

*Hint.* To start, add  $-v$  to both sides of the identity  $v + w = v$ , and take account first of the associative law (2.1.5), and then of the rest of (2.1.4)–(2.1.7). For the second line of (2.1.12), use the rules (2.1.9) and (2.1.11). Then use the first two lines of (2.1.12) to justify the third line...

3. Demonstrate that the following results for any vector space. Take  $a \in \mathbb{F}$ ,  $v \in V$ .

$$a \cdot 0 = 0 \in V,$$

$$a(-v) = -av.$$

*Hint.* Feel free to use the results of (2.1.12).

Let  $V$  be a vector space (over  $\mathbb{F}$ ) and  $W, X \subset V$  linear subspaces. We say

$$(2.1.16) \quad V = W + X$$

provided each  $v \in V$  can be written

$$(2.1.17) \quad v = w + x, \quad w \in W, \quad x \in X.$$

We say

$$(2.1.18) \quad V = W \oplus X$$

provided each  $v \in V$  has a unique representation (2.1.17).

4. Show that

$$V = W \oplus X \iff V = W + X \quad \text{and} \quad W \cap X = 0.$$

5. Take  $V = \mathbb{R}^3$ . Specify in each case (a)–(c) whether  $V = W + X$  and whether  $V = W \oplus X$ .

$$(a) \quad W = \{(x, y, z) : z = 0\}, \quad X = \{(x, y, z) : x = 0\},$$

$$(b) \quad W = \{(x, y, z) : z = 0\}, \quad X = \{(x, y, z) : x = y = 0\},$$

$$(c) \quad W = \{(x, y, z) : z = 0\}, \quad X = \{(x, y, z) : y = z = 0\}.$$

6. If  $W_1, \dots, W_m$  are linear subspaces of  $V$ , extend (2.1.16) to the notion

$$(2.1.19) \quad V = W_1 + \dots + W_m,$$

and extend (2.1.18) to the notion that

$$(2.1.20) \quad V = W_1 \oplus \dots \oplus W_m.$$

## 2.2. Linear transformations and matrices

If  $V$  and  $W$  are vector spaces over  $\mathbb{F}$  ( $\mathbb{R}$  or  $\mathbb{C}$ ), a map

$$(2.2.1) \quad T : V \longrightarrow W$$

is said to be a *linear transformation* provided

$$(2.2.2) \quad T(a_1v_1 + a_2v_2) = a_1Tv_1 + a_2Tv_2, \quad \forall a_j \in \mathbb{F}, v_j \in V.$$

We also write  $T \in \mathcal{L}(V, W)$ . In case  $V = W$ , we also use the notation  $\mathcal{L}(V) = \mathcal{L}(V, V)$ .

Linear transformations arise in a number of ways. For example, an  $m \times n$  matrix  $A$  with entries in  $\mathbb{F}$  defines a linear transformation

$$(2.2.3) \quad A : \mathbb{F}^n \longrightarrow \mathbb{F}^m$$

by

$$(2.2.4) \quad \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} \sum a_{1\ell} b_\ell \\ \vdots \\ \sum a_{m\ell} b_\ell \end{pmatrix}.$$

We also have linear transformations on function spaces, such as multiplication operators

$$(2.2.5) \quad M_f : C^k(I) \longrightarrow C^k(I), \quad M_f g(x) = f(x)g(x),$$

given  $f \in C^k(I)$ ,  $I = [a, b]$ , and the operation of differentiation:

$$(2.2.6) \quad D : C^{k+1}(I) \longrightarrow C^k(I), \quad Df(x) = f'(x).$$

We also have integration:

$$(2.2.7) \quad \mathcal{I} : C^k(I) \longrightarrow C^{k+1}(I), \quad \mathcal{I}f(x) = \int_a^x f(y) dy.$$

Note also that

$$(2.2.8) \quad D : \mathcal{P}_{k+1} \longrightarrow \mathcal{P}_k, \quad \mathcal{I} : \mathcal{P}_k \longrightarrow \mathcal{P}_{k+1},$$

where  $\mathcal{P}_k$  denotes the space of polynomials in  $x$  of degree  $\leq k$ .

Two linear transformations  $T_j \in \mathcal{L}(V, W)$  can be added:

$$(2.2.9) \quad T_1 + T_2 : V \longrightarrow W, \quad (T_1 + T_2)v = T_1v + T_2v.$$

Also  $T \in \mathcal{L}(V, W)$  can be multiplied by a scalar:

$$(2.2.10) \quad aT : V \longrightarrow W, \quad (aT)v = a(Tv).$$

This makes  $\mathcal{L}(V, W)$  a vector space.

We can also compose linear transformations  $S \in \mathcal{L}(W, X)$ ,  $T \in \mathcal{L}(V, W)$ :

$$(2.2.11) \quad ST : V \longrightarrow X, \quad (ST)v = S(Tv).$$

For example, we have

$$(2.2.12) \quad M_f D : C^{k+1}(I) \longrightarrow C^k(I), \quad M_f Dg(x) = f(x)g'(x),$$

given  $f \in C^k(I)$ . When two transformations

$$(2.2.13) \quad A : \mathbb{F}^n \longrightarrow \mathbb{F}^m, \quad B : \mathbb{F}^k \longrightarrow \mathbb{F}^n$$

are represented by matrices, e.g.,  $A$  as in (2.2.4) and

$$(2.2.14) \quad B = \begin{pmatrix} b_{11} & \cdots & b_{1k} \\ \vdots & & \vdots \\ b_{n1} & \cdots & b_{nk} \end{pmatrix},$$

then

$$(2.2.15) \quad AB : \mathbb{F}^k \longrightarrow \mathbb{F}^m$$

is given by matrix multiplication:

$$(2.2.16) \quad AB = \begin{pmatrix} \Sigma a_{1\ell} b_{\ell 1} & \cdots & \Sigma a_{1\ell} b_{\ell k} \\ \vdots & & \vdots \\ \Sigma a_{m\ell} b_{\ell 1} & \cdots & \Sigma a_{m\ell} b_{\ell k} \end{pmatrix}.$$

For example,

$$(2.2.17) \quad \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}.$$

Another way of writing (2.2.16) is to represent  $A$  and  $B$  as

$$(2.2.18) \quad A = (a_{ij}), \quad B = (b_{ij}),$$

and then we have

$$(2.2.19) \quad AB = (d_{ij}), \quad d_{ij} = \sum_{\ell=1}^n a_{i\ell} b_{\ell j}.$$

To establish the identity (2.2.16), we note that it suffices to show the two sides have the same effect on each  $e_j \in \mathbb{F}^k$ ,  $1 \leq j \leq k$ , where  $e_j$  is the column vector in  $\mathbb{F}^k$  whose  $j$ th entry is 1 and whose other entries are 0. First note that

$$(2.2.20) \quad Be_j = \begin{pmatrix} b_{1j} \\ \vdots \\ b_{nj} \end{pmatrix},$$

the  $j$ th column in  $B$ , as one can see via (2.2.4). Similarly, if  $D$  denotes the right side of (2.2.16),  $De_j$  is the  $j$ th column of this matrix, i.e.,

$$(2.2.21) \quad De_j = \begin{pmatrix} \Sigma a_{1\ell} b_{\ell j} \\ \vdots \\ \Sigma a_{m\ell} b_{\ell j} \end{pmatrix}.$$

On the other hand, applying  $A$  to (2.2.20), via (2.2.4), gives the same result, so (2.2.16) holds.

Associated with a linear transformation as in (2.2.1) there are two special linear spaces, the *null space* of  $T$  and the *range* of  $T$ . The null space of  $T$  is

$$(2.2.22) \quad \mathcal{N}(T) = \{v \in V : Tv = 0\},$$

and the range of  $T$  is

$$(2.2.23) \quad \mathcal{R}(T) = \{Tv : v \in V\}.$$

Note that  $\mathcal{N}(T)$  is a linear subspace of  $V$  and  $\mathcal{R}(T)$  is a linear subspace of  $W$ . If  $\mathcal{N}(T) = 0$  we say  $T$  is injective; if  $\mathcal{R}(T) = W$  we say  $T$  is surjective. Note that  $T$  is injective if and only if  $T$  is one-to-one, i.e.,

$$(2.2.24) \quad Tv_1 = Tv_2 \implies v_1 = v_2.$$

If  $T$  is surjective, we also say  $T$  is *onto*. If  $T$  is one-to-one and onto, we say it is an *isomorphism*. In such a case the *inverse*

$$(2.2.25) \quad T^{-1} : W \longrightarrow V$$

is well defined, and it is a linear transformation. We also say  $T$  is invertible, in such a case.

### Exercises

1. With  $D$  and  $\mathcal{I}$  given by (2.2.6)–(2.2.7), compute  $D\mathcal{I}$  and  $\mathcal{I}D$ .
2. In the context of Exercise 1, specify  $\mathcal{N}(D)$ ,  $\mathcal{N}(\mathcal{I})$ ,  $\mathcal{R}(D)$ , and  $\mathcal{R}(\mathcal{I})$ .
3. Consider  $A, B : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ , given by

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Compute  $AB$  and  $BA$ .

4. In the context of Exercise 3, specify

$$\mathcal{N}(A), \quad \mathcal{N}(B), \quad \mathcal{R}(A), \quad \mathcal{R}(B).$$

5. We say two  $n \times n$  matrices  $A$  and  $B$  *commute* provided  $AB = BA$ . Note that  $AB \neq BA$  in Exercise 3. Pick out the pair of commuting matrices from this list:

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

6. Show that (2.2.4) is a special case of matrix multiplication, as defined by the right side of (2.2.16).

7. Show, without using the formula (2.2.16) identifying compositions of linear transformations and matrix multiplication, that matrix multiplication is associative, i.e.,

$$(2.2.26) \quad A(BC) = (AB)C,$$

where  $C : \mathbb{F}^\ell \rightarrow \mathbb{F}^k$  is given by a  $k \times \ell$  matrix and the products in (2.2.26) are defined as matrix products, as in (2.2.19).

8. Show that the asserted identity (2.2.16) identifying compositions of linear transformations with matrix products follows from the result of Exercise 7.

*Hint.* (2.2.4), defining the action of  $A$  on  $\mathbb{F}^n$ , is a matrix product.

9. Let  $A : \mathbb{F}^n \rightarrow \mathbb{F}^m$  be defined by an  $m \times n$  matrix, as in (2.2.3)–(2.2.4).

(a) Show that  $\mathcal{R}(A)$  is the span of the columns of  $A$ .

*Hint.* See (2.2.20).

(b) Show that  $\mathcal{N}(A) = 0$  if and only if the columns of  $A$  are linearly independent.

10. Define the transpose of an  $m \times n$  matrix  $A = (a_{jk})$  to be the  $n \times m$  matrix  $A^t = (a_{kj})$ . Thus, if  $A$  is as in (2.2.3)–(2.2.4),

$$(2.2.27) \quad A^t = \begin{pmatrix} a_{11} & \cdots & a_{m1} \\ \vdots & & \vdots \\ a_{1n} & \cdots & a_{mn} \end{pmatrix}.$$

For example,

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix} \implies A^t = \begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{pmatrix}.$$

Suppose also  $B$  is an  $n \times k$  matrix, as in (2.2.14), so  $AB$  is defined, as in (2.2.15).

Show that

$$(2.2.28) \quad (AB)^t = B^t A^t.$$

11. Let

$$A = (1 \quad 2 \quad 3), \quad B = \begin{pmatrix} 2 \\ 0 \\ 2 \end{pmatrix}.$$

Compute  $AB$  and  $BA$ . Then compute  $A^t B^t$  and  $B^t A^t$ .

### 2.3. Basis and dimension

Given a finite set  $S = \{v_1, \dots, v_k\}$  in a vector space  $V$ , the *span* of  $S$  is the set of vectors in  $V$  of the form

$$(2.3.1) \quad c_1 v_1 + \cdots + c_k v_k,$$

with  $c_j$  arbitrary scalars, ranging over  $\mathbb{F} = \mathbb{R}$  or  $\mathbb{C}$ . This set, denoted  $\text{Span}(S)$  is a linear subspace of  $V$ . The set  $S$  is said to be *linearly dependent* if and only if there exist scalars  $c_1, \dots, c_k$ , not all zero, such that (2.3.1) vanishes. Otherwise we say  $S$  is *linearly independent*.

If  $\{v_1, \dots, v_k\}$  is linearly independent, we say  $S$  is a *basis* of  $\text{Span}(S)$ , and that  $k$  is the *dimension* of  $\text{Span}(S)$ . In particular, if this holds and  $\text{Span}(S) = V$ , we say  $k = \dim V$ . We also say  $V$  has a finite basis, and that  $V$  is finite dimensional.

By convention, if  $V$  has only one element, the zero element, we say  $V = 0$  and  $\dim V = 0$ .

It is easy to see that any finite set  $S = \{v_1, \dots, v_k\} \subset V$  has a maximal subset that is linearly independent, and such a subset has the same span as  $S$ , so  $\text{Span}(S)$  has a basis. To take a complementary perspective,  $S$  will have a minimal subset  $S_0$  with the same span, and any such minimal subset will be a basis of  $\text{Span}(S)$ . Soon we will show that any two bases of a finite-dimensional vector space  $V$  have the same number of elements (so  $\dim V$  is well defined). First, let us relate  $V$  to  $\mathbb{F}^k$ .

So say  $V$  has a basis  $S = \{v_1, \dots, v_k\}$ . We define a linear transformation

$$(2.3.2) \quad \mathcal{J}_S : \mathbb{F}^k \longrightarrow V$$

by

$$(2.3.3) \quad \mathcal{J}_S(c_1 e_1 + \cdots + c_k e_k) = c_1 v_1 + \cdots + c_k v_k,$$

where

$$(2.3.4) \quad e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \dots, e_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

We say  $\{e_1, \dots, e_k\}$  is the standard basis of  $\mathbb{F}^k$ . The linear independence of  $S$  is equivalent to the injectivity of  $\mathcal{J}_S$  and the statement that  $S$  spans  $V$  is equivalent to the surjectivity of  $\mathcal{J}_S$ . Hence the statement that  $S$  is a basis of  $V$  is equivalent to the statement that  $\mathcal{J}_S$  is an isomorphism, with inverse uniquely specified by

$$(2.3.5) \quad \mathcal{J}_S^{-1}(c_1 v_1 + \cdots + c_k v_k) = c_1 e_1 + \cdots + c_k e_k.$$

We begin our demonstration that  $\dim V$  is well defined, with the following concrete result.

**Lemma 2.3.1.** *If  $v_1, \dots, v_{k+1}$  are vectors in  $\mathbb{F}^k$ , then they are linearly dependent.*

**Proof.** We use induction on  $k$ . The result is obvious if  $k = 1$ . We can suppose the last component of some  $v_j$  is nonzero, since otherwise we can regard these vectors as elements of  $\mathbb{F}^{k-1}$  and use the inductive hypothesis. Reordering these vectors, we

can assume the last component of  $v_{k+1}$  is nonzero, and it can be assumed to be 1. Form

$$w_j = v_j - v_{kj}v_{k+1}, \quad 1 \leq j \leq k,$$

where  $v_j = (v_{1j}, \dots, v_{kj})^t$ . Then the last component of each of the vectors  $w_1, \dots, w_k$  is 0, so we can regard these as  $k$  vectors in  $\mathbb{F}^{k-1}$ . By induction, there exist scalars  $a_1, \dots, a_k$ , not all zero, such that

$$a_1w_1 + \dots + a_kw_k = 0,$$

so we have

$$a_1v_1 + \dots + a_kv_k = (a_1v_{k1} + \dots + a_kv_{kk})v_{k+1},$$

the desired linear dependence relation on  $\{v_1, \dots, v_{k+1}\}$ .  $\square$

With this result in hand, we proceed.

**Proposition 2.3.2.** *If  $V$  has a basis  $\{v_1, \dots, v_k\}$  with  $k$  elements and  $\{w_1, \dots, w_\ell\} \subset V$  is linearly independent, then  $\ell \leq k$ .*

**Proof.** Take the isomorphism  $\mathcal{J}_S : \mathbb{F}^k \rightarrow V$  described in (2.3.2)–(2.3.3). The hypotheses imply that  $\{\mathcal{J}_S^{-1}w_1, \dots, \mathcal{J}_S^{-1}w_\ell\}$  is linearly independent in  $\mathbb{F}^k$ , so Lemma 2.3.1 implies  $\ell \leq k$ .  $\square$

**Corollary 2.3.3.** *If  $V$  is finite-dimensional, any two bases of  $V$  have the same number of elements. If  $V$  is isomorphic to  $W$ , these spaces have the same dimension.*

**Proof.** If  $S$  (with  $\#S$  elements) and  $T$  are bases of  $V$ , we have  $\#S \leq \#T$  and  $\#T \leq \#S$ , hence  $\#S = \#T$ . For the latter part, an isomorphism of  $V$  onto  $W$  takes a basis of  $V$  to a basis of  $W$ .  $\square$

The following is an easy but useful consequence.

**Proposition 2.3.4.** *If  $V$  is finite dimensional and  $W \subset V$  a linear subspace, then  $W$  has a finite basis, and  $\dim W \leq \dim V$ .*

**Proof.** Suppose  $\{w_1, \dots, w_\ell\}$  is a linearly independent subset of  $W$ . Proposition 2.3.2 implies  $\ell \leq \dim V$ . If this set spans  $W$ , we are done. If not, there is an element  $w_{\ell+1} \in W$  not in this span, and  $\{w_1, \dots, w_{\ell+1}\}$  is a linearly independent subset of  $W$ . Again  $\ell + 1 \leq \dim V$ . Continuing this process a finite number of times must produce a basis of  $W$ .  $\square$

A similar argument establishes:

**Proposition 2.3.5.** *Suppose  $V$  is finite dimensional,  $W \subset V$  a linear subspace, and  $\{w_1, \dots, w_\ell\}$  a basis of  $W$ . Then  $V$  has a basis of the form  $\{w_1, \dots, w_\ell, u_1, \dots, u_m\}$ , and  $\ell + m = \dim V$ .*

Having this, we can establish the following result, sometimes called the fundamental theorem of linear algebra.



**Proposition 2.3.6.** *Assume  $V$  and  $W$  are vector spaces,  $V$  finite dimensional, and*

$$(2.3.6) \quad A : V \longrightarrow W$$

*a linear map. Then*

$$(2.3.7) \quad \dim \mathcal{N}(A) + \dim \mathcal{R}(A) = \dim V.$$

**Proof.** Let  $\{w_1, \dots, w_\ell\}$  be a basis of  $\mathcal{N}(A) \subset V$ , and complete it to a basis

$$\{w_1, \dots, w_\ell, u_1, \dots, u_m\}$$

of  $V$ . Set  $L = \text{Span}\{u_1, \dots, u_m\}$ , and consider

$$(2.3.8) \quad A_0 : L \longrightarrow W, \quad A_0 = A|_L.$$

Clearly  $w \in \mathcal{R}(A) \Rightarrow w = A(a_1 w_1 + \dots + a_\ell w_\ell + b_1 u_1 + \dots + b_m u_m) = A_0(b_1 u_1 + \dots + b_m u_m)$ , so

$$(2.3.9) \quad \mathcal{R}(A_0) = \mathcal{R}(A).$$

Furthermore,

$$(2.3.10) \quad \mathcal{N}(A_0) = \mathcal{N}(A) \cap L = 0.$$

Hence  $A_0 : L \rightarrow \mathcal{R}(A_0)$  is an isomorphism. Thus  $\dim \mathcal{R}(A) = \dim \mathcal{R}(A_0) = \dim L = m$ , and we have (2.3.7).  $\square$

The following is a significant special case.

**Corollary 2.3.7.** *Let  $V$  be finite dimensional, and let  $A : V \rightarrow V$  be linear. Then*

$$(2.3.11) \quad A \text{ injective} \iff A \text{ surjective} \iff A \text{ isomorphism.}$$

We mention that these equivalences can fail for infinite dimensional spaces. For example, if  $\mathcal{P}$  denotes the space of polynomials in  $x$ , then  $M_x : \mathcal{P} \rightarrow \mathcal{P}$  ( $M_x f(x) = x f(x)$ ) is injective but not surjective, while  $D : \mathcal{P} \rightarrow \mathcal{P}$  ( $D f(x) = f'(x)$ ) is surjective but not injective.

Next we have the following important characterization of injectivity and surjectivity.

**Proposition 2.3.8.** *Assume  $V$  and  $W$  are finite dimensional and  $A : V \rightarrow W$  is linear. Then*

$$(2.3.12) \quad A \text{ surjective} \iff AB = I_W, \text{ for some } B \in \mathcal{L}(W, V),$$

*and*

$$(2.3.13) \quad A \text{ injective} \iff CA = I_V, \text{ for some } C \in \mathcal{L}(W, V).$$

**Proof.** Clearly  $AB = I \Rightarrow A$  surjective and  $CA = I \Rightarrow A$  injective. We establish the converses.

First assume  $A : V \rightarrow W$  is surjective. Let  $\{w_1, \dots, w_\ell\}$  be a basis of  $W$ . Pick  $v_j \in V$  such that  $Av_j = w_j$ . Set

$$(2.3.14) \quad B(a_1 w_1 + \dots + a_\ell w_\ell) = a_1 v_1 + \dots + a_\ell v_\ell.$$

This works in (2.3.12).

Next assume  $A : V \rightarrow W$  is injective. Let  $\{v_1, \dots, v_k\}$  be a basis of  $V$ . Set  $w_j = Av_j$ . Then  $\{w_1, \dots, w_k\}$  is linearly independent, hence a basis of  $\mathcal{R}(A)$ , and we then can produce a basis  $\{w_1, \dots, w_k, u_1, \dots, u_m\}$  of  $W$ . Set

$$(2.3.15) \quad C(a_1w_1 + \dots + a_kw_k + b_1u_1 + \dots + b_mu_m) = a_1v_1 + \dots + a_kv_k.$$

This works in (2.3.13).  $\square$

An  $m \times n$  matrix  $A$  defines a linear transformation  $A : \mathbb{F}^n \rightarrow \mathbb{F}^m$ , as in (2.2.3)–(2.2.4). The columns of  $A$  are

$$(2.3.16) \quad a_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{pmatrix}.$$

As seen in §2.2,

$$(2.3.17) \quad Ae_j = a_j,$$

where  $e_1, \dots, e_n$  is the standard basis of  $\mathbb{F}^n$ . Hence

$$(2.3.18) \quad \mathcal{R}(A) = \text{linear span of the columns of } A,$$

so

$$(2.3.19) \quad \mathcal{R}(A) = \mathbb{F}^m \iff a_1, \dots, a_n \text{ span } \mathbb{F}^m.$$

Furthermore,

$$(2.3.20) \quad A\left(\sum_{j=1}^n c_j e_j\right) = 0 \iff \sum_{j=1}^n c_j a_j = 0,$$

so

$$(2.3.21) \quad \mathcal{N}(A) = 0 \iff \{a_1, \dots, a_n\} \text{ is linearly independent.}$$

We have the following conclusion, in case  $m = n$ .

**Proposition 2.3.9.** *Let  $A$  be an  $n \times n$  matrix, defining  $A : \mathbb{F}^n \rightarrow \mathbb{F}^n$ . Then the following are equivalent:*

$$(2.3.22) \quad \begin{aligned} &A \text{ is invertible,} \\ &\text{The columns of } A \text{ are linearly independent,} \\ &\text{The columns of } A \text{ span } \mathbb{F}^n. \end{aligned}$$

## Exercises

1. Suppose  $\{v_1, \dots, v_k\}$  is a basis of  $V$ . Show that

$$w_1 = v_1, \quad w_2 = v_1 + v_2, \quad \dots, \quad w_j = v_1 + \dots + v_j, \quad \dots, \quad w_k = v_1 + \dots + v_k$$

is also a basis of  $V$ .

2. Let  $V$  be the space of polynomials in  $x$  and  $y$  of degree  $\leq 10$ . Specify a basis of  $V$  and compute  $\dim V$ .

3. Let  $V$  be the space of polynomials in  $x$  of degree  $\leq 5$ , satisfying  $p(-1) = p(0) = p(1) = 0$ . Find a basis of  $V$  and give its dimension.

4. Assume the existence and uniqueness result stated at the beginning of §1.17 in Chapter 1. Let  $a_j$  be continuous functions on an interval  $I$ , with  $a_n$  nowhere vanishing. Show that the space of functions  $x \in C^{(n)}(I)$  solving

$$a_n(t)x^{(n)}(t) + \cdots + a_1(t)x'(t) + a_0(t)x(t) = 0$$

is a vector space of dimension  $n$ .

5. Denote the space of  $m \times n$  matrices with entries in  $\mathbb{F}$  (as in (2.2.4)) by

$$(2.3.23) \quad M(m \times n, \mathbb{F}).$$

If  $m = n$ , denote it by

$$(2.3.24) \quad M(n, \mathbb{F}).$$

Show that

$$\dim M(m \times n, \mathbb{F}) = mn,$$

especially

$$\dim M(n, \mathbb{F}) = n^2.$$

6. If  $V$  and  $W$  are finite dimensional vector spaces,  $n = \dim V$ ,  $m = \dim W$ , what is  $\dim \mathcal{L}(V, W)$ ?

Let  $V$  be a finite dimensional vector space, with linear subspaces  $W$  and  $X$ . Recall the conditions under which  $V = W + X$  or  $V = W \oplus X$ , from §2.1. Let  $\{w_1, \dots, w_k\}$  be a basis of  $W$  and  $\{x_1, \dots, x_\ell\}$  a basis of  $X$ .

7. Show that

$$V = W + X \iff \{w_1, \dots, w_k, x_1, \dots, x_\ell\} \text{ spans } V$$

$$V = W \oplus X \iff \{w_1, \dots, w_k, x_1, \dots, x_\ell\} \text{ is a basis of } V.$$

8. Show that

$$V = W + X \implies \dim W + \dim X \geq \dim V,$$

$$V = W \oplus X \iff W \cap X = 0 \text{ and } \dim W + \dim X = \dim V.$$

9. Produce variants of Exercises 7–8 involving  $V = W_1 + \cdots + W_m$  and  $V = W_1 \oplus \cdots \oplus W_m$ , as in (2.1.19)–(2.1.20).

## 2.4. Matrix representation of a linear transformation

We show how a linear transformation

$$(2.4.1) \quad T : V \longrightarrow W$$

has a representation as an  $m \times n$  matrix, with respect to a basis  $S = \{v_1, \dots, v_n\}$  of  $V$  and a basis  $\Sigma = \{w_1, \dots, w_m\}$  of  $W$ . Namely, define  $a_{ij}$  by

$$(2.4.2) \quad Tv_j = \sum_{i=1}^m a_{ij}w_i, \quad 1 \leq j \leq n.$$

The matrix representation of  $T$  with respect to these bases is then

$$(2.4.3) \quad A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}.$$

Note that the  $j$ th column of  $A$  consists of the coefficients of  $Tv_j$ , when this is written as a linear combination of  $w_1, \dots, w_m$ . Compare (2.2.20).

If we want to record the dependence on the bases  $S$  and  $\Sigma$ , we can write

$$(2.4.4) \quad A = \mathcal{M}_{\Sigma}^{\Sigma}(T) = \mathcal{J}_{\Sigma}^{-1}T\mathcal{J}_S : \mathbb{F}^n \longrightarrow \mathbb{F}^m,$$

given the isomorphism  $\mathcal{J}_S : \mathbb{F}^n \rightarrow V$  as in (2.3.2)–(2.3.3) (with  $n$  instead of  $k$ ) and its counterpart  $\mathcal{J}_{\Sigma} : \mathbb{F}^m \rightarrow W$ , and with the identification of  $A$  with a matrix as in (2.2.3)–(2.2.4).

The definition of matrix multiplication is set up precisely so that, if  $X$  is a vector space with basis  $\Gamma = \{x_1, \dots, x_k\}$  and  $U : X \rightarrow V$  is linear, then  $TU : X \rightarrow W$  has matrix representation

$$(2.4.5) \quad \mathcal{M}_{\Sigma}^{\Sigma}(TU) = AB, \quad B = \mathcal{M}_{\Gamma}^S(U).$$

Indeed, if we complement (2.4.4) with

$$(2.4.6) \quad B = \mathcal{J}_S^{-1}U\mathcal{J}_{\Gamma} = \mathcal{M}_{\Gamma}^S(U),$$

we have

$$(2.4.7) \quad AB = \mathcal{J}_{\Sigma}^{-1}(TU)\mathcal{J}_{\Gamma}.$$

As for the representation of  $AB$  as a matrix product, see the discussion around (2.2.15)–(2.2.21).

For example, if

$$(2.4.8) \quad T : V \longrightarrow V,$$

and we use the basis  $S$  of  $V$  as above, we have an  $n \times n$  matrix  $\mathcal{M}_S^S(T)$ . If we pick another basis  $\tilde{S} = \{\tilde{v}_1, \dots, \tilde{v}_n\}$  of  $V$ , it follows from (2.4.5) that

$$(2.4.9) \quad \mathcal{M}_{\tilde{S}}^{\tilde{S}}(T) = \mathcal{M}_{\tilde{S}}^{\tilde{S}}(I)\mathcal{M}_S^S(T)\mathcal{M}_{\tilde{S}}^S(I).$$

Here

$$(2.4.10) \quad \mathcal{M}_{\tilde{S}}^S(I) = \mathcal{J}_S^{-1}\mathcal{J}_{\tilde{S}} = C = (c_{ij}),$$

where

$$(2.4.11) \quad \tilde{v}_j = \sum_{i=1}^n c_{ij} v_i, \quad 1 \leq j \leq n,$$

and we see (via (2.4.5)) that

$$(2.4.12) \quad \mathcal{M}_{\tilde{S}}^{\tilde{S}}(T) = C^{-1}.$$

To rewrite (2.4.9), we can say that if  $A$  is the matrix representation of  $T$  with respect to the basis  $S$  and  $\tilde{A}$  the matrix representation of  $T$  with respect to the basis  $\tilde{S}$ , then

$$(2.4.13) \quad \tilde{A} = C^{-1}AC.$$

REMARK. We say that  $n \times n$  matrices  $A$  and  $\tilde{A}$ , related as in (2.4.13), are *similar*.

EXAMPLE. Consider the linear transformation

$$(2.4.14) \quad D : \mathcal{P}_2 \longrightarrow \mathcal{P}_2, \quad Df(x) = f'(x).$$

With respect to the basis

$$(2.4.15) \quad v_1 = 1, \quad v_2 = x, \quad v_3 = x^2,$$

$D$  has the matrix representation

$$(2.4.16) \quad A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix},$$

since  $Dv_1 = 0$ ,  $Dv_2 = v_1$ , and  $Dv_3 = 2v_2$ . With respect to the basis

$$(2.4.17) \quad \tilde{v}_1 = 1, \quad \tilde{v}_2 = 1 + x, \quad \tilde{v}_3 = 1 + x + x^2,$$

$D$  has the matrix representation

$$(2.4.18) \quad \tilde{A} = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix},$$

since  $D\tilde{v}_1 = 0$ ,  $D\tilde{v}_2 = \tilde{v}_1$ , and  $D\tilde{v}_3 = 1 + 2x = 2\tilde{v}_2 - \tilde{v}_1$ . The reader is invited to verify (2.4.13) for this example.

---

**Exercises**

1. Consider  $\mathcal{T} : \mathcal{P}_2 \rightarrow \mathcal{P}_2$ , given by  $\mathcal{T}p(x) = x^{-1} \int_0^x p(y) dy$ . Compute the matrix representation  $B$  of  $\mathcal{T}$  with respect to the basis (2.4.15). Compute  $AB$  and  $BA$ , with  $A$  given by (2.4.16).

2. In the setting of Exercise 1, compute  $D\mathcal{T}$  and  $\mathcal{T}D$  on  $\mathcal{P}_2$  and compare their matrix representations, with respect to the basis (2.4.15), with  $AB$  and  $BA$ .

3. In the setting of Exercise 1, take  $a \in \mathbb{R}$  and define

$$(2.4.19) \quad \mathcal{T}_a p(x) = \frac{1}{x-a} \int_a^x p(y) dy, \quad \mathcal{T}_a : \mathcal{P}_2 \longrightarrow \mathcal{P}_2.$$

Compute the matrix representation of  $\mathcal{T}_a$  with respect to the basis (2.4.15).

4. Compute the matrix representation of  $\mathcal{T}_a$ , given by (2.4.19), with respect to the basis of  $\mathcal{P}_2$  given in (2.4.17).

5. Let  $A : \mathbb{C}^2 \rightarrow \mathbb{C}^2$  be given by

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

(with respect to the standard basis). Find a basis of  $\mathbb{C}^2$  with respect to which the matrix representation of  $A$  is

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

6. Let  $V = \{a \cos t + b \sin t : a, b \in \mathbb{C}\}$ , and consider

$$D = \frac{d}{dt} : V \longrightarrow V.$$

Compute the matrix representation of  $D$  with respect to the basis  $\{\cos t, \sin t\}$ .

7. In the setting of Exercise 6, compute the matrix representation of  $D$  with respect to the basis  $\{e^{it}, e^{-it}\}$ .

## 2.5. Determinants and invertibility

Determinants arise in the study of inverting a matrix. To take the  $2 \times 2$  case, solving for  $x$  and  $y$  the system

$$(2.5.1) \quad \begin{aligned} ax + by &= u, \\ cx + dy &= v \end{aligned}$$

can be done by multiplying these equations by  $d$  and  $b$ , respectively, and subtracting, and by multiplying them by  $c$  and  $a$ , respectively, and subtracting, yielding

$$(2.5.2) \quad \begin{aligned} (ad - bc)x &= du - bv, \\ (ad - bc)y &= av - cu. \end{aligned}$$

The factor on the left is

$$(2.5.3) \quad \det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc,$$

and solving (2.5.2) for  $x$  and  $y$  leads to

$$(2.5.4) \quad A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \implies A^{-1} = \frac{1}{\det A} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix},$$

provided  $\det A \neq 0$ .

We now consider determinants of  $n \times n$  matrices. Let  $M(n, \mathbb{F})$  denote the set of  $n \times n$  matrices with entries in  $\mathbb{F} = \mathbb{R}$  or  $\mathbb{C}$ . We write

$$(2.5.5) \quad A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} = (a_1, \dots, a_n),$$

where

$$(2.5.6) \quad a_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix}$$

is the  $j$ th column of  $A$ . The determinant is defined as follows.

**Proposition 2.5.1.** *There is a unique function*

$$(2.5.7) \quad \vartheta : M(n, \mathbb{F}) \longrightarrow \mathbb{F},$$

*satisfying the following three properties:*

- (a)  $\vartheta$  is linear in each column  $a_j$  of  $A$ ,
- (b)  $\vartheta(\tilde{A}) = -\vartheta(A)$  if  $\tilde{A}$  is obtained from  $A$  by interchanging two columns,
- (c)  $\vartheta(I) = 1$ .

*This defines the determinant:*

$$(2.5.8) \quad \vartheta(A) = \det A.$$

*If (c) is replaced by*

$$(c') \quad \vartheta(I) = r,$$

then

$$(2.5.9) \quad \vartheta(A) = r \det A.$$

The proof will involve constructing an explicit formula for  $\det A$  by following the rules (a)–(c). We start with the case  $n = 3$ . We have

$$(2.5.10) \quad \det A = \sum_{j=1}^3 a_{j1} \det(e_j, a_2, a_3),$$

by applying (a) to the first column of  $A$ ,  $a_1 = \sum_j a_{j1} e_j$ . Here and below,  $\{e_j : 1 \leq j \leq n\}$  denotes the standard basis of  $\mathbb{F}^n$ , so  $e_j$  has a 1 in the  $j$ th slot and 0s elsewhere. Applying (a) to the second and third columns gives

$$(2.5.11) \quad \begin{aligned} \det A &= \sum_{j,k=1}^3 a_{j1} a_{k2} \det(e_j, e_k, a_3) \\ &= \sum_{j,k,\ell=1}^3 a_{j1} a_{k2} a_{\ell3} \det(e_j, e_k, e_\ell). \end{aligned}$$

This is a sum of 27 terms, but most of them are 0. Note that rule (b) implies

$$(2.5.12) \quad \det B = 0 \text{ whenever } B \text{ has two identical columns.}$$

Hence  $\det(e_j, e_k, e_\ell) = 0$  unless  $j, k$ , and  $\ell$  are distinct, that is, unless  $(j, k, \ell)$  is a *permutation* of  $(1, 2, 3)$ . Now rule (c) says

$$(2.5.13) \quad \det(e_1, e_2, e_3) = 1,$$

and we see from rule (b) that  $\det(e_j, e_k, e_\ell) = 1$  if one can convert  $(e_j, e_k, e_\ell)$  to  $(e_1, e_2, e_3)$  by an even number of column interchanges, and  $\det(e_j, e_k, e_\ell) = -1$  if it takes an odd number of interchanges. Explicitly,

$$(2.5.14) \quad \begin{aligned} \det(e_1, e_2, e_3) &= 1, & \det(e_1, e_3, e_2) &= -1, \\ \det(e_2, e_3, e_1) &= 1, & \det(e_2, e_1, e_3) &= -1, \\ \det(e_3, e_1, e_2) &= 1, & \det(e_3, e_2, e_1) &= -1. \end{aligned}$$

Consequently (2.5.11) yields

$$(2.5.15) \quad \begin{aligned} \det A &= a_{11}a_{22}a_{33} - a_{11}a_{32}a_{23} \\ &\quad + a_{21}a_{32}a_{13} - a_{21}a_{12}a_{33} \\ &\quad + a_{31}a_{12}a_{23} - a_{31}a_{22}a_{13}. \end{aligned}$$

Note that the second indices occur in  $(1, 2, 3)$  order in each product. We can rearrange these products so that the *first* indices occur in  $(1, 2, 3)$  order:

$$(2.5.16) \quad \begin{aligned} \det A &= a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} \\ &\quad + a_{13}a_{21}a_{32} - a_{12}a_{21}a_{33} \\ &\quad + a_{12}a_{23}a_{31} - a_{13}a_{22}a_{31}. \end{aligned}$$



Now we tackle the case of general  $n$ . Parallel to (2.5.10)–(2.5.11), we have

$$\begin{aligned} \det A &= \sum_j a_{j1} \det(e_j, a_2, \dots, a_n) = \dots \\ (2.5.17) \quad &= \sum_{j_1, \dots, j_n} a_{j_1 1} \cdots a_{j_n n} \det(e_{j_1}, \dots, e_{j_n}), \end{aligned}$$

by applying rule (a) to each of the  $n$  columns of  $A$ . As before, (2.5.12) implies  $\det(e_{j_1}, \dots, e_{j_n}) = 0$  unless  $(j_1, \dots, j_n)$  are all distinct, that is, unless  $(j_1, \dots, j_n)$  is a permutation of the set  $(1, 2, \dots, n)$ . We set

$$(2.5.18) \quad S_n = \text{set of permutations of } (1, 2, \dots, n).$$

That is,  $S_n$  consists of elements  $\sigma$ , mapping the set  $\{1, \dots, n\}$  to itself,

$$(2.5.19) \quad \sigma : \{1, 2, \dots, n\} \longrightarrow \{1, 2, \dots, n\},$$

that are one-to-one and onto. We can compose two such permutations, obtaining the product  $\sigma\tau \in S_n$ , given  $\sigma$  and  $\tau$  in  $S_n$ . A permutation that interchanges just two elements of  $\{1, \dots, n\}$ , say  $j$  and  $k$  ( $j \neq k$ ), is called a *transposition*, and labeled  $(jk)$ . It is easy to see that each permutation of  $\{1, \dots, n\}$  can be achieved by successively transposing pairs of elements of this set. That is, each element  $\sigma \in S_n$  is a product of transpositions. We claim that

$$(2.5.20) \quad \det(e_{\sigma(1)1}, \dots, e_{\sigma(n)n}) = (\text{sgn } \sigma) \det(e_1, \dots, e_n) = \text{sgn } \sigma,$$

where

$$(2.5.21) \quad \begin{aligned} \text{sgn } \sigma &= 1 && \text{if } \sigma \text{ is a product of an even number of transpositions,} \\ &= -1 && \text{if } \sigma \text{ is a product of an odd number of transpositions.} \end{aligned}$$

In fact, the first identity in (2.5.20) follows from rule (b) and the second identity from rule (c).

There is one point to be checked here. Namely, we claim that a given  $\sigma \in S_n$  cannot simultaneously be written as a product of an even number of transpositions and an odd number of transpositions. If  $\sigma$  could be so written,  $\text{sgn } \sigma$  would not be well defined, and it would be impossible to satisfy condition (b), so Proposition 2.5.1 would fail. One neat way to see that  $\text{sgn } \sigma$  is well defined is the following. Let  $\sigma \in S_n$  act on functions of  $n$  variables by

$$(2.5.22) \quad (\sigma f)(x_1, \dots, x_n) = f(x_{\sigma(1)}, \dots, x_{\sigma(n)}).$$

It is readily verified that if also  $\tau \in S_n$ ,

$$(2.5.23) \quad g = \sigma f \implies \tau g = (\tau\sigma)f.$$

Now, let  $P$  be the polynomial

$$(2.5.24) \quad P(x_1, \dots, x_n) = \prod_{1 \leq j < k \leq n} (x_j - x_k).$$

One readily has

$$(2.5.25) \quad (\sigma P)(x) = -P(x), \quad \text{whenever } \sigma \text{ is a transposition,}$$

and hence, by (2.5.23),

$$(2.5.26) \quad (\sigma P)(x) = (\text{sgn } \sigma)P(x), \quad \forall \sigma \in S_n,$$

and  $\text{sgn } \sigma$  is well defined.

The proof of (2.5.20) is complete, and substitution into (2.5.17) yields the formula

$$(2.5.27) \quad \det A = \sum_{\sigma \in S_n} (\operatorname{sgn} \sigma) a_{\sigma(1)1} \cdots a_{\sigma(n)n}.$$

It is routine to check that this satisfies the properties (a)–(c). Regarding (b), note that if  $\vartheta(A)$  denotes the right side of (2.5.27) and  $\tilde{A}$  is obtained from  $A$  by applying a permutation  $\tau$  to the columns of  $A$ , so  $\tilde{A} = (a_{\tau(1)}, \dots, a_{\tau(n)})$ , then

$$(2.5.28) \quad \begin{aligned} \vartheta(\tilde{A}) &= \sum_{\sigma \in S_n} (\operatorname{sgn} \sigma) a_{\sigma(1)\tau(1)} \cdots a_{\sigma(n)\tau(n)} \\ &= \sum_{\sigma \in S_n} (\operatorname{sgn} \sigma) a_{\sigma\tau^{-1}(1)1} \cdots a_{\sigma\tau^{-1}(n)n} \\ &= \sum_{\omega \in S_n} (\operatorname{sgn} \omega\tau) a_{\omega(1)1} \cdots a_{\omega(n)n} \\ &= (\operatorname{sgn} \tau) \vartheta(A), \end{aligned}$$

the last identity because

$$(2.5.29) \quad \operatorname{sgn} \omega\tau = (\operatorname{sgn} \omega)(\operatorname{sgn} \tau), \quad \forall \omega, \tau \in S_n.$$

As for the final part of Proposition 2.5.1, if (c) is replaced by (c'), then (2.5.20) is replaced by

$$(2.5.30) \quad \vartheta(e_{\sigma(1)}, \dots, e_{\sigma(n)}) = r(\operatorname{sgn} \sigma),$$

and (2.5.9) follows.

REMARK. The formula (2.5.27) is taken as a definition of the determinant by some authors. While it is a useful *formula* for the determinant, it is a bad *definition*, which has perhaps led to a bit of fear and loathing among math students.

REMARK. Here is another formula for  $\operatorname{sgn} \sigma$ , which the reader is invited to verify. If  $\sigma \in S_n$ ,

$$(2.5.31) \quad \operatorname{sgn} \sigma = (-1)^{\kappa(\sigma)},$$

where

$$(2.5.32) \quad \begin{aligned} \kappa(\sigma) &= \text{number of pairs } (j, k) \text{ such that } 1 \leq j < k \leq n, \\ &\text{but } \sigma(j) > \sigma(k). \end{aligned}$$

Note that

$$(2.5.33) \quad a_{\sigma(1)1} \cdots a_{\sigma(n)n} = a_{1\tau(1)} \cdots a_{n\tau(n)}, \quad \text{with } \tau = \sigma^{-1},$$

and  $\operatorname{sgn} \sigma = \operatorname{sgn} \sigma^{-1}$ , so, parallel to (2.5.16), we also have

$$(2.5.34) \quad \det A = \sum_{\sigma \in S_n} (\operatorname{sgn} \sigma) a_{1\sigma(1)} \cdots a_{n\sigma(n)}.$$

Comparison with (2.5.27) gives

$$(2.5.35) \quad \det A = \det A^t,$$

where  $A = (a_{jk}) \Rightarrow A^t = (a_{kj})$ . Note that the  $j$ th column of  $A^t$  has the same entries as the  $j$ th row of  $A$ . In light of this, we have:

**Corollary 2.5.2.** *In Proposition 2.5.1, one can replace “columns” by “rows.”*

The following is a key property of the determinant.

**Proposition 2.5.3.** *Given  $A$  and  $B$  in  $M(n, \mathbb{F})$ ,*

$$(2.5.36) \quad \det(AB) = (\det A)(\det B).$$

**Proof.** For fixed  $A$ , apply Proposition 2.5.1 to

$$(2.5.37) \quad \vartheta_1(B) = \det(AB).$$

If  $B = (b_1, \dots, b_n)$ , with  $j$ th column  $b_j$ , then

$$(2.5.38) \quad AB = (Ab_1, \dots, Ab_n).$$

Clearly rule (a) holds for  $\vartheta_1$ . Also, if  $\tilde{B} = (b_{\sigma(1)}, \dots, b_{\sigma(n)})$  is obtained from  $B$  by permuting its columns, then  $A\tilde{B}$  has columns  $(Ab_{\sigma(1)}, \dots, Ab_{\sigma(n)})$ , obtained by permuting the columns of  $AB$  in the same fashion. Hence rule (b) holds for  $\vartheta_1$ . Finally, rule (c') holds for  $\vartheta_1$ , with  $r = \det A$ , and (2.5.36) follows.  $\square$

**Corollary 2.5.4.** *If  $A \in M(n, \mathbb{F})$  is invertible, then  $\det A \neq 0$ .*

**Proof.** If  $A$  is invertible, there exists  $B \in M(n, \mathbb{F})$  such that  $AB = I$ . Then, by (2.5.36),  $(\det A)(\det B) = 1$ , so  $\det A \neq 0$ .  $\square$

The converse of Corollary 2.5.4 also holds. Before proving it, it is convenient to show that the determinant is invariant under a certain class of column operations, given as follows.

**Proposition 2.5.5.** *If  $\tilde{A}$  is obtained from  $A = (a_1, \dots, a_n) \in M(n, \mathbb{F})$  by adding  $ca_\ell$  to  $a_k$  for some  $c \in \mathbb{F}$ ,  $\ell \neq k$ , then*

$$(2.5.39) \quad \det \tilde{A} = \det A.$$

**Proof.** By rule (a),  $\det \tilde{A} = \det A + c \det A^b$ , where  $A^b$  is obtained from  $A$  by replacing the column  $a_k$  by  $a_\ell$ . Hence  $A^b$  has two identical columns, so  $\det A^b = 0$ , and (2.5.39) holds.  $\square$

We now extend Corollary 2.5.4.

**Proposition 2.5.6.** *If  $A \in M(n, \mathbb{F})$ , then  $A$  is invertible if and only if  $\det A \neq 0$ .*

**Proof.** We have half of this from Corollary 2.5.4. To finish, assume  $A$  is not invertible. As seen in §3, this implies the columns  $a_1, \dots, a_n$  of  $A$  are linearly dependent. Hence, for some  $k$ ,

$$(2.5.40) \quad a_k + \sum_{\ell \neq k} c_\ell a_\ell = 0,$$

with  $c_\ell \in \mathbb{F}$ . Now we can apply Proposition 2.5.5 to obtain  $\det A = \det \tilde{A}$ , where  $\tilde{A}$  is obtained by adding  $\sum c_\ell a_\ell$  to  $a_k$ . But then the  $k$ th column of  $\tilde{A}$  is 0, so  $\det A = \det \tilde{A} = 0$ . This finishes the proof of Proposition 2.5.6.  $\square$

Further useful facts about determinants arise in the following exercises.

---

### Exercises

1. Show that

$$(2.5.41) \quad \det \begin{pmatrix} 1 & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2} & \cdots & a_{nn} \end{pmatrix} = \det \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2} & \cdots & a_{nn} \end{pmatrix} = \det A_{11}$$

where  $A_{11} = (a_{jk})_{2 \leq j, k \leq n}$ .

*Hint.* Do the first identity using Proposition 2.5.5. Then exploit uniqueness for  $\det$  on  $M(n-1, \mathbb{F})$ .

2. Deduce that  $\det(e_j, a_2, \dots, a_n) = (-1)^{j-1} \det A_{1j}$  where  $A_{kj}$  is formed by deleting the  $k$ th column and the  $j$ th row from  $A$ .

3. Deduce from the first sum in (2.5.17) that

$$(2.5.42) \quad \det A = \sum_{j=1}^n (-1)^{j-1} a_{j1} \det A_{1j}.$$

More generally, for any  $k \in \{1, \dots, n\}$ ,

$$(2.5.43) \quad \det A = \sum_{j=1}^n (-1)^{j-k} a_{jk} \det A_{kj}.$$

This is called an expansion of  $\det A$  by minors, down the  $k$ th column.

4. Let  $c_{kj} = (-1)^{j-k} \det A_{kj}$ . Show that

$$(2.5.44) \quad \sum_{j=1}^n a_{j\ell} c_{kj} = 0, \quad \text{if } \ell \neq k.$$

Deduce from this and (2.5.43) that  $C = (c_{jk})$  satisfies

$$(2.5.45) \quad CA = (\det A)I.$$

*Hint.* Reason as in Exercises 1–3 that the left side of (2.5.44) is equal to

$$\det(a_1, \dots, a_\ell, \dots, a_\ell, \dots, a_n),$$

with  $a_\ell$  in the  $k$ th column as well as in the  $\ell$ th column. The identity (2.5.45) is known as Cramer's formula. Note how this generalizes (2.5.4).

5. Show that

$$(2.5.46) \quad \det \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ & a_{22} & \cdots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{pmatrix} = a_{11}a_{22}\cdots a_{nn}.$$

*Hint.* Use (2.5.41) and induction. *Alternative:* Use (2.5.27). Show that  $\sigma \in S_n$ ,  $\sigma(k) \leq k \forall k \Rightarrow \sigma(k) \equiv k$ .

The next two exercises deal with the determinant of a linear transformation. Let  $V$  be an  $n$ -dimensional vector space, and

$$(2.5.47) \quad T : V \longrightarrow V$$

a linear transformation. We would like to define

$$(2.5.48) \quad \det T = \det A,$$

where  $A = \mathcal{M}_S^S(T)$  for some basis  $S = \{v_1, \dots, v_n\}$  of  $V$ .

6. Suppose  $\tilde{S} = \{\tilde{v}_1, \dots, \tilde{v}_n\}$  is another basis of  $V$ . Show that

$$\det A = \det \tilde{A},$$

where  $\tilde{A} = \mathcal{M}_{\tilde{S}}^{\tilde{S}}(T)$ . Hence (2.5.48) defines  $\det T$ , independently of the choice of basis of  $V$ .

*Hint.* Use (2.4.13) and (2.5.36).

7. If also  $U \in \mathcal{L}(V)$ , show that

$$\det(UT) = (\det U)(\det T).$$

### Row reduction, matrix products, and Gaussian elimination

In Exercises 8–13, we consider the following three types of row operations on an  $n \times n$  matrix  $A = (a_{jk})$ . If  $\sigma$  is a permutation of  $\{1, \dots, n\}$ , let

$$(2.5.49) \quad \rho_\sigma(A) = (a_{\sigma(j)k}).$$

If  $c = (c_1, \dots, c_n)$ , and all  $c_j$  are nonzero, set

$$(2.5.50) \quad \mu_c(A) = (c_j^{-1}a_{jk}).$$

Finally, if  $c \in \mathbb{F}$  and  $\mu \neq \nu$ , define

$$(2.5.51) \quad \varepsilon_{\mu\nu c}(A) = (b_{jk}), \quad b_{\nu k} = a_{\nu k} - ca_{\mu k}, \quad b_{jk} = a_{jk} \quad \text{for } j \neq \nu.$$

Note that a major part of this section dealt with the effect of such row operations on the determinant of a matrix. More precisely, they directly dealt with column operations, but as remarked after (2.5.35), one has analogues for row operations.

We want to relate these operations to left multiplication by matrices  $P_\sigma, M_c$ , and  $E_{\mu\nu c}$ , defined by the following actions on the standard basis  $\{e_1, \dots, e_n\}$  of  $\mathbb{F}^n$ :

$$(2.5.52) \quad P_\sigma e_j = e_{\sigma(j)}, \quad M_c e_j = c_j e_j,$$

and

$$(2.5.53) \quad E_{\mu\nu c} e_\mu = e_\mu + c e_\nu, \quad E_{\mu\nu c} e_j = e_j \text{ for } j \neq \mu.$$

These relations are established in the following exercises.

8. Show that

$$(2.5.54) \quad A = P_\sigma \rho_\sigma(A), \quad A = M_c \mu_c(A), \quad A = E_{\mu\nu c} \varepsilon_{\mu\nu c}(A).$$

9. Show that  $P_\sigma^{-1} = P_{\sigma^{-1}}$ .

10. Show that, if  $\mu \neq \nu$ , then  $E_{\mu\nu c} = P_\sigma^{-1} E_{21c} P_\sigma$ , for some permutation  $\sigma$ .

11. If  $B = \rho_\sigma(A)$  and  $C = \mu_c(B)$ , show that  $A = P_\sigma M_c C$ . Generalize this to other cases where a matrix  $C$  is obtained from a matrix  $A$  via a sequence of row operations.

12. If  $A$  is an invertible  $n \times n$  matrix, with entries in  $\mathbb{F} = \mathbb{R}$  or  $\mathbb{C}$  (we write  $A \in \text{Gl}(n, \mathbb{F})$ ), then the rows of  $A$  form a basis of  $\mathbb{F}^n$ . Use this to show that  $A$  can be transformed to the identity matrix via a sequence of row operations. Deduce that any  $A \in \text{Gl}(n, \mathbb{F})$  can be written as a finite product of matrices of the form  $P_\sigma, M_c$  and  $E_{\mu\nu c}$ .

13. Suppose  $A$  is an invertible  $n \times n$  matrix, and a sequence of row operations is applied to  $A$ , transforming it to the identity matrix  $I$ . Show that the *same* sequence of row operations, applied to  $I$ , transforms it to  $A^{-1}$ . This method of constructing  $A^{-1}$  is called the method of Gaussian elimination.

EXAMPLE. We take a  $2 \times 2$  matrix  $A$ , write  $A$  and  $I$  side by side, and perform the same sequence of row operations on each of these two matrices, obtaining finally  $I$  and  $A^{-1}$  side by side.

$$A = \begin{pmatrix} 1 & 2 \\ 1 & 3 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 3 & -2 \\ -1 & 1 \end{pmatrix} = A^{-1}.$$

*Hint.* Turning around (2.5.54), we have

$$(2.5.55) \quad \rho_\sigma(A) = P_\sigma^{-1} A, \quad \mu_c(A) = M_c^{-1} A, \quad \varepsilon_{\mu\nu c}(A) = E_{\mu\nu c}^{-1} A.$$

Thus applying a sequence of row operations to  $A$  yields

$$(2.5.56) \quad S_k^{-1} \cdots S_1^{-1} A,$$

where each  $S_j$  is of the form (2.5.52) or (2.5.53). If (2.5.56) is the identity matrix, then

$$(2.5.57) \quad A^{-1} = S_k^{-1} \cdots S_1^{-1}.$$

REMARK. The method of Gaussian elimination is computationally superior to the use of Cramer's formula (2.5.45) for computing matrix inverses, though Cramer's formula has theoretical interest.

A related issue is that, for computing determinants of  $n \times n$  matrices, for  $n \geq 3$ , it is computationally superior to utilize a sequence of column operations, applying rules (a) and (b) and Proposition 2.5.5 (and/or the corresponding row operations), rather than directly using the formula (2.5.27), which contains  $n!$  terms. This "Gaussian elimination" method of calculating  $\det A$  gives, from (2.5.55)–(2.5.56),

$$(2.5.58) \quad \det A = (\det S_1) \cdots (\det S_k),$$

with

$$(2.5.59) \quad \det P_\sigma = \operatorname{sgn} \sigma, \quad \det M_c = c_1 \cdots c_n, \quad \det E_{\mu\nu c} = 1.$$

## 2.6. Eigenvalues and eigenvectors

Let  $T : V \rightarrow V$  be linear. If there is a nonzero  $v \in V$  such that

$$(2.6.1) \quad Tv = \lambda_j v,$$

for some  $\lambda_j \in \mathbb{F}$ , we say  $\lambda_j$  is an eigenvalue of  $T$ , and  $v$  is an eigenvector. Let  $\mathcal{E}(T, \lambda_j)$  denote the set of vectors  $v \in V$  such that (2.6.1) holds. It is clear that  $\mathcal{E}(T, \lambda_j)$  is a linear subspace of  $V$  and

$$(2.6.2) \quad T : \mathcal{E}(T, \lambda_j) \longrightarrow \mathcal{E}(T, \lambda_j).$$

The set of  $\lambda_j \in \mathbb{F}$  such that  $\mathcal{E}(T, \lambda_j) \neq 0$  is denoted  $\text{Spec}(T)$ . Clearly  $\lambda_j \in \text{Spec}(T)$  if and only if  $T - \lambda_j I$  is not injective, so, if  $V$  is finite dimensional,

$$(2.6.3) \quad \lambda_j \in \text{Spec}(T) \iff \det(\lambda_j I - T) = 0.$$

We call  $K_T(\lambda) = \det(\lambda I - T)$  the *characteristic polynomial* of  $T$ .

If  $\mathbb{F} = \mathbb{C}$ , we can use the *fundamental theorem of algebra*, which says every non-constant polynomial with complex coefficients has at least one complex root. (See Appendix 2.C for a proof of this result.) This proves the following.

**Proposition 2.6.1.** *If  $V$  is a finite-dimensional complex vector space and  $T \in \mathcal{L}(V)$ , then  $T$  has at least one eigenvector in  $V$ .*

REMARK. If  $V$  is real and  $K_T(\lambda)$  does have a real root  $\lambda_j$ , then there is a real  $\lambda_j$ -eigenvector.

Sometimes a linear transformation has only one eigenvector, up to a scalar multiple. Consider the transformation  $A : \mathbb{C}^3 \rightarrow \mathbb{C}^3$  given by

$$(2.6.4) \quad A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}.$$

We see that  $\det(\lambda I - A) = (\lambda - 2)^3$ , so  $\lambda = 2$  is a triple root. It is clear that

$$(2.6.5) \quad \mathcal{E}(A, 2) = \text{Span}\{e_1\},$$

where  $e_1 = (1, 0, 0)^t$  is the first standard basis vector of  $\mathbb{C}^3$ .

If one is given  $T \in \mathcal{L}(V)$ , it is of interest to know whether  $V$  has a basis of eigenvectors of  $T$ . The following result is useful.

**Proposition 2.6.2.** *Assume that the characteristic polynomial of  $T \in \mathcal{L}(V)$  has  $k$  distinct roots,  $\lambda_1, \dots, \lambda_k$ , with eigenvectors  $v_j \in \mathcal{E}(T, \lambda_j)$ ,  $1 \leq j \leq k$ . Then  $\{v_1, \dots, v_k\}$  is linearly independent. In particular, if  $k = \dim V$ , these vectors form a basis of  $V$ .*

**Proof.** We argue by contradiction. If  $\{v_1, \dots, v_k\}$  is linearly dependent, take a minimal subset that is linearly dependent and (reordering if necessary) say this set is  $\{v_1, \dots, v_m\}$ , with  $Tv_j = \lambda_j v_j$ , and

$$(2.6.6) \quad c_1 v_1 + \dots + c_m v_m = 0,$$



with  $c_j \neq 0$  for each  $j \in \{1, \dots, m\}$ . Applying  $T - \lambda_m I$  to (2.6.6) gives

$$(2.6.7) \quad c_1(\lambda_1 - \lambda_m)v_1 + \cdots + c_{m-1}(\lambda_{m-1} - \lambda_m)v_{m-1} = 0,$$

a linear dependence relation on the smaller set  $\{v_1, \dots, v_{m-1}\}$ . This contradiction proves the proposition.  $\square$

Further information on when  $T \in \mathcal{L}(V)$  yields a basis of eigenvectors, and on what one can say when it does not, will be given in the following sections.

---

## Exercises

1. Compute the eigenvalues and eigenvectors of each of the following matrices.

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \\ \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix}, \quad \begin{pmatrix} i & i \\ 0 & 1 \end{pmatrix}.$$

In which cases does  $\mathbb{C}^2$  have a basis of eigenvectors?

2. Compute the eigenvalues and eigenvectors of each of the following matrices.

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

3. Let  $A \in M(n, \mathbb{C})$ . We say  $A$  is diagonalizable if and only if there exists an invertible  $B \in M(n, \mathbb{C})$  such that  $B^{-1}AB$  is diagonal:

$$B^{-1}AB = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}.$$

Show that  $A$  is diagonalizable if and only if  $\mathbb{C}^n$  has a basis of eigenvectors of  $A$ . Recall from (2.4.13) that the matrices  $A$  and  $B^{-1}AB$  are said to be similar.

4. More generally, if  $V$  is an  $n$ -dimensional complex vector space, we say  $T \in \mathcal{L}(V)$  is diagonalisable if and only if there exists invertible  $B : \mathbb{C}^n \rightarrow V$  such that  $B^{-1}TB$  is diagonal, with respect to the standard basis of  $\mathbb{C}^n$ . Formulate and establish the natural analogue of Exercise 3.

5. In the setting of (2.6.1)–(2.6.2), given  $S \in \mathcal{L}(V, V)$ , show that

$$ST = TS \implies S : \mathcal{E}(T, \lambda_j) \rightarrow \mathcal{E}(T, \lambda_j).$$

## 2.7. Generalized eigenvectors and the minimal polynomial

As we have seen, the matrix

$$(2.7.1) \quad A = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{pmatrix}$$

has only one eigenvalue, 2, and, up to a scalar multiple, just one eigenvector,  $e_1$ . However, we have

$$(2.7.2) \quad (A - 2I)^2 e_2 = 0, \quad (A - 2I)^3 e_3 = 0.$$

Generally, if  $T \in \mathcal{L}(V)$ , we say a nonzero  $v \in V$  is a generalized  $\lambda_j$ -eigenvector if there exists  $k \in \mathbb{N}$  such that

$$(2.7.3) \quad (T - \lambda_j I)^k v = 0.$$

We denote by  $\mathcal{GE}(T, \lambda_j)$  the set of vectors  $v \in V$  such that (2.7.3) holds, for some  $k$ . It is clear that  $\mathcal{GE}(T, \lambda_j)$  is a linear subspace of  $V$  and

$$(2.7.4) \quad T : \mathcal{GE}(T, \lambda_j) \longrightarrow \mathcal{GE}(T, \lambda_j).$$

The following is a useful comment.

**Lemma 2.7.1.** *For each  $\lambda_j \in \mathbb{F}$  such that  $\mathcal{GE}(T, \lambda_j) \neq 0$ ,*

$$(2.7.5) \quad T - \mu I : \mathcal{GE}(T, \lambda_j) \longrightarrow \mathcal{GE}(T, \lambda_j) \text{ is an isomorphism, } \forall \mu \neq \lambda_j.$$

**Proof.** If  $T - \mu I$  is not an isomorphism in (2.7.5), then  $Tv = \mu v$  for some nonzero  $v \in \mathcal{GE}(T, \lambda_j)$ . But then  $(T - \lambda_j I)^k v = (\mu - \lambda_j)^k v$  for all  $k \in \mathbb{N}$ , and hence this cannot ever be zero, unless  $\mu = \lambda_j$ .  $\square$

Note that if  $V$  is a finite-dimensional complex vector space, then each nonzero space appearing in (2.7.4) contains an eigenvector, by Proposition 2.6.1. Clearly the corresponding eigenvalue must be  $\lambda_j$ . In particular, the set of  $\lambda_j$  for which  $\mathcal{GE}(T, \lambda_j)$  is nonzero coincides with  $\text{Spec}(T)$ , as given in (2.6.3).

We intend to show that if  $V$  is a finite-dimensional complex vector space and  $T \in \mathcal{L}(V)$ , then  $V$  is spanned by generalized eigenvectors of  $T$ . One tool in this demonstration will be the construction of polynomials  $p(\lambda)$  such that  $p(T) = 0$ . Here, if

$$(2.7.6) \quad p(\lambda) = a_n \lambda^n + a_{n-1} \lambda^{n-1} + \cdots + a_1 \lambda + a_0,$$

then

$$(2.7.7) \quad p(T) = a_n T^n + a_{n-1} T^{n-1} + \cdots + a_1 T + a_0 I.$$

Let us denote by  $\mathcal{P}$  the space of polynomials in  $\lambda$ .

**Lemma 2.7.2.** *If  $V$  is finite dimensional and  $T \in \mathcal{L}(V)$ , then there exists a nonzero  $p \in \mathcal{P}$  such that  $p(T) = 0$ .*

**Proof.** If  $\dim V = n$ , then  $\dim \mathcal{L}(V) = n^2$ , so  $\{I, T, \dots, T^{n^2}\}$  is linearly dependent.  $\square$

Let us set

$$(2.7.8) \quad \mathcal{I}_T = \{p \in \mathcal{P} : p(T) = 0\}.$$

We see that  $\mathcal{I} = \mathcal{I}_T$  has the following properties:

$$(2.7.9) \quad \begin{aligned} p, q \in \mathcal{I} &\implies p + q \in \mathcal{I}, \\ p \in \mathcal{I}, q \in \mathcal{P} &\implies pq \in \mathcal{I}. \end{aligned}$$

A set  $\mathcal{I} \subset \mathcal{P}$  satisfying (2.7.9) is called an *ideal*. Here is another construction of a class of ideals in  $\mathcal{P}$ . Given  $\{p_1, \dots, p_k\} \subset \mathcal{P}$ , set

$$(2.7.10) \quad \mathcal{I}(p_1, \dots, p_k) = \{p_1q_1 + \dots + p_kq_k : q_j \in \mathcal{P}\}.$$

We will find it very useful to know that all nonzero ideals in  $\mathcal{P}$ , including  $\mathcal{I}_T$ , have the following property.

**Lemma 2.7.3.** *Let  $\mathcal{I} \subset \mathcal{P}$  be a nonzero ideal, and let  $p_1 \in \mathcal{I}$  have minimal degree amongst all nonzero elements of  $\mathcal{I}$ . Then*

$$(2.7.11) \quad \mathcal{I} = \mathcal{I}(p_1).$$

**Proof.** Take any  $p \in \mathcal{I}$ . We divide  $p(\lambda)$  into  $p_1(\lambda)$  and take the remainder, obtaining

$$(2.7.12) \quad p(\lambda) = q(\lambda)p_1(\lambda) + r(\lambda).$$

Here  $q, r \in \mathcal{P}$ , hence  $r \in \mathcal{I}$ . Also  $r(\lambda)$  has degree less than the degree of  $p_1(\lambda)$ , so by minimality we have  $r \equiv 0$ . This shows  $p \in \mathcal{I}(p_1)$ , and we have (2.7.11).  $\square$

Applying this to  $\mathcal{I}_T$ , we denote by  $m_T(\lambda)$  the polynomial of smallest degree in  $\mathcal{I}_T$  (having leading coefficient 1), and say

$$(2.7.13) \quad m_T(\lambda) \text{ is the minimal polynomial of } T.$$

Thus every  $p \in \mathcal{P}$  such that  $p(T) = 0$  is a multiple of  $m_T(\lambda)$ .

Assuming  $V$  is a *complex* vector space of dimension  $n$ , we can apply the fundamental theorem of algebra to write

$$(2.7.14) \quad m_T(\lambda) = \prod_{j=1}^K (\lambda - \lambda_j)^{k_j},$$

with distinct roots  $\lambda_1, \dots, \lambda_K$ . The following polynomials will also play a role in our study of the generalized eigenspaces of  $T$ . For each  $\ell \in \{1, \dots, K\}$ , set

$$(2.7.15) \quad p_\ell(\lambda) = \prod_{j \neq \ell} (\lambda - \lambda_j)^{k_j} = \frac{m_T(\lambda)}{(\lambda - \lambda_\ell)^{k_\ell}}.$$

We have the following useful result.

**Proposition 2.7.4.** *If  $V$  is an  $n$ -dimensional complex vector space and  $T \in \mathcal{L}(V)$ , then, for each  $\ell \in \{1, \dots, K\}$ ,*

$$(2.7.16) \quad \mathcal{GE}(T, \lambda_\ell) = \mathcal{R}(p_\ell(T)).$$

**Proof.** Given  $v \in V$ ,

$$(2.7.17) \quad (T - \lambda_\ell)^{k_\ell} p_\ell(T)v = m_T(T)v = 0,$$

so  $p_\ell(T) : V \rightarrow \mathcal{GE}(T, \lambda_\ell)$ . Furthermore, each factor

$$(2.7.18) \quad (T - \lambda_j)^{k_j} : \mathcal{GE}(T, \lambda_\ell) \longrightarrow \mathcal{GE}(T, \lambda_\ell), \quad j \neq \ell,$$

in  $p_\ell(T)$  is an isomorphism, by Lemma 2.7.1, so  $p_\ell(T) : \mathcal{GE}(T, \lambda_\ell) \rightarrow \mathcal{GE}(T, \lambda_\ell)$  is an isomorphism.  $\square$

REMARK. We hence see that each  $\lambda_j$  appearing in (2.7.14) is an element of  $\text{Spec } T$ .

We now establish the following spanning property.

**Proposition 2.7.5.** *If  $V$  is an  $n$ -dimensional complex vector space and  $T \in \mathcal{L}(V)$ , then*

$$(2.7.19) \quad V = \mathcal{GE}(T, \lambda_1) + \cdots + \mathcal{GE}(T, \lambda_K).$$

*That is, each  $v \in V$  can be written as  $v = v_1 + \cdots + v_K$ , with  $v_j \in \mathcal{GE}(T, \lambda_j)$ .*

**Proof.** Let  $m_T(\lambda)$  be the minimal polynomial of  $T$ , with the factorization (2.7.14), and define  $p_\ell(\lambda)$  as in (2.7.15), for  $\ell = 1, \dots, K$ . We claim that

$$(2.7.20) \quad \mathcal{I}(p_1, \dots, p_K) = \mathcal{P}.$$

In fact we know from Lemma 2.7.3 that  $\mathcal{I}(p_1, \dots, p_K) = \mathcal{I}(p_0)$  for some  $p_0 \in \mathcal{P}$ . Then any root of  $p_0(\lambda)$  must be a root of each  $p_\ell(\lambda)$ ,  $1 \leq \ell \leq K$ . But these polynomials are constructed so that no  $\mu \in \mathbb{C}$  is a root of all  $K$  of them. Hence  $p_0(\lambda)$  has no root so (again by the fundamental theorem of algebra) it must be constant, i.e.,  $1 \in \mathcal{I}(p_1, \dots, p_K)$ , which gives (2.7.20), and in particular we have that there exist  $q_\ell \in \mathcal{P}$  such that

$$(2.7.21) \quad p_1(\lambda)q_1(\lambda) + \cdots + p_K(\lambda)q_K(\lambda) = 1.$$

We use this as follows to write an arbitrary  $v \in V$  as a linear combination of generalized eigenvectors. Replacing  $\lambda$  by  $T$  in (2.7.21) gives

$$(2.7.22) \quad p_1(T)q_1(T) + \cdots + p_K(T)q_K(T) = I.$$

Hence, for any given  $v \in V$ ,

$$(2.7.23) \quad v = p_1(T)q_1(T)v + \cdots + p_K(T)q_K(T)v = v_1 + \cdots + v_K,$$

with  $v_\ell = p_\ell(T)q_\ell(T)v \in \mathcal{GE}(T, \lambda_\ell)$ , by Proposition 2.7.4.  $\square$

We next produce a basis consisting of generalized eigenvectors.

**Proposition 2.7.6.** *Under the hypotheses of Proposition 2.7.5, let  $\mathcal{GE}(T, \lambda_\ell)$ ,  $1 \leq \ell \leq K$ , denote the generalized eigenspaces of  $T$  (with  $\lambda_\ell$  mutually distinct), and let*

$$(2.7.24) \quad S_\ell = \{v_{\ell 1}, \dots, v_{\ell, d_\ell}\}, \quad d_\ell = \dim \mathcal{GE}(T, \lambda_\ell),$$

*be a basis of  $\mathcal{GE}(T, \lambda_\ell)$ . Then*

$$(2.7.25) \quad S = S_1 \cup \cdots \cup S_K$$

*is a basis of  $V$ .*

**Proof.** It follows from Proposition 2.7.5 that  $S$  spans  $V$ . We need to show that  $S$  is linearly independent. To show this it suffices to show that if  $w_\ell$  are nonzero elements of  $\mathcal{GE}(T, \lambda_\ell)$ , then no nontrivial linear combination can vanish. The demonstration of this is just slightly more elaborate than the corresponding argument in Proposition 2.6.2. If there exist such linearly dependent sets, take one with a minimal number of elements, and rearrange  $\{\lambda_\ell\}$ , to write it as  $\{w_1, \dots, w_m\}$ , so we have

$$(2.7.26) \quad c_1 w_1 + \dots + c_m w_m = 0,$$

and  $c_j \neq 0$  for each  $j \in \{1, \dots, m\}$ . As seen in Lemma 2.7.1,

$$(2.7.27) \quad T - \mu I : \mathcal{GE}(T, \lambda_\ell) \longrightarrow \mathcal{GE}(T, \lambda_\ell) \text{ is an isomorphism, } \forall \mu \neq \lambda_\ell.$$

Take  $k \in \mathbb{N}$  so large that  $(T - \lambda_m I)^k$  annihilates each element of the basis  $S_m$  of  $\mathcal{GE}(T, \lambda_m)$ , and apply  $(T - \lambda_m I)^k$  to (2.7.26). Given (2.7.27), we will obtain a non-trivial linear dependence relation involving  $m - 1$  terms, a contradiction, so the purported linear dependence relation cannot exist. This proves Proposition 2.7.6.  $\square$

EXAMPLE. Let us consider  $A : \mathbb{C}^3 \rightarrow \mathbb{C}^3$ , given by

$$(2.7.28) \quad A = \begin{pmatrix} 2 & 3 & 3 \\ 0 & 2 & 3 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then  $\text{Spec}(A) = \{2, 1\}$ , so  $m_A(\lambda) = (\lambda - 2)^a(\lambda - 1)^b$  for some positive integers  $a$  and  $b$ . Computations give

$$(2.7.29) \quad (A - 2I)(A - I) = \begin{pmatrix} 0 & 3 & 9 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad (A - 2I)^2(A - I) = 0,$$

hence  $m_A(\lambda) = (\lambda - 2)^2(\lambda - 1)$ . Thus we have

$$(2.7.30) \quad p_1(\lambda) = \lambda - 1, \quad p_2(\lambda) = (\lambda - 2)^2,$$

using the ordering  $\lambda_1 = 2, \lambda_2 = 1$ . As for  $q_\ell(\lambda)$  such that (2.7.21) holds, a little trial and error gives  $q_1(\lambda) = -(\lambda - 3), q_2(\lambda) = 1$ , i.e.,

$$(2.7.31) \quad -(\lambda - 1)(\lambda - 3) + (\lambda - 2)^2 = 1.$$

Note that

$$(2.7.32) \quad A - I = \begin{pmatrix} 1 & 3 & 3 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{pmatrix}, \quad (A - 2I)^2 = \begin{pmatrix} 0 & 0 & 6 \\ 0 & 0 & -3 \\ 0 & 0 & 1 \end{pmatrix}.$$

Hence, by (2.7.16),

$$(2.7.33) \quad \mathcal{GE}(A, 2) = \text{Span} \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\}, \quad \mathcal{GE}(A, 1) = \text{Span} \left\{ \begin{pmatrix} 6 \\ -3 \\ 1 \end{pmatrix} \right\}.$$

REMARK. In general, for  $A \in M(3, \mathbb{C})$ , there are the following three possibilities.

(I)  $A$  has 3 distinct eigenvalues,  $\lambda_1, \lambda_2, \lambda_3$ . Then  $\lambda_j$ -eigenvectors  $v_j, 1 \leq j \leq 3$ ,

span  $\mathbb{C}^3$ .

(II)  $A$  has 2 distinct eigenvalues, say  $\lambda_1$  (single) and  $\lambda_2$  (double). Then

$$m_A(\lambda) = (\lambda - \lambda_1)(\lambda - \lambda_2)^k, \quad k = 1 \text{ or } 2.$$

Whatever the value of  $k$ ,  $p_2(\lambda) = \lambda - \lambda_1$ , and hence

$$\mathcal{GE}(A, \lambda_2) = \mathcal{R}(A - \lambda_1 I),$$

which in turn is the span of the columns of  $A - \lambda_1 I$ . We have

$$\mathcal{GE}(A, \lambda_2) = \mathcal{E}(A, \lambda_2) \iff k = 1.$$

In any case,  $\mathbb{C}^3 = \mathcal{E}(A, \lambda_1) \oplus \mathcal{GE}(A, \lambda_2)$ .

(III)  $A$  has a triple eigenvalue,  $\lambda_1$ . Then  $\text{Spec}(A - \lambda_1 I) = \{0\}$ , and

$$\mathcal{GE}(A, \lambda_1) = \mathbb{C}^3.$$

Compare results of the next section.

## Exercises

1. Consider the matrices

$$A_1 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ -1 & 0 & -1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 1 & 2 & 0 \\ 3 & 1 & 3 \\ 0 & -2 & 1 \end{pmatrix}.$$

Compute the eigenvalues and eigenvectors of each  $A_j$ .

2. Find the minimal polynomial of  $A_j$  and find a basis of generalized eigenvectors of  $A_j$ .

3. Consider the transformation  $D : \mathcal{P}_2 \rightarrow \mathcal{P}_2$  given by (2.4.14). Find the eigenvalues and eigenvectors of  $D$ . Find the minimal polynomial of  $D$  and find a basis of  $\mathcal{P}_2$  consisting of generalized eigenvectors of  $D$ .

4. Suppose  $V$  is a finite dimensional complex vector space and  $T : V \rightarrow V$ . Show that  $V$  has a basis of eigenvectors of  $T$  if and only if all the roots of the minimal polynomial  $m_T(\lambda)$  are simple.

5. In the setting of (2.7.3)–(2.7.4), given  $S \in \mathcal{L}(V)$ , show that

$$ST = TS \implies S : \mathcal{GE}(T, \lambda_j) \rightarrow \mathcal{GE}(T, \lambda_j).$$

6. Show that if  $V$  is an  $n$ -dimensional complex vector space,  $S, T \in \mathcal{L}(V)$ , and  $ST = TS$ , then  $V$  has a basis consisting of vectors that are simultaneously generalized eigenvectors of  $T$  and of  $S$ .

*Hint.* Apply Proposition 2.7.6 to  $S : \mathcal{GE}(T, \lambda_j) \rightarrow \mathcal{GE}(T, \lambda_j)$ .

7. Let  $V$  be a complex  $n$ -dimensional vector space, and take  $T \in \mathcal{L}(V)$ , with minimal polynomial  $m_T(\lambda)$ , as in (2.7.13). For  $\ell \in \{1, \dots, K\}$ , set

$$P_\ell(\lambda) = \frac{m_T(\lambda)}{\lambda - \lambda_\ell}.$$

Show that, for each  $\ell \in \{1, \dots, K\}$ , there exists  $w_\ell \in V$  such that  $v_\ell = P_\ell(T)w_\ell \neq 0$ . Then show that  $(T - \lambda_\ell I)v_\ell = 0$ , so one has a proof of Proposition 2.6.1 that does not use determinants.

8. Show that Proposition 2.7.6 refines Proposition 2.7.5 to

$$V = \mathcal{GE}(T, \lambda_1) \oplus \cdots \oplus \mathcal{GE}(T, \lambda_K).$$

9. Given  $A, B \in M(n, \mathbb{C})$ , define  $L_A, R_B : M(n, \mathbb{C}) \rightarrow M(n, \mathbb{C})$  by

$$L_A X = AX, \quad R_B X = XB.$$

Show that if  $\text{Spec } A = \{\lambda_j\}$ ,  $\text{Spec } B = \{\mu_k\}$  ( $= \text{Spec } B^t$ ), then

$$\begin{aligned} \mathcal{GE}(L_A, \lambda_j) &= \text{Span}\{vw^t : v \in \mathcal{GE}(A, \lambda_j), w \in \mathbb{C}^n\}, \\ \mathcal{GE}(R_B, \mu_k) &= \text{Span}\{vw^t : v \in \mathbb{C}^n, w \in \mathcal{GE}(B^t, \mu_k)\}. \end{aligned}$$

Show that

$$\mathcal{GE}(L_A - R_B, \sigma) = \text{Span}\{vw^t : v \in \mathcal{GE}(A, \lambda_j), w \in \mathcal{GE}(B^t, \mu_k), \sigma = \lambda_j - \mu_k\}.$$

10. In the setting of Exercise 9, show that if  $A$  is diagonalizable, then  $\mathcal{GE}(L_A, \lambda_j) = \mathcal{E}(L_A, \lambda_j)$ . Draw analogous conclusions if also  $B$  is diagonalizable.

11. In the setting of Exercise 9, show that if  $\text{Spec } A = \{\lambda_j\}$  and  $\text{Spec } B = \{\mu_k\}$ , then

$$\text{Spec}(L_A - R_B) = \{\lambda_j - \mu_k\}.$$

Deduce that if  $C_A : M(n, \mathbb{C}) \rightarrow M(n, \mathbb{C})$  is defined by

$$C_A X = AX - XA,$$

then

$$\text{Spec } C_A = \{\lambda_j - \lambda_k\}.$$

## 2.8. Triangular matrices

We say an  $n \times n$  matrix  $A = (a_{jk})$  is upper triangular if  $a_{jk} = 0$  for  $j > k$ , and strictly upper triangular if  $a_{jk} = 0$  for  $j \geq k$ . Similarly we have the notion of lower triangular and strictly lower triangular matrices. Here are two examples:

$$(2.8.1) \quad A = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 3 \\ 0 & 0 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{pmatrix};$$

$A$  is upper triangular and  $B$  is strictly upper triangular;  $A^t$  is lower triangular and  $B^t$  strictly lower triangular. Note that  $B^3 = 0$ .

We say  $T \in \mathcal{L}(V)$  is *nilpotent* provided  $T^k = 0$  for some  $k \in \mathbb{N}$ . The following is a useful characterization of nilpotent transformations.

**Proposition 2.8.1.** *Let  $V$  be a finite-dimensional complex vector space,  $N \in \mathcal{L}(V)$ . The following are equivalent:*

$$(2.8.2) \quad N \text{ is nilpotent,}$$

$$(2.8.3) \quad \text{Spec}(N) = \{0\},$$

$$(2.8.4) \quad \text{There is a basis of } V \text{ for which } N \text{ is strictly upper triangular,}$$

$$(2.8.5) \quad \text{There is a basis of } V \text{ for which } N \text{ is strictly lower triangular.}$$

**Proof.** The implications (2.8.4)  $\Rightarrow$  (2.8.2) and (2.8.5)  $\Rightarrow$  (2.8.2) are easy. Also (2.8.4) implies the characteristic polynomial of  $N$  is  $\lambda^n$  (if  $n = \dim V$ ), which is equivalent to (2.8.3), and similarly (2.8.5)  $\Rightarrow$  (2.8.3). We need to establish a couple more implications.

To see that (2.8.2)  $\Rightarrow$  (2.8.3), note that if  $N^k = 0$  we can write

$$(2.8.6) \quad (N - \mu I)^{-1} = -\frac{1}{\mu} \left( I - \frac{1}{\mu} N \right)^{-1} = -\frac{1}{\mu} \sum_{\ell=0}^{k-1} \frac{1}{\mu^\ell} N^\ell,$$

whenever  $\mu \neq 0$ .

Next, given (2.8.3),  $N : V \rightarrow V$  is not an isomorphism, so  $V_1 = N(V)$  has dimension  $\leq n - 1$ . Now  $N_1 = N|_{V_1} \in \mathcal{L}(V_1)$  also has only 0 as an eigenvalue, so  $N_1(V_1) = V_2$  has dimension  $\leq n - 2$ , and so on. Thus  $N^k = 0$  for sufficiently large  $k$ . We have (2.8.3)  $\Rightarrow$  (2.8.2). Now list these spaces as  $V = V_0 \supset V_1 \supset \cdots \supset V_{k-1}$ , with  $V_{k-1} \neq 0$  but  $N(V_{k-1}) = 0$ . Pick a basis for  $V_{k-1}$ , augment it as in Proposition 2.3.5 to produce a basis for  $V_{k-2}$ , and continue, obtaining in this fashion a basis of  $V$ , with respect to which  $N$  is strictly upper triangular. Thus (2.8.3)  $\Rightarrow$  (2.8.4). On the other hand, if we reverse the order of this basis we have a basis with respect to which  $N$  is strictly lower triangular, so also (2.8.3)  $\Rightarrow$  (2.8.5). The proof of Proposition 2.8.1 is complete.  $\square$

REMARK. Having proven Proposition 2.8.1, we see another condition equivalent to (2.8.2)–(2.8.5):

$$(2.8.7) \quad N^k = 0, \quad \forall k \geq \dim V.$$



EXAMPLE. Consider

$$(2.8.8) \quad N = \begin{pmatrix} 0 & 2 & 0 \\ 3 & 0 & 3 \\ 0 & -2 & 0 \end{pmatrix}.$$

We have

$$(2.8.9) \quad N^2 = \begin{pmatrix} 6 & 0 & 6 \\ 0 & 0 & 0 \\ -6 & 0 & -6 \end{pmatrix}, \quad N^3 = 0.$$

Hence we have a chain  $V = V_0 \supset V_1 \supset V_2$  as in the proof of Proposition 2.8.1, with

$$(2.8.10) \quad \begin{aligned} V_2 &= \text{Span} \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, & V_1 &= \text{Span} \left\{ \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\}, \\ V_0 &= \text{Span} \left\{ \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \right\} = \text{Span}\{v_1, v_2, v_3\}, \end{aligned}$$

and we have

$$(2.8.11) \quad Nv_1 = 0, \quad Nv_2 = -v_1, \quad Nv_3 = 3v_2,$$

so the matrix representation of  $N$  with respect to the basis  $\{v_1, v_2, v_3\}$  is

$$(2.8.12) \quad \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{pmatrix}.$$

Generally, if  $A$  is an upper triangular  $n \times n$  matrix with diagonal entries  $d_1, \dots, d_n$ , the characteristic polynomial of  $A$  is

$$(2.8.13) \quad \det(\lambda I - A) = (\lambda - d_1) \cdots (\lambda - d_n),$$

by (2.5.46), so  $\text{Spec}(A) = \{d_j\}$ . If  $d_1, \dots, d_n$  are all distinct it follows that  $\mathbb{F}^n$  has a basis of eigenvectors of  $A$ .

We can show that whenever  $V$  is a finite-dimensional complex vector space and  $T \in \mathcal{L}(V)$ , then  $V$  has a basis with respect to which  $T$  is upper triangular. In fact, we can say a bit more. Recall what was established in Proposition 2.7.6. If  $\text{Spec}(T) = \{\lambda_\ell : 1 \leq \ell \leq K\}$  and  $S_\ell = \{v_{\ell 1}, \dots, v_{\ell, d_\ell}\}$  is a basis of  $\mathcal{GE}(T, \lambda_\ell)$ , then  $S = S_1 \cup \cdots \cup S_K$  is a basis of  $V$ . Now look more closely at

$$(2.8.14) \quad T_\ell : V_\ell \longrightarrow V_\ell, \quad V_\ell = \mathcal{GE}(T, \lambda_\ell), \quad T_\ell = T|_{V_\ell}.$$

The result (2.7.5) says  $\text{Spec}(T_\ell) = \{\lambda_\ell\}$ , i.e.,  $\text{Spec}(T_\ell - \lambda_\ell I) = \{0\}$ , so we can apply Proposition 2.8.1. Thus we can pick a basis  $S_\ell$  of  $V_\ell$  with respect to which  $T_\ell - \lambda_\ell I$  is strictly upper triangular, hence in which  $T_\ell$  takes the form

$$(2.8.15) \quad A_\ell = \begin{pmatrix} \lambda_\ell & & * \\ & \ddots & \\ 0 & & \lambda_\ell \end{pmatrix}.$$

Then, with respect to the basis  $S = S_1 \cup \cdots \cup S_K$ ,  $T$  has a matrix representation  $A$  consisting of blocks  $A_\ell$ , given by (2.8.15). It follows that

$$(2.8.16) \quad K_T(\lambda) = \det(\lambda I - T) = \prod_{\ell=1}^K (\lambda - \lambda_\ell)^{d_\ell}, \quad d_\ell = \dim V_\ell.$$

This matrix representation also makes it clear that  $K_T(T)|_{V_\ell} = 0$  for each  $\ell \in \{1, \dots, K\}$  (cf. (2.8.7)), hence

$$(2.8.17) \quad K_T(T) = 0 \quad \text{on } V.$$

This result is known as the Cayley-Hamilton theorem. Recalling the characterization of the minimal polynomial  $m_T(\lambda)$  given in (2.7.11)–(2.7.13), we see that

$$(2.8.18) \quad K_T(\lambda) \text{ is a polynomial multiple of } m_T(\lambda).$$

### Exercises

1. Consider

$$A_1 = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 2 \\ 3 & 2 & 1 \end{pmatrix}.$$

Compute the characteristic polynomial of each  $A_j$  and verify that these matrices satisfy the Cayley-Hamilton theorem, stated in (2.8.17).

2. Let  $\mathcal{P}_k$  denote the space of polynomials of degree  $\leq k$  in  $x$ , and consider

$$D : \mathcal{P}_k \longrightarrow \mathcal{P}_k, \quad Dp(x) = p'(x).$$

Show that  $D^{k+1} = 0$  on  $\mathcal{P}_k$  and that  $\{1, x, \dots, x^k\}$  is a basis of  $\mathcal{P}_k$  with respect to which  $D$  is strictly upper triangular.

3. Use the identity

$$(I - D)^{-1} = \sum_{\ell=0}^{k+1} D^\ell, \quad \text{on } \mathcal{P}_k,$$

to obtain a solution  $u \in \mathcal{P}_k$  to

$$(2.8.19) \quad u' - u = x^k.$$

4. Use the equivalence of (2.8.19) with

$$\frac{d}{dx}(e^{-x}u) = x^k e^{-x}$$

to obtain a formula for

$$\int x^k e^{-x} dx.$$

For an alternative approach, see (1.1.45)–(1.1.52) of Chapter 1; see also exercises at the end of §3.4 of Chapter 3.

5. The proof of Proposition 2.8.1 given above includes the chain of implications

$$(2.8.4) \Rightarrow (2.8.2) \Rightarrow (2.8.3) \Rightarrow (2.8.4).$$

Use Proposition 2.7.4 to show directly that

$$(2.8.3) \Rightarrow (2.8.2).$$

6. Establish the following variant of Proposition 2.7.4. Let  $K_T(\lambda)$  be the characteristic polynomial of  $T$ , as in (2.8.16), and set

$$P_\ell(\lambda) = \prod_{j \neq \ell} (\lambda - \lambda_j)^{d_j} = \frac{K_T(\lambda)}{(\lambda - \lambda_\ell)^{d_\ell}}.$$

Show that

$$\mathcal{GE}(T, \lambda_\ell) = \mathcal{R}(P_\ell(T)).$$

## 2.9. Inner products and norms

Vectors in  $\mathbb{R}^n$  have a dot product, given by

$$(2.9.1) \quad v \cdot w = v_1 w_1 + \cdots + v_n w_n,$$

where  $v = (v_1, \dots, v_n)$ ,  $w = (w_1, \dots, w_n)$ . Then the norm of  $v$ , denoted  $\|v\|$ , is given by

$$(2.9.2) \quad \|v\|^2 = v \cdot v = v_1^2 + \cdots + v_n^2.$$

The geometrical significance of  $\|v\|$  as the distance of  $v$  from the origin is a version of the Pythagorean theorem. If  $v, w \in \mathbb{C}^n$ , we use

$$(2.9.3) \quad (v, w) = v \cdot \bar{w} = v_1 \bar{w}_1 + \cdots + v_n \bar{w}_n,$$

and then

$$(2.9.4) \quad \|v\|^2 = (v, v) = |v_1|^2 + \cdots + |v_n|^2;$$

here, if  $v_j = x_j + iy_j$ , with  $x_j, y_j \in \mathbb{R}$ , we have  $\bar{v}_j = x_j - iy_j$ , and  $|v_j|^2 = x_j^2 + y_j^2$ .

The objects (2.9.1) and (2.9.3) are special cases of *inner products*. Generally, an inner product on a vector space (over  $\mathbb{F} = \mathbb{R}$  or  $\mathbb{C}$ ) assigns to vectors  $v, w \in V$  the quantity  $(v, w) \in \mathbb{F}$ , in a fashion that obeys the following three rules:

$$(2.9.5) \quad (a_1 v_1 + a_2 v_2, w) = a_1 (v_1, w) + a_2 (v_2, w),$$

$$(2.9.6) \quad (v, w) = \overline{(w, v)},$$

$$(2.9.7) \quad (v, v) > 0, \quad \text{unless } v = 0.$$

If  $\mathbb{F} = \mathbb{R}$ , then (2.9.6) just means  $(v, w) = (w, v)$ . Note that (2.9.5)–(2.9.6) together imply

$$(2.9.8) \quad (v, b_1 w_1 + b_2 w_2) = \bar{b}_1 (v, w_1) + \bar{b}_2 (v, w_2).$$

A vector space equipped with an inner product is called an inner product space. Inner products arise naturally in various contexts. For example,

$$(2.9.9) \quad (f, g) = \int_a^b f(x) \overline{g(x)} dx$$

defines an inner product on  $C([a, b])$ . It also defines an inner product on  $\mathcal{P}$ , the space of polynomials in  $x$ . Different choices of  $a$  and  $b$  yield different inner products on  $\mathcal{P}$ . More generally, one considers inner products of the form

$$(2.9.10) \quad (f, g) = \int_a^b f(x) \overline{g(x)} w(x) dx,$$

on various function spaces, where  $w$  is a positive, integrable “weight” function.

Given an inner product on  $V$ , one says the object  $\|v\|$  defined by

$$(2.9.11) \quad \|v\| = \sqrt{(v, v)}$$

is the *norm* on  $V$  associated with the inner product. Generally, a norm on  $V$  is a function  $v \mapsto \|v\|$  satisfying

$$(2.9.12) \quad \|av\| = |a| \cdot \|v\|, \quad \forall a \in \mathbb{F}, v \in V,$$

$$(2.9.13) \quad \|v\| > 0, \quad \text{unless } v = 0,$$

$$(2.9.14) \quad \|v + w\| \leq \|v\| + \|w\|.$$

Here  $|a|$  denotes the absolute value of  $a \in \mathbb{F}$ . The property (9.14) is called the *triangle inequality*. A vector space equipped with a norm is called a normed vector space.

If  $\|v\|$  is given by (2.9.11), from an inner product satisfying (2.9.5)–(2.9.7), it is clear that (2.9.12)–(2.9.13) hold, but (2.9.14) requires a demonstration. Note that

$$\begin{aligned} \|v+w\|^2 &= (v+w, v+w) \\ (2.9.15) \quad &= \|v\|^2 + (v, w) + (w, v) + \|w\|^2 \\ &= \|v\|^2 + 2\operatorname{Re}(v, w) + \|w\|^2, \end{aligned}$$

while

$$(2.9.16) \quad (\|v\| + \|w\|)^2 = \|v\|^2 + 2\|v\| \cdot \|w\| + \|w\|^2.$$

Thus to establish (2.9.14) it suffices to prove the following, known as Cauchy's inequality:

**Proposition 2.9.1.** *For any inner product on a vector space  $V$ , with  $\|v\|$  defined by (2.9.11),*

$$(2.9.17) \quad |(v, w)| \leq \|v\| \|w\|, \quad \forall v, w \in V.$$

**Proof.** We start with

$$(2.9.18) \quad 0 \leq \|v-w\|^2 = \|v\|^2 - 2\operatorname{Re}(v, w) + \|w\|^2,$$

which implies

$$2\operatorname{Re}(v, w) \leq \|v\|^2 + \|w\|^2, \quad \forall v, w \in V.$$

Replacing  $v$  by  $\alpha v$  for arbitrary  $\alpha \in \mathbb{F}$  of absolute value 1 yields  $2\operatorname{Re}\alpha(v, w) \leq \|v\|^2 + \|w\|^2$ . This implies

$$(2.9.19) \quad 2|(v, w)| \leq \|v\|^2 + \|w\|^2, \quad \forall v, w \in V.$$

Replacing  $v$  by  $tv$  and  $w$  by  $t^{-1}w$  for arbitrary  $t \in (0, \infty)$ , we have

$$(2.9.20) \quad 2|(v, w)| \leq t^2\|v\|^2 + t^{-2}\|w\|^2, \quad \forall v, w \in V, t \in (0, \infty).$$

If we take  $t^2 = \|w\|/\|v\|$ , we obtain the desired inequality (2.9.17). (This assumes  $v$  and  $w$  are both nonzero, but (2.9.17) is trivial if  $v$  or  $w$  is 0.)  $\square$

There are other norms on vector spaces besides those that are associated with inner products. For example, on  $\mathbb{F}^n$ , we have

$$(2.9.21) \quad \|v\|_1 = |v_1| + \cdots + |v_n|, \quad \|v\|_\infty = \max_{1 \leq k \leq n} |v_k|,$$

and many others, but we will not dwell on this here.

If  $V$  is a finite-dimensional inner product space, a basis  $\{u_1, \dots, u_n\}$  of  $V$  is called an *orthonormal basis* of  $V$  provided

$$(2.9.22) \quad (u_j, u_k) = \delta_{jk}, \quad 1 \leq j, k \leq n,$$

i.e.,

$$(2.9.23) \quad \|u_j\| = 1, \quad j \neq k \Rightarrow (u_j, u_k) = 0.$$

(When  $(u_j, u_k) = 0$ , we say  $u_j$  and  $u_k$  are *orthogonal*.) When (2.9.22) holds, we have

$$(2.9.24) \quad v = a_1 u_1 + \cdots + a_n u_n, \quad w = b_1 u_1 + \cdots + b_n u_n \Rightarrow (v, w) = a_1 \bar{b}_1 + \cdots + a_n \bar{b}_n.$$

It is often useful to construct orthonormal bases. The construction we now describe is called the Gram-Schmidt construction.

**Proposition 2.9.2.** *Let  $\{v_1, \dots, v_n\}$  be a basis of  $V$ , an inner product space. Then there is an orthonormal basis  $\{u_1, \dots, u_n\}$  of  $V$  such that*

$$(2.9.25) \quad \text{Span}\{u_j : j \leq \ell\} = \text{Span}\{v_j : j \leq \ell\}, \quad 1 \leq \ell \leq n.$$

**Proof.** To begin, take

$$(2.9.26) \quad u_1 = \frac{1}{\|v_1\|} v_1.$$

Now define the linear transformation  $P_1 : V \rightarrow V$  by  $P_1 v = (v, u_1) u_1$  and set

$$\tilde{v}_2 = v_2 - P_1 v_2 = v_2 - (v_2, u_1) u_1.$$

We see that  $(\tilde{v}_2, u_1) = (v_2, u_1) - (v_2, u_1) = 0$ . Also  $\tilde{v}_2 \neq 0$  since  $u_1$  and  $v_2$  are linearly independent. Hence we set

$$(2.9.27) \quad u_2 = \frac{1}{\|\tilde{v}_2\|} \tilde{v}_2.$$

Inductively, suppose we have an orthonormal set  $\{u_1, \dots, u_m\}$  with  $m < n$  and (2.9.25) holding for  $1 \leq \ell \leq m$ . Then define  $P_m : V \rightarrow V$  (the orthogonal projection of  $V$  onto  $\text{Span}(u_1, \dots, u_m)$ ) by

$$(2.9.28) \quad P_m v = (v, u_1) u_1 + \cdots + (v, u_m) u_m,$$

and set

$$(2.9.29) \quad \tilde{v}_{m+1} = v_{m+1} - P_m v_{m+1} = v_{m+1} - (v_{m+1}, u_1) u_1 - \cdots - (v_{m+1}, u_m) u_m.$$

We see that

$$(2.9.30) \quad j \leq m \Rightarrow (\tilde{v}_{m+1}, u_j) = (v_{m+1}, u_j) - (v_{m+1}, u_j) = 0.$$

Also, since  $v_{m+1} \notin \text{Span}\{v_1, \dots, v_m\} = \text{Span}\{u_1, \dots, u_m\}$ , it follows that  $\tilde{v}_{m+1} \neq 0$ . Hence we set

$$(2.9.31) \quad u_{m+1} = \frac{1}{\|\tilde{v}_{m+1}\|} \tilde{v}_{m+1}.$$

This completes the construction.  $\square$

EXAMPLE. Take  $V = \mathcal{P}_2$ , with basis  $\{1, x, x^2\}$ , and inner product given by

$$(2.9.32) \quad (p, q) = \int_{-1}^1 p(x) \overline{q(x)} dx.$$

The Gram-Schmidt construction gives first

$$(2.9.33) \quad u_1(x) = \frac{1}{\sqrt{2}}.$$

Then

$$\tilde{v}_2(x) = x,$$

since by symmetry  $(x, u_1) = 0$ . Now  $\int_{-1}^1 x^2 dx = 2/3$ , so we take

$$(2.9.34) \quad u_2(x) = \sqrt{\frac{3}{2}}x.$$

Next

$$\tilde{v}_3(x) = x^2 - (x^2, u_1)u_1 = x^2 - \frac{1}{3},$$

since by symmetry  $(x^2, u_2) = 0$ . Now  $\int_{-1}^1 (x^2 - 1/3)^2 dx = 8/45$ , so we take

$$(2.9.35) \quad u_3(x) = \sqrt{\frac{45}{8}}\left(x^2 - \frac{1}{3}\right).$$

### Exercises

1. Let  $V$  be a finite dimensional inner product space, and let  $W$  be a linear subspace of  $V$ . Show that any orthonormal basis  $\{w_1, \dots, w_k\}$  of  $W$  can be enlarged to an orthonormal basis  $\{w_1, \dots, w_k, u_1, \dots, u_\ell\}$  of  $V$ , with  $k + \ell = \dim V$ .

2. As in Exercise 1, let  $V$  be a finite dimensional inner product space, and let  $W$  be a linear subspace of  $V$ . Define the orthogonal complement

$$(2.9.36) \quad W^\perp = \{v \in V : (v, w) = 0, \forall w \in W\}.$$

Show that

$$W^\perp = \text{Span}\{u_1, \dots, u_\ell\},$$

in the context of Exercise 1. Deduce that

$$(2.9.37) \quad (W^\perp)^\perp = W.$$

3. In the context of Exercise 2, show that

$$\dim V = n, \dim W = k \implies \dim W^\perp = n - k.$$

4. Construct an orthonormal basis of the  $(n - 1)$ -dimensional vector space

$$V = \left\{ \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \in \mathbb{R}^n : v_1 + \dots + v_n = 0 \right\}.$$

5. Take  $V = \mathcal{P}_2$ , with basis  $\{1, x, x^2\}$ , and inner product

$$(p, q) = \int_0^1 p(x)\overline{q(x)} dx,$$

in contrast to (2.9.32). Construct an orthonormal basis of this inner product space.

6. Take  $V$ , with basis  $\{1, \cos x, \sin x\}$ , and inner product

$$(f, g) = \int_0^\pi f(x)\overline{g(x)} dx.$$

Construct an orthonormal basis of this inner product space.



### 2.10. Norm, trace, and adjoint of a linear transformation

If  $V$  and  $W$  are normed linear spaces and  $T \in \mathcal{L}(V, W)$ , we define

$$(2.10.1) \quad \|T\| = \sup \{\|Tv\| : \|v\| \leq 1\}.$$

Equivalently,  $\|T\|$  is the smallest quantity  $K$  such that

$$(2.10.2) \quad \|Tv\| \leq K\|v\|, \quad \forall v \in V.$$

We call  $\|T\|$  the *operator norm* of  $T$ . If  $V$  and  $W$  are finite dimensional, it can be shown that  $\|T\| < \infty$  for all  $T \in \mathcal{L}(V, W)$ . We omit the general argument, but we will make some estimates below when  $V$  and  $W$  are inner product spaces.

Note that if also  $S : W \rightarrow X$ , another normed vector space, then

$$(2.10.3) \quad \|STv\| \leq \|S\| \|Tv\| \leq \|S\| \|T\| \|v\|, \quad \forall v \in V,$$

and hence

$$(2.10.4) \quad \|ST\| \leq \|S\| \|T\|.$$

In particular, we have by induction that

$$(2.10.5) \quad T : V \rightarrow V \implies \|T^n\| \leq \|T\|^n.$$

This will be useful when we discuss the exponential of a linear transformation, in Chapter 3.

We turn to the notion of the *trace* of a transformation  $T \in \mathcal{L}(V)$ , given  $\dim V < \infty$ . We start with the trace of an  $n \times n$  matrix, which is simply the sum of the diagonal elements:

$$(2.10.6) \quad A = (a_{jk}) \in M(n, \mathbb{F}) \implies \operatorname{Tr} A = \sum_{j=1}^n a_{jj}.$$

Note that if also  $B = (b_{jk}) \in M(n, \mathbb{F})$ , then

$$(2.10.7) \quad \begin{aligned} AB = C = (c_{jk}), \quad c_{jk} &= \sum_{\ell} a_{j\ell} b_{\ell k}, \\ BA = D = (d_{jk}), \quad d_{jk} &= \sum_{\ell} b_{j\ell} a_{\ell k}, \end{aligned}$$

and hence

$$(2.10.8) \quad \operatorname{Tr} AB = \sum_{j,\ell} a_{j\ell} b_{\ell j} = \operatorname{Tr} BA.$$

Hence, if  $B$  is invertible,

$$(2.10.9) \quad \operatorname{Tr} B^{-1}AB = \operatorname{Tr} ABB^{-1} = \operatorname{Tr} A.$$

Thus if  $T \in \mathcal{L}(V)$ , we can choose a basis  $S = \{v_1, \dots, v_n\}$  of  $V$ , if  $\dim V = n$ , and define

$$(2.10.10) \quad \operatorname{Tr} T = \operatorname{Tr} A, \quad A = \mathcal{M}_S^S(T),$$

and (2.10.9) implies this is independent of the choice of basis.

Next we define the *adjoint* of  $T \in \mathcal{L}(V, W)$ , when  $V$  and  $W$  are finite-dimensional inner product spaces, as the transformation  $T^* \in \mathcal{L}(W, V)$  with the property

$$(2.10.11) \quad (Tv, w) = (v, T^*w), \quad \forall v \in V, w \in W.$$

If  $\{v_1, \dots, v_n\}$  is an orthonormal basis of  $V$  and  $\{w_1, \dots, w_m\}$  an orthonormal basis of  $W$ , then

$$(2.10.12) \quad A = (a_{ij}), \quad a_{ij} = (Tv_j, w_i),$$

is the matrix representation of  $T$ , as in (2.4.2), and the matrix representation of  $T^*$  is

$$(2.10.13) \quad A^* = (\bar{a}_{ji}).$$

Now we define the Hilbert-Schmidt norm of  $T \in \mathcal{L}(V, W)$  when  $V$  and  $W$  are finite-dimensional inner product spaces. Namely, we set

$$(2.10.14) \quad \|T\|_{HS}^2 = \text{Tr } T^*T.$$

In terms of the matrix representation (2.10.12) of  $T$ , we have

$$(2.10.15) \quad T^*T = (b_{jk}), \quad b_{jk} = \sum_{\ell} \bar{a}_{\ell j} a_{\ell k},$$

hence

$$(2.10.16) \quad \|T\|_{HS}^2 = \sum_j b_{jj} = \sum_{j,k} |a_{jk}|^2.$$

Equivalently, using an arbitrary orthonormal basis  $\{v_1, \dots, v_n\}$  of  $V$ , we have

$$(2.10.17) \quad \|T\|_{HS}^2 = \sum_{j=1}^n \|Tv_j\|^2.$$

Using (2.10.17), we can show that the operator norm of  $T$  is dominated by the Hilbert-Schmidt norm:

$$(2.10.18) \quad \|T\| \leq \|T\|_{HS}.$$

In fact, pick a unit  $v_1 \in V$  such that  $\|Tv_1\|$  is maximized on  $\{v : \|v\| \leq 1\}$ , extend this to an orthonormal basis  $\{v_1, \dots, v_n\}$ , and use

$$\|T\|^2 = \|Tv_1\|^2 \leq \sum_{j=1}^n \|Tv_j\|^2 = \|T\|_{HS}^2.$$

Also we can dominate each term on the right side of (2.10.17) by  $\|T\|^2$ , so

$$(2.10.19) \quad \|T\|_{HS} \leq \sqrt{n}\|T\|, \quad n = \dim V.$$

Another consequence of (2.10.17)–(2.10.18) is

$$(2.10.20) \quad \|ST\|_{HS} \leq \|S\| \|T\|_{HS} \leq \|S\|_{HS} \|T\|_{HS},$$

for  $S$  as in (2.10.3). In particular, parallel to (2.10.5), we have

$$(2.10.21) \quad T : V \rightarrow V \implies \|T^n\|_{HS} \leq \|T\|_{HS}^n.$$

---

**Exercises**

1. Suppose  $V$  and  $W$  are finite dimensional inner product spaces and  $T \in \mathcal{L}(V, W)$ . Show that

$$T^{**} = T.$$

2. In the context of Exercise 1, show that

$$T \text{ injective} \iff T^* \text{ surjective.}$$

More generally, show that

$$\mathcal{N}(T) = \mathcal{R}(T^*)^\perp.$$

(See Exercise 2 of §2.9 for a discussion of the orthogonal complement  $W^\perp$ .)

3. Say  $A$  is a  $k \times n$  real matrix and the  $k$  columns are linearly independent. Show that  $A$  has  $k$  linearly independent rows. (Similarly treat complex matrices.)

*Hint.* The hypothesis is equivalent to  $A : \mathbb{R}^k \rightarrow \mathbb{R}^n$  being injective. What does that say about  $A^* : \mathbb{R}^n \rightarrow \mathbb{R}^k$ ?

4. If  $A$  is a  $k \times n$  real (or complex) matrix, we define the *column rank* of  $A$  to be the dimension of the span of the columns of  $A$ . We similarly define the *row rank* of  $A$ . Show that the row rank of  $A$  is equal to its column rank.

*Hint.* Reduce this to showing  $\dim \mathcal{R}(A) = \dim \mathcal{R}(A^*)$ . Apply Exercise 2 (and Exercise 3 of §2.9).

5. Suppose  $A$  is an  $n \times n$  matrix and  $\|A\| < 1$ . Show that

$$(I - A)^{-1} = I + A + A^2 + \cdots + A^k + \cdots,$$

a convergent infinite series.

6. If  $A$  is an  $n \times n$  complex matrix, show that

$$\lambda \in \text{Spec}(A) \implies |\lambda| \leq \|A\|.$$

7. Show that, for any real  $\theta$ , the matrix

$$A = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$$

has operator norm 1. Compute its Hilbert-Schmidt norm.

8. Given  $a > b > 0$ , show that the matrix

$$B = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$$

has operator norm  $a$ . Compute its Hilbert-Schmidt norm.

9. Show that if  $V$  is an  $n$ -dimensional complex inner product space, then, for  $T \in \mathcal{L}(V)$ ,

$$\det T^* = \overline{\det T}.$$

10. If  $V$  is an  $n$ -dimensional inner product space, show that, for  $T \in \mathcal{L}(V)$ ,

$$\|T\| = \sup\{|(Tu, v)| : \|u\|, \|v\| \leq 1\}.$$

Show that

$$\|T^*\| = \|T\|.$$

### 2.11. Self-adjoint and skew-adjoint transformations

If  $V$  is a finite-dimensional inner product space,  $T \in \mathcal{L}(V)$  is said to be self-adjoint if  $T = T^*$  and skew-adjoint if  $T = -T^*$ . If  $\{u_1, \dots, u_n\}$  is an orthonormal basis of  $V$  and  $A$  the matrix representation of  $T$  with respect to this basis, given by

$$(2.11.1) \quad A = (a_{ij}), \quad a_{ij} = (Tu_j, u_i),$$

then  $T^*$  is represented by  $A^* = (\bar{a}_{ji})$ , so  $T$  is self-adjoint if and only if  $a_{ij} = \bar{a}_{ji}$  and  $T$  is skew-adjoint if and only if  $a_{ij} = -\bar{a}_{ji}$ .

The eigenvalues and eigenvectors of these two classes of operators have special properties, as we proceed to show.

**Lemma 2.11.1.** *If  $\lambda_j$  is an eigenvalue of a self-adjoint  $T \in \mathcal{L}(V)$ , then  $\lambda_j$  is real.*

**Proof.** Say  $Tv_j = \lambda_j v_j$ ,  $v_j \neq 0$ . Then

$$(2.11.2) \quad \lambda_j \|v_j\|^2 = (Tv_j, v_j) = \bar{\lambda}_j \|v_j\|^2,$$

so  $\lambda_j = \bar{\lambda}_j$ . □

This allows us to prove the following result for both real and complex vector spaces.

**Proposition 2.11.2.** *If  $V$  is a finite-dimensional inner product space and  $T \in \mathcal{L}(V)$  is self-adjoint, then  $V$  has an orthonormal basis of eigenvectors of  $T$ .*

**Proof.** Proposition 2.6.1 (and the comment following it in case  $\mathbb{F} = \mathbb{R}$ ) implies there is a unit  $v_1 \in V$  such that  $Tv_1 = \lambda_1 v_1$ , and we know  $\lambda_1 \in \mathbb{R}$ . Say  $\dim V = n$ . Let

$$(2.11.3) \quad W = \{w \in V : (v_1, w) = 0\}.$$

Then  $\dim W = n - 1$ , as we can see by completing  $\{v_1\}$  to an orthonormal basis of  $V$ . We claim

$$(2.11.4) \quad T = T^* \implies T : W \rightarrow W.$$

Indeed,

$$(2.11.5) \quad w \in W \implies (v_1, Tw) = (Tv_1, w) = \lambda_1 (v_1, w) = 0 \implies Tw \in W.$$

An inductive argument gives an orthonormal basis of  $W$  consisting of eigenvalues of  $T$ , so Proposition 2.11.2 is proven. □

The following could be deduced from Proposition 2.11.2, but we prove it directly.

**Proposition 2.11.3.** *Assume  $T \in \mathcal{L}(V)$  is self-adjoint. If  $Tv_j = \lambda_j v_j$ ,  $Tv_k = \lambda_k v_k$ , and  $\lambda_j \neq \lambda_k$ , then  $(v_j, v_k) = 0$ .*

**Proof.** Then we have

$$\lambda_j (v_j, v_k) = (Tv_j, v_k) = (v_j, Tv_k) = \lambda_k (v_j, v_k).$$

□

If  $\mathbb{F} = \mathbb{C}$ , we have

$$(2.11.6) \quad T \text{ skew-adjoint} \iff iT \text{ self-adjoint},$$

so Proposition 2.11.2 has an extension to skew-adjoint transformations if  $\mathbb{F} = \mathbb{C}$ . The case  $\mathbb{F} = \mathbb{R}$  requires further study.

For concreteness, take  $V = \mathbb{R}^n$ , with its standard inner product, and consider a skew-adjoint transformation  $A : \mathbb{R}^n \rightarrow \mathbb{R}^n$ . In this case, skew-adjointness is equivalent to skew-symmetry:

$$(2.11.7) \quad A = (a_{ij}), \quad a_{ij} = -a_{ji}. \quad (\text{We say } A \in \text{Skew}(n).)$$

Now we can consider

$$(2.11.8) \quad A : \mathbb{C}^n \longrightarrow \mathbb{C}^n,$$

given by the same matrix as in (2.11.7), which is a matrix with real entries. Thus the characteristic polynomial  $K_A(\lambda) = \det(\lambda I - A)$  is a polynomial of degree  $n$  with real coefficients, so its non-real roots occur in complex conjugate pairs. Thus the nonzero elements of  $\text{Spec}(A)$  are

$$(2.11.9) \quad \text{Spec}'(A) = \{i\lambda_1, \dots, i\lambda_m, -i\lambda_1, \dots, -i\lambda_m\},$$

with  $\lambda_j \neq \lambda_k$  if  $j \neq k$ ; for the sake of concreteness, say each  $\lambda_j > 0$ . By Proposition 2.11.2,  $\mathbb{C}^n$  has an orthonormal basis of eigenvalues of  $A$ , and of course each such basis element belongs to  $\mathcal{E}(A, i\lambda_j)$  or to  $\mathcal{E}(A, -i\lambda_j)$ , for some  $j \in \{1, \dots, m\}$ , or to  $\mathcal{E}(A, 0) = \mathcal{N}(A)$ . For each  $j \in \{1, \dots, m\}$ , let

$$(2.11.10) \quad \{v_{j1}, \dots, v_{j,d_j}\}$$

be an orthonormal basis of  $\mathcal{E}(A, i\lambda_j)$ . Say

$$(2.11.11) \quad v_{jk} = \xi_{jk} + i\eta_{jk}, \quad \xi_{jk}, \eta_{jk} \in \mathbb{R}^n.$$

Then we can take

$$(2.11.12) \quad \bar{v}_{jk} = \xi_{jk} - i\eta_{jk} \in \mathbb{C}^n,$$

and

$$(2.11.13) \quad \{\bar{v}_{j1}, \dots, \bar{v}_{j,d_j}\}$$

is an orthonormal basis of  $\mathcal{E}(A, -i\lambda_j)$ . Note that

$$(2.11.14) \quad A\xi_{jk} = -\lambda_j\eta_{jk}, \quad A\eta_{jk} = \lambda_j\xi_{jk}, \quad 1 \leq k \leq d_j.$$

Note also that

$$(2.11.15) \quad \text{Span}_{\mathbb{C}}\{\xi_{jk}, \eta_{jk} : 1 \leq k \leq d_j\} = \mathcal{E}(A, i\lambda_j) + \mathcal{E}(A, -i\lambda_j),$$

while we can also take

$$(2.11.16) \quad \text{Span}_{\mathbb{R}}\{\xi_{jk}, \eta_{jk} : 1 \leq k \leq d_j\} = \mathcal{H}(A, \lambda_j) \subset \mathbb{R}^n,$$

a linear subspace of  $\mathbb{R}^n$ , of dimension  $2d_j$ . Furthermore, applying Proposition 2.11.3 to  $iA$ , we see that

$$(2.11.17) \quad (v_{jk}, \bar{v}_{jk}) = 0 \implies \|\xi_{jk}\|^2 = \|\eta_{jk}\|^2, \quad \text{and } (\xi_{jk}, \eta_{jk}) = 0,$$

hence

$$(2.11.18) \quad \|\xi_{jk}\| = \|\eta_{jk}\| = \frac{1}{\sqrt{2}}.$$

Making further use of

$$(2.11.19) \quad (v_{ij}, \bar{v}_{k\ell}) = 0, \quad (v_{ij}, v_{k\ell}) = \delta_{ik}\delta_{j\ell},$$

we see that

$$(2.11.20) \quad \left\{ \sqrt{2}\xi_{jk}, \sqrt{2}\eta_{jk} : 1 \leq k \leq d_j, 1 \leq j \leq m \right\}$$

is an orthonormal set in  $\mathbb{R}^n$ , whose linear span over  $\mathbb{C}$  coincides with the span of all the nonzero eigenspaces of  $A$  in  $\mathbb{C}^n$ .

Next we compare  $\mathcal{N}_{\mathbb{C}}(A) \subset \mathbb{C}^n$  with  $\mathcal{N}_{\mathbb{R}}(A) \subset \mathbb{R}^n$ . It is clear that, if  $v_j = \xi_j + i\eta_j$ ,  $\xi_j, \eta_j \in \mathbb{R}^n$ ,

$$(2.11.21) \quad v_j \in \mathcal{N}_{\mathbb{C}}(A) \iff \xi_j, \eta_j \in \mathcal{N}_{\mathbb{R}}(A),$$

since  $A$  is a real matrix. Thus, if  $\{\xi_1, \dots, \xi_\mu\}$  is an orthonormal basis for  $\mathcal{N}_{\mathbb{R}}(A)$ , it is also an orthonormal basis for  $\mathcal{N}_{\mathbb{C}}(A)$ . Therefore we have the following conclusion:

**Proposition 2.11.4.** *If  $A : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is skew-adjoint, then  $\mathbb{R}^n$  has an orthonormal basis in which the matrix representation of  $A$  consists of blocks*

$$(2.11.22) \quad \begin{pmatrix} 0 & \lambda_j \\ -\lambda_j & 0 \end{pmatrix},$$

plus perhaps a zero matrix, when  $\mathcal{N}(A) \neq 0$ .

## Exercises

1. Verify Proposition 2.11.2 for  $V = \mathbb{R}^3$  and

$$T = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

2. Verify Proposition 2.11.4 for

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

3. In the setting of Proposition 2.11.2, suppose  $S, T \in \mathcal{L}(V)$  are both self-adjoint and suppose they *commute*, i.e.,  $ST = TS$ . Show that  $V$  has an orthonormal basis of vectors that are simultaneously eigenvectors of  $S$  and of  $T$ .

4. If  $V$  is a finite dimensional inner product space, we say  $T \in \mathcal{L}(V)$  is positive definite if and only if  $T = T^*$  and

$$(2.11.23) \quad (Tv, v) > 0 \quad \text{for all nonzero } v \in V.$$

Show that  $T \in \mathcal{L}(V)$  is positive definite if and only if  $T = T^*$  and all its eigenvalues are  $> 0$ . We say  $T$  is positive semidefinite if and only if  $T = T^*$  and

$$(Tv, v) \geq 0, \quad \forall v \in V.$$

Show that  $T \in \mathcal{L}(V)$  is positive semidefinite if and only if  $T = T^*$  and all its eigenvalues are  $\geq 0$ .

5. If  $T \in \mathcal{L}(V)$  is positive semidefinite, show that

$$\|T\| = \max\{\lambda : \lambda \in \text{Spec } T\}.$$

6. If  $S \in \mathcal{L}(V)$ , show that  $S^*S$  is positive semidefinite, and

$$\|S\|^2 = \|S^*S\|.$$



## 2.12. Unitary and orthogonal transformations

Let  $V$  be a finite-dimensional inner product space (over  $\mathbb{F}$ ) and  $T \in \mathcal{L}(V)$ . Suppose

$$(2.12.1) \quad T^{-1} = T^*.$$

If  $\mathbb{F} = \mathbb{C}$  we say  $T$  is *unitary*, and if  $\mathbb{F} = \mathbb{R}$  we say  $T$  is *orthogonal*. We denote by  $U(n)$  the set of unitary transformations on  $\mathbb{C}^n$  and by  $O(n)$  the set of orthogonal transformations on  $\mathbb{R}^n$ . Note that (2.12.1) implies

$$(2.12.2) \quad |\det T|^2 = (\det T)(\det T^*) = 1,$$

i.e.,  $\det T \in \mathbb{F}$  has absolute value 1. In particular,

$$(2.12.3) \quad T \in O(n) \implies \det T = \pm 1.$$

We set

$$(2.12.4) \quad \begin{aligned} SO(n) &= \{T \in O(n) : \det T = 1\}, \\ SU(n) &= \{T \in U(n) : \det T = 1\}. \end{aligned}$$

As with self-adjoint and skew-adjoint transformations, the eigenvalues and eigenvectors of unitary transformations have special properties, as we now demonstrate.

**Lemma 2.12.1.** *If  $\lambda_j$  is an eigenvalue of a unitary  $T \in \mathcal{L}(V)$ , then  $|\lambda_j| = 1$ .*

**Proof.** Say  $Tv_j = \lambda_j v_j$ ,  $v_j \neq 0$ . Then

$$(2.12.5) \quad \|v_j\|^2 = (T^*Tv_j, v_j) = (Tv_j, Tv_j) = |\lambda_j|^2 \|v_j\|^2.$$

□

Next, parallel to Proposition 2.11.2, we show unitary transformations have eigenvectors forming a basis.

**Proposition 2.12.2.** *If  $V$  is a finite-dimensional complex inner product space and  $T \in \mathcal{L}(V)$  is unitary, then  $V$  has an orthonormal basis of eigenvectors of  $T$ .*

**Proof.** Proposition 6.1 implies there is a unit  $v_1 \in V$  such that  $Tv_1 = \lambda_1 v_1$ . Say  $\dim V = n$ . Let

$$(2.12.6) \quad W = \{w \in V : (v_1, w) = 0\}.$$

As in the analysis of (2.11.3) we have  $\dim W = n - 1$ . We claim

$$(2.12.7) \quad T \text{ unitary} \implies T : W \rightarrow W.$$

Indeed,

$$(2.12.8) \quad w \in W \implies (v_1, Tw) = (T^{-1}v_1, w) = \lambda_1^{-1}(v_1, w) = 0 \implies Tw \in W.$$

Now, as in Proposition 2.11.2, an inductive argument gives an orthonormal basis of  $W$  consisting of eigenvectors of  $T$ , so Proposition 2.12.2 is proven. □

Next we have a result parallel to Proposition 2.11.3:

**Proposition 2.12.3.** *Assume  $T \in \mathcal{L}(V)$  is unitary. If  $Tv_j = \lambda_j v_j$  and  $Tv_k = \lambda_k v_k$ , and  $\lambda_j \neq \lambda_k$ , then  $(v_j, v_k) = 0$ .*

**Proof.** Then we have

$$\lambda_j(v_j, v_k) = (Tv_j, v_k) = (v_j, T^{-1}v_k) = \lambda_k(v_j, v_k),$$

since  $\overline{\lambda_k}^{-1} = \lambda_k$ . □

We next examine the structure of orthogonal transformations, in a fashion parallel to our study in §2.11 of skew-adjoint transformations on  $\mathbb{R}^n$ . Thus let

$$(2.12.9) \quad A : \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

be orthogonal, so

$$(2.12.10) \quad AA^* = I,$$

which for real matrices is equivalent to  $AA^t = I$ . Now we can consider

$$A : \mathbb{C}^n \longrightarrow \mathbb{C}^n,$$

given by the same matrix as in (2.12.9), a matrix with real entries. Thus the characteristic polynomial  $K_A(\lambda) = \det(\lambda I - A)$  is a polynomial of degree  $n$  with real coefficients, so its non-real roots occur in complex conjugate pairs. Thus the elements of  $\text{Spec}(A)$  other than  $\pm 1$  are given by

$$(2.12.11) \quad \text{Spec}^\#(A) = \{\omega_1, \dots, \omega_m, \overline{\omega}_1, \dots, \overline{\omega}_m\}, \quad \overline{\omega}_j = \omega_j^{-1},$$

with the various listed eigenvalues mutually distinct. For the sake of concreteness, say  $\text{Im } \omega_j > 0$  for each  $j \in \{1, \dots, m\}$ . By Proposition 2.12.2,  $\mathbb{C}^n$  has an orthonormal basis of eigenvectors of  $A$ , and of course each such basis element belongs to  $\mathcal{E}(A, \omega_j)$ , or to  $\mathcal{E}(A, \overline{\omega}_j)$ , for some  $j \in \{1, \dots, m\}$ , or to  $\mathcal{E}(A, 1)$  or  $\mathcal{E}(A, -1)$ . For each  $j \in \{1, \dots, m\}$ , let

$$(2.12.12) \quad \{v_{j1}, \dots, v_{j, d_j}\}$$

be an orthonormal basis of  $\mathcal{E}(A, \omega_j)$ . Say

$$(2.12.13) \quad v_{jk} = \xi_{jk} + i\eta_{jk}, \quad \xi_{jk}, \eta_{jk} \in \mathbb{R}^n.$$

Then we can take

$$(2.12.14) \quad \overline{v}_{jk} = \xi_{jk} - i\eta_{jk} \in \mathbb{C}^n,$$

and

$$(2.12.15) \quad \{\overline{v}_{j1}, \dots, \overline{v}_{j, d_j}\}$$

is an orthonormal basis of  $\mathcal{E}(A, \overline{\omega}_j)$ . Writing

$$(2.12.16) \quad \omega_j = c_j + is_j, \quad c_j, s_j \in \mathbb{R},$$

we have

$$(2.12.17) \quad \begin{aligned} A\xi_{jk} &= c_j\xi_{jk} - s_j\eta_{jk}, \\ A\eta_{jk} &= s_j\xi_{jk} + c_j\eta_{jk}, \end{aligned}$$

for  $1 \leq k \leq d_j$ . Note that

$$(2.12.18) \quad \text{Span}_{\mathbb{C}}\{\xi_{jk}, \eta_{jk} : 1 \leq k \leq d_j\} = \mathcal{E}(A, \omega_j) + \mathcal{E}(A, \overline{\omega}_j),$$

while we can also take

$$(2.12.19) \quad \text{Span}_{\mathbb{R}}\{\xi_{jk}, \eta_{jk} : 1 \leq k \leq d_j\} = \mathcal{H}(A, \omega_j) \subset \mathbb{R}^n,$$

a linear subspace of  $\mathbb{R}^n$ , of dimension  $2d_j$ .

Parallel to the arguments involving (2.11.17)–(2.11.20), we have that

$$(2.12.20) \quad \left\{ \frac{1}{\sqrt{2}}\xi_{jk}, \frac{1}{\sqrt{2}}\eta_{jk} : 1 \leq k \leq d_j, 1 \leq j \leq m \right\}$$

is an orthonormal set in  $\mathbb{R}^n$ , whose linear span over  $\mathbb{C}$  coincides with the span of all the eigenspaces of  $A$  with eigenvalues  $\neq \pm 1$ , in  $\mathbb{C}^n$ .

We have the following conclusion:

**Proposition 2.12.4.** *If  $A : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is orthogonal, then  $\mathbb{R}^n$  has an orthonormal basis in which the matrix representation of  $A$  consists of blocks*

$$(2.12.21) \quad \begin{pmatrix} c_j & s_j \\ -s_j & c_j \end{pmatrix}, \quad c_j^2 + s_j^2 = 1,$$

plus perhaps an identity matrix block, if  $\mathcal{E}(A, 1) \neq 0$ , and a block that is  $-I$ , if  $\mathcal{E}(A, -1) \neq 0$ .

EXAMPLE 1. Picking  $c, s \in \mathbb{R}$  such that  $c^2 + s^2 = 1$ , we see that

$$B = \begin{pmatrix} c & s \\ s & -c \end{pmatrix}$$

is orthogonal, with  $\det B = -1$ . Note that  $\text{Spec}(B) = \{1, -1\}$ . Thus there is an orthonormal basis of  $\mathbb{R}^2$  in which the matrix representation of  $B$  is

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

EXAMPLE 2. If  $A : \mathbb{R}^3 \rightarrow \mathbb{R}^3$  is orthogonal, then there is an orthonormal basis  $\{u_1, u_2, u_3\}$  of  $\mathbb{R}^3$  in which

$$(2.12.22) \quad A = \begin{pmatrix} c & -s & \\ s & c & \\ & & 1 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} c & -s & \\ s & c & \\ & & -1 \end{pmatrix},$$

depending on whether  $\det A = 1$  or  $\det A = -1$ . (Note we have switched signs on  $s$ , which is harmless. This lines our notation up with that used in §3.2 of Chapter 3.) Since  $c^2 + s^2 = 1$ , it follows that there is an angle  $\theta$ , uniquely determined up to an additive multiple of  $2\pi$ , such that

$$(2.12.23) \quad c = \cos \theta, \quad s = \sin \theta.$$

(See §1.1 of Chapter 1, and also §3.2 of Chapter 3.) If  $\det A = 1$  in (2.12.22) we say  $A$  is a rotation about the axis  $u_3$ , through an angle  $\theta$ .

---

**Exercises**

1. Let  $V$  be a real inner product space. Consider nonzero vectors  $u, v \in V$ . Show that the *angle*  $\theta$  between these vectors is uniquely defined by the formula

$$(u, v) = \|u\| \cdot \|v\| \cos \theta, \quad 0 \leq \theta \leq \pi.$$

Show that  $0 < \theta < \pi$  if and only if  $u$  and  $v$  are linearly independent. Show that

$$\|u + v\|^2 = \|u\|^2 + \|v\|^2 + 2\|u\| \cdot \|v\| \cos \theta.$$

This identity is known as the Law of Cosines.

For  $V$  as above,  $u, v, w \in V$ , we define the angle between the line segment from  $w$  to  $u$  and the line segment from  $w$  to  $v$  to be the angle between  $u - w$  and  $v - w$ . (We assume  $w \neq u$  and  $w \neq v$ .)

2. Take  $V = \mathbb{R}^2$ , with its standard orthonormal basis  $i = (1, 0)$ ,  $j = (0, 1)$ . Let

$$u = (1, 0), \quad v = (\cos \varphi, \sin \varphi), \quad 0 \leq \varphi < 2\pi.$$

Show that, according to the definition of Exercise 1, the angle  $\theta$  between  $u$  and  $v$  is given by

$$\theta = \begin{cases} \varphi & \text{if } 0 \leq \varphi \leq \pi, \\ 2\pi - \varphi & \text{if } \pi \leq \varphi < 2\pi. \end{cases}$$

3. Let  $V$  be a real inner product space and let  $R \in \mathcal{L}(V)$  be orthogonal. Show that if  $u, v \in V$  are nonzero and  $\tilde{u} = Ru$ ,  $\tilde{v} = Rv$ , then the angle between  $u$  and  $v$  is equal to the angle between  $\tilde{u}$  and  $\tilde{v}$ . Show that if  $\{e_j\}$  is an orthonormal basis of  $V$ , there exists an orthogonal transformation  $R$  on  $V$  such that  $Ru = \|u\|e_1$  and  $Rv$  is in the linear span of  $e_1$  and  $e_2$ .

4. Consider a triangle as in Figure 2.12.1. Show that

$$h = c \sin A,$$

and also

$$h = a \sin C.$$

Use these calculations to show that

$$\frac{\sin A}{a} = \frac{\sin C}{c} = \frac{\sin B}{b}.$$

This identity is known as the Law of Sines.

**Exercises on cross products**

Exercises 5–8 deal with cross products of vectors in  $\mathbb{R}^3$ .

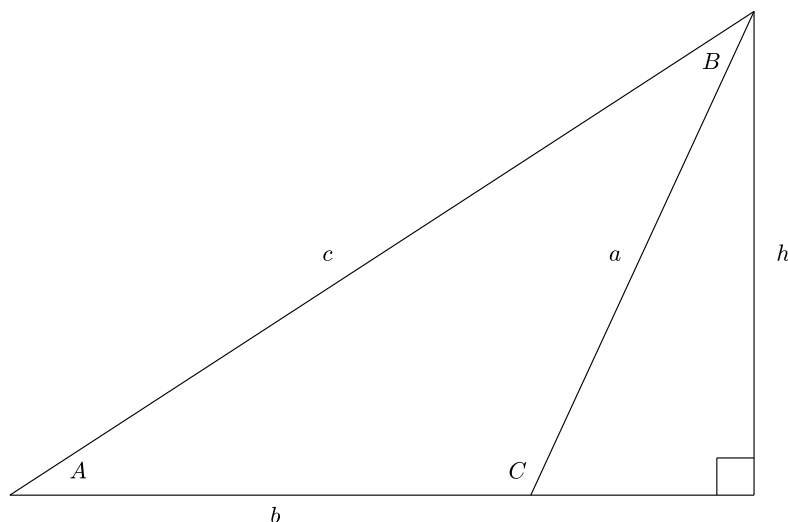


Figure 2.12.1. Law of sines

5. If  $u, v \in \mathbb{R}^3$ , show that the formula

$$(2.12.24) \quad w \cdot (u \times v) = \det \begin{pmatrix} w_1 & u_1 & v_1 \\ w_2 & u_2 & v_2 \\ w_3 & u_3 & v_3 \end{pmatrix}$$

for  $u \times v = \Pi(u, v)$  defines uniquely a bilinear map  $\Pi : \mathbb{R}^3 \times \mathbb{R}^3 \rightarrow \mathbb{R}^3$ . Show that it satisfies

$$i \times j = k, \quad j \times k = i, \quad k \times i = j,$$

where  $\{i, j, k\}$  is the standard basis of  $\mathbb{R}^3$ .

*Note.* To say  $\Pi$  is bilinear is to say  $\Pi(u, v)$  is linear in both  $u$  and  $v$ .

6. Recall that  $T \in SO(3)$  provided that  $T$  is a real  $3 \times 3$  matrix satisfying  $T^t T = I$  and  $\det T > 0$ , (hence  $\det T = 1$ ). Show that

$$(2.12.25) \quad T \in SO(3) \implies Tu \times Tv = T(u \times v).$$

*Hint.* Multiply the  $3 \times 3$  matrix in Exercise 5 on the left by  $T$ .

7. Show that, if  $\theta$  is the angle between  $u$  and  $v$  in  $\mathbb{R}^3$ , then

$$(2.12.26) \quad \|u \times v\| = \|u\| \cdot \|v\| \cdot |\sin \theta|.$$

*Hint.* Check (2.12.26) for  $u = i$ ,  $v = ai + bj$ , and use Exercise 6 to show this suffices.

8. More generally, show that for all  $u, v, w, x \in \mathbb{R}^3$ ,

$$(2.12.27) \quad (u \times v) \cdot (w \times x) = \det \begin{pmatrix} u \cdot w & v \cdot w \\ u \cdot x & v \cdot x \end{pmatrix}.$$

*Hint.* Using Exercise 6, show that it suffices to check this for

$$w = i, \quad x = ai + bj, \quad \text{so } w \times x = bk.$$

Then the left side of (2.12.27) is equal to

$$\begin{aligned} (u \times v) \cdot bk &= \det \begin{pmatrix} 0 & u \cdot i & v \cdot i \\ 0 & u \cdot j & v \cdot j \\ b & u \cdot k & v \cdot k \end{pmatrix} \\ &= b \det \begin{pmatrix} u \cdot i & v \cdot i \\ u \cdot j & v \cdot j \end{pmatrix} \\ &= \det \begin{pmatrix} u \cdot i & v \cdot i \\ u \cdot (ai + bj) & v \cdot (ai + bj) \end{pmatrix}, \end{aligned}$$

which is equal to the right side of (2.12.27).

9. Show that  $\kappa : \mathbb{R}^3 \rightarrow \text{Skew}(3)$ , the set of antisymmetric real  $3 \times 3$  matrices, given by

$$(2.12.28) \quad \kappa(y) = \begin{pmatrix} 0 & -y_3 & y_2 \\ y_3 & 0 & -y_1 \\ -y_2 & y_1 & 0 \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix},$$

satisfies

$$(2.12.29) \quad \kappa(y)x = y \times x.$$

Show that, with  $[A, B] = AB - BA$ ,

$$(2.12.30) \quad \begin{aligned} \kappa(x \times y) &= [\kappa(x), \kappa(y)], \\ \text{Tr}(\kappa(x)\kappa(y)^t) &= 2x \cdot y. \end{aligned}$$

10. Demonstrate the following result, which contains both Proposition 2.11.2 and Proposition 2.12.2. Let  $V$  be a finite dimensional inner product space. We say  $T : V \rightarrow V$  is *normal* provided  $T$  and  $T^*$  commute, i.e.,

$$(2.12.31) \quad TT^* = T^*T.$$

**Proposition.** *If  $V$  is a finite dimensional complex inner product space and  $T \in \mathcal{L}(V)$  is normal, then  $V$  has an orthonormal basis of eigenvectors of  $T$ .*

*Hint.* Write  $T = A + iB$ ,  $A$  and  $B$  self adjoint. Then (2.12.31)  $\Rightarrow AB = BA$ . Apply Exercise 3 of §2.11.

## 2.A. The Jordan canonical form

Let  $V$  be an  $n$ -dimensional complex vector space, and suppose  $T : V \rightarrow V$ . The following result gives the Jordan canonical form for  $T$ .

**Proposition 2.A.1.** *There is a basis of  $V$  with respect to which  $T$  is represented as a direct sum of blocks of the form*

$$(2.A.1) \quad \begin{pmatrix} \lambda_j & 1 & & \\ & \lambda_j & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_j \end{pmatrix}.$$

In light of Proposition 2.7.6 on generalized eigenspaces, together with Proposition 2.8.1 characterizing nilpotent operators and the discussion around (2.8.14), to prove Proposition 2.A.1 it suffices to establish such a Jordan canonical form for a nilpotent transformation  $N : V \rightarrow V$ . (Then  $\lambda_j = 0$ .) We turn to this task.

Given  $v_0 \in V$ , let  $m$  be the smallest integer such that  $N^m v_0 = 0$ ;  $m \leq n$ . If  $m = n$ , then  $\{v_0, Nv_0, \dots, N^{m-1}v_0\}$  gives a basis of  $V$  putting  $N$  in Jordan canonical form, with one block of the form (2.A.1) (with  $\lambda_j = 0$ ). In any case, we call  $\{v_0, \dots, N^{m-1}v_0\}$  a *string*. To obtain a Jordan canonical form for  $N$ , it will suffice to find a basis of  $V$  consisting of a family of strings. We will establish that this can be done by induction on  $\dim V$ . It is clear for  $\dim V \leq 1$ .

So, given a nilpotent  $N : V \rightarrow V$ , we can assume inductively that  $V_1 = N(V)$  has a basis that is a union of strings:

$$(2.A.2) \quad \{v_j, Nv_j, \dots, N^{\ell_j}v_j\}, \quad 1 \leq j \leq d.$$

Furthermore, each  $v_j$  has the form  $v_j = Nw_j$  for some  $w_j \in V$ . Hence we have the following strings in  $V$ :

$$(2.A.3) \quad \{w_j, v_j = Nw_j, Nv_j, \dots, N^{\ell_j}v_j\}, \quad 1 \leq j \leq d.$$

Note that the vectors in (2.A.3) are linearly independent. To see this, apply  $N$  to a linear combination and invoke the independence of the vectors in (2.A.2).

Now, pick a set  $\{\zeta_1, \dots, \zeta_\nu\} \subset V$  which, together with the vectors in (2.A.3) form a basis of  $V$ . Then each  $N\zeta_j$  can be written  $N\zeta_j = N\zeta'_j$  for some  $\zeta'_j$  in the linear span of the vectors in (2.A.3), so

$$(2.A.4) \quad z_1 = \zeta_1 - \zeta'_1, \dots, z_\nu = \zeta_\nu - \zeta'_\nu$$

also together with (2.A.3) forms a basis of  $V$ , and furthermore  $z_j \in \mathcal{N}(N)$ . Hence the strings

$$(2.A.5) \quad \{w_j, v_j, \dots, N^{\ell_j}v_j\}, \quad 1 \leq j \leq d, \quad \{z_1\}, \dots, \{z_\nu\}$$

provide a basis of  $V$ , giving  $N$  its Jordan canonical form.

There is some choice in producing bases putting  $T \in \mathcal{L}(V)$  in block form. So we ask, in what sense is the Jordan form canonical? The answer is that the sizes of the various blocks is independent of the choices made. To show this, again it suffices to consider the case of a nilpotent  $N : V \rightarrow V$ . Let  $\beta(k)$  denote the number of blocks of size  $k \times k$  in a Jordan decomposition of  $N$ , and let  $\beta = \sum_k \beta(k)$  denote

the total number of blocks. Note that  $\dim \mathcal{N}(N) = \beta$ . Also  $\dim \mathcal{N}(N^2)$  exceeds  $\dim \mathcal{N}(N)$  by  $\beta - \beta(1)$ . In fact, generally,

$$\begin{aligned} \dim \mathcal{N}(N) &= \beta, \\ \dim \mathcal{N}(N^2) &= \dim \mathcal{N}(N) + \beta - \beta(1), \\ (2.A.6) \quad &\vdots \\ \dim \mathcal{N}(N^{k+1}) &= \dim \mathcal{N}(N^k) + \beta - \beta(1) - \cdots - \beta(k). \end{aligned}$$

These identities specify  $\beta$  and then inductively each  $\beta(k)$  in terms of  $\dim \mathcal{N}(N^j)$ ,  $1 \leq j \leq k+1$ .



## 2.B. Schur's upper triangular representation

Let  $V$  be an  $n$ -dimensional complex vector space, equipped with an inner product, and let  $T \in \mathcal{L}(V)$ . The following is an important alternative to Proposition 2.A.1.

**Proposition 2.B.1.** *There is an orthonormal basis of  $V$  with respect to which  $T$  has an upper triangular form.*

Note that an upper triangular form with respect to some basis was achieved in (2.8.15), but there the basis was not guaranteed to be orthonormal. We will obtain Proposition 2.B.1 as a consequence of

**Proposition 2.B.2.** *There is a sequence of vector spaces  $V_j$  of dimension  $j$  such that*

$$(2.B.1) \quad V = V_n \supset V_{n-1} \supset \cdots \supset V_1$$

and

$$(2.B.2) \quad T : V_j \rightarrow V_j.$$

We show how Proposition 2.B.2 implies Proposition 2.B.1. In fact, given (2.B.1)–(2.B.2), pick  $u_n \perp V_{n-1}$ , a unit vector, then pick a unit  $u_{n-1} \in V_{n-1}$  such that  $u_{n-1} \perp V_{n-2}$ , and so forth, to achieve the conclusion of Proposition 2.B.1.

Meanwhile, Proposition 2.B.2 is a simple inductive consequence of the following result.

**Lemma 2.B.3.** *Given  $T \in \mathcal{L}(V)$  as above, there is a linear subspace  $V_{n-1}$ , of dimension  $n-1$ , such that  $T : V_{n-1} \rightarrow V_{n-1}$ .*

**Proof.** We apply Proposition 2.6.1 to  $T^*$  to obtain a nonzero  $v_1 \in V$  such that  $T^*v_1 = \lambda v_1$ , for some  $\lambda \in \mathbb{C}$ . Then the conclusion of Lemma 2.B.3 holds with  $V_{n-1} = (v_1)^\perp$ .  $\square$

## 2.C. The fundamental theorem of algebra

The following result is known as the fundamental theorem of algebra. It played a crucial role in §2.6, to guarantee the existence of eigenvalues of a complex  $n \times n$  matrix.

**Theorem 2.C.1.** *If  $p(z)$  is a nonconstant polynomial (with complex coefficients), then  $p(z)$  must have a complex root.*

**Proof.** We have, for some  $n \geq 1$ ,  $a_n \neq 0$ ,

$$(2.C.1) \quad \begin{aligned} p(z) &= a_n z^n + \cdots + a_1 z + a_0 \\ &= a_n z^n (1 + R(z)), \quad |z| \rightarrow \infty, \end{aligned}$$

where

$$|R(z)| \leq \frac{C}{|z|}, \quad \text{for } |z| \text{ large.}$$

This implies

$$(2.C.2) \quad \lim_{|z| \rightarrow \infty} |p(z)| = \infty.$$

Picking  $R \in (0, \infty)$  such that

$$(2.C.3) \quad \inf_{|z| \geq R} |p(z)| > |p(0)|,$$

we deduce that

$$(2.C.4) \quad \inf_{|z| \leq R} |p(z)| = \inf_{z \in \mathbb{C}} |p(z)|.$$

Since  $D_R = \{z : |z| \leq R\}$  is closed and bounded and  $p$  is continuous, there exists  $z_0 \in D_R$  such that

$$(2.C.5) \quad |p(z_0)| = \inf_{z \in \mathbb{C}} |p(z)|.$$

(For further discussion of this point, see Appendix 4.B of Chapter 4.) The theorem hence follows from:  $\square$

**Lemma 2.C.2.** *If  $p(z)$  is a nonconstant polynomial and (2.C.5) holds, then  $p(z_0) = 0$ .*

**Proof.** Suppose to the contrary that

$$(2.C.6) \quad p(z_0) = a \neq 0.$$

We can write

$$(2.C.7) \quad p(z_0 + \zeta) = a + q(\zeta),$$

where  $q(\zeta)$  is a (nonconstant) polynomial in  $\zeta$ , satisfying  $q(0) = 0$ . Hence, for some  $k \geq 1$  and  $b \neq 0$ , we have  $q(\zeta) = b\zeta^k + \cdots + b_n \zeta^n$ , i.e.,

$$(2.C.8) \quad q(\zeta) = b\zeta^k + \zeta^{k+1}r(\zeta), \quad |r(\zeta)| \leq C, \quad \zeta \rightarrow 0,$$

so, with  $\zeta = \varepsilon\omega$ ,  $\omega \in S^1 = \{\omega : |\omega| = 1\}$ ,

$$(2.C.9) \quad p(z_0 + \varepsilon\omega) = a + b\omega^k \varepsilon^k + (\varepsilon\omega)^{k+1}r(\varepsilon\omega), \quad \varepsilon \searrow 0.$$

Pick  $\omega \in S^1$  such that

$$(2.C.10) \quad \frac{b}{|b|}\omega^k = -\frac{a}{|a|},$$

which is possible since  $a \neq 0$  and  $b \neq 0$ . Then

$$(2.C.11) \quad p(z_0 + \varepsilon\omega) = a\left(1 - \left|\frac{b}{a}\right|\varepsilon^k\right) + (\varepsilon\omega)^{k+1}r(\varepsilon\omega),$$

with  $r(\zeta)$  as in (2.C.8), which contradicts (2.C.5) for  $\varepsilon > 0$  small enough. Thus (2.C.6) is impossible. This proves Lemma 2.C.2, hence Theorem 2.C.1.  $\square$

Now that we have shown that  $p(z)$  in (2.C.1) must have one root, we can show it has  $n$  roots (counting multiplicity).

**Proposition 2.C.3.** *For a polynomial  $p(z)$  of degree  $n$ , as in (2.C.1), there exist  $r_1, \dots, r_n \in \mathbb{C}$  such that*

$$(2.C.12) \quad p(z) = a_n(z - r_1) \cdots (z - r_n).$$

**Proof.** We have shown that  $p(z)$  has one root; call it  $r_1$ . Dividing  $p(z)$  by  $z - r_1$ , we have

$$(2.C.13) \quad p(z) = (z - r_1)\tilde{p}(z) + q,$$

where  $\tilde{p}(z) = a_n z^{n-1} + \cdots + \tilde{a}_0$  and  $q$  is a polynomial of degree  $< 1$ , i.e., a constant. Setting  $z = r_1$  in (2.C.13) yields  $q = 0$ , i.e.,

$$(2.C.14) \quad p(z) = (z - r_1)\tilde{p}(z).$$

Since  $\tilde{p}(z)$  is a polynomial of degree  $n - 1$ , the result (2.C.12) follows by induction on  $n$ .  $\square$

REMARK 1. The numbers  $r_j$ ,  $1 \leq j \leq n$ , in (2.C.12) are the roots of  $p(z)$ . If  $k$  of them coincide (say with  $r_\ell$ ), we say  $r_\ell$  is a root of multiplicity  $k$ . If  $r_\ell$  is distinct from  $r_j$  for all  $j \neq \ell$ , we say  $r_\ell$  is a simple root.

REMARK 2. In complex analysis texts, like [4] and [47], one can find proofs of the fundamental theorem of algebra that use more advanced techniques than the proof given above, and are shorter.

## Linear systems of differential equations

This chapter connects the linear algebra developed in Chapter 2 with Differential Equations. We define the matrix exponential in §3.1 and show how it produces the solution to first order systems of differential equations with constant coefficients. We show how the use of eigenvectors and generalized eigenvectors helps to compute matrix exponentials. In §3.2 we look again at connections between exponential and trigonometric functions, complementing results of Chapter 1, §1.1.

In §3.3 we discuss how to reduce a higher order differential equation to a first order system, and show how the “companion matrix” of a polynomial arises in doing this. We show in §3.4 how the matrix exponential allows us to write down an integral formula (Duhamel’s formula) for the solution to a non-homogeneous first order system, and illustrate how this in concert with the reduction process just mentioned, allows us to write down the solution to a non-homogeneous second order differential equation.

Section 3.5 discusses how to derive first order systems describing the behavior of simple circuits, consisting of resistors, inductors, and capacitors. Here we treat a more general class of circuits than done in Chapter 1, §1.13.

Section 3.6 deals with second order systems. While it is the case that second order  $n \times n$  systems can always be converted into first order  $(2n) \times (2n)$  systems, many such systems have special structure, worthy of separate study. Material on self adjoint transformations from Chapter 2 plays an important role in this section.

In §3.7 we discuss the Frenet-Serret equations, for a curve in three-dimensional Euclidean space. These equations involve the curvature and torsion of a curve, and also a frame field along the curve, called the Frenet frame, which forms an orthonormal basis of  $\mathbb{R}^3$  at each point on the curve. Regarding these equations as a system of differential equations, we discuss the problem of finding a curve with given curvature and torsion. Doing this brings in a number of topics from the

previous sections, and from Chapter 2, such as the use of properties of orthogonal matrices.

Having introduced equations with variable coefficients in §3.7, we concentrate on their treatment in subsequent sections. In §3.8 we study the solution operator  $S(t, s)$  to a homogeneous system, show how it extends the notion of matrix exponential, and extend Duhamel's formula to the variable coefficient setting. In §3.9 we show how the method of variation of parameters, introduced in Chapter 1, ties in with and becomes a special case of Duhamel's formula.

Section 3.10 treats power series expansions for a first order linear system with analytic coefficients, and §3.11 extends the study to equations with regular singular points. These sections provide a systematic treatment of material touched on in Chapter 1, §1.15. In these sections we use elementary power series techniques. Additional insight can be gleaned from the theory of functions of a complex variable. Readers who have seen some complex variable theory can consult [4], pp. 299–312, [19], pp. 70–83, or [47], Chapter 7, for material on this.

Appendix 3.A treats logarithms of matrices, a construction inverse to the matrix exponential introduced in §3.1, establishing results that are of use in §§3.8 and 3.11. Building on material from §1.18 of Chapter 1, Appendix 3.B develops the Laplace transform in the matrix setting, as a tool for solving nonhomogeneous linear systems. It also draws a connection between this method and Duhamel's formula. Appendix 3.C provides a brief introduction to the class of complex analytic functions, whose relevance for power series techniques in ODE was touched on in §3.10.

### 3.1. The matrix exponential

Here we discuss a key concept in matrix analysis, the matrix exponential. Given  $A \in M(n, \mathbb{F})$ ,  $\mathbb{F} = \mathbb{R}$  or  $\mathbb{C}$ , we define  $e^A$  by the same power series used in Chapter 1 to define  $e^A$  for  $A \in \mathbb{R}$ :

$$(3.1.1) \quad e^A = \sum_{k=1}^{\infty} \frac{1}{k!} A^k.$$

Note that  $A$  can be a real or complex  $n \times n$  matrix. In either case, recall from §2.10 of Chapter 2 that  $\|A^k\| \leq \|A\|^k$ . Hence the standard ratio test implies (3.1.1) is absolutely convergent for each  $A \in M(n, \mathbb{F})$ . Hence

$$(3.1.2) \quad e^{tA} = \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k$$

is a convergent power series in  $t$ , for all  $t \in \mathbb{R}$  (indeed for  $t \in \mathbb{C}$ ). As for all such convergent power series, we can differentiate term by term. We have

$$(3.1.3) \quad \begin{aligned} \frac{d}{dt} e^{tA} &= \sum_{k=1}^{\infty} k \frac{t^{k-1}}{k!} A^k \\ &= \sum_{k=1}^{\infty} \frac{t^{k-1}}{(k-1)!} A^{k-1} A. \end{aligned}$$

We can factor  $A$  out on either the left or the right, obtaining

$$(3.1.4) \quad \frac{d}{dt} e^{tA} = e^{tA} A = A e^{tA}.$$

Hence  $x(t) = e^{tA} x_0$  solves the first-order system

$$(3.1.5) \quad \frac{dx}{dt} = Ax, \quad x(0) = x_0.$$

This is the *unique* solution to (3.1.5). To see this, let  $x(t)$  be any solution to (3.1.5), and consider

$$(3.1.6) \quad u(t) = e^{-tA} x(t).$$

Then  $u(0) = x(0) = x_0$  and

$$(3.1.7) \quad \frac{d}{dt} u(t) = -e^{-tA} Ax(t) + e^{-tA} x'(t) = 0,$$

so  $u(t) \equiv u(0) = x_0$ . The same argument yields

$$(3.1.8) \quad \frac{d}{dt} (e^{tA} e^{-tA}) = 0, \quad \text{hence } e^{tA} e^{-tA} \equiv I.$$

Hence  $x(t) = e^{tA} x_0$ , as asserted.

Using a variant of the computation (3.1.7), we show that the matrix exponential has the following property, which generalizes the identity  $e^{s+t} = e^s e^t$  for real  $s, t$ , established in Chapter 1.

**Proposition 3.1.1.** *Given  $A \in M(n, \mathbb{C})$ ,  $s, t \in \mathbb{R}$ ,*

$$(3.1.9) \quad e^{(s+t)A} = e^{sA} e^{tA}.$$

**Proof.** Using the Leibniz formula for the derivative of a product, plus (3.1.4), we have

$$(3.1.10) \quad \frac{d}{dt} \left( e^{(s+t)A} e^{-tA} \right) = e^{(s+t)A} A e^{-tA} - e^{(s+t)A} A e^{-tA} = 0.$$

Hence  $e^{(s+t)A} e^{-tA}$  is independent of  $t$ , so

$$(3.1.11) \quad e^{(s+t)A} e^{-tA} = e^{sA}, \quad \forall s, t \in \mathbb{R}.$$

Taking  $s = 0$  yields  $e^{tA} e^{-tA} = I$  (as we have already seen in (3.1.8)) or  $e^{-tA} = (e^{tA})^{-1}$ , so we can multiply both sides of (3.1.11) on the right by  $e^{tA}$  and obtain (3.1.9).  $\square$

Now, generally, for  $A, B \in M(n, \mathbb{F})$ ,

$$(3.1.12) \quad e^A e^B \neq e^{A+B}.$$

However, we do have the following.

**Proposition 3.1.2.** *Given  $A, B \in M(n, \mathbb{C})$ ,*

$$(3.1.13) \quad AB = BA \implies e^{A+B} = e^A e^B.$$

**Proof.** We compute

$$(3.1.14) \quad \begin{aligned} & \frac{d}{dt} \left( e^{t(A+B)} e^{-tB} e^{-tA} \right) \\ &= e^{t(A+B)} (A+B) e^{-tB} e^{-tA} - e^{t(A+B)} B e^{-tB} e^{-tA} - e^{t(A+B)} e^{-tB} A e^{-tA}. \end{aligned}$$

Now  $AB = BA \implies AB^k = B^k A$ , hence

$$(3.1.15) \quad e^{-tB} A = \sum_{k=0}^{\infty} \frac{(-t)^k}{k!} B^k A = A e^{-tB},$$

so (3.1.14) vanishes. Hence  $e^{t(A+B)} e^{-tB} e^{-tA}$  is independent of  $t$ , so

$$(3.1.16) \quad e^{t(A+B)} e^{-tB} e^{-tA} = I,$$

the value at  $t = 0$ . Multiplying through on the right by  $e^{tA}$  and then by  $e^{tB}$  gives

$$(3.1.17) \quad e^{t(A+B)} = e^{tA} e^{tB}.$$

Setting  $t = 1$  gives (3.1.13).  $\square$

We now look at examples of matrix exponentials. We start with some computations via the infinite series (3.1.2). Take

$$(3.1.18) \quad A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Then

$$(3.1.19) \quad A^k = \begin{pmatrix} 1 & 0 \\ 0 & 2^k \end{pmatrix}, \quad B^2 = B^3 = \dots = 0,$$

so

$$(3.1.20) \quad e^{tA} = \begin{pmatrix} e^t & 0 \\ 0 & e^{2t} \end{pmatrix}, \quad e^{tB} = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}.$$

Note that  $A$  and  $B$  do not commute, and neither do  $e^{tA}$  and  $e^{tB}$ , for general  $t \neq 0$ . On the other hand, if we take

$$(3.1.21) \quad C = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = I + B,$$

since  $I$  and  $B$  commute, we have without further effort that

$$(3.1.22) \quad e^{tC} = e^{tI}e^{tB} = \begin{pmatrix} e^t & te^t \\ 0 & e^t \end{pmatrix}.$$

We turn to constructions of matrix exponentials via use of eigenvalues and eigenvectors. Suppose  $v_j$  is an eigenvector of  $A$  with eigenvalue  $\lambda_j$ ,

$$(3.1.23) \quad Av_j = \lambda_j v_j.$$

Then  $A^k v_j = \lambda_j^k v_j$ , and hence

$$(3.1.24) \quad e^{tA} v_j = \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k v_j = \sum_{k=0}^{\infty} \frac{t^k}{k!} \lambda_j^k v_j = e^{t\lambda_j} v_j.$$

This enables us to construct  $e^{tA}v$  for each  $v \in \mathbb{C}^n$  if  $A \in M(n, \mathbb{C})$  and  $\mathbb{C}^n$  has a basis of eigenvectors,  $\{v_j : 1 \leq j \leq n\}$ . In such a case, write  $v$  as a linear combination of the eigenvectors,

$$(3.1.25) \quad v = c_1 v_1 + \cdots + c_n v_n,$$

and then

$$(3.1.26) \quad \begin{aligned} e^{tA}v &= c_1 e^{tA}v_1 + \cdots + c_n e^{tA}v_n \\ &= c_1 e^{t\lambda_1}v_1 + \cdots + c_n e^{t\lambda_n}v_n. \end{aligned}$$

We illustrate this process with some examples.

EXAMPLE 1. Take

$$(3.1.27) \quad A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

One has  $\det(\lambda I - A) = \lambda^2 - 1$ , hence eigenvalues

$$(3.1.28) \quad \lambda_1 = 1, \quad \lambda_2 = -1,$$

with corresponding eigenvectors

$$(3.1.29) \quad v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad v_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Hence

$$(3.1.30) \quad e^{tA}v_1 = e^t v_1, \quad e^{tA}v_2 = e^{-t} v_2.$$

To write out  $e^{tA}$  as a  $2 \times 2$  matrix, note that the first and second columns of this matrix are given respectively by

$$(3.1.31) \quad e^{tA} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad e^{tA} \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$



To compute this, we write  $(1, 0)^t$  and  $(0, 1)^t$  as linear combinations of the eigenvectors. We have

$$(3.1.32) \quad \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Hence

$$(3.1.33) \quad \begin{aligned} e^{tA} \begin{pmatrix} 1 \\ 0 \end{pmatrix} &= \frac{1}{2} e^{tA} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{1}{2} e^{tA} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \\ &= \frac{e^t}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{e^{-t}}{2} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{2}(e^t + e^{-t}) \\ \frac{1}{2}(e^t - e^{-t}) \end{pmatrix}, \end{aligned}$$

and similarly

$$(3.1.34) \quad e^{tA} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{2}(e^t - e^{-t}) \\ \frac{1}{2}(e^t + e^{-t}) \end{pmatrix}.$$

Recalling that

$$(3.1.35) \quad \cosh t = \frac{e^t + e^{-t}}{2}, \quad \sinh t = \frac{e^t - e^{-t}}{2},$$

we have

$$(3.1.36) \quad e^{tA} = \begin{pmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{pmatrix}.$$

EXAMPLE 2. Take

$$(3.1.37) \quad A = \begin{pmatrix} 0 & -2 \\ 1 & 2 \end{pmatrix}.$$

One has  $\det(\lambda I - A) = \lambda^2 - 2\lambda + 2$ , hence eigenvalues

$$(3.1.38) \quad \lambda_1 = 1 + i, \quad \lambda_2 = 1 - i,$$

with corresponding eigenvectors

$$(3.1.39) \quad v_1 = \begin{pmatrix} -2 \\ 1 + i \end{pmatrix}, \quad v_2 = \begin{pmatrix} -2 \\ 1 - i \end{pmatrix}.$$

We have

$$(3.1.40) \quad e^{tA} v_1 = e^{(1+i)t} v_1, \quad e^{tA} v_2 = e^{(1-i)t} v_2.$$

We can write

$$(3.1.41) \quad \begin{aligned} \begin{pmatrix} 1 \\ 0 \end{pmatrix} &= -\frac{i+1}{4} \begin{pmatrix} -2 \\ 1+i \end{pmatrix} + \frac{i-1}{4} \begin{pmatrix} -2 \\ 1-i \end{pmatrix}, \\ \begin{pmatrix} 0 \\ 1 \end{pmatrix} &= -\frac{i}{2} \begin{pmatrix} -2 \\ 1+i \end{pmatrix} + \frac{i}{2} \begin{pmatrix} -2 \\ 1-i \end{pmatrix}, \end{aligned}$$

to obtain

$$(3.1.42) \quad \begin{aligned} e^{tA} \begin{pmatrix} 1 \\ 0 \end{pmatrix} &= -\frac{i+1}{4} e^{(1+i)t} \begin{pmatrix} -2 \\ 1+i \end{pmatrix} + \frac{i-1}{4} e^{(1-i)t} \begin{pmatrix} -2 \\ 1-i \end{pmatrix} \\ &= \frac{e^t}{4} \begin{pmatrix} (2i+2)e^{it} + (2-2i)e^{-it} \\ -2ie^{it} + 2ie^{-it} \end{pmatrix}, \end{aligned}$$

and

$$(3.1.43) \quad \begin{aligned} e^{tA} \begin{pmatrix} 0 \\ 1 \end{pmatrix} &= -\frac{i}{2} e^{(1+i)t} \begin{pmatrix} -2 \\ 1+i \end{pmatrix} + \frac{i}{2} e^{(1-i)t} \begin{pmatrix} -2 \\ 1-i \end{pmatrix} \\ &= \frac{e^t}{2} \begin{pmatrix} 2ie^{it} - 2ie^{-it} \\ (1-i)e^{it} + (1+i)e^{-it} \end{pmatrix}. \end{aligned}$$

We can write these in terms of trigonometric functions, using the fundamental Euler identities

$$(3.1.44) \quad e^{it} = \cos t + i \sin t, \quad e^{-it} = \cos t - i \sin t,$$

established in §1.1 of Chapter 1. (See §3.2 of this chapter for more on this.) These yield

$$(3.1.45) \quad \cos t = \frac{e^{it} + e^{-it}}{2}, \quad \sin t = \frac{e^{it} - e^{-it}}{2i},$$

and an inspection of the formulas above gives

$$(3.1.46) \quad e^{tA} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = e^t \begin{pmatrix} \cos t - \sin t \\ \sin t \end{pmatrix}, \quad e^{tA} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = e^t \begin{pmatrix} -2 \sin t \\ \cos t + \sin t \end{pmatrix},$$

hence

$$(3.1.47) \quad e^{tA} = e^t \begin{pmatrix} \cos t - \sin t & -2 \sin t \\ \sin t & \cos t + \sin t \end{pmatrix}.$$

As was shown in Chapter 2, §2.6, if  $A \in M(n, \mathbb{C})$  has  $n$  distinct eigenvalues, then  $\mathbb{C}^n$  has a basis of eigenvectors. If  $A$  has multiple eigenvalues,  $\mathbb{C}^n$  might or might not have a basis of eigenvectors, though as shown in §2.7 of Chapter 2, there will be a basis of generalized eigenvectors. If  $v$  is a generalized eigenvector of  $A$ , say

$$(3.1.48) \quad (A - \lambda I)^m v = 0,$$

then

$$(3.1.49) \quad e^{t(A-\lambda I)} v = \sum_{k < m} \frac{t^k}{k!} (A - \lambda I)^k v,$$

so

$$(3.1.50) \quad e^{tA} v = e^{t\lambda} \sum_{k < m} \frac{t^k}{k!} (A - \lambda I)^k v.$$

EXAMPLE 3. Consider the  $3 \times 3$  matrix  $A$  used in (2.7.28) of Chapter 2:

$$(3.1.51) \quad A = \begin{pmatrix} 2 & 3 & 3 \\ 0 & 2 & 3 \\ 0 & 0 & 1 \end{pmatrix}.$$

Here 2 is a double eigenvalue and 1 a simple eigenvalue. Calculations done in Chapter 2, §2.7, yield

$$(3.1.52) \quad (A - 2I) \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = 0, \quad (A - 2I) \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = 3 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad (A - I) \begin{pmatrix} 6 \\ -3 \\ 1 \end{pmatrix} = 0.$$

Hence

$$(3.1.53) \quad e^{tA} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} e^{2t} \\ 0 \\ 0 \end{pmatrix},$$

$$(3.1.54) \quad e^{tA} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = e^{2t} \sum_{k=0}^{\infty} \frac{t^k}{k!} (A - 2I)^k \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

$$(3.1.55) \quad = e^{2t} \left[ \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 3t \\ 0 \\ 0 \end{pmatrix} \right],$$

and

$$(3.1.56) \quad e^{tA} \begin{pmatrix} 6 \\ -3 \\ 1 \end{pmatrix} = e^t \begin{pmatrix} 6 \\ -3 \\ 1 \end{pmatrix}.$$

Note that

$$(3.1.57) \quad e^{tA} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = e^{tA} \begin{pmatrix} 6 \\ -3 \\ 1 \end{pmatrix} - 6e^{tA} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + 3e^{tA} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

Putting these calculations together yields

$$(3.1.58) \quad e^{tA} = \begin{pmatrix} e^{2t} & 3te^{2t} & 6e^t - 6e^{2t} + 9te^{2t} \\ 0 & e^{2t} & -3e^t + 3e^{2t} \\ 0 & 0 & e^t \end{pmatrix}.$$

EXAMPLE 4. Consider the  $3 \times 3$  matrix

$$(3.1.59) \quad A = \begin{pmatrix} 1 & 2 & 0 \\ 3 & 1 & 3 \\ 0 & -2 & 1 \end{pmatrix}.$$

A computation gives  $\det(\lambda I - A) = (\lambda - 1)^3$ . Hence for  $N = A - I$  we have  $\text{Spec}(N) = \{0\}$ , so we know  $N$  is nilpotent (by Proposition 2.8.1 of Chapter 2). In fact, a calculation gives

$$(3.1.60) \quad N = \begin{pmatrix} 0 & 2 & 0 \\ 3 & 0 & 3 \\ 0 & -2 & 0 \end{pmatrix}, \quad N^2 = \begin{pmatrix} 6 & 0 & 6 \\ 0 & 0 & 0 \\ -6 & 0 & -6 \end{pmatrix}, \quad N^3 = 0.$$

Hence

$$(3.1.61) \quad \begin{aligned} e^{tA} &= e^t \left[ I + tN + \frac{t^2}{2} N^2 \right] \\ &= e^t \begin{pmatrix} 1 + 3t^2 & 2t & +3t^2 \\ 3t & 1 & 3t \\ -3t^2 & -2t & 1 - 3t^2 \end{pmatrix}. \end{aligned}$$

---

### Exercises

1. Use the method of eigenvalues and eigenvectors given in (3.1.23)–(3.1.26) to compute  $e^{tA}$  for each of the following:

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

2. Use the method given in (3.1.48)–(3.1.50) and illustrated in (3.1.51)–(3.1.61) to compute  $e^{tA}$  for each of the following:

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 2 & 4 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ -1 & 0 & -1 \end{pmatrix}.$$

3. Show that

$$e^{t(A+B)} = e^{tA} e^{tB}, \quad \forall t \implies AB = BA.$$

*Hint.* Set  $X(t) = e^{t(A+B)}$ ,  $Y(t) = e^{tA} e^{tB}$ . Show that

$$X \equiv Y \implies X'(t) - Y'(t) = B e^{t(A+B)} - e^{tA} B e^{tB} \equiv 0,$$

and hence that

$$X \equiv Y \implies B e^{tA} = e^{tA} B, \quad \forall t.$$

4. Given  $A \in M(n, \mathbb{C})$ , suppose  $\Phi(t)$  is an  $n \times n$  matrix valued solution to

$$\frac{d}{dt} \Phi(t) = A \Phi(t).$$

Show that

$$\Phi(t) = e^{tA} B,$$

where  $B = \Phi(0)$ . Deduce that  $\Phi(t)$  is invertible for all  $t \in \mathbb{R}$  if and only if  $\Phi(0)$  is invertible, and that in such a case

$$e^{(t-s)A} = \Phi(t) \Phi(s)^{-1}.$$

(For a generalization, see (3.8.13).)

5. Let  $A, B \in M(n, \mathbb{C})$  and assume  $B$  is invertible. Show that

$$(B^{-1}AB)^k = B^{-1}A^k B,$$

and use this to show that

$$e^{tB^{-1}AB} = B^{-1}e^{tA}B.$$

6. Show that if  $A$  is diagonal, i.e.,

$$A = \begin{pmatrix} a_{11} & & \\ & \ddots & \\ & & a_{nn} \end{pmatrix},$$

then

$$e^{tA} = \begin{pmatrix} e^{ta_{11}} & & \\ & \ddots & \\ & & e^{ta_{nn}} \end{pmatrix}.$$

Exercises 7–10 bear on the identity

$$(3.1.62) \quad \det e^{tA} = e^{t \operatorname{Tr} A},$$

given  $A \in M(n, \mathbb{C})$ .

7. Show that if (3.1.62) holds for  $A = A_1$  and if  $A_2 = B^{-1}A_1B$ , then (3.1.62) holds for  $A = A_2$ .

8. Show that (3.1.62) holds whenever  $A$  is diagonalizable.

*Hint.* Use Exercises 5–6.

9. Assume  $A \in M(n, \mathbb{C})$  is upper triangular:

$$(3.1.63) \quad A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ & \ddots & \vdots \\ & & a_{nn} \end{pmatrix}.$$

Show that  $e^{tA}$  is upper triangular, of the form

$$e^{tA} = \begin{pmatrix} e_{11}(t) & \cdots & e_{1n}(t) \\ & \ddots & \vdots \\ & & e_{nn}(t) \end{pmatrix}, \quad e_{jj}(t) = e^{ta_{jj}}.$$

10. Deduce that (3.1.62) holds when  $A$  has the form (3.1.63). Then deduce that (3.1.62) holds for all  $A \in M(n, \mathbb{C})$ .

11. Let  $A(t)$  be a smooth function of  $t$  with values in  $M(n, \mathbb{C})$ . Show that

$$(3.1.64) \quad A(0) = 0 \implies \left. \frac{d}{dt} e^{A(t)} \right|_{t=0} = A'(0).$$

*Hint.* Take the power series expansion of  $e^{A(t)}$ , in powers of  $A(t)$ .

12. Let  $A(t)$  be a smooth  $M(n, \mathbb{C})$ -valued function of  $t \in I$  and assume

$$(3.1.65) \quad A(s)A(t) = A(t)A(s), \quad \forall s, t \in I.$$

Show that

$$(3.1.66) \quad \frac{d}{dt}e^{A(t)} = A'(t)e^{A(t)} = e^{A(t)}A'(t).$$

*Hint.* Show that if (3.1.65) holds,

$$\frac{d}{dt}e^{A(t)} = \frac{d}{ds}e^{A(s)-A(t)}e^{A(t)} \Big|_{s=t},$$

and apply Exercise 11.

13. Here is an alternative approach to Proposition 3.1.2. Assume

$$(3.1.67) \quad A, B \in M(n, \mathbb{C}), \quad AB = BA.$$

Show that

$$(3.1.68) \quad (A + B)^m = \sum_{j=0}^m \binom{m}{j} A^j B^{m-j}, \quad \binom{m}{j} = \frac{m!}{j!(m-j)!}.$$

From here, show that

$$(3.1.69) \quad \begin{aligned} e^{A+B} &= \sum_{m=0}^{\infty} \frac{1}{m!} (A + B)^m \\ &= \sum_{m=0}^{\infty} \sum_{j=0}^m \frac{1}{j!(m-j)!} A^j B^{m-j}. \end{aligned}$$

Then take  $n = m - j$  and show this is

$$(3.1.70) \quad \begin{aligned} &= \sum_{j=0}^{\infty} \sum_{n=0}^{\infty} \frac{1}{j!n!} A^j B^n \\ &= \sum_{j=0}^{\infty} \frac{1}{j!} A^j \sum_{n=0}^{\infty} \frac{1}{n!} B^n \\ &= e^A e^B, \end{aligned}$$

so

$$(3.1.71) \quad e^{A+B} = e^A e^B.$$

14. As an alternative to the proof of (3.1.4), given in (3.1.3), which depends on term by term differentiation of power series, verify that, for  $A \in M(n, \mathbb{C})$ ,

$$(3.1.72) \quad \begin{aligned} \frac{d}{dt}e^{tA} &= \lim_{h \rightarrow 0} \frac{1}{h} (e^{(t+h)A} - e^{tA}) \\ &= e^{tA} \lim_{h \rightarrow 0} \frac{1}{h} (e^{hA} - I) \\ &= e^{tA} A \\ &= A e^{tA}, \end{aligned}$$

the second identity in (3.1.72) by (3.1.71), the third by the definition (3.1.2), and the fourth by commutativity.

### 3.2. Exponentials and trigonometric functions

In Chapter 1 we have seen how to use complex exponentials to give a self-contained treatment of basic results on the trigonometric functions  $\cos t$  and  $\sin t$ . Here we present a variant, using matrix exponentials. We begin by looking at

$$(3.2.1) \quad x(t) = e^{tJ}x_0, \quad J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

which solves

$$(3.2.2) \quad x'(t) = Jx(t), \quad x(0) = x_0 \in \mathbb{R}^2.$$

We first note that the planar curve  $x(t)$  moves about on a circle centered about the origin. Indeed,

$$(3.2.3) \quad \begin{aligned} \frac{d}{dt}\|x(t)\|^2 &= \frac{d}{dt}(x(t) \cdot x(t)) = x'(t) \cdot x(t) + x(t) \cdot x'(t) \\ &= Jx(t) \cdot x(t) + x(t) \cdot Jx(t) \\ &= 0, \end{aligned}$$

since  $J^t = -J$ . Thus  $\|x(t)\| = \|x_0\|$  is constant. Furthermore the velocity  $v(t) = x'(t)$  has constant magnitude; in fact

$$(3.2.4) \quad \|v(t)\|^2 = v(t) \cdot v(t) = Jx(t) \cdot Jx(t) = \|x(t)\|^2,$$

since  $J^t J = -J^2 = I$ .

For example,

$$(3.2.5) \quad \begin{pmatrix} c(t) \\ s(t) \end{pmatrix} = e^{tJ} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

is a curve, moving on the unit circle  $x_1^2 + x_2^2 = 1$ , at unit speed, with initial position  $x(0) = (1, 0)^t$  and initial velocity  $v(0) = (0, 1)^t$ . Now in trigonometry the functions  $\cos t$  and  $\sin t$  are defined to be the  $x_1$  and  $x_2$  coordinates of such a parametrization of the unit circle, so we have

$$(3.2.6) \quad \begin{pmatrix} \cos t \\ \sin t \end{pmatrix} = e^{tJ} \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

The differential equation (3.2.2) then gives

$$(3.2.7) \quad \frac{d}{dt} \cos t = -\sin t, \quad \frac{d}{dt} \sin t = \cos t.$$

Using

$$e^{tJ} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = e^{tJ} J \begin{pmatrix} 1 \\ 0 \end{pmatrix} = J e^{tJ} \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

we have a formula for  $e^{tJ} \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ , which together with (3.2.6) yields

$$(3.2.8) \quad e^{tJ} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix} = (\cos t)I + (\sin t)J.$$

Then the identity  $e^{(s+t)J} = e^{sJ} e^{tJ}$  yields the following identities, when matrix multiplication is carried out:

$$(3.2.9) \quad \begin{aligned} \cos(s+t) &= (\cos s)(\cos t) - (\sin s)(\sin t), \\ \sin(s+t) &= (\cos s)(\sin t) + (\sin s)(\cos t). \end{aligned}$$



We now show how the treatment of  $\sin t$  and  $\cos t$  presented above is really quite close to that given in Chapter 1, §1.1. To start, we note that if  $\mathbb{C}$  is regarded as a real vector space, with basis  $e_1 = 1$ ,  $e_2 = i$ , and hence identified with  $\mathbb{R}^2$ , via

$$(3.2.10) \quad z = x + iy \leftrightarrow \begin{pmatrix} x \\ y \end{pmatrix},$$

then the matrix representation for the linear transformation  $z \mapsto iz$  is given by  $J$ :

$$(3.2.11) \quad iz = -y + ix, \quad J \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -y \\ x \end{pmatrix}.$$

More generally, the linear transformation  $z \mapsto (c + is)z$  has matrix representation

$$(3.2.12) \quad \begin{pmatrix} c & -s \\ s & c \end{pmatrix}.$$

Taking this into account, we see that the identity (3.2.8) is equivalent to

$$(3.2.13) \quad e^{it} = \cos t + i \sin t,$$

which is Euler's formula, as in (1.1.39) of Chapter 1.

Here is another approach to the evaluation of  $e^{tJ}$ . We compute the eigenvalues and eigenvectors of  $J$ :

$$(3.2.14) \quad \lambda_1 = i, \quad \lambda_2 = -i; \quad v_1 = \begin{pmatrix} 1 \\ -i \end{pmatrix}, \quad v_2 = \begin{pmatrix} 1 \\ i \end{pmatrix}.$$

Then, using the fact that  $e^{tJ}v_k = e^{t\lambda_k}v_k$ , we have

$$(3.2.15) \quad e^{tJ} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{2}e^{it} \begin{pmatrix} 1 \\ -i \end{pmatrix} + \frac{1}{2}e^{-it} \begin{pmatrix} 1 \\ i \end{pmatrix}.$$

Comparison with (3.2.6) gives

$$(3.2.16) \quad \cos t = \frac{1}{2}(e^{it} + e^{-it}), \quad \sin t = \frac{1}{2i}(e^{it} - e^{-it}),$$

again leading to (3.2.13).

### Exercises

1. Recall  $\text{Skew}(n)$  and  $SO(n)$ , defined by (2.11.7) and (2.12.4) of Chapter 2. Show that

$$(3.2.17) \quad A \in \text{Skew}(n) \implies e^{tA} \in SO(n), \quad \forall t \in \mathbb{R}.$$

Note how this generalizes (3.2.3).

2. Given an  $n \times n$  matrix  $A$ , let us set

$$(3.2.18) \quad \cos tA = \frac{1}{2}(e^{itA} + e^{-itA}), \quad \sin tA = \frac{1}{2i}(e^{itA} - e^{-itA}).$$

Show that

$$(3.2.19) \quad \frac{d}{dt} \cos tA = -A \sin tA, \quad \frac{d}{dt} \sin tA = A \cos tA.$$

3. In the context of Exercise 2, show that

$$(3.2.20) \quad \cos tA = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k)!} (tA)^{2k}, \quad \sin tA = \sum_{k=0}^{\infty} \frac{(-1)^k}{(2k+1)!} (tA)^{2k+1}.$$

4. Show that

$$Av = \lambda v \implies (\cos tA)v = (\cos t\lambda)v, \\ (\sin tA)v = (\sin t\lambda)v.$$

5. Compute  $\cos tA$  and  $\sin tA$  in each of the following cases:

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}.$$

6. Suppose  $A \in M(n, \mathbb{C})$  and

$$B = \begin{pmatrix} 0 & -A \\ A & 0 \end{pmatrix} \in M(2n, \mathbb{C}).$$

Show that

$$e^{tB} = \begin{pmatrix} \cos tA & -\sin tA \\ \sin tA & \cos tA \end{pmatrix}.$$

### 3.3. First-order systems derived from higher-order equations

There is a standard process to convert an  $n$ th order differential equation

$$(3.3.1) \quad \frac{d^n y}{dt^n} + a_{n-1} \frac{d^{n-1} y}{dt^{n-1}} + \cdots + a_1 \frac{dy}{dt} + a_0 y = 0$$

to a first-order system. Set

$$(3.3.2) \quad x_0(t) = y(t), \quad x_1(t) = y'(t), \dots, \quad x_{n-1}(t) = y^{(n-1)}(t).$$

Then  $x = (x_0, \dots, x_{n-1})^t$  satisfies

$$(3.3.3) \quad \begin{aligned} x'_0 &= x_1 \\ &\vdots \\ x'_{n-2} &= x_{n-1} \\ x'_{n-1} &= -a_{n-1}x_{n-1} - \cdots - a_0x_0, \end{aligned}$$

or equivalently

$$(3.3.4) \quad \frac{dx}{dt} = Ax,$$

with

$$(3.3.5) \quad A = \begin{pmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ -a_0 & -a_1 & \cdots & -a_{n-2} & -a_{n-1} \end{pmatrix}.$$

The matrix  $A$  given by (3.3.5) is called the *companion matrix* of the polynomial

$$(3.3.6) \quad p(\lambda) = \lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0.$$

Note that a direct search of solutions to (3.3.1) of the form  $e^{\lambda t}$  leads one to solve  $p(\lambda) = 0$ . Thus the following result is naturally suggested.

**Proposition 3.3.1.** *If  $p(\lambda)$  is a polynomial of the form (3.3.6), with companion matrix  $A$ , given by (3.3.5), then*

$$(3.3.7) \quad p(\lambda) = \det(\lambda I - A).$$

**Proof.** We look at

$$(3.3.8) \quad \lambda I - A = \begin{pmatrix} \lambda & -1 & \cdots & 0 & 0 \\ 0 & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda & -1 \\ a_0 & a_1 & \cdots & a_{n-2} & \lambda + a_{n-1} \end{pmatrix},$$

and compute its determinant by expanding by minors down the first column. We see that

$$(3.3.9) \quad \det(\lambda I - A) = \lambda \det(\lambda I - \tilde{A}) + (-1)^{n-1} a_0 \det B,$$

where

$$(3.3.10) \quad \begin{aligned} \tilde{A} & \text{ is the companion matrix of } \lambda^{n-1} + a_{n-1}\lambda^{n-2} + \cdots + a_1, \\ B & \text{ is lower triangular, with } -1\text{'s on the diagonal.} \end{aligned}$$

By induction on  $n$ , we have  $\det(\lambda I - \tilde{A}) = \lambda^{n-1} + a_{n-1}\lambda^{n-2} + \cdots + a_1$ , while  $\det B = (-1)^{n-1}$ . Substituting this into (3.3.9) gives (3.3.7).  $\square$

### Converse construction

We next show that each solution to a first order  $n \times n$  system of the form (3.3.4) (for general  $A \in M(n, \mathbb{F})$ ) also satisfies an  $n$ th order scalar ODE. Indeed, if (3.3.4) holds, then

$$(3.3.11) \quad x^{(k)} = Ax^{(k-1)} = \cdots = A^k x.$$

Now if  $p(\lambda)$  is given by (3.3.7), and say

$$(3.3.12) \quad p(\lambda) = \lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0,$$

then, by the Cayley-Hamilton theorem (cf. (2.8.17) of Chapter 2),

$$(3.3.13) \quad p(A) = A^n + a_{n-1}A^{n-1} + \cdots + a_1A + a_0I = 0.$$

Hence

$$(3.3.14) \quad \begin{aligned} x^{(n)} & = A^n x \\ & = -a_{n-1}A^{n-1}x - \cdots - a_1Ax - a_0x \\ & = -a_{n-1}x^{(n-1)} - \cdots - a_1x' - a_0x, \end{aligned}$$

so we have the asserted  $n$ th order scalar equation:

$$(3.3.15) \quad x^{(n)} + a_{n-1}x^{(n-1)} + \cdots + a_1x' + a_0x = 0.$$

REMARK. If the minimal polynomial  $q(\lambda)$  of  $A$  has degree  $m$ , less than  $n$ , we can replace  $p$  by  $q$  and derive analogues of (3.3.14)–(3.3.15), giving a single differential equation of degree  $m$  for  $x$ .

---

## Exercises

- Using the method (3.3.12)–(3.3.15), convert

$$\frac{dx}{dt} = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix} x$$

into a second order scalar equation.

2. Using the method (3.3.2)–(3.3.3), convert

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

into a  $2 \times 2$  first order system.

In Exercises 3–4, assume that  $\lambda_1$  is a root of multiplicity  $k \geq 2$  for the polynomial  $p(\lambda)$  given by (3.3.6).

3. Verify that  $e^{\lambda_1 t}, te^{\lambda_1 t}, \dots, t^{k-1}e^{\lambda_1 t}$  are solutions to (3.3.1).

4. Deduce that, for each  $j = 0, \dots, k-1$ , the system (3.3.3) has a solution of the form

$$(3.3.16) \quad x(t) = (t^j + \alpha t^{j-1} + \dots + \beta)e^{t\lambda_1}v,$$

(with  $v$  depending on  $j$ ).

5. For given  $A \in M(n, \mathbb{C})$ , suppose  $x' = Ax$  has a solution of the form (3.3.16). Show that  $\lambda_1$  must be a root of multiplicity  $\geq j+1$  of the minimal polynomial of  $A$ .

*Hint.* Take into account the remark below (3.3.15).

6. Using Exercises 3–5, show that the minimal polynomial of the companion matrix  $A$  in (3.3.5) must be the characteristic polynomial  $p(\lambda)$ .

### 3.4. Non-homogeneous equations and Duhamel's formula

In §§3.1–3.3 we have focused on homogeneous equations,  $x' - Ax = 0$ . Here we consider the non-homogeneous equation

$$(3.4.1) \quad \frac{dx}{dt} - Ax = f(t), \quad x(0) = x_0 \in \mathbb{C}^n.$$

Here  $A \in M(n, \mathbb{C})$  and  $f(t)$  takes values in  $\mathbb{C}^n$ . The key to solving this is to recognize that the left side of (3.4.1) is equal to

$$(3.4.2) \quad e^{tA} \frac{d}{dt} \left( e^{-tA} x(t) \right),$$

as follows from the product formula for the derivative and the defining property of  $e^{tA}$ , given in (3.1.4). Thus (3.4.1) is equivalent to

$$(3.4.3) \quad \frac{d}{dt} \left( e^{-tA} x(t) \right) = e^{-tA} f(t), \quad x(0) = x_0,$$

and integration yields

$$(3.4.4) \quad e^{-tA} x(t) = x_0 + \int_0^t e^{-sA} f(s) ds.$$

Applying  $e^{tA}$  to both sides then gives the solution:

$$(3.4.5) \quad x(t) = e^{tA} x_0 + \int_0^t e^{(t-s)A} f(s) ds.$$

This is called Duhamel's formula.

EXAMPLE. We combine methods of this section and §3.3 (and also §3.2) to solve

$$(3.4.6) \quad y'' + y = f(t), \quad y(0) = y_0, \quad y'(0) = y_1.$$

As in §3.3, set  $x = (x_0, x_1) = (y, y')$ , to obtain the system

$$(4.7) \quad \frac{d}{dt} \begin{pmatrix} x_0 \\ x_1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \end{pmatrix} + \begin{pmatrix} 0 \\ f(t) \end{pmatrix}.$$

Recognizing the  $2 \times 2$  matrix above as  $-J$ , and recalling from §3.2 that

$$(3.4.7) \quad e^{(s-t)J} = \begin{pmatrix} \cos(s-t) & -\sin(s-t) \\ \sin(s-t) & \cos(s-t) \end{pmatrix},$$

we obtain

$$(3.4.8) \quad \begin{pmatrix} x_0(t) \\ x_1(t) \end{pmatrix} = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} \begin{pmatrix} y_0 \\ y_1 \end{pmatrix} + \int_0^t \begin{pmatrix} \cos(s-t) & -\sin(s-t) \\ \sin(s-t) & \cos(s-t) \end{pmatrix} \begin{pmatrix} 0 \\ f(s) \end{pmatrix} ds,$$

and hence

$$(3.4.9) \quad y(t) = (\cos t)y_0 + (\sin t)y_1 + \int_0^t \sin(t-s) f(s) ds.$$

Use of Duhamel's formula is a good replacement for the method of variation of parameters, discussed in §1.14 of Chapter 1. See §3.9 of this chapter for more on this. See also §3.B for connections with the Laplace transform.

Next, we briefly discuss a variant of the method of undetermined coefficients, introduced for single second-order equations in §1.10 of Chapter 1. We consider the following special case, the first-order  $n \times n$  system

$$(3.4.10) \quad \frac{dx}{dt} - Ax = (\cos \sigma t)v,$$

given  $\sigma \in \mathbb{R}$ ,  $v \in \mathbb{R}^n$ , and  $A \in M(n, \mathbb{R})$  (or we could use complex coefficients). We assume

$$(3.4.11) \quad i\sigma, -i\sigma \notin \text{Spec } A,$$

and look for a solution to (3.4.10) of the form

$$(3.4.12) \quad x_p(t) = (\cos \sigma t)a + (\sin \sigma t)b, \quad a, b \in \mathbb{R}^n.$$

Substitution into (3.4.10) leads to success with

$$(3.4.13) \quad \begin{aligned} a &= -A(A^2 + \sigma I)^{-1}v, \\ b &= -\sigma(A^2 + \sigma I)^{-1}v. \end{aligned}$$

If (3.4.11) does not hold, (3.4.13) fails, and (3.4.10) might not have a solution of the form (3.4.12). Of course, (3.4.5) will work; (3.4.10) will have a solution of the form

$$(3.4.14) \quad x(t) = \int_0^t (\cos \sigma s)e^{(t-s)A} v ds.$$

When (3.4.11) holds and (3.4.12) works, the general solution to (3.4.10) is

$$(3.4.15) \quad x(t) = e^{tA}u_0 + (\cos \sigma t)a + (\sin \sigma t)b, \quad u_0 \in \mathbb{R}^n,$$

$u_0$  related to  $x(0)$  by

$$(3.4.16) \quad x(0) = u_0 + a.$$

If all the eigenvalues of  $A$  have negative real part,  $e^{tA}u_0$  will decay to 0 as  $t \rightarrow +\infty$ . Then  $e^{tA}u_0$  is called the *transient* part of the solution. The other part,  $(\cos \sigma t)a + (\sin \sigma t)b$ , is called the *steady state* solution.

## Exercises

1. Given  $A \in M(n, \mathbb{C})$ , set

$$E_k(t) = \sum_{j=0}^k \frac{t^j}{j!} A^j.$$

Verify that  $E'_k(t) = AE_{k-1}(t)$  and that

$$\frac{d}{dt}(E_k(t)e^{-tA}) = -\frac{t^k}{k!} A^{k+1} e^{-tA}.$$

2. Verify that, if  $A$  is invertible,

$$\int_0^t s^k e^{-sA} ds = -k! A^{-(k+1)} [E_k(t)e^{-tA} - I].$$

3. Solve the initial value problem

$$\frac{dx}{dt} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} x + \begin{pmatrix} e^t \\ e^{-t} \end{pmatrix}, \quad x(0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

4. Solve the initial value problem

$$\frac{dx}{dt} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} x + \begin{pmatrix} e^t \\ e^{-t} \end{pmatrix}, \quad x(0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

5. Solve the initial value problem

$$\frac{dx}{dt} = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix} x + \begin{pmatrix} e^t \\ e^{-t} \end{pmatrix}, \quad x(0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

6. Produce analogues of (3.4.7)–(3.4.9) for

$$y'' - 3y' + 2y = f(t), \quad y(0) = y_0, \quad y'(0) = y_1.$$

In Exercises 7–8, take  $X, Y \in M(n, \mathbb{C})$  and

$$(3.4.17) \quad U(t, s) = e^{t(X+sY)}, \quad U_s(t, s) = \frac{\partial}{\partial s} U(t, s).$$

7. Show that  $U_s$  satisfies

$$\frac{\partial U_s}{\partial t} = (X + sY)U_s + YU, \quad U_s(0, s) = 0.$$

8. Use Duhamel's formula to show that

$$U_s(t, s) = \int_0^t e^{(t-\tau)(X+sY)} Y e^{\tau(X+sY)} d\tau.$$

Deduce that

$$(3.4.18) \quad \frac{d}{ds} e^{X+sY} \Big|_{s=0} = e^X \int_0^1 e^{-\tau X} Y e^{\tau X} d\tau.$$

9. Assume  $X(t)$  is a smooth function of  $t \in I$  with values in  $M(n, \mathbb{C})$ . Show that, for  $t \in I$ ,

$$(3.4.19) \quad \frac{d}{dt} e^{X(t)} = e^{X(t)} \int_0^1 e^{-\tau X(t)} X'(t) e^{\tau X(t)} d\tau.$$

10. In the context of Exercise 9, assume

$$t, t' \in I \implies X(t)X(t') = X(t')X(t).$$

In such a case, simplify (3.4.19), and compare the result with that of Exercise 12 in §3.1.



### 3.5. Simple electrical circuits

Here we extend the scope of the treatment of electrical circuits in §1.13 of Chapter 1. Rules worked out by Kirchhoff and others in the 1800s allow one to write down a system of linear differential equations describing the voltages and currents running along a variety of electrical circuits, containing resistors, capacitors, and inductors.

There are two types of basic laws. The first type consists of two rules known as Kirchhoff's laws:

- (A) The sum of the voltage drops around any closed loop is zero.
- (B) The sum of the currents at any node is zero.

The second type of law specifies the voltage drop across each circuit element:

$$\begin{aligned} (a) \quad \text{Resistor:} \quad & V = IR, \\ (b) \quad \text{Inductor:} \quad & V = L \frac{dI}{dt}, \\ (c) \quad \text{Capacitor:} \quad & V = \frac{Q}{C}. \end{aligned}$$

In each case,  $V$  is the voltage drop (in volts),  $I$  is the current (in amps),  $R$  is the resistance (in ohms),  $L$  is the inductance (in henrys),  $C$  is the capacitance (in farads), and  $Q$  is the charge (in coulombs). We refer to §1.13 of Chapter 1 for basic information about these units. The rule (c) is supplemented by the following formula for the current across a capacitor:

$$(c2) \quad I = \frac{dQ}{dt}.$$

In (b) and (c2), time is measured in seconds.

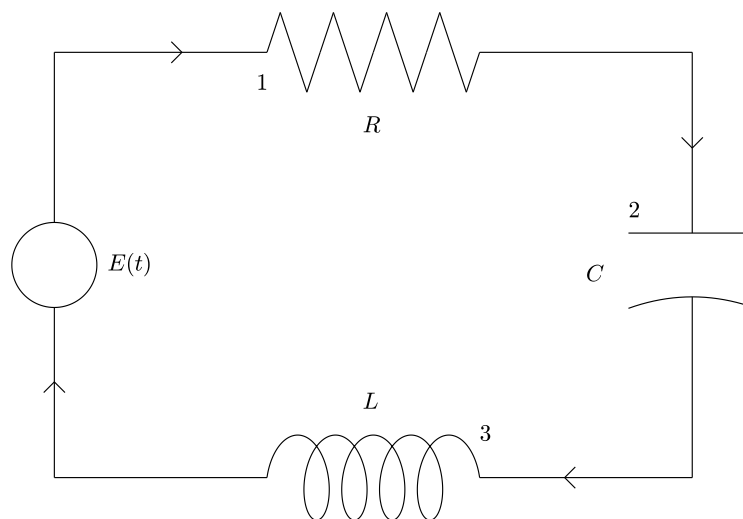
Rules (A), (B), and (a) give algebraic relations among the various voltages and currents, while rules (b) and (c)–(c2) give differential equations, namely

$$(3.5.1) \quad L \frac{dI}{dt} = V \quad (\text{Inductor}),$$

$$(3.5.2) \quad C \frac{dV}{dt} = I \quad (\text{Capacitor}).$$

Note that (3.5.2) results from applying  $d/dt$  to (c) and then using (c2). If a circuit has  $k$  capacitors and  $\ell$  inductors, we get an  $m \times m$  system of first order differential equations, with  $m = k + \ell$ .

We illustrate the formulation of such differential equations for circuits presented in Figure 3.5.1 and Figure 3.5.2. In each case, the circuit elements are numbered. We denote by  $V_j$  the voltage drop across element  $j$  and by  $I_j$  the current across element  $j$ .



**Figure 3.5.1.** RLC circuit

Figure 3.5.1 depicts a classical RLC circuit, such as treated in §1.13 of Chapter 1. Rules (A), (B), and (a) give

$$(3.5.3) \quad \begin{aligned} V_1 + V_2 + V_3 &= E(t), \\ I_1 &= I_2 = I_3, \\ V_1 &= RI_1. \end{aligned}$$

Equations (3.5.1)–(3.5.3) yield a system of two ODEs, for  $I_3$  and  $V_2$ :

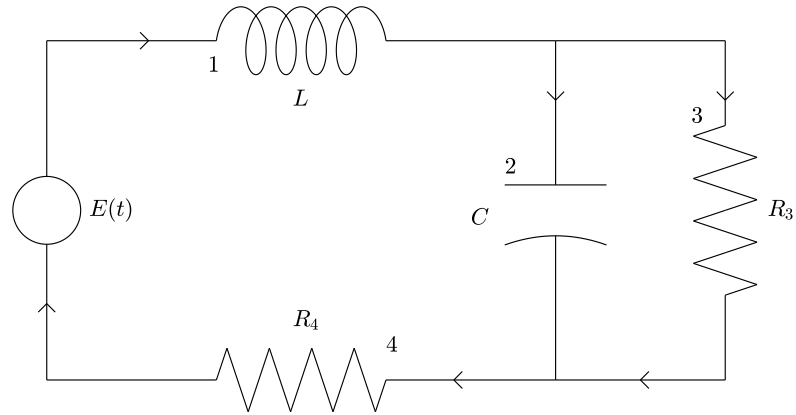
$$(3.5.4) \quad L \frac{dI_3}{dt} = V_3, \quad C \frac{dV_2}{dt} = I_2.$$

We need to express  $V_3$  and  $I_2$  in terms of  $I_3$ ,  $V_2$ , and  $E(t)$ , using (3.5.3). In fact, we have

$$(3.5.5) \quad \begin{aligned} V_3 &= E(t) - V_1 - V_2 = E(t) - RI_1 - V_2 = E(t) - RI_3 - V_2, \\ I_2 &= I_3, \end{aligned}$$

so we get the system

$$(3.5.6) \quad \begin{aligned} L \frac{dI_3}{dt} &= -RI_3 - V_2 + E(t), \\ C \frac{dV_2}{dt} &= I_3, \end{aligned}$$



**Figure 3.5.2.** Another circuit

or, in matrix form,

$$(3.5.7) \quad \frac{d}{dt} \begin{pmatrix} I_3 \\ V_2 \end{pmatrix} = \begin{pmatrix} -R/L & -1/L \\ 1/C & 0 \end{pmatrix} \begin{pmatrix} I_3 \\ V_2 \end{pmatrix} + \frac{1}{L} \begin{pmatrix} E(t) \\ 0 \end{pmatrix}.$$

Note that the characteristic polynomial of the matrix

$$(3.5.8) \quad A = \begin{pmatrix} -R/L & -1/L \\ 1/C & 0 \end{pmatrix}$$

is

$$(3.5.9) \quad \lambda^2 + \frac{R}{L}\lambda + \frac{1}{LC},$$

with roots

$$(3.5.10) \quad \lambda = -\frac{R}{2L} \pm \frac{1}{2L} \sqrt{R^2 - 4\frac{L}{C}}.$$

Let us now look at the slightly more complicated circuit depicted in Figure 3.5.2. Again we get a  $2 \times 2$  system of differential equations. Rules (A), (B), and (a) give

$$(3.5.11) \quad \begin{aligned} V_1 + V_2 + V_4 &= E(t), & V_2 &= V_3, \\ I_1 &= I_2 + I_3 = I_4, \\ V_3 &= R_3 I_3, & V_4 &= R_4 I_4. \end{aligned}$$

Equations (3.5.1)–(3.5.2) yield differential equations for  $I_1$  and  $V_2$ :

$$(3.5.12) \quad C \frac{dV_2}{dt} = I_2, \quad L \frac{dI_1}{dt} = V_1.$$

We need to express  $I_2$  and  $V_1$  in terms of  $V_2$ ,  $I_1$ , and  $E(t)$ , using (3.5.11). In fact, we have

$$(3.5.13) \quad \begin{aligned} I_2 &= I_1 - I_3 = I_1 - \frac{1}{R_3} V_3 = I_1 - \frac{1}{R_3} V_2, \\ V_1 &= E(t) - V_2 - V_4 = E(t) - V_2 - R_4 I_4 = E(t) - V_2 - R_4 I_1, \end{aligned}$$

so we get the system

$$(3.5.14) \quad \begin{aligned} C \frac{dV_2}{dt} &= -\frac{1}{R_3} V_2 + I_1, \\ L \frac{dI_1}{dt} &= -V_2 - R_4 I_1 + E(t), \end{aligned}$$

or, in matrix form,

$$(3.5.15) \quad \frac{d}{dt} \begin{pmatrix} V_2 \\ I_1 \end{pmatrix} = \begin{pmatrix} -1/R_3 C & 1/C \\ -1/L & -R_4/L \end{pmatrix} \begin{pmatrix} V_2 \\ I_1 \end{pmatrix} + \frac{1}{L} \begin{pmatrix} 0 \\ E(t) \end{pmatrix}.$$

## Exercises

1. Work out the  $3 \times 3$  system of differential equations describing the behavior of the circuit depicted in Figure 3.5.3. Assume

$$E(t) = 5 \sin 12t \quad \text{volts.}$$

2. Using methods developed in §3.4, solve the  $2 \times 2$  system (3.5.7) when

$$R = 5 \text{ ohms}, \quad L = 4 \text{ henrys}, \quad C = 1 \text{ farad},$$

and

$$E(t) = 5 \cos 2t \quad \text{volts,}$$

with initial data

$$I_3(0) = 0 \text{ amps}, \quad V_2(0) = 5 \text{ volts.}$$

3. Solve the  $2 \times 2$  system (3.5.15) when

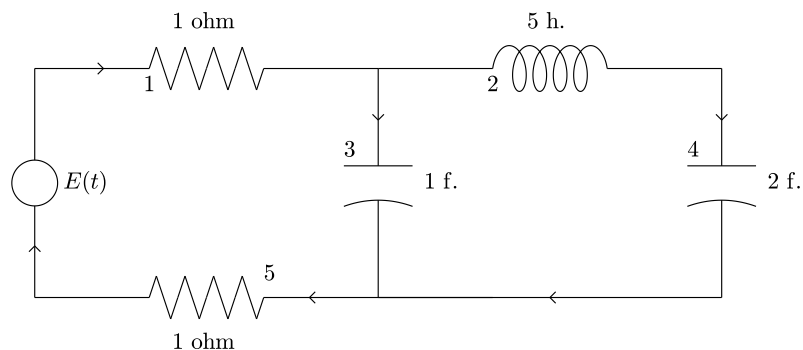
$$R_3 = 1 \text{ ohm}, \quad R_4 = 4 \text{ ohms}, \quad L = 4 \text{ henrys}, \quad C = 2 \text{ farads,}$$

and

$$E(t) = 2 \cos 2t \quad \text{volts.}$$

4. Use the method of (3.4.10)–(3.4.13) to find the steady state solution to (3.5.7), when

$$E(t) = A \cos \sigma t.$$



**Figure 3.5.3.** Circuit with two capacitors

Take  $A$ ,  $\sigma$ ,  $R$  and  $L$  fixed and allow  $C$  to vary. Show that the amplitude of the steady state solution is maximal (we say resonance is achieved) when

$$LC = \frac{1}{\sigma^2},$$

recovering calculations of (1.13.7)–(1.13.13) in Chapter 1.

5. Work out the analogue of Exercise 4 with the system (3.5.7) replaced by (3.5.15). Is the condition for resonance the same as in Exercise 4?

6. Draw an electrical circuit that leads to a  $4 \times 4$  system of differential equations, and write down said system.




Figure 3.6.1. Spring system

### 3.6. Second-order systems

Interacting physical systems often give rise to second-order systems of differential equations. Consider for example a system of  $n$  objects, of mass  $m_1, \dots, m_n$ , connected to each other and to two walls by  $n + 1$  springs, with spring constants  $k_1, \dots, k_{n+1}$ , as in Figure 3.6.1. We assume the masses slide without friction. Denote by  $x_j$  the position of the  $j$ th mass and by  $y_j$  the degree to which the  $j$ th spring is stretched. The equations of motion are

$$(3.6.1) \quad m_j x_j'' = -k_j y_j + k_{j+1} y_{j+1}, \quad 1 \leq j \leq n,$$

and for certain constants  $a_j$ ,

$$(3.6.2) \quad \begin{aligned} y_j &= x_j - x_{j+1} + a_j, \quad 2 \leq j \leq n, \\ y_1 &= x_1 + a_1, \quad y_{n+1} = -x_n + a_{n+1}. \end{aligned}$$

Substituting (3.6.2) into (3.6.1) yields an  $n \times n$  system, which we can write in matrix form as

$$(3.6.3) \quad Mx'' = -Kx + b,$$

where  $x = (x_1, \dots, x_n)^t$ ,  $b = (-k_1 a_1 + k_2 a_2, \dots, -k_n a_n + k_{n+1} a_{n+1})^t$ ,

$$(3.6.4) \quad M = \begin{pmatrix} m_1 & & \\ & \ddots & \\ & & m_n \end{pmatrix},$$

and

$$(3.6.5) \quad K = \begin{pmatrix} k_1 + k_2 & -k_2 & & & \\ -k_2 & k_2 + k_3 & \ddots & & \\ & & \ddots & & \\ & & & k_{n-1} + k_n & -k_n \\ & & & -k_n & k_n + k_{n+1} \end{pmatrix}.$$

We assume  $m_j > 0$  and  $k_j > 0$  for each  $j$ . Then clearly  $M$  is a positive definite matrix and  $K$  is a real symmetric matrix.

**Proposition 3.6.1.** *If each  $k_j > 0$ , then  $K$ , given by (3.6.5), is positive definite.*

**Proof.** We have

$$(3.6.6) \quad x \cdot Kx = \sum_{j=1}^n (k_j + k_{j+1})x_j^2 - 2 \sum_{j=2}^n k_j x_{j-1} x_j.$$

Now

$$(3.6.7) \quad 2x_{j-1}x_j \leq x_{j-1}^2 + x_j^2,$$

so

$$(3.6.8) \quad \begin{aligned} x \cdot Kx &\geq \sum_{j=1}^n k_j x_j^2 + \sum_{j=2}^{n+1} k_j x_{j-1}^2 \\ &\quad - \sum_{j=2}^n k_j x_j^2 - \sum_{j=2}^n k_j x_{j-1}^2 \\ &\geq k_1 x_1^2 + k_{n+1} x_n^2. \end{aligned}$$

Furthermore note that the inequality in (3.6.7) is strict unless  $x_{j-1} = x_j$  so the inequality in (3.6.8) is strict unless  $x_{j-1} = x_j$  for each  $j \in \{2, \dots, n\}$ , i.e., unless  $x_1 = \dots = x_n$ . This proves that  $x \cdot Kx > 0$  whenever  $x \in \mathbb{R}^n$ ,  $x \neq 0$ .  $\square$

To be more precise, we can sharpen (3.6.7) to

$$(3.6.9) \quad 2x_{j-1}x_j = x_{j-1}^2 + x_j^2 - (x_j - x_{j-1})^2,$$

and then (3.6.8) is sharpened to

$$(3.6.10) \quad x \cdot Kx = k_1 x_1^2 + k_{n+1} x_n^2 + \sum_{j=2}^n k_j (x_j - x_{j-1})^2.$$

If we set

$$(3.6.11) \quad \kappa = \min\{k_j : 1 \leq j \leq n+1\},$$

then (6.10) implies

$$(3.6.12) \quad x \cdot Kx \geq \kappa \left( x_1^2 + x_n^2 + \sum_{j=2}^n (x_j - x_{j-1})^2 \right).$$

The system (3.6.3) is inhomogeneous, but it is readily converted into the homogeneous system

$$(3.6.13) \quad Mz'' = -Kz, \quad z = x - K^{-1}b.$$

This in turn can be rewritten

$$(3.6.14) \quad z'' = -M^{-1}Kz.$$

Note that

$$(3.6.15) \quad L = M^{-1/2}KM^{-1/2} \implies M^{-1}K = M^{-1/2}LM^{1/2},$$

where

$$(3.6.16) \quad M^{1/2} = \begin{pmatrix} m_1^{1/2} & & \\ & \ddots & \\ & & m_n^{1/2} \end{pmatrix}.$$

**Proposition 3.6.2.** *The matrix  $L$  is positive definite.*

**Proof.**  $x \cdot Lx = (M^{-1/2}x) \cdot K(M^{-1/2}x) > 0$  whenever  $x \neq 0$ . □

According to (3.6.15),  $M^{-1}K$  and  $L$  are similar, so we have:

**Corollary 3.6.3.** *For  $M$  and  $K$  of the form (3.6.4)–(3.6.5), with  $m_j, k_j > 0$ , the matrix  $M^{-1}K$  is diagonalizable, and all its eigenvalues are positive.*

It follows that  $\mathbb{R}^n$  has a basis  $\{v_1, \dots, v_n\}$  satisfying

$$(3.6.17) \quad M^{-1}Kv_j = \lambda_j^2 v_j, \quad \lambda_j > 0.$$

Then the initial value problem

$$(3.6.18) \quad Mz'' = -Kz, \quad z(0) = z_0, \quad z'(0) = z_1$$

has the solution

$$(3.6.19) \quad z(t) = \sum_{j=1}^n \left( \alpha_j \cos \lambda_j t + \frac{\beta_j}{\lambda_j} \sin \lambda_j t \right) v_j,$$

where the coefficients  $\alpha_j$  and  $\beta_j$  are given by

$$(3.6.20) \quad z_0 = \sum \alpha_j v_j, \quad z_1 = \sum \beta_j v_j.$$

An alternative approach to the system (3.6.14) is to set

$$(3.6.21) \quad u = M^{1/2}z,$$

for which (3.6.14) becomes

$$(3.6.22) \quad u'' = -Lu,$$



with  $L$  given by (3.6.15). Then  $\mathbb{R}^n$  has an orthonormal basis  $\{w_j : 1 \leq j \leq n\}$ , satisfying

$$(3.6.23) \quad Lw_j = \lambda_j^2 w_j, \quad \text{namely } w_j = M^{1/2} v_j,$$

with  $v_j$  as in (3.6.17). Note that we can set

$$(3.6.24) \quad L = A^2, \quad Aw_j = \lambda_j w_j,$$

and (3.6.22) becomes

$$(3.6.25) \quad u'' + A^2 u = 0.$$

One way to convert (3.6.25) to a first order  $(2n) \times (2n)$  system is to set

$$(3.6.26) \quad v = Au, \quad w = u'.$$

Then (3.6.25) becomes

$$(3.6.27) \quad \frac{d}{dt} \begin{pmatrix} v \\ w \end{pmatrix} = X \begin{pmatrix} v \\ w \end{pmatrix}, \quad X = \begin{pmatrix} 0 & A \\ -A & 0 \end{pmatrix}.$$

It is useful to note that when  $X$  is given by (3.6.27), then

$$(3.6.28) \quad e^{tX} = \begin{pmatrix} \cos tA & \sin tA \\ -\sin tA & \cos tA \end{pmatrix},$$

where  $\cos tA$  and  $\sin tA$  are given in Exercises 6–7 in §3.2. One way to see this is to let  $\Phi(t)$  denote the right side of (3.6.28) and use (3.2.19) to see that

$$(3.6.29) \quad \frac{d}{dt} \Phi(t) = X \Phi(t), \quad \Phi(0) = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}.$$

Then Exercise 4 of §3.1 implies  $e^{tX} \equiv \Phi(t)$ . These calculations imply that the solution to (3.6.25), with initial data  $u(0) = u_0$ ,  $u'(0) = u_1$ , is given by

$$(3.6.30) \quad u(t) = (\cos tA)u_0 + A^{-1}(\sin tA)u_1.$$

Compare (3.6.18)–(3.6.20). This works for each invertible  $A \in M(n, \mathbb{C})$ .

We move to the inhomogeneous variant of (3.6.14), which as above we can transform to the following inhomogeneous variant of (3.6.25):

$$(3.6.31) \quad u'' + A^2 u = f(t), \quad u(0) = u_0, \quad u'(0) = u_1.$$

Using the substitution (3.6.26), we get

$$(3.6.32) \quad \frac{d}{dt} \begin{pmatrix} v \\ w \end{pmatrix} = X \begin{pmatrix} v \\ w \end{pmatrix} + \begin{pmatrix} 0 \\ f(t) \end{pmatrix}, \quad \begin{pmatrix} v(0) \\ w(0) \end{pmatrix} = \begin{pmatrix} Au_0 \\ u_1 \end{pmatrix}.$$

Duhamel's formula applies to give

$$(3.6.33) \quad \begin{pmatrix} v(t) \\ w(t) \end{pmatrix} = e^{tX} \begin{pmatrix} Au_0 \\ u_1 \end{pmatrix} + \int_0^t e^{(t-s)X} \begin{pmatrix} 0 \\ f(s) \end{pmatrix} ds.$$

Using the formula (3.6.28) for  $e^{tX}$ , we see that the resulting formula for  $v(t)$  in (3.6.33) is equivalent to

$$(3.6.34) \quad u(t) = (\cos tA)u_0 + A^{-1}(\sin tA)u_1 + \int_0^t A^{-1} \sin(t-s)A f(s) ds.$$

This is the analogue of Duhamel's formula for the solution to (3.6.31).

We now return to the coupled spring problem and modify (3.6.1)–(3.6.2) to allow for friction. Thus we replace (3.6.1) by

$$(3.6.35) \quad m_j x_j'' = -k_j y_j + k_{j+1} y_j - d_j x_j',$$

where  $y_j$  are as in (3.6.2) and  $d_j > 0$  are friction coefficients. Then (3.6.3) is replaced by

$$(3.6.36) \quad Mx'' = -Kx - Dx' + b,$$

with  $b$  as in (3.6.3),  $M$  and  $K$  as in (3.6.4)–(3.6.5), and

$$(3.6.37) \quad D = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_n \end{pmatrix}, \quad d_j > 0.$$

As in (3.6.13), we can convert (3.6.36) to the homogeneous system

$$(3.6.38) \quad Mz'' = -Kz - Dz', \quad z = x - K^{-1}b.$$

If we set  $u = M^{1/2}z$ , as in (3.6.21), then, parallel to (3.6.22)–(3.6.24), we get

$$(3.6.39) \quad u'' + Bu' + A^2u = 0,$$

where  $A^2$  is as in (3.6.24), with  $L = M^{-1/2}KM^{-1/2}$ , as in (3.6.22), and

$$(3.6.40) \quad B = M^{-1/2}DM^{-1/2} = \begin{pmatrix} d_1/m_1 & & \\ & \ddots & \\ & & d_n/m_n \end{pmatrix}.$$

The substitution (3.6.26) converts the  $n \times n$  second order system (3.6.39) to the  $(2n) \times (2n)$  first order system

$$(3.6.41) \quad \frac{d}{dt} \begin{pmatrix} v \\ w \end{pmatrix} = \begin{pmatrix} 0 & A \\ -A & -B \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix}.$$

We can write (3.6.41) as

$$(3.6.42) \quad \frac{d}{dt} \begin{pmatrix} v \\ w \end{pmatrix} = (X + Y) \begin{pmatrix} v \\ w \end{pmatrix},$$

with  $X$  as in (3.6.27) and

$$(3.6.43) \quad Y = \begin{pmatrix} 0 & 0 \\ 0 & -B \end{pmatrix}.$$

Note that

$$(3.6.44) \quad XY = \begin{pmatrix} 0 & -AB \\ 0 & 0 \end{pmatrix}, \quad YX = \begin{pmatrix} 0 & 0 \\ BA & 0 \end{pmatrix},$$

so these matrices do not commute. Thus  $e^{t(X+Y)}$  might be difficult to calculate even when  $A$  and  $B$  commute. Such commutativity would hold, for example, if  $m_1 = \cdots = m_n$  and  $d_1 = \cdots = d_n$ , in which case  $B$  is a scalar multiple of the identity matrix.

When the positive, self adjoint operators  $A$  and  $B$  do commute, we can make the following direct attack on the system (3.6.39). We know (cf. Exercise 3 in §2.11, Chapter 2) that  $\mathbb{R}^n$  has an orthonormal basis  $\{w_1, \dots, w_n\}$  for which

$$(3.6.45) \quad Aw_j = \lambda_j w_j, \quad Bw_j = 2\mu_j w_j, \quad \lambda_j, \mu_j > 0.$$

Then we can write a solution to (3.6.39) as

$$(3.6.46) \quad u(t) = \sum u_j(t)w_j,$$

where the real-valued coefficients  $u_j(t)$  satisfy the equations

$$(3.6.47) \quad u_j'' + 2\mu_j u_j' + \lambda_j^2 u_j = 0,$$

with solutions that are linear combinations:

$$(3.6.48) \quad \begin{aligned} e^{-\mu_j t} \left( \alpha_j \cos \sqrt{\lambda_j^2 - \mu_j^2} t + \beta_j \sin \sqrt{\lambda_j^2 - \mu_j^2} t \right), & \quad \lambda_j > \mu_j, \\ e^{-\mu_j t} \left( \alpha_j e^{\sqrt{\mu_j^2 - \lambda_j^2} t} + \beta_j e^{-\sqrt{\mu_j^2 - \lambda_j^2} t} \right), & \quad \lambda_j < \mu_j, \\ e^{-\mu_j t} (\alpha_j + \beta_j t), & \quad \lambda_j = \mu_j. \end{aligned}$$

These three cases correspond to modes that are said to be underdamped, overdamped, and critically damped, respectively.

In cases where  $A$  and  $B$  do not commute, analysis of (3.6.39) is less explicit, but we can establish the following decay result.

**Proposition 3.6.4.** *If  $A, B \in M(n, \mathbb{C})$  are positive definite, then all of the eigenvalues of  $Z = \begin{pmatrix} 0 & A \\ -A & -B \end{pmatrix}$  have negative real part.*

**Proof.** Let's say  $(v, w)^t \neq 0$  and  $Z(v, w)^t = \lambda(v, w)^t$ . Then

$$(3.6.49) \quad Aw = \lambda v, \quad Av + Bw = -\lambda w,$$

and

$$(3.6.50) \quad (Z(v, w)^t, (v, w)^t) = -(Bw, w) + [(Aw, v) - (Av, w)],$$

while also

$$(3.6.51) \quad (Z(v, w)^t, (v, w)^t) = \lambda(\|v\|^2 + \|w\|^2).$$

The two terms on the right side of (3.6.50) are real and purely imaginary, respectively, so we obtain

$$(3.6.52) \quad (\operatorname{Re} \lambda)(\|v\|^2 + \|w\|^2) = -(Bw, w).$$

If  $(v, w)^t \neq 0$ , we deduce that either  $\operatorname{Re} \lambda < 0$  or  $w = 0$ . If  $w = 0$ , then (3.6.49) gives  $Av = 0$ , hence  $v = 0$ . Hence  $w \neq 0$ , and  $\operatorname{Re} \lambda < 0$ , as asserted.  $\square$

---

**Exercises**

1. Find the eigenvalues and eigenvectors of

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

2. Use the results of Exercise 1 to find the eigenvalues and eigenvectors of  $K$ , given by (3.6.5), in case  $n = 3$  and

$$k_1 = k_2 = k_3 = k_4 = k.$$

3. Find the general solution to

$$u'' + Bu' + A^2u = 0,$$

in case  $A^2 = K$ , with  $K$  as in Exercise 2, and  $B = I$ .

4. Find the general solution to

$$u'' + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} u' + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} u = 0.$$

5. Generalizing the treatment of (3.6.25), consider

$$(3.6.53) \quad u'' + Lu = 0, \quad L \in M(N, \mathbb{C}).$$

Assume  $\mathbb{C}^N$  has a basis of eigenvectors  $v_j$ , such that  $Lv_j = \lambda_j^2 v_j$ ,  $\lambda_j \in \mathbb{C}$ ,  $\lambda_j \neq 0$ . Show that the general solution to (3.6.53) has the form

$$(3.6.54) \quad u(t) = \sum_{j=1}^N (\alpha_j e^{\lambda_j t} + \beta_j e^{-\lambda_j t}) v_j, \quad \alpha_j, \beta_j \in \mathbb{C}.$$

How is this modified if some  $\lambda_j = 0$ ?

6. Find the general solution to

$$u'' + \begin{pmatrix} 2 & 1 \\ 0 & -1 \end{pmatrix} u = 0,$$

and to

$$u'' + \begin{pmatrix} & & -1 \\ & 1 & \\ 1 & & \end{pmatrix} u = 0.$$

### 3.7. Curves in $\mathbb{R}^3$ and the Frenet-Serret equations

Given a curve  $c(t) = (x(t), y(t), z(t))$  in 3-space, we define its velocity and acceleration by

$$(3.7.1) \quad v(t) = c'(t), \quad a(t) = v'(t) = c''(t).$$

We also define its speed  $s'(t)$  and arclength by

$$(3.7.2) \quad s'(t) = \|v(t)\|, \quad s(t) = \int_{t_0}^t s'(\tau) d\tau,$$

assuming we start at  $t = t_0$ . We define the unit tangent vector to the curve as

$$(3.7.3) \quad T(t) = \frac{v(t)}{\|v(t)\|}.$$

Henceforth we assume the curve is parametrized by arclength.

We define the *curvature*  $\kappa(s)$  of the curve and the normal  $N(s)$  by

$$(3.7.4) \quad \kappa(s) = \left\| \frac{dT}{ds} \right\|, \quad \frac{dT}{ds} = \kappa(s)N(s).$$

Note that

$$(3.7.5) \quad T(s) \cdot T(s) = 1 \implies T'(s) \cdot T(s) = 0,$$

so indeed  $N(s)$  is orthogonal to  $T(s)$ . We then define the binormal  $B(s)$  by

$$(3.7.6) \quad B(s) = T(s) \times N(s).$$

For each  $s$ , the vectors  $T(s)$ ,  $N(s)$  and  $B(s)$  are mutually orthogonal unit vectors, known as the Frenet frame for the curve  $c(s)$ . See Figure 3.7.1 for an illustration. Rules governing the cross product yield

$$(3.7.7) \quad T(s) = N(s) \times B(s), \quad N(s) = B(s) \times T(s).$$

(For material on the cross product, see the exercises at the end of §2.12 of Chapter 2.)

The *torsion* of a curve measures the change in the plane generated by  $T(s)$  and  $N(s)$ , or equivalently it measures the rate of change of  $B(s)$ . Note that, parallel to (3.7.5),

$$B(s) \cdot B(s) = 1 \implies B'(s) \cdot B(s) = 0.$$

Also, differentiating (3.7.6) and using (3.7.4), we have

$$(3.7.8) \quad B'(s) = T'(s) \times N(s) + T(s) \times N'(s) = T(s) \times N'(s) \implies B'(s) \cdot T(s) = 0.$$

We deduce that  $B'(s)$  is parallel to  $N(s)$ . We define the torsion by

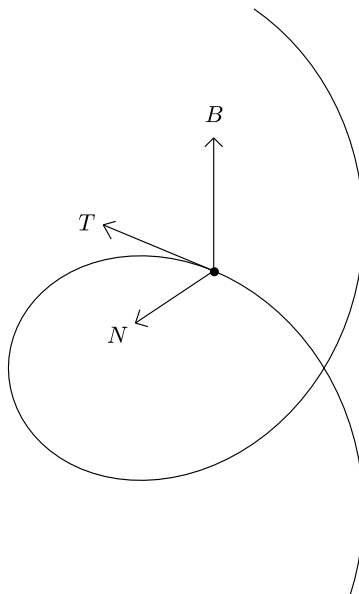
$$(3.7.9) \quad \frac{dB}{ds} = -\tau(s)N(s).$$

We complement the formulas (3.7.4) and (3.7.9) for  $dT/ds$  and  $dB/ds$  with one for  $dN/ds$ . Since  $N(s) = B(s) \times T(s)$ , we have

$$(3.7.10) \quad \frac{dN}{ds} = \frac{dB}{ds} \times T + B \times \frac{dT}{ds} = \tau N \times T + \kappa B \times N,$$

or

$$(3.7.11) \quad \frac{dN}{ds} = -\kappa(s)T(s) + \tau(s)B(s).$$



**Figure 3.7.1.** Frenet frame at a point on a 3D curve

Together, (3.7.4), (3.7.9) and (3.7.11) are known as the Frenet-Serret formulas.

EXAMPLE. Pick  $a, b > 0$  and consider the helix

$$(3.7.12) \quad c(t) = (a \cos t, a \sin t, bt).$$

Then  $v(t) = (-a \sin t, a \cos t, b)$  and  $\|v(t)\| = \sqrt{a^2 + b^2}$ , so we can pick  $s = t\sqrt{a^2 + b^2}$  to parametrize by arc length. We have

$$(3.7.13) \quad T(s) = \frac{1}{\sqrt{a^2 + b^2}}(-a \sin t, a \cos t, b),$$

hence

$$(3.7.14) \quad \frac{dT}{ds} = \frac{1}{a^2 + b^2}(-a \cos t, -a \sin t, 0).$$

By (3.7.4), this gives

$$(3.7.15) \quad \kappa(s) = \frac{a}{a^2 + b^2}, \quad N(s) = (-\cos t, -\sin t, 0).$$

Hence

$$(3.7.16) \quad B(s) = T(s) \times N(s) = \frac{1}{\sqrt{a^2 + b^2}}(b \sin t, -b \cos t, a).$$

Then

$$(3.7.17) \quad \frac{dB}{ds} = \frac{1}{a^2 + b^2} (b \cos t, b \sin t, 0),$$

so, by (3.7.9),

$$(3.7.18) \quad \tau(s) = \frac{b}{a^2 + b^2}.$$

In particular, for the helix (3.7.12), we see that the curvature and torsion are *constant*.

Let us collect the Frenet-Serret equations

$$(3.7.19) \quad \begin{aligned} \frac{dT}{ds} &= \kappa N \\ \frac{dN}{ds} &= -\kappa T + \tau B \\ \frac{dB}{ds} &= -\tau N \end{aligned}$$

for a smooth curve  $c(s)$  in  $\mathbb{R}^3$ , parametrized by arclength, with unit tangent  $T(s)$ , normal  $N(s)$ , and binormal  $B(s)$ , given by

$$(3.7.20) \quad N(s) = \frac{1}{\kappa(s)} T'(s), \quad B(s) = T(s) \times N(s),$$

assuming  $\kappa(s) = \|T'(s)\| > 0$ .

The basic existence and uniqueness theory, which will be presented in Chapter 4, applies to (3.7.19). If  $\kappa(s)$  and  $\tau(s)$  are given smooth functions on an interval  $I = (a, b)$  and  $s_0 \in I$ , then, given  $T_0, N_0, B_0 \in \mathbb{R}^3$ , (7.19) has a unique solution on  $s \in I$  satisfying

$$(3.7.21) \quad T(s_0) = T_0, \quad N(s_0) = N_0, \quad B(s_0) = B_0.$$

In fact, the case when  $\kappa(s)$  and  $\tau(s)$  are analytic will be subsumed in the material of §3.10 of this chapter. We now establish the following.

**Proposition 3.7.1.** *Assume  $\kappa$  and  $\tau$  are given smooth functions on  $I$ , with  $\kappa > 0$  on  $I$ . Assume  $\{T_0, N_0, B_0\}$  is an orthonormal basis of  $\mathbb{R}^3$ , such that  $B_0 = T_0 \times N_0$ . Then there exists a smooth, unit-speed curve  $c(s)$ ,  $s \in I$ , for which the solution to (3.7.19) and (3.7.21) is the Frenet frame.*

To construct the curve, take  $T(s)$ ,  $N(s)$ , and  $B(s)$  to solve (3.7.19) and (3.7.21), pick  $p \in \mathbb{R}^3$  and set

$$(3.7.22) \quad c(s) = p + \int_{s_0}^s T(\sigma) d\sigma,$$

so  $T(s) = c'(s)$  is the velocity of this curve. To deduce that  $\{T(s), N(s), B(s)\}$  is the Frenet frame for  $c(s)$ , for all  $s \in I$ , we need to know:

$$(3.7.23) \quad \{T(s), N(s), B(s)\} \text{ orthonormal, with } B(s) = T(s) \times N(s), \quad \forall s \in I.$$

In order to pursue the analysis further, it is convenient to form the  $3 \times 3$  matrix-valued function

$$(3.7.24) \quad F(s) = (T(s), N(s), B(s)),$$

whose *columns* consist respectively of  $T(s)$ ,  $N(s)$ , and  $B(s)$ . Then (3.7.23) is equivalent to

$$(3.7.25) \quad F(s) \in SO(3), \quad \forall s \in I,$$

with  $SO(3)$  defined as in (2.12.4) of Chapter 2. The hypothesis on  $\{T_0, N_0, B_0\}$  stated in Proposition 3.7.1 is equivalent to  $F_0 = (T_0, N_0, B_0) \in SO(3)$ . Now  $F(s)$  satisfies the differential equation

$$(3.7.26) \quad F'(s) = F(s)A(s), \quad F(s_0) = F_0,$$

where

$$(3.7.27) \quad A(s) = \begin{pmatrix} 0 & -\kappa(s) & 0 \\ \kappa(s) & 0 & -\tau(s) \\ 0 & \tau(s) & 0 \end{pmatrix}.$$

Note that

$$(3.7.28) \quad \frac{dF^*}{ds} = A(s)^* F(s)^* = -A(s)F(s)^*,$$

since  $A(s)$  in (3.7.27) is skew-adjoint. Hence

$$(3.7.29) \quad \begin{aligned} \frac{d}{ds} F(s)F(s)^* &= \frac{dF}{ds} F(s)^* + F(s) \frac{dF^*}{ds} \\ &= F(s)A(s)F(s)^* - F(s)A(s)F(s)^* \\ &= 0. \end{aligned}$$

Thus, whenever (3.7.26)–(3.7.27) hold,

$$(3.7.30) \quad F_0 F_0^* = I \implies F(s)F(s)^* \equiv I,$$

and we have (3.7.23).

Let us specialize the system (3.7.19), or equivalently (3.7.26), to the case where  $\kappa$  and  $\tau$  are *constant*, i.e.,

$$(3.7.31) \quad F'(s) = F(s)A, \quad A = \begin{pmatrix} 0 & -\kappa & 0 \\ \kappa & 0 & -\tau \\ 0 & \tau & 0 \end{pmatrix},$$

with solution

$$(3.7.32) \quad F(s) = F_0 e^{(s-s_0)A}.$$

We have already seen in that a helix of the form (3.7.12) has curvature  $\kappa$  and torsion  $\tau$ , with

$$(3.7.33) \quad \kappa = \frac{a}{a^2 + b^2}, \quad \tau = \frac{b}{a^2 + b^2},$$

and hence

$$(3.7.34) \quad a = \frac{\kappa}{\kappa^2 + \tau^2}, \quad b = \frac{\tau}{\kappa^2 + \tau^2}.$$

In (3.7.12),  $s$  and  $t$  are related by  $t = s\sqrt{\kappa^2 + \tau^2}$ .

We can also see such a helix arise via a direct calculation of  $e^{sA}$ , which we now produce. First, a straightforward calculation gives, for  $A$  as in (3.7.31),

$$(3.7.35) \quad \det(\lambda I - A) = \lambda(\lambda^2 + \kappa^2 + \tau^2),$$



hence

$$(3.7.36) \quad \text{Spec}(A) = \{0, \pm i\sqrt{\kappa^2 + \tau^2}\}.$$

An inspection shows that we can take

$$(3.7.37) \quad v_1 = \frac{1}{\sqrt{\kappa^2 + \tau^2}} \begin{pmatrix} \tau \\ 0 \\ \kappa \end{pmatrix}, \quad v_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad v_3 = \frac{1}{\sqrt{\kappa^2 + \tau^2}} \begin{pmatrix} -\kappa \\ 0 \\ \tau \end{pmatrix},$$

and then

$$(3.7.38) \quad Av_1 = 0, \quad Av_2 = \sqrt{\kappa^2 + \tau^2} v_3, \quad Av_3 = -\sqrt{\kappa^2 + \tau^2} v_2.$$

In particular, with respect to the basis  $\{v_2, v_3\}$  of  $V = \text{Span}\{v_2, v_3\}$ ,  $A|_V$  has the matrix representation

$$(3.7.39) \quad B = \sqrt{\kappa^2 + \tau^2} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

We see that

$$(3.7.40) \quad e^{sA}v_1 = v_1,$$

while, in light of the calculations giving (3.2.8),

$$(3.7.41) \quad \begin{aligned} e^{sA}v_2 &= (\cos s\sqrt{\kappa^2 + \tau^2})v_2 + (\sin s\sqrt{\kappa^2 + \tau^2})v_3, \\ e^{sA}v_3 &= -(\sin s\sqrt{\kappa^2 + \tau^2})v_2 + (\cos s\sqrt{\kappa^2 + \tau^2})v_3. \end{aligned}$$

### Exercises

1. Consider a curve  $c(t)$  in  $\mathbb{R}^3$ , not necessarily parametrized by arclength. Show that the acceleration  $a(t)$  is given by

$$(3.7.42) \quad a(t) = \frac{d^2s}{dt^2}T + \kappa\left(\frac{ds}{dt}\right)^2 N.$$

*Hint.* Differentiate  $v(t) = (ds/dt)T(t)$  and use the chain rule  $dT/dt = (ds/dt)(dT/ds)$ , plus (3.7.4).

2. Show that

$$(3.7.43) \quad \kappa B = \frac{v \times a}{\|v\|^3}.$$

*Hint.* Take the cross product of both sides of (3.7.42) with  $T$ , and use (3.7.6).

3. In the setting of Exercises 1–2, show that

$$(3.7.44) \quad \kappa^2 \tau \|v\|^6 = -a \cdot (v \times a').$$

Deduce from (3.7.43)–(3.7.44) that

$$(3.7.45) \quad \tau = \frac{(v \times a) \cdot a'}{\|v \times a\|^2}.$$

*Hint.* Proceed from (3.7.43) to

$$\frac{d}{dt}(\kappa\|v\|^3)B + \kappa\|v\|^3\frac{dB}{dt} = \frac{d}{dt}(v \times a) = v \times a',$$

and use  $dB/dt = -\tau(ds/dt)N$ , as a consequence of (3.7.9). Then dot with  $a$ , and use  $a \cdot N = \kappa\|v\|^2$ , from (3.7.42), to get (3.7.44).

4. Consider the curve  $c(t)$  in  $\mathbb{R}^3$  given by

$$c(t) = (a \cos t, b \sin t, t),$$

where  $a$  and  $b$  are given positive constants. Compute the curvature, torsion, and Frenet frame.

*Hint.* Use (3.7.43) to compute  $\kappa$  and  $B$ . Then use  $N = B \times T$ . Use (3.7.45) to compute  $\tau$ .

5. Suppose  $c$  and  $\tilde{c}$  are two curves, both parametrized by arc length over  $0 \leq s \leq L$ , and both having the same curvature  $\kappa(s) > 0$  and the same torsion  $\tau(s)$ . Show that there exist  $x_0 \in \mathbb{R}^3$  and  $A \in O(3)$  such that

$$\tilde{c}(s) = Ac(s) + x_0, \quad \forall s \in [0, L].$$

*Hint.* To begin, show that if their Frenet frames coincide at  $s = 0$ , i.e.,  $\tilde{T}(0) = T(0)$ ,  $\tilde{N}(0) = N(0)$ ,  $\tilde{B}(0) = B(0)$ , then  $\tilde{T} \equiv T$ ,  $\tilde{N} \equiv N$ ,  $\tilde{B} \equiv B$ .

6. Suppose  $c$  is a curve in  $\mathbb{R}^3$  with curvature  $\kappa > 0$ . Show that there exists a plane in which  $c(t)$  lies for all  $t$  if and only if  $\tau \equiv 0$ .

*Hint.* When  $\tau \equiv 0$ , the plane should be parallel to the orthogonal complement of  $B$ .

### 3.8. Variable coefficient systems

Here we consider a variable coefficient  $n \times n$  first order system

$$(3.8.1) \quad \frac{dx}{dt} = A(t)x, \quad x(t_0) = x_0 \in \mathbb{C}^n,$$

and its inhomogeneous analogue. The general theory, which will be presented in Chapter 4, implies that if  $A(t)$  is a continuous function of  $t \in I = (a, b)$  and  $t_0 \in I$ , then (3.8.1) has a unique solution  $x(t)$  for  $t \in I$ , depending linearly on  $x_0$ , so

$$(3.8.2) \quad x(t) = S(t, t_0)x_0, \quad S(t, t_0) \in \mathcal{L}(\mathbb{C}^n).$$

See §3.10 of this chapter for power series methods of constructing  $S(t, t_0)$ , when  $A(t)$  is analytic. As we have seen,

$$(3.8.3) \quad A(t) \equiv A \implies S(t, t_0) = e^{(t-t_0)A}.$$

However, for variable coefficient equations there is not such a simple formula, and the matrix entries of  $S(t, s)$  can involve a multiplicity of new special functions, such as Bessel functions, Airy functions, Legendre functions, and many more. We will not dwell on this here, but we will note how  $S(t, t_0)$  is related to a “complete set” of solutions to (3.8.1).

Suppose  $x_1(t), \dots, x_n(t)$  are  $n$  solutions to (3.8.1) (but with different initial conditions). Fix  $t_0 \in I$ , and assume

$$(3.8.4) \quad x_1(t_0), \dots, x_n(t_0) \text{ are linearly independent in } \mathbb{C}^n,$$

or equivalently these vectors form a basis of  $\mathbb{C}^n$ . Given such solutions  $x_j(t)$ , we form the  $n \times n$  matrix

$$(3.8.5) \quad M(t) = (x_1(t), \dots, x_n(t)),$$

whose  $j$ th column is  $x_j(t)$ . This matrix function solves

$$(3.8.6) \quad \frac{dM}{dt} = A(t)M(t).$$

The condition (3.8.4) is equivalent to the statement that  $M(t_0)$  is invertible. We claim that if  $M$  solves (3.8.6) and  $M(t_0)$  is invertible then  $M(t)$  is invertible for all  $t \in I$ . To see this, we use the fact that the invertibility of  $M(t)$  is equivalent to the non-vanishing of the quantity

$$(3.8.7) \quad W(t) = \det M(t),$$

called the *Wronskian* of  $\{x_1(t), \dots, x_n(t)\}$ . It is also notable that  $W(t)$  solves a differential equation. In general we have

$$(3.8.8) \quad \frac{d}{dt} \det M(t) = (\det M(t)) \operatorname{Tr}(M(t)^{-1}M'(t)).$$

(See Exercises 1–3 below.) Let  $\tilde{I} \subset I$  be the maximal interval containing  $t_0$  on which  $M(t)$  is invertible. Then (3.8.8) holds for  $t \in \tilde{I}$ . When (3.8.6) holds, we have  $\operatorname{Tr}(M(t)^{-1}M'(t)) = \operatorname{Tr}(M(t)^{-1}A(t)M(t)) = \operatorname{Tr} A(t)$ , so the Wronskian solves the differential equation

$$(3.8.9) \quad \frac{dW}{dt} = (\operatorname{Tr} A(t)) W(t).$$

Hence

$$(3.8.10) \quad W(t) = e^{b(t,s)}W(s), \quad b(t,s) = \int_s^t \operatorname{Tr} A(\tau) d\tau.$$

This implies  $\tilde{I} = I$  and hence gives the asserted invertibility. From here we obtain the following.

**Proposition 3.8.1.** *If  $M(t)$  solves (3.8.6) for  $t \in I$  and  $M(t_0)$  is invertible, then*

$$(3.8.11) \quad S(t, t_0) = M(t)M(t_0)^{-1}, \quad \forall t \in I.$$

**Proof.** We have seen that  $M(t)$  is invertible for all  $t \in I$ . If  $x(t)$  solves (3.8.1), set

$$(3.8.12) \quad y(t) = M(t)^{-1}x(t),$$

and apply  $d/dt$  to  $x(t) = M(t)y(t)$ , obtaining

$$(3.8.13) \quad \begin{aligned} \frac{dx}{dt} &= M'(t)y(t) + M(t)y'(t) \\ &= A(t)M(t)y(t) + M(t)y'(t) \\ &= A(t)x + M(t)y'(t). \end{aligned}$$

If  $x(t)$  solves (3.8.1), this yields

$$(3.8.14) \quad \frac{dy}{dt} = 0,$$

hence  $y(t) = y(t_0)$  for all  $t \in I$ , i.e.,

$$(3.8.15) \quad M(t)^{-1}x(t) \equiv M(t_0)^{-1}x(t_0).$$

Applying  $M(t)$  to both sides gives (3.8.11).  $\square$

Note also that, for  $s, t \in I$ ,

$$(3.8.16) \quad S(t, s) = M(t)M(s)^{-1}$$

gives  $S(t, s)x(s) = x(t)$  for each solution  $x(t)$  to (3.8.1). We also have

$$(3.8.17) \quad S(t, t_0) = S(t, s)S(s, t_0), \quad S(t, s) = S(s, t)^{-1}.$$

There is a more general version of the Duhamel formula (3.4.5) for the solution to an inhomogeneous differential equation

$$(3.8.18) \quad \frac{dx}{dt} = A(t)x + f(t), \quad x(t_0) = x_0.$$

To solve (3.8.18), set  $x(t) = M(t)y(t)$ , as in (3.8.12). This time, (3.8.13) yields  $M(t)y'(t) = f(t)$ , or

$$\frac{dy}{dt} = M(t)^{-1}f(t), \quad y(t_0) = M(t_0)^{-1}x_0,$$

so

$$(3.8.19) \quad x(t) = M(t)M(t_0)^{-1}x_0 + M(t) \int_{t_0}^t M(s)^{-1}f(s) ds,$$

for invertible  $M(t)$  as in (3.8.5)–(3.8.6). Equivalently,

$$(3.8.20) \quad x(t) = S(t, t_0)x_0 + \int_{t_0}^t S(t, s)f(s) ds.$$

We note that there is a simple formula for the solution operator  $S(t, s)$  to (3.8.1) in case the following commutativity hypothesis holds:

$$(3.8.21) \quad A(t)A(t') = A(t')A(t), \quad \forall t, t' \in I.$$

We claim that if

$$(3.8.22) \quad B(t, s) = - \int_s^t A(\tau) d\tau,$$

then

$$(3.8.23) \quad (3.8.21) \implies \frac{d}{dt} (e^{B(t,s)} x(t)) = e^{B(t,s)} (x'(t) - A(t)x(t)),$$

from which it follows that

$$(3.8.24) \quad (3.8.21) \implies S(t, s) = e^{\int_s^t A(\tau) d\tau}.$$

(This identity fails in the absence of the hypothesis (3.8.21).)

To establish (3.8.23), we note that (3.8.21) implies

$$(3.8.25) \quad B(t, s)B(t', s) = B(t', s)B(t, s), \quad \forall s, t, t' \in I.$$

Next,

$$(3.8.25) \implies \lim_{h \rightarrow 0} \frac{1}{h} (e^{B(t+h,s)} - e^{B(t,s)}) \\ (3.8.26) \quad = \lim_{h \rightarrow 0} \frac{1}{h} e^{B(t,s)} (e^{B(t+h,s)-B(t,s)} - I) \\ = -e^{B(t,s)} A(t) \\ \implies \frac{d}{dt} e^{B(t,s)} = -e^{B(t,s)} A(t),$$

from which (3.8.23) follows.

Here is an application of (3.8.24). Let  $x(s)$  be a planar curve, on an interval about  $s = 0$ , parametrized by arc-length, with unit tangent  $T(s) = x'(s)$ . Then the Frenet-Serret equations (3.7.1) simplify to  $T' = \kappa N$ , with  $N = JT$ , i.e., to

$$(3.8.27) \quad T'(s) = \kappa(s)JT(s),$$

with  $J$  as in (3.2.1). Clearly the commutativity hypothesis (3.8.21) holds for  $A(s) = \kappa(s)J$ , so we deduce that

$$(3.8.28) \quad T(s) = e^{\lambda(s)J} T(0), \quad \lambda(s) = \int_0^s \kappa(\tau) d\tau.$$

Recall that  $e^{tJ}$  is given by (3.2.8), i.e.,

$$(3.8.29) \quad e^{tJ} = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}.$$

We return to the general (noncommutative) case, and record the following estimate on the operator norm of  $S(t, s)$ .

**Proposition 3.8.2.** *Assume that for  $t \in I = (a, b)$ ,*

$$(3.8.30) \quad \operatorname{Re}(A(t)v, v) \leq K\|v\|^2, \quad \forall v \in \mathbb{C}^n.$$

*Then*

$$(3.8.31) \quad a < s < t < b \implies \|S(t, s)\| \leq e^{K(t-s)}.$$

**Proof.** If  $v(t) = S(t, s)v(s)$ , then  $w(t) = e^{-K(t-s)}v(t)$  solves

$$(3.8.32) \quad \frac{dw}{dt} = C(t)w(t), \quad C(t) = A(t) - K.$$

It suffices to show that  $\|w(t)\|$  is monotonically decreasing. Indeed,

$$(3.8.33) \quad \begin{aligned} \frac{d}{dt}\|w(t)\|^2 &= 2\operatorname{Re}(w'(t), w(t)) \\ &= 2\operatorname{Re}(C(t)w(t), w(t)) \\ &\leq 0, \end{aligned}$$

and we have the result.  $\square$

We now combine Proposition 3.8.2 and Duhamel's formula to estimate the difference between solutions  $x(t)$  and  $y(t)$  to

$$(3.8.34) \quad \begin{aligned} \frac{dx}{dt} &= A(t)x, \quad x(t_0) = v, \\ \frac{dy}{dt} &= B(t)y, \quad y(t_0) = v. \end{aligned}$$

Let us assume that

$$(3.8.35) \quad \begin{aligned} \operatorname{Re}(A(t)w, w) &\leq K\|w\|^2, \\ \operatorname{Re}(B(t)w, w) &\leq K\|w\|^2, \end{aligned}$$

for all  $w \in \mathbb{C}^n$ . Now, given (3.8.34), our goal is to estimate  $z(t) = x(t) - y(t)$ , for  $t \geq t_0$ . We have

$$(3.8.36) \quad \begin{aligned} \frac{dz}{dt} &= A(t)x - B(t)y \\ &= A(t)z + [A(t) - B(t)]y, \quad z(t_0) = 0, \end{aligned}$$

so, with  $S(t, s)$  denoting the solution operator to (3.8.1), we have

$$(3.8.37) \quad z(t) = \int_{t_0}^t S(t, s)[A(s) - B(s)]y(s) ds.$$

Proposition 3.8.2 gives

$$(3.8.38) \quad \begin{aligned} \|y(s)\| &\leq e^{K(s-t_0)}\|v\|, \\ \|S(t, s)\| &\leq e^{K(t-s)}, \quad \text{for } t_0 \leq s \leq t, \end{aligned}$$

so we have the following conclusion.

**Proposition 3.8.3.** *Assume that  $x$  and  $y$  solve (3.8.34) and that (3.8.35) holds. Then*

$$(3.8.39) \quad \|x(t) - y(t)\| \leq e^{K(t-t_0)} \left( \int_{t_0}^t \|A(s) - B(s)\| ds \right) \|v\|,$$

for  $t \geq t_0$ .

---

**Exercises**

Exercises 1–3 lead to a proof of the formula (3.8.8) for the derivative of  $\det M(t)$ .

1. Let  $A \in M(n, \mathbb{C})$ . Show that, as  $s \rightarrow 0$ ,

$$\begin{aligned}\det(I + sA) &= (1 + sa_{11}) \cdots (1 + sa_{nn}) + O(s^2) \\ &= 1 + s \operatorname{Tr} A + O(s^2),\end{aligned}$$

hence

$$\frac{d}{ds} \det(I + sA)|_{s=0} = \operatorname{Tr} A.$$

2. Let  $B(s)$  be a smooth matrix valued function of  $s$ , with  $B(0) = I$ . Use Exercise 1 to show that

$$\frac{d}{ds} \det B(s)|_{s=0} = \operatorname{Tr} B'(0).$$

*Hint.* Write  $B(s) = I + sB'(0) + O(s^2)$ .

3. Let  $C(s)$  be a smooth matrix valued function, and assume  $C(0)$  is invertible. Use Exercise 2 plus

$$\det C(s) = (\det C(0)) \det B(s), \quad B(s) = C(0)^{-1}C(s)$$

to show that

$$\frac{d}{ds} \det C(s)|_{s=0} = (\det C(0)) \operatorname{Tr} C(0)^{-1}C'(0).$$

Use this to prove (3.8.8).

*Hint.* Fix  $t$  and set  $C(s) = M(t + s)$ , so

$$\frac{d}{dt} \det M(t) = \frac{d}{ds} \det C(s)|_{s=0}.$$

4. Show that, if  $M(t)$  is smooth and invertible on an interval  $I$ , then

$$\frac{d}{dt} M(t)^{-1} = -M(t)^{-1}M'(t)M(t)^{-1}.$$

Then apply  $d/dt$  to (3.8.12), to obtain an alternative derivation of Proposition 3.8.1.

*Hint.* Set  $U(t) = M(t)^{-1}$  and differentiate the identity  $U(t)M(t) = I$ .

5. Set up and solve the Wronskian equation (3.8.9) in the following cases:

$$A(t) = \begin{pmatrix} 1 & t \\ t & 1 \end{pmatrix}, \quad \begin{pmatrix} t & 1 \\ 1 & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ t & 0 \end{pmatrix}.$$

Exercises 6–7 generalize (3.8.27)–(3.8.29) from the case of zero torsion (cf. Exercise 6 of §3.7) to the case

$$(3.8.40) \quad \tau(t) = \beta\kappa(t), \quad \beta \text{ constant.}$$



6. Assume  $x(t)$  is a curve in  $\mathbb{R}^3$  for which (3.8.40) holds. Show that  $x(t) = x(0) + \int_0^t T(s) ds$ , with

$$(3.8.41) \quad (T(t), N(t), B(t)) = (T(0), N(0), B(0))e^{\sigma(t)K},$$

where

$$(3.8.42) \quad K = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & -\beta \\ 0 & \beta & 0 \end{pmatrix}, \quad \sigma(t) = \int_0^t \kappa(s) ds.$$

*Hint.* Use (3.7.26)–(3.7.27) and (3.8.21)  $\Rightarrow$  (3.8.22)–(3.8.24).

7. Let  $e_1, e_2, e_3$  denote the standard basis of  $\mathbb{R}^3$ , and let

$$v_1 = (1 + \beta^2)^{-1/2}(\beta e_1 + e_3), \quad v_2 = e_2, \quad v_3 = (1 + \beta^2)^{-1/2}(e_1 - \beta e_3).$$

Show that  $v_1, v_2, v_3$  form an orthonormal basis of  $\mathbb{R}^3$  and, with  $K$  as in (3.8.42),

$$Kv_1 = 0, \quad Kv_2 = -(1 + \beta^2)^{1/2}v_3, \quad Kv_3 = (1 + \beta^2)^{1/2}v_2.$$

Deduce that

$$\begin{aligned} e^{\sigma K}v_1 &= v_1, \\ e^{\sigma K}v_2 &= (\cos \eta)v_2 - (\sin \eta)v_3, \\ e^{\sigma K}v_3 &= (\sin \eta)v_2 + (\cos \eta)v_3, \end{aligned}$$

where  $\eta = (1 + \beta^2)^{1/2}\sigma$ .

8. Given  $B \in M(n, \mathbb{C})$ , write down the solution to

$$\frac{dx}{dt} = e^{tB}x, \quad x(0) = x_0.$$

*Hint.* Use (3.8.24).

Exercises 9–10 deal with a linear equation with periodic coefficients:

$$(3.8.43) \quad \frac{dx}{dt} = A(t)x, \quad A(t+1) = A(t).$$

Say  $A(t) \in M(n, \mathbb{C})$ .

9. Assume  $M(t)$  solves (3.8.6), with  $A(t)$  as in (3.8.43), and  $M(0) = I$ . Show that

$$(3.8.44) \quad M(1) = C \implies M(t+1) = M(t)C.$$

10. In the setting of Exercise 9, we know  $M(t)$  is invertible for all  $t$ , so  $C$  is invertible. Results of Appendix 3.A yield  $X \in M(n, \mathbb{C})$  such that

$$(3.8.45) \quad e^X = C.$$

Show that

$$(3.8.46) \quad P(t) = M(t)e^{-tX} \implies P(t+1) = P(t).$$

The representation

$$(3.8.47) \quad M(t) = P(t)e^{tX}$$

is called the *Floquet representation* of  $M(t)$ .

### 3.9. Variation of parameters and Duhamel's formula

An inhomogeneous equation

$$(3.9.1) \quad y'' + a(t)y' + b(t)y = f(t)$$

can be solved via the method of variation of parameters, if one is given a complete set  $u_1(t), u_2(t)$  of solutions to the homogeneous equation

$$(3.9.2) \quad u_j'' + a(t)u_j' + b(t)u_j = 0.$$

The method (derived already in §1.12 of Chapter 1 when  $a(t)$  and  $b(t)$  are constant) consists of seeking a solution to (3.9.1) in the form

$$(3.9.3) \quad y(t) = v_1(t)u_1(t) + v_2(t)u_2(t),$$

and finding equations for  $v_j(t)$  which can be solved and which work to yield a solution to (3.9.1). We have

$$(3.9.4) \quad y' = v_1u_1' + v_2u_2' + v_1'u_1 + v_2'u_2.$$

We impose the condition

$$(3.9.5) \quad v_1'u_1 + v_2'u_2 = 0.$$

Then  $y'' = v_1'u_1' + v_2'u_2' + v_1u_1'' + v_2u_2''$ , and plugging in (3.9.2) gives

$$(3.9.6) \quad y'' = v_1'u_1' + v_2'u_2' - (au_1' + bu_1)v_1 - (au_2' + bu_2)v_2,$$

hence

$$(3.9.7) \quad y'' + ay' + by = v_1'u_1' + v_2'u_2'.$$

Thus we have a solution to (3.9.1) in the form (3.9.3) provided  $v_1'$  and  $v_2'$  solve

$$(3.9.8) \quad \begin{aligned} v_1'u_1 + v_2'u_2 &= 0, \\ v_1'u_1' + v_2'u_2' &= f. \end{aligned}$$

This linear system for  $v_1', v_2'$  has the explicit solution

$$(3.9.9) \quad v_1' = -\frac{u_2}{W}f, \quad v_2' = \frac{u_1}{W}f,$$

where  $W(t)$  is the Wronskian:

$$(3.9.10) \quad W = u_1u_2' - u_2u_1' = \det \begin{pmatrix} u_1 & u_2 \\ u_1' & u_2' \end{pmatrix}.$$

Then

$$(3.9.11) \quad \begin{aligned} v_1(t) &= -\int_{t_0}^t \frac{u_2(s)}{W(s)} f(s) ds + C_1, \\ v_2(t) &= \int_{t_0}^t \frac{u_1(s)}{W(s)} f(s) ds + C_2. \end{aligned}$$

So

$$(3.9.12) \quad y(t) = C_1u_1(t) + C_2u_2(t) + \int_{t_0}^t [u_2(t)u_1(s) - u_1(t)u_2(s)] \frac{f(s)}{W(s)} ds.$$

Note that

$$u_2(t)u_1(s) - u_1(t)u_2(s) = \det \begin{pmatrix} u_1(s) & u_2(s) \\ u_1(t) & u_2(t) \end{pmatrix}.$$

We can connect the formula (3.9.12) with that produced in §3.8 as follows. If  $y(t)$  solves (3.9.1), then  $x(t) = (y(t), y'(t))^t$  solves the first order system

$$(3.9.13) \quad \frac{dx}{dt} = A(t)x + \begin{pmatrix} 0 \\ f(t) \end{pmatrix},$$

where

$$(3.9.14) \quad A(t) = \begin{pmatrix} 0 & 1 \\ -b(t) & -a(t) \end{pmatrix},$$

and a complete set of solutions to the homogeneous version of (3.9.13) is given by

$$(3.9.15) \quad x_j(t) = \begin{pmatrix} u_j(t) \\ u'_j(t) \end{pmatrix}, \quad j = 1, 2.$$

Thus we can set

$$(3.9.16) \quad M(t) = \begin{pmatrix} u_1(t) & u_2(t) \\ u'_1(t) & u'_2(t) \end{pmatrix},$$

and as in (3.8.19) we have

$$(3.9.17) \quad \begin{pmatrix} y(t) \\ y'(t) \end{pmatrix} = M(t)M(t_0)^{-1} \begin{pmatrix} y_0 \\ y_1 \end{pmatrix} + M(t) \int_{t_0}^t M(s)^{-1} \begin{pmatrix} 0 \\ f(s) \end{pmatrix} ds,$$

solving (3.9.13) with  $y(t_0) = y_0$ ,  $y'(t_0) = y_1$ . Note that

$$(3.9.18) \quad M(s)^{-1} = \frac{1}{W(s)} \begin{pmatrix} u'_2(s) & -u_2(s) \\ -u'_1(s) & u_1(s) \end{pmatrix},$$

with  $W(s)$ , the Wronskian, as in (3.9.10). Thus the last term on the right side of (3.9.17) is equal to

$$(3.9.19) \quad \begin{pmatrix} u_1(t) & u_2(t) \\ u'_1(t) & u'_2(t) \end{pmatrix} \int_{t_0}^t \frac{1}{W(s)} \begin{pmatrix} -u_2(s)f(s) \\ u_1(s)f(s) \end{pmatrix} ds,$$

and performing this matrix multiplication yields the integrand in (3.9.12). Thus we see that Duhamel's formula provides an alternative approach to the method of variation of parameters.

## Exercises

1. Use the method of variation of parameters to solve

$$(3.9.20) \quad y'' + y = \tan t.$$

2. Convert (3.9.20) to a  $2 \times 2$  first order system and use Duhamel's formula to solve it. Compare the result with your work on Exercise 1. Compare also with (3.4.6)–(3.4.9).

3. Do analogues of Exercises 1–2 for each of the following equations.

(a)  $y'' + y = e^t$ ,

(b)  $y'' + y = \sin t$ ,

(c)  $y'' + y = t$ ,

(d)  $y'' + y = t^2$ .

4. Show that the Wronskian, defined by (3.9.10), satisfies the equation

(3.9.21) 
$$\frac{dW}{dt} = -a(t)W,$$

if  $u_1$  and  $u_2$  solve (3.9.2). Relate this to (3.8.9).

5. Show that one solution to

(3.9.22) 
$$u'' + 2tu' + 2u = 0$$

is

(3.9.23) 
$$u_1(t) = e^{-t^2}.$$

Set up and solve the differential equation for  $W(t) = u_1u_2' - u_2u_1'$ . Then solve the associated first order equation for  $u_2$ , to produce a linearly independent solution  $u_2$  to (3.9.22), in terms of an integral.

6. Do Exercise 5 with (3.9.22) replaced by

$$u'' + 2u' + u = 0,$$

one of whose solutions is

$$u_1(t) = e^{-t}.$$

### 3.10. Power series expansions

Here we produce solutions to initial value problems

$$(3.10.1) \quad \frac{dx}{dt} = A(t)x + f(t), \quad x(0) = x_0,$$

in terms of a power series expansion,

$$(3.10.2) \quad x(t) = x_0 + x_1t + x_2t^2 + \cdots = \sum_{k=0}^{\infty} x_k t^k,$$

under the hypothesis that the  $n \times n$  matrix-valued function  $A(t)$  and vector-valued function  $f(t)$  are given by power series,

$$(3.10.3) \quad A(t) = \sum_{k=0}^{\infty} A_k t^k, \quad f(t) = \sum_{k=0}^{\infty} f_k t^k,$$

convergent for  $|t| < R_0$ . The coefficients  $x_k$  in (3.10.2) will be obtained recursively, as follows. Given  $x(t)$  of the form (3.10.2), we have

$$(3.10.4) \quad \frac{dx}{dt} = \sum_{k=1}^{\infty} k x_k t^{k-1} = \sum_{k=0}^{\infty} (k+1) x_{k+1} t^k,$$

and

$$(3.10.5) \quad \begin{aligned} A(t)x &= \sum_{j=0}^{\infty} A_j t^j \sum_{\ell=0}^{\infty} x_{\ell} t^{\ell} \\ &= \sum_{k=0}^{\infty} \left( \sum_{j=0}^k A_{k-j} x_j \right) t^k, \end{aligned}$$

so the power series on the left and right sides of (3.10.1) agree if and only if, for each  $k \geq 0$ ,

$$(3.10.6) \quad (k+1)x_{k+1} = \sum_{j=0}^k A_{k-j} x_j + f_k.$$

In particular, the first three recursions are

$$(3.10.7) \quad \begin{aligned} x_1 &= A_0 x_0 + f_0, \\ 2x_2 &= A_1 x_0 + A_0 x_1 + f_1, \\ 3x_3 &= A_2 x_0 + A_1 x_1 + A_0 x_2 + f_2. \end{aligned}$$

To start the recursion, the initial condition in (3.10.1) specifies  $x_0$ .

We next address the issue of convergence of the power series thus produced for  $x(t)$ . We will establish the following

**Proposition 3.10.1.** *Under the hypotheses given above, the power series (3.10.2) converges to the solution  $x(t)$  to (3.10.1), for  $|t| < R_0$ .*

**Proof.** The hypotheses on (3.10.3) imply that for each  $R < R_0$ , there exist  $a, b \in (0, \infty)$  such that

$$(3.10.8) \quad \|A_k\| \leq aR^{-k}, \quad \|f_k\| \leq bR^{-k}, \quad \forall k \in \mathbb{Z}^+.$$

We will show that, given  $r \in (0, R)$ , there exists  $C \in (0, \infty)$  such that

$$(3.10.9) \quad \|x_j\| \leq Cr^{-j}, \quad \forall j \in \mathbb{Z}^+.$$

Such estimates imply that the power series (3.10.2) converges for  $|t| < r$ , for each  $r < R_0$ , hence for  $|t| < R_0$ .

We will prove (3.10.9) by induction. The inductive step is to assume it holds for all  $j \leq k$  and to deduce that it holds for  $j = k + 1$ . This deduction proceeds as follows. We have, by (3.10.6), (3.10.8), and (3.10.9) for  $j \leq k$ ,

$$(3.10.10) \quad \begin{aligned} (k+1)\|x_{k+1}\| &\leq \sum_{j=0}^k \|A_{k-j}\| \cdot \|x_j\| + \|f_k\| \\ &\leq aC \sum_{j=0}^k R^{j-k} r^{-j} + bR^{-k} \\ &= aCr^{-k} \sum_{j=0}^k \left(\frac{r}{R}\right)^{k-j} + bR^{-k}. \end{aligned}$$

Now, given  $0 < r < R$ ,

$$(3.10.11) \quad \sum_{j=0}^k \left(\frac{r}{R}\right)^{k-j} < \sum_{j=0}^{\infty} \left(\frac{r}{R}\right)^j = \frac{1}{1 - \frac{r}{R}} = M(R, r) < \infty.$$

Hence

$$(3.10.12) \quad (k+1)\|x_{k+1}\| \leq aCM(R, r)r^{-k} + br^{-k}.$$

We place on  $C$  the constraint that

$$(3.10.13) \quad C \geq b,$$

and obtain

$$(3.10.14) \quad \|x_{k+1}\| \leq \frac{aM(R, r) + 1}{k+1} r \cdot Cr^{-k-1}.$$

This gives the desired result

$$(3.10.15) \quad \|x_{k+1}\| \leq Cr^{-k-1},$$

as long as

$$(3.10.16) \quad \frac{aM(R, r) + 1}{k+1} r \leq 1.$$

Thus, to finish the argument, we pick  $K \in \mathbb{N}$  such that

$$(3.10.17) \quad K + 1 \geq [aM(R, r) + 1]r.$$

(Recall that we have  $a, R, r$ , and  $M(R, r)$ .) Then we pick  $C \in (0, \infty)$  large enough that (3.10.9) holds for all  $j \in \{0, 1, \dots, K\}$ , i.e., we take (in addition to (3.10.13))

$$(3.10.18) \quad C \geq \max_{0 \leq j \leq K} r^j \|x_j\|.$$

Then for all  $k \geq K$ , the inductive step yielding (3.10.15) from the validity of (3.10.9) for all  $j \leq k$  holds, and the inductive proof of (3.10.9) is complete.  $\square$

For notational simplicity, we have discussed power series expansions about  $t = 0$  so far, but the same considerations apply to power series about a more general point  $t_0$ . Thus we could replace (3.10.1) by

$$(3.10.19) \quad \frac{dx}{dt} = A(t)x + f(t), \quad x(t_0) = x_0,$$

with  $A(t)$  and  $f(t)$  given by power series

$$(3.10.20) \quad A(t) = \sum_{k=0}^{\infty} A_k(t-t_0)^k, \quad f(t) = \sum_{k=0}^{\infty} f_k(t-t_0)^k,$$

for  $|t - t_0| < R_0$ , and find  $x(t)$  in the form

$$(3.10.21) \quad x(t) = \sum_{k=0}^{\infty} x_k(t-t_0)^k.$$

The recursive formula for the coefficients  $x_k$  is again given by (3.10.6), and (3.10.8)–(3.10.18) apply without further change.

It is worth noting that, in (3.10.20)–(3.10.21),

$$(3.10.22) \quad A_k = \frac{1}{k!} A^{(k)}(t_0), \quad f_k = \frac{1}{k!} f^{(k)}(t_0), \quad x_k = \frac{1}{k!} x^{(k)}(t_0).$$

These formulas, say for  $f_k$ , arise as follows. Setting  $t = t_0$  in (3.10.20) gives  $f_0 = f(t_0)$ . Generally, if the power series for  $f(t)$  converges for  $|t - t_0| < R_0$ , so does the power series

$$(3.10.23) \quad f'(t) = \sum_{k=1}^{\infty} k f_k(t-t_0)^{k-1},$$

and more generally,

$$(3.10.24) \quad f^{(n)}(t) = \sum_{k=n}^{\infty} k(k-1)\cdots(k-n+1) f_k(t-t_0)^{k-n},$$

and setting  $t = t_0$  in (3.10.24) gives  $f^{(n)}(t_0) = n! f_n$ .

As an aside, we mention that a convenient way to prove (3.10.23) is to define  $g(t)$  to be the power series on the right side of (3.10.23) and show that

$$(3.10.25) \quad \int_{t_0}^t g(s) ds = \sum_{k=1}^{\infty} f_k(t-t_0)^k = f(t) - f(t_0).$$

Compare the proof of Proposition 1.C.4 in Chapter 1.

We next establish the following important fact about functions given by convergent power series.

**Proposition 3.10.2.** *If  $f(t)$  is given by a power series as in (3.10.20), convergent for  $|t - t_0| < R_0$ , then  $f$  can also be expanded in a power series in  $t - t_1$ , for each  $t_1 \in (t_0 - R_0, t_0 + R_0)$ , with radius of convergence  $R_0 - |t_0 - t_1|$ .*



**Proof.** For notational simplicity, we take  $t_0 = 0$ . Thus we assume  $|t_1| < R_0$ . For  $|s| < R_0 - |t_1|$ , we have

$$\begin{aligned} f(t_1 + s) &= \sum_{k=0}^{\infty} f_k(t_1 + s)^k \\ (3.10.26) \quad &= \sum_{k=0}^{\infty} \sum_{j=0}^k f_k \binom{k}{j} s^j t_1^{k-j}, \end{aligned}$$

the second identity by the binomial formula. Call the last double series  $\sum_k \sum_j \alpha_{jk}$ . Note that

$$\begin{aligned} \sum_{j,k} \|\alpha_{jk}\| &= \sum_{k=0}^{\infty} \sum_{j=0}^k \|f_k\| \binom{k}{j} |s|^j |t_1|^{k-j} \\ (3.10.27) \quad &= \sum_{k=0}^{\infty} \|f_k\| (|s| + |t_1|)^k \\ &< \infty, \end{aligned}$$

given  $|s| + |t_1| < R_0$ . In such a case, we can reverse the order of summation, and write

$$(3.10.28) \quad \sum_{k=0}^{\infty} \sum_{j=0}^k \alpha_{jk} = \sum_{j=0}^{\infty} \sum_{k=j}^{\infty} \alpha_{jk},$$

all series being absolutely convergent. Hence

$$(3.10.29) \quad f(t_1 + s) = \sum_{j=0}^{\infty} \left( \sum_{k=j}^{\infty} f_k \binom{k}{j} t_1^{k-j} \right) s^j$$

is an absolutely convergent series, as long as  $|s| < R_0 - |t_1|$ .  $\square$

A (vector-valued) function  $f$  defined on an interval  $I = (a, b)$  is said to be an *analytic function* on  $I$  if and only if for each  $t_1 \in I$ , there is an  $r_1 > 0$  such that for  $|t - t_1| < r_1$ ,  $f(t)$  is given by a convergent power series in  $t - t_1$ . Parallel to (3.10.22), such a power series is necessarily of the form

$$(3.10.30) \quad f(t) = \sum_{n=0}^{\infty} \frac{f^{(n)}(t_1)}{n!} (t - t_1)^n.$$

It follows from Proposition 3.10.2 that if  $f(t)$  is given by a convergent power series in  $t - t_0$  for  $|t - t_0| < R_0$ , then  $f$  is analytic in the interval  $(t_0 - R_0, t_0 + R_0)$ .

The following is a useful fact about analytic functions.

**Lemma 3.10.3.** *If  $f$  is analytic on  $(a, b)$  and  $a < \alpha < \beta < b$ , then there exists  $\delta > 0$  such that, for each  $t_1 \in [\alpha, \beta]$ , the power series (3.10.30) converges whenever  $|t - t_1| < \delta$ .*

**Proof.** By hypothesis, each  $p \in [\alpha, \beta]$  is the center of an open interval  $I_p$  on which  $f$  is given by a convergent power series about  $p$ . Let  $(1/2)I_p$  denote the open interval centered at  $p$  whose length is half that of  $I_p$ . Since  $[\alpha, \beta]$  is a closed, bounded interval, results of Appendix 4.B in Chapter 4 (see Proposition 4.B.9) imply that

there exists a finite set  $\{p_1, \dots, p_K\} \subset [\alpha, \beta]$  such that the intervals  $(1/2)I_{p_j}$  cover  $[\alpha, \beta]$ . Take

$$(3.10.31) \quad \delta = \frac{1}{4} \min_{1 \leq j \leq K} \ell(I_{p_j}).$$

Then each  $t_1 \in [\alpha, \beta]$  is contained in some  $(1/2)I_{p_j}$ , and hence  $|t_1 - p_j| \leq (1/4)\ell(I_{p_j})$ , so

$$(3.10.32) \quad (t_1 - \delta, t_1 + \delta) \subset I_{p_j}.$$

Hence the convergence of the power series for  $f$  about  $t_1$  on  $(t_1 - \delta, t_1 + \delta)$  follows from Proposition 3.10.2.  $\square$

With these tools in hand, we have the following result.

**Proposition 3.10.4.** *Assume  $A(t)$  and  $f(t)$  are analytic on an interval  $(a, b)$  and  $t_0 \in (a, b)$ . Then the initial value problem (3.10.19) has a unique solution  $x(t)$ , analytic on  $(a, b)$ .*

**Proof.** Let  $\tilde{I} \subset (a, b)$  be the maximal interval on which the solution  $x(t)$  exists and is analytic, say  $\tilde{I} = (\alpha, \beta)$ . If  $\beta < b$ , take  $\delta$  as in Lemma 10.3 for  $[t_0, \beta]$ . Take  $t_1 = \beta - \delta/2$ . (If necessary, shrink  $\delta$  to arrange that  $t_1 > t_0$ .) Then  $A(t)$  and  $f(t)$  have convergent power series about  $t_1$ , with radius of convergence  $\geq \delta$ , so

$$(3.10.33) \quad x(t) \text{ extends analytically to } [t_1, t_1 + \delta), \text{ and } t_1 + \delta > \beta.$$

This contradiction proves that  $\beta = b$ , and a similar argument gives  $\alpha = a$ , so in fact  $\tilde{I} = (a, b)$ .  $\square$

To conclude this section, we mention a connection with the study of functions of a complex variable, which the reader could pursue further, consulting texts on complex analysis, such as [4], or Chapter 7 of [47]. Here is the general set-up. Let  $\Omega \subset \mathbb{C}$  be an open set, and  $f : \Omega \rightarrow \mathbb{C}$ . We say  $f$  is *complex differentiable* at  $z_0 \in \Omega$  provided

$$(3.10.34) \quad \lim_{w \rightarrow 0} \frac{f(z_0 + w) - f(z_0)}{w}$$

exists. Here,  $w \rightarrow 0$  in  $\mathbb{C}$ . If this limit exists, we call it

$$(3.10.35) \quad f'(z_0).$$

We say  $f$  is complex differentiable on  $\Omega$  if it is complex differentiable at each  $z_0 \in \Omega$ .

The relevance of this concept to the material of this section is the following. If  $f(t)$  is given by the power series (3.10.20), absolutely convergent for real  $t \in (t_0 - R_0, t_0 + R_0)$ , then

$$(3.10.36) \quad f(z) = \sum_{k=0}^{\infty} f_k (z - t_0)^k$$

is absolutely convergent for  $z \in \mathbb{C}$  satisfying  $|z - t_0| < R_0$ , i.e., on the disk

$$(3.10.37) \quad D_{R_0}(t_0) = \{z \in \mathbb{C} : |z - t_0| < R_0\},$$

and it is complex differentiable on this disk. Furthermore,  $f'$  is complex differentiable on this disk, etc., including the  $k$ th order derivative  $f^{(k)}$ , and

$$(3.10.38) \quad f_k = \frac{f^{(k)}(t_0)}{k!}.$$

More is true, namely the following converse.

**Theorem 3.10.5.** *Assume  $f$  is complex differentiable on the open set  $\Omega \subset \mathbb{C}$ . Let  $t_0 \in \Omega$  and assume  $D_{R_0}(t_0) \subset \Omega$ . Then  $f$  is given by a power series, of the form (3.10.36), absolutely convergent on  $D_{R_0}(t_0)$ .*

This is one of the central basic results of complex analysis. A proof can be found in Chapter 5 of [4], and in Chapter 2 of [47]. In view of Theorem 3.10.5, complex differentiable functions are also called “complex analytic.”

EXAMPLE. Consider

$$(3.10.39) \quad f(z) = \frac{1}{z^2 + 1}.$$

This is well defined except at  $\pm i$ , where the denominator vanishes, and one can readily verify that  $f$  is complex differentiable on  $\mathbb{C} \setminus \{i, -i\}$ . It follows from Theorem 3.10.5 that if  $t_0 \in \mathbb{C} \setminus \{i, -i\}$ , this function is given by a power series expansion about  $t_0$ , absolutely convergent on  $D_R(t_0)$ , where

$$(3.10.40) \quad R = \min \{|t_0 - i|, |t_0 + i|\}.$$

In particular,

$$(3.10.41) \quad t_0 \in \mathbb{R} \implies R = \sqrt{t_0^2 + 1}$$

gives the radius of convergence of the power series expansion of  $1/(z^2 + 1)$  about  $t_0$ . This is easy to see directly for  $t_0 = 0$ :

$$(3.10.42) \quad \frac{1}{z^2 + 1} = \sum_{k=0}^{\infty} (-1)^k z^{2k}.$$

However, for other  $t_0 \in \mathbb{R}$ , it is not so easy to see directly that this function has a power series expansion about  $t_0$  with radius of convergence given by (3.10.41). The reader might give this a try.

To interface this example with Proposition 3.10.4, we note that, by this proposition, plus the results just derived on  $1/(z^2 + 1)$ , the equation

$$(3.10.43) \quad \frac{dx}{dt} = \begin{pmatrix} 1 & (t^2 + 1)^{-1} \\ 0 & -1 \end{pmatrix} x + \begin{pmatrix} e^t \\ e^{-t} \end{pmatrix}, \quad x(0) = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

has a solution that is analytic on  $(-\infty, \infty)$ . The power series expansion for the solution  $x(t)$  about  $t_0$  converges for  $|t| < 1$  if  $t_0 = 0$  (this is an easy consequence of Proposition 3.10.4 and (3.10.42)), and for other  $t_0 \in \mathbb{R}$ , it converges for  $|t - t_0| < \sqrt{t_0^2 + 1}$  (as a consequence of Proposition 3.10.4 and Theorem 3.10.5).

See Appendix 3.C for a further discussion of complex analytic functions.

---

**Exercises**

1. Consider the function  $g : \mathbb{C} \setminus \{0\} \rightarrow \mathbb{C}$  given by

$$(3.10.44) \quad g(z) = e^{-1/z^2}.$$

Show that  $g$  is complex differentiable on  $\mathbb{C} \setminus \{0\}$ . Use Theorem 3.10.5 to deduce that  $h : \mathbb{R} \rightarrow \mathbb{R}$ , given by

$$(3.10.45) \quad h(t) = \begin{cases} e^{-1/t^2}, & t \neq 0, \\ 0, & t = 0, \end{cases}$$

is analytic on  $\mathbb{R} \setminus \{0\}$ . Show that  $h$  is not analytic on any interval containing 0. Compute

$$h^{(k)}(0).$$

2. Consider the Airy equation

$$(3.10.46) \quad y'' = ty, \quad y(0) = y_0, \quad y'(0) = y_1,$$

Introduced in (1.15.9) of Chapter 1. Show that this yields the first order system

$$(3.10.47) \quad \frac{dx}{dt} = (A_0 + A_1 t)x, \quad x(0) = x_0,$$

with

$$(3.10.48) \quad A_0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad x_0 = \begin{pmatrix} y_0 \\ y_1 \end{pmatrix}.$$

Note that

$$(3.10.49) \quad A_0^2 = A_1^2 = 0,$$

and

$$(3.10.50) \quad A_0 A_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad A_1 A_0 = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

3. For a system of the form (3.10.47), whose solution has a power series of the form (3.10.2), the recursion (3.10.6) becomes

$$(3.10.51) \quad (k+1)x_{k+1} = A_0 x_k + A_1 x_{k-1},$$

with the convention that  $x_{-1} = 0$ . Assume (3.10.49) holds. Show that

$$(3.10.52) \quad x_{k+3} = \frac{1}{k+3} \left( \frac{1}{k+2} A_0 A_1 + \frac{1}{k+1} A_1 A_0 \right) x_k.$$

Note that when  $A_0$  and  $A_1$  are given by (3.10.48), this becomes

$$(3.10.53) \quad x_{k+3} = \frac{1}{k+3} \begin{pmatrix} \frac{1}{k+2} & \\ & \frac{1}{k+1} \end{pmatrix} x_k.$$

Establish directly from (3.10.52) that the series  $\sum x_k t^k$  is absolutely convergent for all  $t$ . *Hint.* Separately tackle the three series

$$(3.10.54) \quad \sum_{\ell=0}^{\infty} x_{3\ell+j} t^{3\ell+j}, \quad j = 0, 1, 2.$$

Use (3.10.52) and the ratio test to show that each one converges for all  $t$ .

### 3.11. Regular singular points

Here we consider equations of the form

$$(3.11.1) \quad t \frac{dx}{dt} = A(t)x,$$

where  $x$  takes values in  $\mathbb{C}^n$ , and  $A(t)$ , with values in  $M(n, \mathbb{C})$ , has a power series convergent for  $t$  in some interval  $(-T_0, T_0)$ ,

$$(3.11.2) \quad A(t) = A_0 + A_1 t + A_2 t^2 + \cdots.$$

The system (3.11.1) is said to have a regular singular point at  $t = 0$ . One source of such systems is the following class of second order equations:

$$(3.11.3) \quad t^2 u''(t) + tb(t)u'(t) + c(t)u(t) = 0,$$

where  $b(t)$  and  $c(t)$  have convergent power series for  $|t| < T_0$ . In such a case, one can set

$$(3.11.4) \quad x(t) = \begin{pmatrix} u(t) \\ v(t) \end{pmatrix}, \quad v(t) = tu'(t),$$

obtaining (3.11.1) with

$$(3.11.5) \quad A(t) = \begin{pmatrix} 0 & 1 \\ -c(t) & 1 - b(t) \end{pmatrix}.$$

A paradigm example, studied in Section 1.16 of Chapter 1, is the Bessel equation

$$(3.11.6) \quad \frac{d^2 u}{dt^2} + \frac{1}{t} \frac{du}{dt} + \left(1 - \frac{\nu^2}{t^2}\right)u = 0,$$

which via (3.11.4) takes the form (3.11.1), with

$$(3.11.7) \quad A(t) = A_0 + A_2 t^2, \quad A_0 = \begin{pmatrix} 0 & 1 \\ \nu^2 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

It follows from Proposition 3.10.4 that, given  $t_0 \in (0, T_0)$ , the equation (3.11.1), with initial condition  $x(t_0) = x_0$ , has a unique solution analytic on  $(0, T_0)$ . Our goal here is to analyze the behavior of  $x(t)$  as  $t \searrow 0$ .

A starting point for the analysis of (3.11.1) is the case  $A(t) \equiv A_0$ , i.e.,

$$(3.11.8) \quad t \frac{dx}{dt} = A_0 x.$$

The change of variable  $z(s) = x(e^s)$  yields

$$(3.11.9) \quad \frac{dz}{ds} = A_0 z(s),$$

with solution

$$(3.11.10) \quad z(s) = e^{sA_0} v, \quad v = z(0),$$

hence

$$(3.11.11) \quad x(t) = e^{(\log t)A_0} v = t^{A_0} v, \quad t > 0,$$

the latter identity defining  $t^{A_0}$ , for  $t > 0$ . Compare results on the ‘‘Euler equations’’ in Exercises 1–3, §1.15, Chapter 1.

Note that if  $v \in \mathcal{E}(A_0, \lambda)$ , then  $t^{A_0}v = t^\lambda v$ , which either blows up or vanishes as  $t \searrow 0$ , if  $\operatorname{Re} \lambda < 0$  or  $\operatorname{Re} \lambda > 0$ , respectively, or oscillates rapidly as  $t \searrow 0$ , if  $\lambda$  is purely imaginary but not zero. On the other hand,

$$(3.11.12) \quad v \in \mathcal{N}(A_0) \implies t^{A_0}v \equiv v.$$

It is useful to have the following extension of this result to the setting of (3.11.1).

**Lemma 3.11.1.** *If  $v \in \mathcal{N}(A_0)$ , then (3.11.1) has a solution given by a convergent power series on some interval about the origin,*

$$(3.11.13) \quad x(t) = x_0 + x_1 t + x_2 t^2 + \cdots, \quad x_0 = v,$$

as long as the eigenvalues of  $A_0$  satisfy a mild condition, given in (3.11.18) below.

**Proof.** We produce a recursive formula for the coefficients  $x_k$  in (3.11.13), in the spirit of the calculations of §3.10. We have

$$(3.11.14) \quad t \frac{dx}{dt} = \sum_{k \geq 1} k x_k t^k,$$

and

$$(3.11.15) \quad \begin{aligned} A(t)x &= \sum_{j \geq 0} A_j t^j \sum_{\ell \geq 0} x_\ell t^\ell \\ &= A_0 x_0 + \sum_{k \geq 1} \sum_{\ell=0}^k A_{k-\ell} x_\ell t^k. \end{aligned}$$

Equating the power series in (3.11.14) and (3.11.15) would be impossible without our hypothesis that  $A_0 x_0 = 0$ , but having that, we obtain the recursive formulas, for  $k \geq 1$ ,

$$(3.11.16) \quad k x_k = A_0 x_k + \sum_{\ell=0}^{k-1} A_{k-\ell} x_\ell,$$

i.e.,

$$(3.11.17) \quad (kI - A_0)x_k = \sum_{\ell=0}^{k-1} A_{k-\ell} x_\ell.$$

Clearly we can solve uniquely for  $x_k$  provided

$$(3.11.18) \quad \forall k \in \mathbb{N} = \{1, 2, 3, \dots\}, \quad k \notin \operatorname{Spec} A_0.$$

This is the condition on  $\operatorname{Spec} A_0$  mentioned in the lemma. As long as this holds, we can solve for the coefficients  $x_k$  for all  $k \in \mathbb{N}$ , obtaining (3.11.13). Estimates on these coefficients implying that (3.11.13) has a positive radius of convergence are quite similar to those made in §3.10, and will not be repeated here.  $\square$

Our next goal is to extend this analysis to solutions to (3.11.1), for general  $A(t)$ , of the form (3.11.2), without such an hypothesis of membership in  $\mathcal{N}(A_0)$  as in Lemma 3.11.1. We will seek a matrix-valued power series

$$(3.11.19) \quad U(t) = I + U_1 t + U_2 t^2 + \cdots$$

such that under the change of variable

$$(3.11.20) \quad x(t) = U(t)y(t),$$

(11.1) becomes

$$(3.11.21) \quad t \frac{dy}{dt} = A_0 y.$$

This will work as long as  $A_0$  does not have two eigenvalues that differ by a nonzero integer, in which case a more elaborate construction will be needed.

To implement (3.11.20) and achieve (3.11.21), we have from (3.11.20) and (3.11.1) that

$$(3.11.22) \quad A(t)U(t)y = t \frac{dx}{dt} = tU(t) \frac{dy}{dt} + tU'(t)y,$$

which gives (3.11.21) provided  $U(t)$  satisfies

$$(3.11.23) \quad t \frac{dU}{dt} = A(t)U(t) - U(t)A_0.$$

Now (3.11.23) has the same form as (3.11.1), i.e.,

$$(3.11.24) \quad t \frac{dU}{dt} = \mathcal{A}(t)U(t),$$

where  $U$  takes values in  $M(n, \mathbb{C})$  and  $\mathcal{A}(t)$  takes values in  $\mathcal{L}(M(n, \mathbb{C}))$ ;

$$(3.11.25) \quad \begin{aligned} \mathcal{A}(t)U &= A(t)U(t) - U(t)A_0 \\ &= (\mathcal{A}_0 + \mathcal{A}_1 t + \mathcal{A}_2 t^2 + \cdots)U. \end{aligned}$$

In particular,

$$(3.11.26) \quad \mathcal{A}_0 U = A_0 U - U A_0 = [A_0, U] = C_{A_0} U,$$

the last identity defining  $C_{A_0} \in \mathcal{L}(M(n, \mathbb{C}))$ . Note that

$$(3.11.27) \quad U(0) = I \in \mathcal{N}(C_{A_0}),$$

so Lemma 3.11.1 applies to (3.11.24), i.e., to (3.11.23). In this setting, the recursion for  $U_k$ ,  $k \geq 1$ , analogous to (3.11.16)–(3.11.17), takes the form

$$(3.11.28) \quad kU_k = [A_0, U_k] + \sum_{j=0}^{k-1} A_{k-j} U_j,$$

i.e.,

$$(3.11.29) \quad (kI - C_{A_0})U_k = \sum_{j=0}^{k-1} A_{k-j} U_j.$$

Recall  $U_0 = I$ . The condition for solvability of (3.11.29) for all  $k \in \mathbb{N} = \{1, 2, 3, \dots\}$  is that no positive integer belong to  $\text{Spec } C_{A_0}$ . Results of Chapter 2, §2.7 (cf. Exercise 9) yield the following:

$$(3.11.30) \quad \text{Spec } A_0 = \{\lambda_j\} \implies \text{Spec } C_{A_0} = \{\lambda_j - \lambda_k\}.$$

Thus the condition that  $\text{Spec } C_{A_0}$  contain no positive integer is equivalent to the condition that  $A_0$  have no two eigenvalues that differ by a nonzero integer. We thus have the following result.



**Proposition 3.11.2.** *Assume  $A_0$  has no two eigenvalues that differ by a nonzero integer. Then there exists  $T_0 > 0$  and  $U(t)$  as in (3.11.19) with power series convergent for  $|t| < T_0$ , such that the general solution to (3.11.1) for  $t \in (0, T_0)$  has the form*

$$(3.11.31) \quad x(t) = U(t)t^{A_0}v, \quad v \in \mathbb{C}^n.$$

Let us see how Proposition 3.11.2 applies to the Bessel equation (3.11.6), which we have recast in the form (3.11.1) with  $A(t) = A_0 + A_2t^2$ , as in (3.11.7). Note that

$$(3.11.32) \quad A_0 = \begin{pmatrix} 0 & 1 \\ \nu^2 & 0 \end{pmatrix} \implies \text{Spec } A_0 = \{\nu, -\nu\}.$$

Thus  $\text{Spec } C_{A_0} = \{2\nu, 0, -2\nu\}$ , and Proposition 3.11.2 applies whenever  $\nu$  is not an integer or half-integer. Now as shown in §1.16 of Chapter 1, there is not an obstruction to series expansions consistent with (3.11.31) when  $\nu$  is a half-integer. This is due to the special structure of (3.11.7), and suggests a more general result, of the following sort. Suppose only even powers of  $t$  appear in the series for  $A(t)$ :

$$(3.11.33) \quad A(t) = A_0 + A_2t^2 + A_4t^4 + \cdots.$$

Then we look for  $U(t)$ , solving (3.11.23), in the form

$$(3.11.34) \quad U(t) = I + U_2t^2 + U_4t^4 + \cdots.$$

In such a case, only even powers of  $t$  occur in the power series for (3.11.23), and in place of (3.11.28)–(3.11.29), one gets the following recursion formulas for  $U_{2k}$ ,  $k \geq 1$ :

$$(3.11.35) \quad 2kU_{2k} = [A_0, U_{2k}] + \sum_{j=0}^{k-1} A_{2k-2j}U_{2j},$$

i.e.,

$$(3.11.36) \quad (2kI - C_{A_0})U_{2k} = \sum_{j=0}^{k-1} A_{2k-2j}U_{2j}.$$

This is solvable for  $U_{2k}$  as long as  $2k \notin \text{Spec } C_{A_0}$ , and we have the following.

**Proposition 3.11.3.** *Assume  $A(t)$  satisfies (3.11.33), and  $A_0$  has no two eigenvalues that differ by a nonzero even integer. Then there exists  $T_0 > 0$  and  $U(t)$  as in (3.11.34), with power series convergent for  $|t| < T_0$ , such that the general solution to (3.11.1) for  $t \in (0, T_0)$  has the form (3.11.31).*

We return to the Bessel equation (3.11.6) and consider the case  $\nu = 0$ . That is, we consider (3.11.1) with  $A(t)$  as in (3.11.7), and

$$(3.11.37) \quad A_0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Proposition 3.11.3 applies to this case (so does Proposition 3.11.2), and the general solution to (3.11.6) is the first entry in (3.11.31), where  $U(t)$  has the form (3.11.34). Note that in case (3.11.37),  $A_0^2 = 0$ , so for  $t > 0$ ,

$$(3.11.38) \quad t^{A_0} = \begin{pmatrix} 1 & \log t \\ 0 & 1 \end{pmatrix}.$$

Thus there are two linearly independent solutions to

$$(3.11.39) \quad \frac{d^2 u}{dt^2} + \frac{1}{t} \frac{du}{dt} + u = 0,$$

for  $t > 0$ , one having the form

$$(3.11.40) \quad \sum_{k \geq 0} a_k t^{2k},$$

with coefficients  $a_k$  given recursively, and another having the form

$$(3.11.41) \quad \sum_{k \geq 0} (b_k + c_k \log t) t^{2k},$$

again with coefficients  $b_k$  and  $c_k$  given recursively. The solution of the form (3.11.40) is as in (1.16.16) of Chapter 1 (with  $\nu = 0$ ), while the solution of the form (3.11.41) can be shown to be consistent with  $Y_0(t)$  in (1.16.34)–(1.16.36) of Chapter 1.

Proceeding beyond the purview of Propositions 3.11.2 and 3.11.3, we now treat the case when  $A_0$  satisfies the following conditions. First,

$$(3.11.42) \quad \text{Spec } C_{A_0} \text{ contains exactly one positive integer, } \ell,$$

and second

$$(3.11.43) \quad A_0 \text{ is diagonalizable,}$$

which implies

$$(3.11.44) \quad C_{A_0} \text{ is diagonalizable;}$$

cf. Chapter 2, §2.7, Exercise 8. Later we discuss weakening these conditions.

As in Proposition 3.11.2, we use a transformation of the form (3.11.20), i.e.,  $x(t) = U(t)y(t)$ , with  $U(t)$  as in (3.11.19), but this time our goal is to obtain for  $y$ , not the equation (3.11.21), but rather one of the form

$$(3.11.45) \quad t \frac{dy}{dt} = (A_0 + B_\ell t^\ell) y,$$

with the additional property of a special structure on the commutator  $[A_0, B_\ell]$ , given in (3.11.55) below. To get (3.11.45), we use (3.11.22) to obtain for  $U(t)$  the equation

$$(3.11.46) \quad t \frac{dU}{dt} = A(t)U(t) - U(t)(A_0 + B_\ell t^\ell),$$

in place of (3.11.23). Taking  $A(t)$  as in (3.11.2) and  $U(t)$  as in (3.11.19), we have

$$(3.11.47) \quad t \frac{dU}{dt} = \sum_{k \geq 1} k U_k t^k,$$

$$(3.11.48) \quad A(t)U(t) - U(t)A_0 = \sum_{k \geq 1} [A_0, U_k] t^k + \sum_{k \geq 1} \left( \sum_{j=0}^{k-1} A_{k-j} U_j \right) t^k,$$

and hence solving (3.11.46) requires for  $U_k$ ,  $k \geq 1$ , that

$$(3.11.49) \quad k U_k = [A_0, U_k] + \sum_{j=0}^{k-1} A_{k-j} U_j - \Gamma_k,$$

where

$$(3.11.50) \quad \begin{aligned} \Gamma_k &= 0, & k < \ell, \\ B_\ell, & & k = \ell, \\ U_{k-\ell} B_\ell, & & k > \ell. \end{aligned}$$

Equivalently,

$$(3.11.51) \quad (kI - C_{A_0})U_k = \sum_{j=0}^{k-1} A_{k-j}U_j - \Gamma_k.$$

As before, (3.11.51) has a unique solution for each  $k < \ell$ , since  $C_{A_0} - kI$  is invertible on  $M(n, \mathbb{C})$ . For  $k = \ell$ , the equation is

$$(3.11.52) \quad (\ell I - C_{A_0})U_\ell = \sum_{j=0}^{\ell-1} A_{\ell-j}U_j - B_\ell.$$

This time  $C_{A_0} - \ell I$  is not invertible. However, if (3.11.44) holds,

$$(3.11.53) \quad M(n, \mathbb{C}) = \mathcal{N}(C_{A_0} - \ell I) \oplus \mathcal{R}(C_{A_0} - \ell I).$$

Consequently, given  $\sum_{j=0}^{\ell-1} A_{\ell-j}U_j \in M(n, \mathbb{C})$ , we can take

$$(3.11.54) \quad B_\ell \in \mathcal{N}(C_{A_0} - \ell I)$$

so that the right side of (3.11.52) belongs to  $\mathcal{R}(C_{A_0} - \ell I)$ , and then we can find a solution  $U_\ell$ . We can uniquely specify  $U_\ell$  by requiring  $U_\ell \in \mathcal{R}(C_{A_0} - \ell I)$ , though that is of no great consequence. Having such  $B_\ell$  and  $U_\ell$ , we can proceed to solve (3.11.51) for each  $k > \ell$ . Estimates on the coefficients  $U_k$  guaranteeing a positive radius of convergence for the power series (3.11.19) again follow by techniques of §3.10. We have reduced the problem of representing the general solution to (3.11.1) for  $t \in (0, T_0)$  to that of representing the general solution to (3.11.45), given that (3.11.54) holds. The following result accomplishes this latter task. Note that (3.11.54) is equivalent to

$$(3.11.55) \quad [A_0, B_\ell] = \ell B_\ell, \quad \text{i.e., } A_0 B_\ell = B_\ell (A_0 + \ell I).$$

**Lemma 3.11.4.** *Given  $A_0, B_\ell \in M(n, \mathbb{C})$  satisfying (3.11.55), the general solution to (3.11.45) on  $t > 0$  is given by*

$$(3.11.56) \quad y(t) = t^{A_0} t^{B_\ell} v, \quad v \in \mathbb{C}^n.$$

**Proof.** As mentioned earlier in this section, results of §3.10 imply that for each  $v \in \mathbb{C}^n$ , there is a unique solution to (3.11.45) on  $t > 0$  satisfying  $y(1) = v$ . It remains to show that the right side of (3.11.56) satisfies (3.11.45). Indeed, if  $y(t)$  is given by (3.11.56), then, for  $t > 0$ ,

$$(3.11.57) \quad t \frac{dy}{dt} = A_0 t^{A_0} t^{B_\ell} v + t^{A_0} B_\ell t^{B_\ell} v.$$

Now (3.11.55) implies, for each  $m \in \mathbb{N}$ ,

$$(3.11.58) \quad \begin{aligned} A_0^m B_\ell &= A_0^{m-1} B_\ell (A_0 + \ell I) = \cdots \\ &= B_\ell (A_0 + \ell I)^m, \end{aligned}$$

which in turn implies

$$(3.11.59) \quad e^{sA_0} B_\ell = B_\ell e^{s(A_0 + \ell I)} = B_\ell e^{s\ell} e^{sA_0},$$

hence

$$(3.11.60) \quad t^{A_0} B_\ell = B_\ell t^\ell t^{A_0},$$

so (3.11.57) gives

$$(3.11.61) \quad t \frac{dy}{dt} = (A_0 + B_\ell t^\ell) t^{A_0} t^{B_\ell} v,$$

as desired.  $\square$

The construction involving (3.11.45)–(3.11.55) plus Lemma 3.11.4 yields the following.

**Proposition 3.11.5.** *Assume  $A_0 \in M(n, \mathbb{C})$  has the property (3.11.42) and is diagonalizable. Then there exist  $T_0 > 0$ ,  $U(t)$  as in (3.11.19), and  $B_\ell \in M(n, \mathbb{C})$ , satisfying (3.11.55), such that the general solution to (3.11.1) on  $t \in (0, T_0)$  is*

$$(3.11.62) \quad x(t) = U(t) t^{A_0} t^{B_\ell} v, \quad v \in \mathbb{C}^n.$$

The following is an important property of  $B_\ell$ .

**Proposition 3.11.6.** *In the setting of Proposition 3.11.5,  $B_\ell$  is nilpotent.*

**Proof.** This follows readily from (3.11.55), which implies that for each  $\lambda_j \in \text{Spec } A_0$ ,

$$(3.11.63) \quad B_\ell : \mathcal{GE}(A_0, \lambda_j) \longrightarrow \mathcal{GE}(A_0, \lambda_j + \ell).$$

$\square$

REMARK. Note that if  $B_\ell^{m+1} = 0$ , then, for  $t > 0$ ,

$$(3.11.64) \quad t^{B_\ell} = \sum_{k=0}^m \frac{1}{k!} (\log t)^k B_\ell^k.$$

Let us apply these results to the Bessel equation (3.11.6) in case  $\nu = n$  is a positive integer. We are hence looking at (3.11.1) when

$$(3.11.65) \quad A(t) = A_0 + A_2 t^2, \quad A_0 = \begin{pmatrix} 0 & 1 \\ n^2 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

We have

$$(3.11.66) \quad \text{Spec } A_0 = \{n, -n\}, \quad \text{Spec } C_{A_0} = \{2n, 0, -2n\}.$$

Clearly  $A_0$  is diagonalizable. The recursion (3.11.51) for  $U_k$  takes the form

$$(3.11.67) \quad (kI - C_{A_0})U_k = \Sigma_k + \Gamma_k,$$

where

$$(3.11.68) \quad \begin{aligned} \Sigma_k &= 0, & k < 2, \\ & A_2, & k = 2, \\ & A_2 U_{k-2}, & k > 2, \end{aligned}$$

and

$$(3.11.69) \quad \begin{aligned} \Gamma_k &= 0, & k < 2n, \\ B_{2n}, & & k = 2n, \\ U_{k-2n}B_{2n}, & & k > 2n. \end{aligned}$$

In particular, the critical equation (3.11.52) is

$$(3.11.70) \quad (2nI - C_{A_0})U_{2n} = A_2U_{2n-2} + B_{2n},$$

and we solve this after picking

$$(3.11.71) \quad B_{2n} \in \mathcal{N}(C_{A_0} - 2nI),$$

such that the right side of (3.11.70) belongs to  $\mathcal{R}(C_{A_0} - 2nI)$ . We have from Chapter 2, §2.7, Exercise 9, that (since  $A_0$  is diagonalizable)

$$(3.11.72) \quad \mathcal{N}(C_{A_0} - 2nI) = \text{Span}\{vw^t : v \in \mathcal{E}(A_0, n), w \in \mathcal{E}(A_0^t, -n)\}.$$

For  $A_0$  in (3.11.64),  $v$  is a multiple of  $(1, n)^t$  and  $w^t$  is a multiple of  $(-n, 1)$ , so

$$(3.11.73) \quad B_{2n} = \beta_n B_{2n}^\#, \quad \beta_n \in \mathbb{C}, \quad B_{2n}^\# = \begin{pmatrix} -n & 1 \\ -n^2 & n \end{pmatrix}.$$

Note that  $B_{2n}^2 = 0$ . Consequently the general solution to (3.11.1) in this case takes the form

$$(3.11.74) \quad x(t) = U(t)t^{A_0}t^{B_{2n}}v,$$

with

$$(3.11.75) \quad t^{B_{2n}} = I + (\log t)B_{2n}.$$

Note that

$$(3.11.76) \quad \mathcal{N}(B_{2n}) = \text{Span} \begin{pmatrix} 1 \\ n \end{pmatrix} = \mathcal{E}(A_0, n),$$

so

$$(3.11.77) \quad U(t)t^{A_0}t^{B_{2n}} \begin{pmatrix} 1 \\ n \end{pmatrix} = U(t)t^{A_0} \begin{pmatrix} 1 \\ n \end{pmatrix} = U(t)t^n \begin{pmatrix} 1 \\ n \end{pmatrix}$$

is a regular solution to (3.11.1). Its first component is, up to a constant multiple, the solution  $J_n(t)$  (in case  $\nu = n$ ) given in (1.16.16) of Chapter 1. The recursion gives results similar to (1.16.7)–(1.16.9) of Chapter 1, and  $U(t)$  has infinite radius of convergence. Note also that

$$(3.11.78) \quad A_0 \begin{pmatrix} 1 \\ -n \end{pmatrix} = -n \begin{pmatrix} 1 \\ -n \end{pmatrix}, \quad B_{2n} \begin{pmatrix} 1 \\ -n \end{pmatrix} = -2n\beta_n \begin{pmatrix} 1 \\ -n \end{pmatrix},$$

which in concert with (3.11.75)–(3.11.76) gives

$$(3.11.79) \quad \begin{aligned} U(t)t^{A_0}t^{B_{2n}} \begin{pmatrix} 1 \\ -n \end{pmatrix} &= U(t)t^{A_0} \begin{pmatrix} 1 \\ -n \end{pmatrix} - 2n\beta_n(\log t)U(t)t^{A_0} \begin{pmatrix} 1 \\ -n \end{pmatrix} \\ &= U(t)t^{-n} \begin{pmatrix} 1 \\ -n \end{pmatrix} - 2n\beta_n(\log t)U(t)t^n \begin{pmatrix} 1 \\ -n \end{pmatrix}. \end{aligned}$$

The first component gives a solution to (3.11.6), with  $\nu = n$ , complementary to  $J_n(t)$ , for  $t > 0$ . Compare the formula for  $Y_n(t)$  in (1.16.36) of Chapter 1.

REMARK. When (3.11.1) is a  $2 \times 2$  system, either Proposition 3.11.2 or Proposition 3.11.5 will be applicable. Indeed, if  $A_0 \in M(2, \mathbb{C})$  and its two eigenvalues differ by a nonzero integer  $\ell$ , then  $A_0$  is diagonalizable, and

$$\text{Spec } C_{A_0} = \{\ell, 0, -\ell\},$$

so (3.11.42) holds.

To extend the scope of Proposition 3.11.5, let us first note that the hypothesis (3.11.43) that  $A_0$  is diagonalizable was used only to pass to (3.11.53), so we can replace this hypothesis by

$$(3.11.80) \quad \ell \in \mathbb{N} \cap \text{Spec } C_{A_0} \implies M(n, \mathbb{C}) = \mathcal{N}(C_{A_0} - \ell I) \oplus \mathcal{R}(C_{A_0} - \ell I).$$

We now show that we can drop hypothesis (3.11.42). In general, if  $\text{Spec } C_{A_0} \cap \mathbb{N} \neq \emptyset$ , we have a finite set

$$(3.11.81) \quad \text{Spec } C_{A_0} \cap \mathbb{N} = \{\ell_j : 1 \leq j \leq m\};$$

say  $\ell_1 < \dots < \ell_m$ . In this more general setting, we use a transformation of the form (3.11.20), i.e.,  $x(t) = U(t)y(t)$ , with  $U(t)$  as in (3.11.19), to obtain for  $y$  an equation of the form

$$(3.11.82) \quad t \frac{dy}{dt} = (A_0 + B_{\ell_1} t^{\ell_1} + \dots + B_{\ell_m} t^{\ell_m})y,$$

with commutator properties on  $[A_0, B_{\ell_j}]$  analogous to (3.11.55) (see (3.11.85) below). In this setting, in place of (3.11.46), we aim for

$$(3.11.83) \quad t \frac{dU}{dt} = A(t)U(t) - U(t)(A_0 + B_{\ell_1} t^{\ell_1} + \dots + B_{\ell_m} t^{\ell_m}).$$

We continue to have (3.11.47)–(3.11.51), with a natural replacement for  $\Gamma_k$ , which the reader can supply. The equation (3.11.51) for  $k \notin \{\ell_1, \dots, \ell_m\}$  is uniquely solvable for  $U_k$  because  $kI - C_{A_0}$  is invertible. For  $k = \ell_j$ ,  $1 \leq j \leq m$ , one can pick

$$(3.11.84) \quad B_{\ell_j} \in \mathcal{N}(C_{A_0} - \ell_j I),$$

and solve the appropriate variant of (3.11.52), using (3.11.80). Note that (3.11.84) is equivalent to

$$(3.11.85) \quad [A_0, B_{\ell_j}] = \ell_j B_{\ell_j}, \quad \text{i.e., } A_0 B_{\ell_j} = B_{\ell_j} (A_0 + \ell_j I).$$

**Proposition 3.11.7.** *Assume  $A_0 \in M(n, \mathbb{C})$  has the property (3.11.80). For each  $\ell_j$  as in (3.11.81), take  $B_{\ell_j}$  as indicated above, and set*

$$(3.11.86) \quad B = B_{\ell_1} + \dots + B_{\ell_m}.$$

*Then there exist  $T_0 > 0$  and  $U(t)$  as in (3.11.19) such that the general solution to (3.11.1) on  $t \in (0, T_0)$  is*

$$(3.11.87) \quad x(t) = U(t)t^{A_0}t^B v, \quad v \in \mathbb{C}^n.$$

**Proof.** It suffices to show that the general solution to (3.11.82) is

$$(3.11.88) \quad y(t) = t^{A_0}t^B v, \quad v \in \mathbb{C}^n,$$

given that  $B_{\ell_j}$  satisfy (3.11.85). In turn, it suffices to show that if  $y(t)$  is given by (3.11.88), then (3.11.82) holds. To verify this, write

$$(3.11.89) \quad t \frac{dy}{dt} = A_0 t^{A_0} t^B v + t^{A_0} (B_{\ell_1} + \cdots + B_{\ell_m}) t^B v.$$

Now (11.85) yields

$$(3.11.90) \quad A_0^k B_{\ell_j} = B_{\ell_j} (A_0 + \ell_j I)^k, \quad \text{hence } t^{A_0} B_{\ell_j} = B_{\ell_j} t^{\ell_j} t^{A_0},$$

which together with (3.11.89) yields (3.11.82), as needed.  $\square$

Parallel to Proposition 3.11.6, we have the following.

**Proposition 3.11.8.** *In the setting of Proposition 3.11.7,  $B$  is nilpotent.*

**Proof.** By (3.11.85), we have, for each  $\lambda_j \in \text{Spec } A_0$ ,

$$(3.11.91) \quad B : \mathcal{GE}(A_0, \lambda_j) \longrightarrow \mathcal{GE}(A_0, \lambda_j + \ell_1) \oplus \cdots \oplus \mathcal{GE}(A_0, \lambda_j + \ell_m),$$

which readily implies nilpotence.  $\square$

### Exercises

1. In place of (3.11.3), consider second order equations of the form

$$(3.11.92) \quad tu''(t) + b(t)u'(t) + c(t)u(t) = 0,$$

where  $b(t)$  and  $c(t)$  have convergent power series in  $t$  for  $|t| < T_0$ . In such a case, show that setting

$$(3.11.93) \quad x(t) = \begin{pmatrix} u(t) \\ u'(t) \end{pmatrix}$$

yields a system of the form (3.11.1) with

$$(3.11.94) \quad A(t) = \begin{pmatrix} 0 & t \\ -c(t) & -b(t) \end{pmatrix}.$$

Contrast this with what you would get by multiplying (3.11.92) by  $t$  and using the formula (3.11.4) for  $x(t)$ .

2. Make note of how to extend the study of (3.11.1) to

$$(3.11.95) \quad (t - t_0) \frac{dx}{dt} = A(t)x,$$

when  $A(t) = \sum_{k \geq 0} A_k (t - t_0)^k$  for  $|t - t_0| < T_0$ . We say  $t_0$  is a regular singular point for (3.11.95).

3. The following is known as the hypergeometric equation:

$$(3.11.96) \quad t(1-t)u''(t) + [\gamma - (\alpha + \beta + 1)t]u'(t) - \alpha\beta u(t) = 0.$$

Show that  $t_0 = 0$  and  $t_0 = 1$  are regular singular points and construct solutions near these points, given  $\alpha$ ,  $\beta$ , and  $\gamma$ .

4. The following is known as the confluent hypergeometric equation:

$$(3.11.97) \quad tu''(t) + (\gamma - t)u'(t) - \alpha u(t) = 0.$$

Show that  $t_0 = 0$  is a regular singular point and construct solutions near this point, given  $\alpha$  and  $\gamma$ .

5. Let  $B(t)$  be analytic in  $t$  for  $|t| > a$ . We say that the equation

$$(3.11.98) \quad \frac{dy}{dt} = B(t)y$$

has a regular singular point at infinity provided that the change of variable

$$(3.11.99) \quad x(t) = y\left(\frac{1}{t}\right)$$

transforms (3.11.98) to an equation with a regular singular point at  $t = 0$ . Specify for which  $B(t)$  this happens.

6. Show that the hypergeometric equation (3.11.96) has a regular singular point at infinity.

7. What can you say about the behavior as  $t \searrow 0$  of solutions to (3.11.1) when

$$A(t) = A_0 + A_1 t, \quad A_0 = \begin{pmatrix} 2 & 1 \\ & 1 \\ & & 0 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 0 & 0 & 0 \\ & 1 & \\ & & 0 \end{pmatrix}?$$

8. What can you say about the behavior as  $t \searrow 0$  of solutions to (11.1) when

$$A(t) = A_0 + A_1 t, \quad A_0 = \begin{pmatrix} 1 & 1 \\ & 1 \\ & & 0 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 0 & 0 & 0 \\ & 1 & \\ & & 0 \end{pmatrix}?$$

9. In the context of Lemma 3.11.4, i.e., with  $A_0$  and  $B_\ell$  satisfying (3.11.55), show that

$$B_\ell \text{ and } e^{2\pi i A_0} \text{ commute.}$$

More generally, in the context of Proposition 3.11.7, with  $A_0$  and  $B$  satisfying (3.11.85)–(3.11.86), show that

$$B \text{ and } e^{2\pi i A_0} \text{ commute.}$$

Deduce that for all  $t > 0$

$$(3.11.100) \quad t^{A_0} e^{2\pi i A_0} t^B e^{2\pi i B} = t^{A_0} t^B C, \quad C = e^{2\pi i A_0} e^{2\pi i B}.$$

10. In the setting of Exercise 9, pick  $E \in M(n, \mathbb{C})$  such that

$$(3.11.101) \quad e^{2\pi i E} = C.$$

(Cf. Appendix 3.A.) Set

$$(3.11.102) \quad Q(t) = t^{A_0} t^B t^{-E}, \quad t > 0.$$



Show that there exists  $m \in \mathbb{Z}^+$  such that  $t^m Q(t)$  is a polynomial in  $t$ , with coefficients in  $M(n, \mathbb{C})$ . Deduce that, in the setting of Proposition 3.11.7, the general solution to (3.11.1) on  $t \in (0, T_0)$  is

$$(3.11.103) \quad x(t) = U(t)Q(t)t^E v, \quad c \in \mathbb{C}^n,$$

with  $U(t)$  as in (3.11.19),  $E$  as in (3.11.101), and  $Q(t)$  as in (3.11.102), so that  $t^m Q(t)$  is a polynomial in  $t$ .

### 3.A. Logarithms of matrices

Given  $C \in M(n, \mathbb{C})$ , we say  $X \in M(n, \mathbb{C})$  is a logarithm of  $C$  provided

$$(3.A.1) \quad e^X = C.$$

In this appendix, we aim to prove the following:

**Proposition 3.A.1.** *If  $C \in M(n, \mathbb{C})$  is invertible, there exists  $X \in M(n, \mathbb{C})$  satisfying (3.A.1).*

Let us start with the case  $n = 1$ , i.e.,  $C \in \mathbb{C}$ . In case  $C$  is a positive real number, we can take  $X = \log C$ , defined as in Chapter 1, §1.1; cf. (1.1.21)–(1.1.27). More generally, for  $C \in \mathbb{C} \setminus 0$ , we can write

$$(3.A.2) \quad C = |C|e^{i\theta}, \quad X = \log |C| + i\theta.$$

Note that the logarithm  $X$  of  $C$  is not uniquely defined. If  $X \in \mathbb{C}$  solves (3.A.1), so does  $X + 2\pi ik$  for each  $k \in \mathbb{Z}$ . As is customary, for  $C \in \mathbb{C} \setminus 0$ , we will denote any such solution by  $\log C$ .

Let us now take an invertible  $C \in M(n, \mathbb{C})$  with  $n > 1$ , and look for a logarithm, i.e., a solution to (3.A.1). Such a logarithm is easy to produce if  $C$  is diagonalizable, i.e., if for some invertible  $B \in M(n, \mathbb{C})$ ,

$$(3.A.3) \quad B^{-1}CB = D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}.$$

Then

$$(3.A.4) \quad Y = \begin{pmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_n \end{pmatrix}, \quad \mu_k = \log \lambda_k \implies e^Y = D,$$

and so

$$(3.A.5) \quad e^{BYB^{-1}} = BDB^{-1} = C.$$

Similar arguments, in concert with results of Chapter 2, §§2.7–2.8, show that to prove Proposition 3.A.1 it suffices to construct a logarithm of

$$(3.A.6) \quad C = \lambda(I + N), \quad \lambda \in \mathbb{C} \setminus 0, \quad N^n = 0.$$

In turn, if we can solve for  $Y$  the equation

$$(3.A.7) \quad e^Y = I + N,$$

given  $N$  nilpotent, then

$$(3.A.8) \quad \mu = \log \lambda \implies e^{\mu I + Y} = \lambda(I + N),$$

so it suffices to solve (3.A.7) for  $Y \in M(n, \mathbb{C})$ .

We will produce a solution  $Y$  in the form of a power series in  $N$ . To prepare for this, we first strike off on a slight tangent and produce a series solution to

$$(3.A.9) \quad e^{X(t)} = I + tA, \quad A \in M(n, \mathbb{C}), \quad \|tA\| < 1.$$

Taking a cue from the power series for  $\log(1+t)$  given in Chapter 1, (1.1.56), we establish the following.

**Proposition 3.A.2.** *In case  $\|tA\| < 1$ , (3.A.9) is solved by*

$$(3.A.10) \quad \begin{aligned} X(t) &= \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} t^k A^k \\ &= tA - \frac{t^2}{2} A^2 + \frac{t^3}{3} A^3 - \dots \end{aligned}$$

**Proof.** If  $X(t)$  is given by (3.A.10), we have

$$(3.A.11) \quad \begin{aligned} \frac{dX}{dt} &= A - tA^2 + t^2 A^3 - \dots \\ &= A(I - tA + t^2 A^2 - \dots) \\ &= A(I + tA)^{-1}. \end{aligned}$$

Hence

$$(3.A.12) \quad \frac{d}{dt} e^{-X(t)} = -e^{-X(t)} A(I + tA)^{-1},$$

for  $|t| < 1/\|A\|$ ; cf. §3.1, Exercise 12. It follows that

$$(3.A.13) \quad \frac{d}{dt} \left( e^{-X(t)} (I + tA) \right) = e^{-X(t)} \left( -A(I + tA)^{-1} (I + tA) + A \right) = 0,$$

so

$$(3.A.14) \quad e^{-X(t)} (I + tA) \equiv e^{-X(0)} = I,$$

which implies (3.A.9). □

The task of solving (3.A.7) and hence completing the proof of Proposition 3.A.1 is accomplished by the following result.

**Proposition 3.A.3.** *If  $N \in M(n, \mathbb{C})$  is nilpotent, then for all  $t \in \mathbb{R}$ ,*

$$(3.A.15) \quad e^{Y(t)} = I + tN$$

*is solved by*

$$(3.A.16) \quad Y(t) = \sum_{k=1}^{n-1} \frac{(-1)^{k-1}}{k} t^k N^k.$$

**Proof.** If  $Y(t)$  is given by (3.A.16), we see that  $Y(t)$  is nilpotent and that  $e^{Y(t)}$  is a polynomial in  $t$ . Thus both sides of (3.A.15) are polynomials in  $t$ , and Proposition 3.A.2 implies they are equal for  $|t| < 1/\|N\|$ , so (3.A.15) holds for all  $t \in \mathbb{R}$ . □

### 3.B. The matrix Laplace transform

In §1.18 of Chapter 1 we defined the Laplace transform

$$(3.B.1) \quad \mathcal{L}f(s) = \int_0^\infty f(t)e^{-st} dt, \quad \operatorname{Re} s > \alpha,$$

for a function  $f : [0, \infty) \rightarrow \mathbb{C}$ , integrable on  $[0, R]$  for each  $R < \infty$ , and satisfying

$$(3.B.2) \quad \int_0^\infty \|f(t)\|e^{-\beta t} dt < \infty, \quad \forall \beta > \alpha.$$

Such a transform is also well defined for

$$(3.B.3) \quad f : [0, \infty) \longrightarrow V,$$

when  $V$  is a finite-dimensional normed vector space, such as  $\mathbb{C}^n$ , or  $M(n, \mathbb{C})$ , and basic results developed in Chapter 1 continue to apply.

Such a Laplace transform provides a tool to treat  $n \times n$  first-order systems of differential equations, of the form

$$(3.B.4) \quad f'(t) = Af(t) + g(t), \quad f(0) = v,$$

given

$$(3.B.5) \quad A \in M(n, \mathbb{C}), \quad v \in \mathbb{C}^n, \quad g : [0, \infty) \rightarrow \mathbb{C}^n,$$

with  $g$  piecewise continuous and satisfying

$$(3.B.6) \quad \|g(t)\| \leq Ce^{\alpha t}, \quad \text{for } t \geq 0.$$

We seek a solution  $f : [0, \infty) \rightarrow \mathbb{C}^n$  that is piecewise continuous and satisfies a similar bound. If (3.B.4) holds,  $f'(t)$  also has this property, and integration by parts in (3.B.1) yields

$$(3.B.7) \quad \mathcal{L}f'(s) = s\mathcal{L}f(s) - f(0),$$

so applying  $\mathcal{L}$  to (3.B.4) yields

$$(3.B.8) \quad s\mathcal{L}f(s) - v = A\mathcal{L}f(s) + \mathcal{L}g(s),$$

or

$$(3.B.9) \quad (sI - A)\mathcal{L}f(s) = v + \mathcal{L}g(s).$$

Hence, for  $\operatorname{Re} s$  sufficiently large,

$$(3.B.10) \quad \mathcal{L}f(s) = (sI - A)^{-1}(v + \mathcal{L}g(s)).$$

In this way, solving (3.B.4) is translated to solving the recognition problem, enunciated in §1.18 of Chapter 1, i.e., finding the function  $f$  that satisfies (3.B.10).

The approach to this introduced in §1.18 of Chapter 1 was to build up a collection of known Laplace transforms. Here, building on (1.18.20) of Chapter 1, we start with the matrix exponential function,

$$(3.B.11) \quad E_A(t) = e^{tA}, \quad A \in M(n, \mathbb{C}).$$

We claim that

$$(3.B.12) \quad \mathcal{L}E_A(s) = (sI - A)^{-1}, \quad \text{for } \operatorname{Re} s > \alpha,$$

provided

$$(3.B.13) \quad \|e^{tA}\| \leq ce^{\alpha t}, \quad t \geq 0.$$

To see this, let  $L \in M(n, \mathbb{C})$  and note that the identity  $(d/dt)e^{-tL} = -Le^{-tL}$  implies

$$(3.B.14) \quad L \int_0^T e^{-tL} dt = I - e^{-TL},$$

for each  $T < \infty$ . If  $L$  satisfies

$$(3.B.15) \quad \|e^{-tL}\| \leq Ce^{-\delta t}, \quad \forall t > 0,$$

for some  $\delta > 0$ , then we can take  $T \rightarrow \infty$  in (3.B.14), and deduce that

$$(3.B.16) \quad L \int_0^\infty e^{-tL} dt = I, \quad \text{i.e.,} \quad \int_0^\infty e^{-tL} dt = L^{-1}.$$

This applies to  $L = sI - A$  as long as (3.B.13) holds and  $\operatorname{Re} s > \alpha$ , since

$$(3.B.17) \quad \|e^{t(A-sI)}\| = e^{-t \operatorname{Re} s} \|e^{tA}\|,$$

so we have (3.B.12).

The result (3.B.12) gives

$$(3.B.18) \quad (sI - A)^{-1}v = \mathcal{L}(E_A v)(s),$$

which treats part of the right side of (3.B.10). It remains to identify the inverse Laplace transform of

$$(3.B.19) \quad (sI - A)^{-1} \mathcal{L}g(s) = \mathcal{L}E_A(s) \mathcal{L}g(s).$$

One approach to this applies the following result, established in Proposition 1.18.2 of Chapter 1, in the scalar case. The extension to the matrix case is straightforward.

**Proposition 3.B.1.** *Let  $g : [0, \infty) \rightarrow \mathbb{C}^n$  and  $B : [0, \infty) \rightarrow M(n, \mathbb{C})$  be piecewise continuous and satisfy exponential bounds of the form (3.B.6). Take the convolution,*

$$(3.B.20) \quad B * g(t) = \int_0^t B(t - \tau)g(\tau) d\tau.$$

Then, for  $\operatorname{Re} s > \alpha$ ,

$$(3.B.21) \quad \mathcal{L}(B * g)(s) = \mathcal{L}B(s) \mathcal{L}g(s).$$

Applying Proposition 3.B.1 to (3.B.19), we have

$$(3.B.22) \quad (sI - A)^{-1} \mathcal{L}g(s) = \mathcal{L}(E_A * g)(s),$$

with

$$(3.B.23) \quad E_A * g(t) = \int_0^t e^{(t-\tau)A} g(\tau) d\tau.$$

Combining this with (3.B.18) and (3.B.10), we derive for the solution to (3.B.4) the formula

$$(3.B.24) \quad f(t) = e^{tA}v + \int_0^t e^{(t-\tau)A} g(\tau) d\tau.$$

---

This is of course the Duhamel formula (3.4.5), derived in §3.4 by different (and arguably more elementary) means. One advantage of the derivation in §3.4 is that it does not require a global bound on the function  $g$ , of the form (3.B.6). Indeed,  $g$  can blow up in finite time,  $T$ , and (3.B.24) will still work for  $t \in [0, T)$ . Another advantage is that the method of §3.4 generalizes to variable coefficient systems, as seen in §3.9.

On the other hand, having a collection of Laplace transforms and inverse Laplace transforms can be useful for computing the convolution product. Hence the connection between the two given by Proposition 3.B.1 is a double-edged tool.

### 3.C. Complex analytic functions

As stated in (3.10.34), if  $\Omega \subset \mathbb{C}$  is open and  $f : \Omega \rightarrow \mathbb{C}$ , then  $f$  is said to be complex differentiable at  $z_0 \in \Omega$ , with derivative  $f'(z_0)$ , provided

$$(3.C.1) \quad \lim_{w \rightarrow 0} \frac{f(z_0 + w) - f(z_0)}{w} = f'(z_0).$$

One also denotes the limit by  $df/dz$ . Other terms used for such functions are “complex analytic” and “holomorphic.” Here we sketch some results that lead to lots of examples of holomorphic functions.

First, clearly  $f_1(z) = z$  is holomorphic, with  $f'_1(z) = 1$ . To go from here, we have the following:

$$(3.C.2) \quad f, g : \Omega \rightarrow \text{holomorphic } \mathbb{C} \implies fg \text{ holomorphic.},$$

where  $fg(z) = f(z)g(z)$ . In fact, one can write

$$(3.C.3) \quad \begin{aligned} & \frac{f(z_0 + w)g(z_0 + w) - f(z_0)g(z_0)}{w} \\ &= \frac{f(z_0 + w) - f(z_0)}{w} g(z_0 + w) + f(z_0) \frac{g(z_0 + w) - g(z_0)}{w}, \end{aligned}$$

and take  $w \rightarrow 0$  to deduce that

$$(3.C.4) \quad \frac{d}{dz}(fg)(z) = f'(z)g(z) + f(z)g'(z).$$

We can apply this to  $f_2(z) = z^2 = z \cdot z$  to get  $f'_2(z) = 2z$ , and, inductively,

$$(3.C.5) \quad \frac{d}{dz} z^n = n z^{n-1}, \quad z \in \mathbb{C}, \quad n \in \mathbb{N}.$$

Next, we claim that  $1/z$  is holomorphic on  $\mathbb{C} \setminus 0$ . Indeed, for  $z_0 \neq 0$ ,  $|w| < |z_0|$ ,

$$(3.C.6) \quad \frac{1}{w} \left( \frac{1}{z_0 + w} - \frac{1}{z_0} \right) = -\frac{1}{z_0(z_0 + w)},$$

and taking  $w \rightarrow 0$  yields

$$(3.C.7) \quad \frac{d}{dz} \frac{1}{z} = -\frac{1}{z^2}, \quad z \in \mathbb{C} \setminus 0.$$

Again an inductive application of (3.C.4) yields that  $1/z^n$  is holomorphic on  $\mathbb{C} \setminus 0$ , and

$$(3.C.8) \quad \frac{d}{dz} \frac{1}{z^n} = -\frac{n}{z^{n+1}}, \quad z \in \mathbb{C} \setminus 0, \quad n \in \mathbb{N}.$$

We turn to the exponential function  $\text{Exp}(z) = e^z$ , introduced in Chapter 1. We claim that this is holomorphic in  $\mathbb{C}$ , and

$$(3.C.9) \quad \frac{d}{dz} e^z = e^z, \quad z \in \mathbb{C},$$

extending from  $\mathbb{R}$  to  $\mathbb{C}$  the formula for the derivative established there. To see this, use the identity  $e^{z_0+w} = e^{z_0}e^w$ , established in Chapter 1, to write

$$(3.C.10) \quad \begin{aligned} \frac{1}{w}(e^{z_0+w} - e^{z_0}) &= e^{z_0} \frac{e^w - 1}{w} \\ &= e^{z_0} \sum_{k=1}^{\infty} \frac{1}{k!} w^{k-1}, \end{aligned}$$

and note that the last sum is equal to

$$(3.C.11) \quad \sum_{\ell=0}^{\infty} \frac{1}{(\ell+1)!} w^{\ell} = 1 + \frac{w}{2} + \cdots \rightarrow 1, \quad \text{as } w \rightarrow 0,$$

yielding (3.C.9).

We can get lots more holomorphic functions by combining the examples above with the following general result, known as the *chain rule* for holomorphic functions.

**Proposition 3.C.1.** *Assume  $\Omega, \mathcal{O} \subset \mathbb{C}$  are open and*

$$(3.C.12) \quad f: \Omega \rightarrow \mathcal{O}, \quad g: \mathcal{O} \rightarrow \mathbb{C}$$

*are holomorphic. Then  $f = g \circ f: \Omega \rightarrow \mathbb{C}$ , defined by*

$$(3.C.13) \quad h(z) = g(f(z)),$$

*is holomorphic, and*

$$(3.C.14) \quad \frac{d}{dz} g \circ f(z) = g'(f(z))f'(z), \quad z \in \Omega.$$

**Proof.** We can write the definition (3.C.1) as

$$(3.C.15) \quad f(z_0) = f(z_0) + f'(z_0)w + r(z_0, w), \quad z_0 \in \Omega,$$

where  $r(z_0, w)/w \rightarrow 0$  as  $w \rightarrow 0$  (we say  $r(z_0, w) = o(w)$ ). Similarly for  $g$ . Then, for  $z_0 \in \Omega$ ,

$$(3.C.16) \quad \begin{aligned} h(z_0 + w) &= g(f(z_0 + w)) \\ &= g(f(z_0) + f'(z_0)w + r) \\ &= g(f(z_0)) + g'(f(z_0))(f'(z_0)w + r) + r_1(z_0, w), \end{aligned}$$

with also  $r_1(z_0, w) = o(w)$ . Hence

$$(3.C.17) \quad h(z_0 + w) = h(z_0) + g'(f(z_0))f'(z_0)w + r_2(z_0, w),$$

with  $r_2(z_0, w) = o(w)$ , and we have (3.C.14).  $\square$

Putting together these results yields such holomorphic functions as

$$(3.C.18) \quad \frac{1}{z^2 + 1}, \quad e^{z/(z^2+1)}, \quad z \neq \pm i,$$

and a host of others, which the reader can play around with.





## Nonlinear systems of differential equations

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 4.1 begins with a demonstration of existence and uniqueness (for  $t$  close to  $t_0$ ) of solutions to

$$(4.0.1) \quad \frac{dx}{dt} = F(t, x), \quad x(t_0) = x_0.$$

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 (4.1.2) for a definition.) We study the issue of global existence, including positive results when  $F(t, x)$  is linear in  $x$ . Section 4.2 studies the smoothness of the solution to (4.0.1) as a function of  $x_0$ , given various additional hypotheses on  $F$ , and related issues.

Section 4.3 reveals a geometric flavor to (4.0.1), described in the language of vector fields and the flows they generate. A *vector field* on  $\Omega \subset \mathbb{R}^n$  is a map  $F : \Omega \rightarrow \mathbb{R}^n$ . This is a special case of (4.0.1), where  $F$  is independent of  $t$ . The path  $x(t)$  in  $\Omega$  satisfying (4.0.1) 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  $\Phi^t$ , 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 4.4 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 4.5 introduces a class of differential equations arising from Newton's

law  $F = ma$ . This resumes the study introduced in §1.5 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 §4.6, and derive Newton’s epoch-making analysis of the planetary motion problem.

In §4.7 we introduce another (though ultimately related) class of problems that lead to differential equations, namely variational problems. The general setup is to consider

$$(4.0.2) \quad I(u) = \int_a^b L(u(t), u'(t)) dt,$$

for paths  $u : [a, b] \rightarrow \Omega \subset \mathbb{R}^n$ , 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 §4.5. We illustrate this in §4.7 by obtaining a derivation of the pendulum equation, alternative to that given in §1.6 of Chapter 1. We proceed to more sophisticated uses of the variational method. In §4.8 we discuss the “brachistochrone problem,” tossed about by the early leading lights of calculus, one of the foundational variational problems. In §4.9 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 §§4.7 and 4.9, and developed further in §4.10.

The majority of the systems studied in this chapter are not amenable to solution in terms of explicit formulas. In §4.11 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 §4.11 we present several techniques that allow for accurate approximation of solutions to (4.0.1), the most important being Runge-Kutta difference schemes.

In §4.12 we return to the study of qualitative features of phase portraits, initiated in §4.3. 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 4.13–4.14 are devoted to some systems of differential equations arising to model the populations of interacting species. In §4.13 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 §4.12. In §4.14 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, §4.15.

This chapter ends with a number of appendices, some providing useful background in calculus, and others taking up further topics in nonlinear systems of ODE. In Appendix 4.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 4.B discusses some basic results about convergence, including the notion of *compactness*.

In Appendix 4.C we show that if the linearization of a vector field  $F$  at a critical point behaves like a saddle, so does  $F$ . Appendix 4.D takes a further look at the behavior of a flow near a critical point of a vector field. It produces a blown up phase portrait of such a flow, by taking spherical polar coordinates centered at the critical point.

Appendix 4.E 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 4.F we show that a spherically symmetric planet produces the same gravitational field as if all its mass were concentrated at its center. In Appendix 4.G we prove the Brouwer fixed-point theorem (in dimension 2), a use of which arises in §4.15. The proof we give makes use of material developed in §4.4.

In Appendix 4.H we discuss geodesic equations on surfaces. This discussion continues results on minima and other critical paths for the energy functional introduced in §4.7, making contact with the length functional here, which is of interest for the differential geometry of surfaces. Appendix 4.I deals with rigid body motion in  $\mathbb{R}^n$ . We set up a Lagrangian and treat this as a variational problem. This approach leads to a geodesic equation on the rotation group  $SO(n)$ , endowed with a certain left-invariant metric. We reduce this to a system of ODE for  $Z : I \rightarrow \text{Skew}(n)$ , with a quadratic nonlinearity. We specialize to  $n = 3$  and produce Euler’s equation for the free motion of a rigid body in  $\mathbb{R}^3$ , and show this is solvable in terms of elliptic integrals.

### 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,

$$(4.1.1) \quad \frac{dx}{dt} = F(t, x), \quad x(t_0) = x_0.$$

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

$$(4.1.2) \quad \|F(t, x) - F(t, y)\| \leq L\|x - y\|,$$

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

$$(4.1.3) \quad \|D_x F(t, x)\| \leq L,$$

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 4.A. The implication (4.1.3)  $\Rightarrow$  (4.1.2) follows readily from (4.A.9). Our first goal is to prove the following.

**Proposition 4.1.1.** *Assume  $F : I \times \Omega \rightarrow \mathbb{R}^n$  is bounded and continuous and satisfies the Lipschitz condition (4.1.2), and let  $x_0 \in \Omega$ . Then there exists  $T_0 > 0$  and a unique  $C^1$  solution to (4.1.1) for  $|t - t_0| < T_0$ .*

The first step in proving this is to rewrite (4.1.1) as an integral equation:

$$(4.1.4) \quad x(t) = x_0 + \int_{t_0}^t F(s, x(s)) ds.$$

The equivalence of (4.1.1) and (4.1.4) follows from the Fundamental Theorem of Calculus. It suffices to find a continuous solution  $x$  to (4.1.4) on  $[t_0 - T_0, t_0 + T_0]$ , since then the right side of (4.1.4) will be  $C^1$  in  $t$ .

We will apply a technique known as Picard iteration to construct a solution to (4.1.4). We set  $x_0(t) \equiv x_0$  and then define  $x_n(t)$  inductively by

$$(4.1.5) \quad x_{n+1}(t) = x_0 + \int_{t_0}^t F(s, x_n(s)) ds.$$

We show that this converges uniformly to a solution to (4.1.4), for  $|t - t_0| \leq T_0$ , if  $T_0$  is taken small enough. To get this, we quantify some hypotheses made above. We assume

$$(4.1.6) \quad \overline{B_R(x_0)} = \{x \in \mathbb{R}^n : \|x - x_0\| \leq R\} \subset \Omega$$

and

$$(4.1.7) \quad \|F(s, x)\| \leq M, \quad \forall s \in I, x \in \overline{B_R(x_0)}.$$

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 (4.1.5). We have

$$(4.1.8) \quad \|x_{n+1}(t) - x_0\| \leq \int_{t_0}^t \|F(s, x_n(s))\| ds \leq M|t - t_0|,$$

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

$$(4.1.9) \quad T_0 \leq \frac{R}{M}.$$

As long as (4.1.9) holds and  $[t_0 - T_0, t_0 + T_0] \subset I$ , we get an infinite sequence  $x_n(t)$  of functions, related by (4.1.5).

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

$$(4.1.10) \quad \begin{aligned} \|x_{n+1}(t) - x_n(t)\| &= \left\| \int_{t_0}^t [F(s, x_n(s)) - F(s, x_{n-1}(s))] ds \right\| \\ &\leq \int_{t_0}^t \|F(s, x_n(s)) - F(s, x_{n-1}(s))\| ds \\ &\leq L \int_{t_0}^t \|x_n(s) - x_{n-1}(s)\| ds, \end{aligned}$$

the last inequality by (4.1.2). Hence

$$(4.1.11) \quad \max_{|t-t_0| \leq T_0} \|x_{n+1}(t) - x_n(t)\| \leq LT_0 \max_{|s-t_0| \leq T_0} \|x_n(s) - x_{n-1}(s)\|.$$

The additional constraint we impose on  $T_0$  is

$$(4.1.12) \quad T_0 \leq \frac{\alpha}{L}, \quad \alpha \in (0, 1).$$

Noting that

$$(4.1.13) \quad \max_{|t-t_0| \leq T_0} \|x_1(t) - x_0\| \leq R,$$

we see that

$$(4.1.14) \quad \max_{|t-t_0| \leq T_0} \|x_{n+1}(t) - x_n(t)\| \leq \alpha^n R.$$

Consequently, the infinite series

$$(4.1.15) \quad x(t) = x_0 + \sum_{n=0}^{\infty} (x_{n+1}(t) - x_n(t))$$

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

$$(4.1.16) \quad \max_{|t-t_0| \leq T_0} \|x(t) - x_n(t)\| \leq \alpha^{n-1} R.$$

It readily follows that

$$(4.1.17) \quad \int_{t_0}^t F(s, x_n(s)) ds \longrightarrow \int_{t_0}^t F(s, x(s)) ds,$$

so (4.1.4) follows from (4.1.5) in the limit  $n \rightarrow \infty$ .

To finish the proof of Proposition 4.1.1, we establish uniqueness. Suppose  $y(t)$  also satisfies (4.1.4) for  $|t - t_0| \leq T_0$ . Then

$$\begin{aligned} \|x(t) - y(t)\| &= \left\| \int_{t_0}^t [F(s, x(s)) - F(s, y(s))] ds \right\| \\ (4.1.18) \quad &\leq \int_{t_0}^t \|F(s, x(s)) - F(s, y(s))\| ds \\ &\leq L \int_{t_0}^t \|x(s) - y(s)\| ds, \end{aligned}$$

and hence

$$(4.1.19) \quad \max_{|t-t_0| \leq T_0} \|x(t) - y(t)\| \leq T_0 L \max_{|s-t_0| \leq T_0} \|x(s) - y(s)\|.$$

As long as (4.1.12) holds,  $T_0 L \leq \alpha < 1$ , so (4.1.19) clearly implies  $\max_{|t-t_0| \leq T_0} \|x(t) - y(t)\| = 0$ , which gives the asserted uniqueness.

Note that the Lipschitz hypothesis (4.1.2) was needed only for  $x, y \in \overline{B_R(x_0)}$ . Thus we can extend Proposition 4.1.1 to the following setting:

$$(4.1.20) \quad \text{For each closed, bounded } K \subset \Omega, \text{ there exists } L_K < \infty \text{ such that} \\ \|F(t, x) - F(t, y)\| \leq L_K \|x - y\|, \quad \forall x, y \in K, \quad t \in I.$$

We can also replace the bound on  $F$  by

$$(4.1.21) \quad \text{For each } K \text{ as above, there exists } M_K < \infty \text{ such that} \\ \|F(t, x)\| \leq M_K, \quad \forall x \in K, \quad t \in I.$$

Results of Appendix 4.B imply that there exists  $R_K > 0$  such that

$$(4.1.22) \quad \tilde{K} = \overline{\bigcup_{x \in K} B_{R_K}(x)} \text{ is a compact subset of } \Omega.$$

It follows that for each  $x_0 \in K$ , the solution to (4.1.1) exists on the interval

$$(4.1.23) \quad \{t \in I : |t - t_0| \leq \min(R_K/M_{\tilde{K}}, \alpha/L_{\tilde{K}})\}.$$

Now that we have local solutions to (4.1.1), it is of interest to investigate when global solutions exist. Here is an example of breakdown:

$$(4.1.24) \quad \frac{dx}{dt} = x^2, \quad x(0) = 1.$$

Here  $I = \mathbb{R}$ ,  $n = 1$ ,  $\Omega = \mathbb{R}$ , and  $F(x) = x^2$  is smooth, satisfying the local bounds (4.1.21)–(4.1.23). The equation (4.1.24) has the unique solution

$$(4.1.25) \quad x(t) = \frac{1}{1-t}, \quad t \in (-\infty, 1),$$

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 4.1.2.** *Let  $F$  be as in Proposition 4.1.1, but with the Lipschitz and boundedness hypotheses relaxed to (4.1.20)–(4.1.21). Assume  $[a, b]$  is contained in the open interval  $I$  and assume  $x(t)$  solves (4.1.1) 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 (4.1.1) for  $t \in (a_1, b_1)$ .

**Proof.** We deduce from (4.1.23) that there exists  $\delta > 0$  such that for each  $x_1 \in K$ ,  $t_1 \in [a, b]$ , the solution to

$$(4.1.26) \quad \frac{dx}{dt} = F(t, x), \quad x(t_1) = x_1$$

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 (4.1.1). Then solving (4.1.26) continues  $x(t)$  past  $t = b$ . Similarly one can continue  $x(t)$  past  $t = a$ .  $\square$

Here is an example of a global existence result that can be deduced from Proposition 4.1.2. Consider the  $2 \times 2$  system for  $x = (y, v)$ :

$$(4.1.27) \quad \begin{aligned} \frac{dy}{dt} &= v, \\ \frac{dv}{dt} &= -y^3. \end{aligned}$$

Here we have  $\Omega = \mathbb{R}^2$ ,  $F(t, x) = F(t, y, v) = (v, -y^3)$ . If (4.1.27) holds for  $t \in (a, b)$ , we have

$$(4.1.28) \quad \frac{d}{dt} \left( \frac{v^2}{2} + \frac{y^4}{4} \right) = v \frac{dv}{dt} + y^3 \frac{dy}{dt} = 0,$$

so each  $x(t) = (y(t), v(t))$  solving (4.1.27) 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 (4.1.27).

We can also apply Proposition 4.1.2 to establish global existence of solutions to linear systems,

$$(4.1.29) \quad \frac{dx}{dt} = A(t)x, \quad x(0) = x_0,$$

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 4.1.3.** *If  $\|A(t)\| \leq K$  for  $t \in I$ , then the solution to (4.1.29) satisfies*

$$(4.1.30) \quad \|x(t)\| \leq e^{K|t|} \|x_0\|.$$

**Proof.** It suffices to prove (4.1.30) for  $t \geq 0$ . Then  $y(t) = e^{-Kt}x(t)$  satisfies

$$(4.1.31) \quad \frac{dy}{dt} = C(t)y, \quad y(0) = x_0,$$

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

$$(4.1.32) \quad \operatorname{Re}(C(t)u, u) \leq 0, \quad \forall u \in \mathbb{C}^n.$$

Then (4.1.30) is a consequence of the following estimate, of interest in its own right.  $\square$

**Lemma 4.1.4.** *If  $y(t)$  solves (4.1.31) and (4.1.32) holds for  $C(t)$ , then*

$$(4.1.33) \quad \|y(t)\| \leq \|y(0)\| \quad \text{for } t \geq 0.$$



**Proof.** We have

$$(4.1.34) \quad \begin{aligned} \frac{d}{dt} \|y(t)\|^2 &= (y'(t), y(t)) + (y(t), y'(t)) \\ &= 2 \operatorname{Re} (C(t)y(t), y(t)) \\ &\leq 0. \end{aligned}$$

□

Thanks to Proposition 4.1.3, we have for  $s, t \in I$ , the solution operator for (4.1.29),

$$(4.1.35) \quad S(t, s) \in M(n, \mathbb{C}), \quad S(t, s)x(s) = x(t),$$

introduced in §3.8 of Chapter 3. As noted there, we have the Duhamel formula

$$(4.1.36) \quad x(t) = S(t, t_0) + \int_{t_0}^t S(t, s)f(s) ds,$$

for the solution to

$$(4.1.37) \quad \frac{dx}{dt} = A(t)x + f(t), \quad x(t_0) = x_0.$$

If  $F(t, x)$  depends explicitly on  $t$ , we call (4.1.1) a non-autonomous system. If  $F$  does not depend explicitly on  $t$ , we say (4.1.1) is autonomous. The following device converts a non-autonomous system to an autonomous one. Take the  $n \times n$  system (4.1.1). Then the  $(n+1) \times (n+1)$  system

$$(4.1.38) \quad \frac{dx}{dt} = F(y, x), \quad \frac{dy}{dt} = 1, \quad x(t_0) = x_0, \quad y(t_0) = t_0$$

has the autonomous form

$$(4.1.39) \quad \frac{dz}{dt} = G(z), \quad z(t_0) = (x_0, t_0),$$

for  $z = (x, y)$ , with  $G(z) = (F(y, x), 1)$ , and the solution to (4.1.38) is  $(x(t), t)$ , where  $x(t)$  solves (4.1.1). Thus for many purposes it suffices to consider autonomous systems.

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

$$(4.1.40) \quad \frac{d^k x}{dt^k} = F(t, x, \dots, x^{(k-1)}), \quad x(t_0) = x_0, \dots, x^{(k-1)}(t_0) = x_{k-1},$$

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

$$(4.1.41) \quad y = \begin{pmatrix} y_0 \\ \vdots \\ y_{k-1} \end{pmatrix}, \quad y_j(t) \in \mathbb{R}^n, \quad 0 \leq j \leq k-1.$$

The system is

$$(4.1.42) \quad \begin{aligned} \frac{dy_0}{dt} &= y_1, \\ &\vdots \\ \frac{dy_{k-2}}{dt} &= y_{k-1}, \\ \frac{dy_{k-1}}{dt} &= F(t, y_0, \dots, y_{k-1}), \end{aligned}$$

with initial data

$$(4.1.43) \quad y_j(t_0) = x_j, \quad 0 \leq j \leq k-1.$$

If  $y(t)$  solves (4.1.42)–(4.1.43), then  $x(t) = y_0(t)$  solves (4.1.40), and we have

$$(4.1.44) \quad x^{(j)}(t) = y_j(t), \quad 0 \leq j \leq k-1.$$

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 (4.1.5) to write out

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

Compare the results with the formula (4.1.25).

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 (4.1.5) to write out

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

Compare and contrast the results with calculations from §3.10 of Chapter 3.

5. Let  $x_n(t)$  be an approximate solution to (4.1.1), and assume that

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

Let  $x_{n+1}(t)$  be defined by (4.1.5), and assume the Lipschitz condition (4.1.2) holds. Show that

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

6. Modify the system (4.1.27) to

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

Show that solutions satisfy

$$\frac{d}{dt} \left( \frac{v^2}{2} + \frac{y^4}{4} \right) \leq 0,$$

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

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

$$x(t) = \frac{1}{4}t^2, \quad t \geq 0, \\ 0, \quad t \leq 0$$

is another solution, on  $t \in (-\infty, \infty)$ . Why does this not contradict the uniqueness part of Proposition 4.1.1? 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  $t \geq 0$  if and only if  $\beta \leq 1$ .

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

$$(4.1.45) \quad \frac{dx}{dt} = F(x), \quad x(t_0) = x_0,$$

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

$$(4.1.46) \quad \frac{d}{dt} \|x(t)\|^2 = 2x(t) \cdot F(x(t)).$$

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

$$(4.1.47) \quad \frac{d}{dt} \|x(t)\|^\alpha = \alpha \|x(t)\|^{\alpha-2} x(t) \cdot F(x(t)).$$

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

$$(4.1.48) \quad \|F(x)\| \leq C(1 + \|x\|)^\beta, \quad \forall x \in \mathbb{R}^n, \quad C < \infty, \quad \beta < 1.$$

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

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

Use this to establish that the solution to (4.1.45) exists for all  $t \in \mathbb{R}$ .

### Gronwall's inequality and consequences

Exercises 11–13 below will extend the conclusion of Exercise 10 to the case  $\beta = 1$  in (4.1.48). One approach is via the following result, known as *Gronwall's inequality*.

**Proposition 4.1.5.** *Assume*

$$(4.1.49) \quad g \in C^1(\mathbb{R}), \quad g' \geq 0.$$

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

$$(4.1.50) \quad \begin{aligned} u(t) &\leq A + \int_{t_0}^t g(u(s)) \, ds, \\ v(t) &\geq A + \int_{t_0}^t g(v(s)) \, ds. \end{aligned}$$

Then

$$(4.1.51) \quad u(t) \leq v(t), \quad \text{for } t \in I, \, t \geq t_0.$$

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

$$(4.1.52) \quad \begin{aligned} w(t) &\leq \int_{t_0}^t [g(u(s)) - g(v(s))] \, ds \\ &= \int_{t_0}^t M(s)w(s) \, ds, \end{aligned}$$

where

$$(4.1.53) \quad M(s) = \int_0^1 g'(\tau u(s) + (1 - \tau)v(s)) \, d\tau.$$

Hence we have

$$(4.1.54) \quad w(t) \leq \int_{t_0}^t M(s)w(s) \, ds, \quad M(s) \geq 0, \quad M \in C(I),$$

and we claim this implies

$$(4.1.55) \quad w(t) \leq 0, \quad \forall t \in I, \, t \geq t_0.$$

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  $\int_{t_0}^{t_1} M(s)w(s) ds \leq 0$ , we deduce from (4.1.54) that

$$(4.1.56) \quad w(t) \leq \int_{t_1}^t M(s)w(s) ds, \quad \forall t \in [t_1, b].$$

Hence, with

$$(4.1.57) \quad K = \max_{[t_1, b]} M(s) < \infty,$$

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

$$(4.1.58) \quad \max_{[t_1, a]} w(t) \leq (a - t_1)K \max_{[t_1, a]} w(s).$$

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

$$(4.1.59) \quad w(t) \leq 0, \quad \forall t \in [t_1, a],$$

contradicting the maximality of  $t_1$ . Hence actually  $t_1 = b$ , and we have the implication (4.1.54)  $\Rightarrow$  (1.53), completing the proof of Proposition 4.1.5.  $\square$

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

$$(4.1.60) \quad \frac{dv}{dt} \leq Cv, \quad v(t_0) = v_0,$$

where  $C \in (0, \infty)$  and  $t_0 \in I$ . Using Proposition 4.1.5, show that

$$(4.1.61) \quad v(t) \leq e^{C(t-t_0)}v_0, \quad \forall t \in [t_0, b).$$

12. In the setting of Exercise 11, avoid use of Proposition 4.1.5 as follows. Write (4.1.60) as

$$(4.1.62) \quad \frac{dv}{dt} = Cv - g(t), \quad v(t_0) = v_0, \quad g \geq 0,$$

with solution

$$(4.1.63) \quad v(t) = e^{C(t-t_0)}v_0 - \int_{t_0}^t e^{C(t-s)}g(s) ds.$$

Deduce (4.1.61) from this.

13. Return to the setting of Exercise 9, and replace the hypothesis (4.1.48) by

$$(4.1.64) \quad \|F(x)\| \leq C(1 + \|x\|), \quad \forall x \in \mathbb{R}^n.$$

Show that the solution to (4.1.45) exists for all  $t \in \mathbb{R}$ .

*Hint.* Take  $v(t) = 1 + \|x(t)\|^2$  and use (4.1.46). Show that Exercise 11 (or 12) applies.

## 4.2. Dependence of solutions on initial data and other parameters

We study how the solution to a system of differential equations

$$(4.2.1) \quad \frac{dx}{dt} = F(x), \quad x(0) = y$$

depends on the initial condition  $y$ . As shown in §4.1, there is no loss of generality in considering the autonomous system (4.2.1). We will assume  $F : \Omega \rightarrow \mathbb{R}^n$  is smooth,  $\Omega \subset \mathbb{R}^n$  open and convex, and denote the solution to (4.2.1) 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 4.A for more on this derivative.)

To start, we assume  $F$  is of class  $C^1$ , i.e.,  $DF$  is continuous on  $\Omega$ , 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 (4.1.6), is contained in  $\Omega$ . We seek an  $n \times n$  matrix  $W(t, y)$  such that, for  $w_0 \in \mathbb{R}^n$ ,  $\|w_0\| \leq R$ ,

$$(4.2.2) \quad x(t, y + w_0) = x(t, y) + W(t, y)w_0 + r(t, y, w_0),$$

where

$$(4.2.3) \quad r(t, y, w_0) = o(\|w_0\|),$$

which means

$$(4.2.4) \quad \lim_{w_0 \rightarrow 0} \frac{r(t, y, w_0)}{\|w_0\|} = 0.$$

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

$$(4.2.5) \quad D_y x(t, y) = W(t, y).$$

In other words,

$$(4.2.6) \quad x(t, y + w_0) = x(t, y) + D_y x(t, y)w_0 + o(\|w_0\|).$$

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 (4.2.1). Using the chain rule, we get the following equation,

$$(4.2.7) \quad \frac{dW}{dt} = DF(x)W, \quad W(0, y) = I,$$

called the linearization of (4.2.1). Here,  $I$  is the  $n \times n$  identity matrix. Equivalently, given  $w_0 \in \mathbb{R}^n$ ,

$$(4.2.8) \quad w(t, y) = W(t, y)w_0$$

is expected to solve

$$(4.2.9) \quad \frac{dw}{dt} = DF(x)w, \quad w(0) = w_0.$$

Now, we do not yet know that  $x(t, y)$  is differentiable, but we do know from results of §4.1 that (4.2.7) and (4.2.9) are uniquely solvable. It remains to show that, with such a choice of  $W(t, y)$ , (4.2.2)–(4.2.3) hold.

To rephrase the task, set

$$(4.2.10) \quad x(t) = x(t, y), \quad x_1(t) = x(t, y + w_0), \quad z(t) = x_1(t) - x(t),$$

and let  $w(t)$  solve (4.2.9). The task of verifying (4.2.2)–(4.2.3) is equivalent to the task of verifying

$$(4.2.11) \quad \|z(t) - w(t)\| = o(\|w_0\|).$$

To show this, we will obtain for  $z(t)$  an equation similar to (4.2.9). To begin, (4.2.10) implies

$$(4.2.12) \quad \frac{dz}{dt} = F(x_1) - F(x), \quad z(0) = w_0.$$

Now the fundamental theorem of calculus gives

$$(4.2.13) \quad F(x_1) - F(x) = G(x_1, x)(x_1 - x),$$

with

$$(4.2.14) \quad G(x_1, x) = \int_0^1 DF(\tau x_1 + (1 - \tau)x) d\tau.$$

If  $F$  is  $C^1$ , then  $G$  is continuous. Then (4.2.12)–(4.2.13) yield

$$(4.2.15) \quad \frac{dz}{dt} = G(x_1, x)z, \quad z(0) = w_0.$$

Given that

$$(4.2.16) \quad \|DF(u)\| \leq L, \quad \forall u \in \Omega,$$

which we have by continuity of  $DF$ , after possibly shrinking  $\Omega$  slightly, we deduce from Proposition 4.1.3 that

$$(4.2.17) \quad \|z(t)\| \leq e^{tL}\|w_0\|,$$

that is,

$$(4.2.18) \quad \|x(t, y) - x(t, y + w_0)\| \leq e^{tL}\|w_0\|.$$

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 (4.2.15) as

$$(4.2.19) \quad \frac{dz}{dt} = G(x + z, x)z = DF(x)z + R(x, z), \quad z(0) = w_0,$$

where

$$(4.2.20) \quad F \in C^1(\Omega) \implies \|R(x, z)\| = o(\|z\|) = o(\|w_0\|).$$

Now comparing (4.2.19) with (4.2.9), we have

$$(4.2.21) \quad \frac{d}{dt}(z - w) = DF(x)(z - w) + R(x, z), \quad (z - w)(0) = 0.$$

Then Duhamel's formula gives

$$(4.2.22) \quad z(t) - w(t) = \int_0^t S(t, s)R(x(s), z(s)) ds,$$

where  $S(t, s)$  is the solution operator for  $d/dt - B(t)$ , with  $B(t) = DF(x(t))$ , which as in (4.2.17), satisfies

$$(4.2.23) \quad \|S(t, s)\| \leq e^{|t-s|L}.$$

We hence have (4.2.11), i.e.,

$$(4.2.24) \quad \|z(t) - w(t)\| = o(\|w_0\|).$$

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 (4.2.7). Hence we have:

**Proposition 4.2.1.** *If  $F \in C^1(\Omega)$  and if solutions to (4.2.1) 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 (4.2.7).*

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 (4.2.7) 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 (4.2.1) and (4.2.7), to get a system of differential equations for  $(x, W)$ :

$$(4.2.25) \quad \begin{aligned} \frac{dx}{dt} &= F(x), \\ \frac{dW}{dt} &= DF(x)W, \end{aligned}$$

with initial conditions

$$(4.2.26) \quad x(0) = y, \quad W(0) = I.$$

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

Similarly, we can consider dependence of the solution to

$$(4.2.27) \quad \frac{dx}{dt} = F(\tau, x), \quad x(0) = y$$

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

$$(4.2.28) \quad \frac{dx}{dt} = F(z, x), \quad \frac{dz}{dt} = 0, \quad x(0) = y, \quad z(0) = \tau.$$

Then we get smoothness of  $x(t, \tau, y)$  jointly in  $(\tau, y)$ . 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 4.2.2 to the following:

$$(4.2.29) \quad F \in C^k(\Omega) \implies x \in C^k \text{ jointly in } (t, y).$$



---

**Exercises**

1. Suppose  $\tau \in \mathbb{R}$  in (4.2.27). 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

$$(4.2.30) \quad \frac{dx}{dt} = -y + \tau(x^2 + y^2), \quad \frac{dy}{dt} = x, \quad x(0) = 1, \quad y(0) = 0.$$

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

$$(4.2.31) \quad \begin{aligned} x_\tau(t) &= X_0(t) + \tau X_1(t) + \tau^2 X_2(t) + \cdots, \\ y_\tau(t) &= Y_0(t) + \tau Y_1(t) + \tau^2 Y_2(t) + \cdots. \end{aligned}$$

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  $\tau$  sufficiently small, solutions to (4.2.30) are periodic in  $t$ .

5. Let  $p(\tau)$  denote the period of the solution to (4.2.30). Using (4.2.31), show that  $p(\tau)$  is smooth in  $\tau$  for  $|\tau|$  small. Note that  $p(0) = 2\pi$ . Compute  $p'(0)$ . Compare results in Appendix 4.E.

6. Suppose  $y$  in (4.2.1) is a critical point of  $F$ , i.e.,  $F(y) = 0$ . Show that (4.2.7) becomes

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

hence

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

### 4.3. Vector fields, orbits, and flows

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

$$(4.3.1) \quad F : \Omega \longrightarrow \mathbb{R}^n,$$

such as encountered in (4.2.1). We say  $F$  is a  $C^k$  vector field if  $F$  is a  $C^k$  map. A  $C^\infty$  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. (4.1.1)), but in this section we restrict attention to the autonomous case.

The solution to (4.2.1), i.e., to

$$(4.3.2) \quad \frac{dx}{dt} = F(x), \quad x(0) = y,$$

will be denoted

$$(4.3.3) \quad x(t) = \Phi_F^t(y).$$

Results of §§4.1–4.2 imply that for each closed bounded  $K \subset \Omega$  there exists an interval  $I = (-T_0, T_1)$  about 0 such that, for each  $t \in I$ ,

$$(4.3.4) \quad \Phi_F^t : K \longrightarrow \Omega,$$

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

$$(4.3.5) \quad \Phi_F^0(y) \equiv y,$$

i.e.,  $\Phi_F^0$  is the identity map. We also have

$$(4.3.6) \quad \Phi_F^{s+t}(y) = \Phi_F^t \circ \Phi_F^s(y),$$

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

$$(4.3.7) \quad t \mapsto \Phi_F^t(y)$$

is called the *orbit* through  $y$ .

Another way to state the defining property of  $\Phi_F^t$  is that (4.3.5) holds and

$$(4.3.8) \quad \frac{d}{dt} \Phi_F^t(x) = F(\Phi_F^t(x)).$$

We next obtain interesting information on the  $t$ -derivative of

$$(4.3.9) \quad v^t(x) = v(\Phi_F^t(x)),$$

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 4.A, especially (4.A.8)) plus (4.3.8) yields

$$(4.3.10) \quad \frac{d}{dt} v^t(x) = F(\Phi_F^t(x)) \cdot \nabla v(\Phi_F^t(x)).$$

In particular,

$$(4.3.11) \quad \left. \frac{d}{ds} v(\Phi_F^s(x)) \right|_{s=0} = F(x) \cdot \nabla v(x).$$

Here  $\nabla v$  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 (4.3.10) is

$$(4.3.12) \quad \begin{aligned} \frac{d}{dt} v^t(x) &= \left. \frac{d}{ds} v^t(\Phi_F^s(x)) \right|_{s=0} \\ &= F(x) \cdot \nabla v^t(x), \end{aligned}$$

the first equality following from (4.3.6) and the second from (4.3.11), with  $v$  replaced by  $v^t$ .

One significant consequence of (4.3.12), which will lead to the important result (4.3.24) below, is that, for  $v \in C_0^1(\Omega)$ ,

$$(4.3.13) \quad \begin{aligned} \frac{d}{dt} \int_{\Omega} v(\Phi_F^t(x)) dx &= \int_{\Omega} F(x) \cdot \nabla v^t(x) dx \\ &= - \int_{\Omega} \operatorname{div} F(x) v(\Phi_F^t(x)) dx. \end{aligned}$$

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

$$(4.3.14) \quad \operatorname{div} F(x) = \frac{\partial F_1}{\partial x_1}(x) + \dots + \frac{\partial F_n}{\partial x_n}(x).$$

The last equality in (4.3.13) follows by integration by parts,

$$(4.3.15) \quad \int_{\Omega} F_k(x) \frac{\partial v^t}{\partial x_k} dx = - \int_{\Omega} \frac{\partial F_k}{\partial x_k} v^t(x) dx,$$

followed by summation over  $k$ . We reiterate the content of (4.3.13):

$$(4.3.16) \quad \frac{d}{dt} \int_{\Omega} v(\Phi_F^t(x)) dx = - \int_{\Omega} \operatorname{div} F(x) v(\Phi_F^t(x)) dx.$$

So far, we have (4.3.16) for  $v \in C_0^1(\Omega)$ . We can extend this by noting that (4.3.16) implies

$$(4.3.17) \quad \begin{aligned} \int_{\Omega} v(\Phi_F^t(x)) dx - \int_{\Omega} v(x) dx \\ = - \int_0^t \int_{\Omega} \operatorname{div} F(x) v(\Phi_F^s(x)) dx ds. \end{aligned}$$

Basic results on the integral allow one to pass from  $v \in C_0^1(\Omega)$  in (4.3.17) 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  $B_\delta = \{x \in \mathbb{R}^n : \operatorname{dist}(x, B) \leq \delta\}$ . There exists  $\delta_0 > 0$  such that  $B_\delta \subset \Omega$  for  $\delta \in (0, \delta_0]$ . For such  $\delta$ , one can produce  $v_\delta \in C_0^1(\Omega)$  such that

$$(4.3.18) \quad v_\delta = 1 \text{ on } B, \quad 0 \leq v_\delta \leq 1, \quad v_\delta = 0 \text{ on } \mathbb{R}^n \setminus B_\delta.$$

Then

$$(4.3.19) \quad \left| \int \chi_B(x) dx - \int v_\delta(x) dx \right| \leq \text{vol}(B_\delta \setminus B) \rightarrow 0, \quad \text{as } \delta \rightarrow 0,$$

so, as  $\delta \rightarrow 0$ ,

$$(4.3.20) \quad \int_{\Omega} v_\delta(x) dx \longrightarrow \int_{\Omega} \chi_B(x) dx.$$

Similar arguments give

$$(4.3.21) \quad \int_{\Omega} v_\delta(\Phi_F^t(x)) dx \longrightarrow \int_{\Omega} \chi_B(\Phi_F^t(x)) dx,$$

and

$$(4.3.22) \quad \int_0^t \int_{\Omega} \text{div } F(x) v_\delta(\Phi_F^s(x)) dx ds \longrightarrow \int_0^t \int_{\Omega} \text{div } F(x) \chi_B(\Phi_F^s(x)) dx ds.$$

These results allow one to take  $v = \chi_B$  in (4.3.17).

Now one can pass from (4.3.17) back to (4.3.16), via the fundamental theorem of calculus. Note that

$$(4.3.23) \quad \text{Vol } \Phi_F^t(B) = \int \chi_B(\Phi_F^{-t}(x)) dx.$$

We can apply (4.3.16) with  $t$  replaced by  $-t$ , and  $v$  by  $\chi_B$ , and deduce the following.

**Proposition 4.3.1.** *If  $F$  is a  $C^1$  vector field, generating the flow  $\Phi_F^t$ , well defined on  $\Omega$  for  $t \in I$ , and  $B \subset \Omega$  is smoothly bounded, then, for  $t \in I$ ,*

$$(4.3.24) \quad \frac{d}{dt} \text{Vol } \Phi_F^t(B) = \int_{\Phi_F^t(B)} \text{div } F(x) dx.$$

This result is behind the notation  $\text{div } F$ , i.e., the *divergence* of  $F$ . Vector fields  $F$  with positive divergence generate flows  $\Phi_F^t$  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 4.3.2.** *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*

$$(4.3.25) \quad V(x) \rightarrow +\infty \quad \text{as } \|x\| \rightarrow \infty$$

and

$$(4.3.26) \quad \|x\| \geq R \implies \nabla V(x) \cdot F(x) \leq 0.$$

*Then the flow  $\Phi_F^t$  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

$$(4.3.27) \quad \|x(t)\| \geq R \implies \frac{d}{dt}V(x(t)) = \nabla V(x(t)) \cdot F(x(t)) \leq 0.$$

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

$$(4.3.28) \quad \left\{ x \in \mathbb{R}^n : V(x) \leq \max V(y), y \in B_R(0) \cup \{x_0\} \right\}.$$

From here, Proposition 4.1.2 yields forward completeness.  $\square$

One way to display the behavior of the flow generated by a vector field  $F$  on a domain  $\Omega$  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 Figure 4.3.1, the phase portrait of the flow associated to the  $2 \times 2$  system

$$(4.3.29) \quad \begin{aligned} \frac{d\theta}{dt} &= \psi, \\ \frac{d\psi}{dt} &= -\frac{g}{\ell} \sin \theta, \end{aligned}$$

which arises from the pendulum equation (cf. Chapter 1, (1.6.9))

$$(4.3.30) \quad \frac{d^2\theta}{dt^2} + \frac{g}{\ell} \sin \theta = 0,$$

by adding the variable  $\psi = d\theta/dt$ . Here  $g, \ell > 0$ . The system (4.3.29) has the form (4.3.2) with  $x = (\theta, \psi)$  and

$$(4.3.31) \quad F(\theta, \psi) = \begin{pmatrix} \psi \\ -\frac{g}{\ell} \sin \theta \end{pmatrix}.$$

Note that Figure 4.3.1 looks like Figure 1.6.2 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

$$(4.3.32) \quad \mathcal{E}(\theta, \psi) = \frac{\psi^2}{2} - \frac{g}{\ell} \cos \theta,$$

since if  $(\theta(t), \psi(t))$  solves (4.3.29),

$$(4.3.33) \quad \frac{d}{dt}\mathcal{E}(\theta, \psi) = \psi\psi' + \frac{g}{\ell}(\sin \theta)\theta' = 0.$$

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

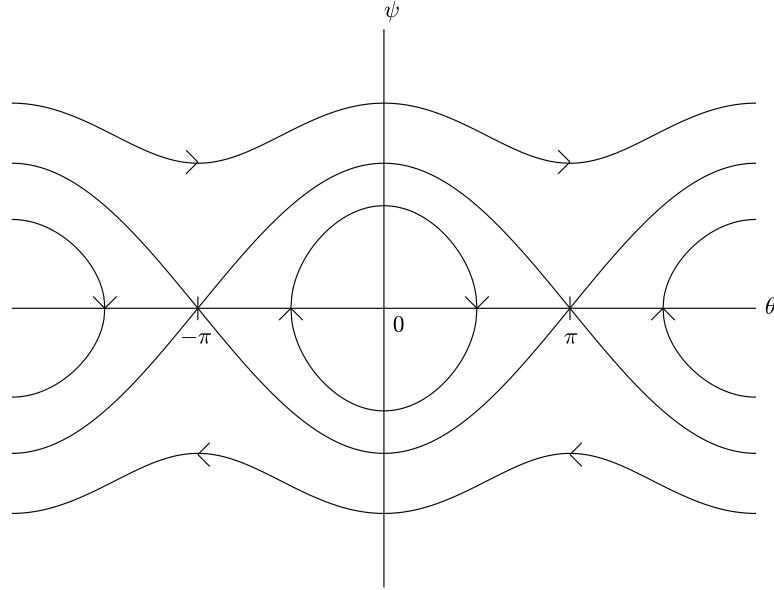
$$(4.3.34) \quad \frac{d}{dt}\mathcal{E}(\theta, \psi) = \nabla \mathcal{E}(\theta, \psi) \cdot F(\theta, \psi), \quad \text{where } \nabla \mathcal{E}(\theta, \psi) = \begin{pmatrix} \partial \mathcal{E} / \partial \theta \\ \partial \mathcal{E} / \partial \psi \end{pmatrix}.$$

Now the formula (4.3.31) gives

$$(4.3.35) \quad F(\theta, \psi) = -J \nabla \mathcal{E}(\theta, \psi),$$

where

$$(4.3.36) \quad J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$



**Figure 4.3.1.** Pendulum phase plane

so the vanishing of  $d\mathcal{E}(\theta, \psi)/dt$  follows from (4.3.34)–(4.3.35) and the skew-symmetry of  $J$ , which implies

$$(4.3.37) \quad v \cdot Jv = 0, \quad \forall v \in \mathbb{R}^2.$$

A vector field of the form (4.3.35) is a special case of a *Hamiltonian* vector field, a class of vector fields that will be discussed further in §§4.5, 4.7, and 4.10.

We mention some noteworthy features of the phase portrait in Figure 4.3.1, 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 (4.3.31), the set of critical points is

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

Figure 4.3.1 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. Figure 4.3.2.

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

- $$(4.3.38) \quad \begin{array}{l} \text{(a)} \quad \text{two orbits that approach } p \text{ as } t \rightarrow +\infty, \\ \text{(b)} \quad \text{two orbits that approach } p \text{ as } t \rightarrow -\infty, \\ \text{(c)} \quad \text{orbits that miss } p, \text{ looking like saddles.} \end{array}$$

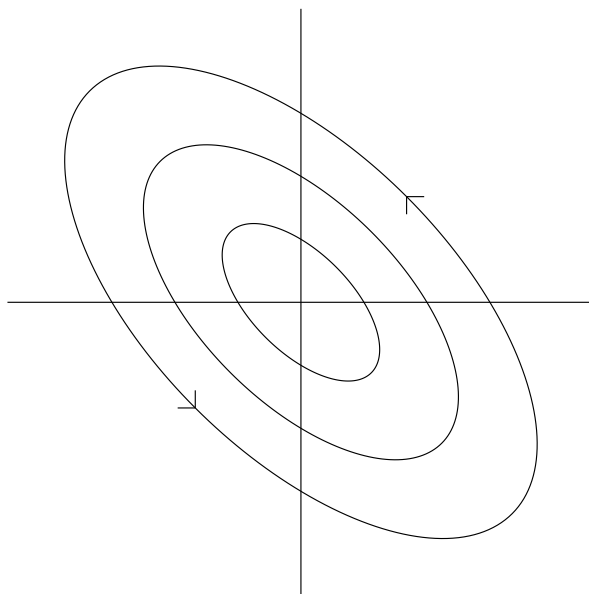


Figure 4.3.2. Center

We say these critical points are *saddles*. Cf. Figure 4.3.3. 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

$$(4.3.39) \quad L = DF(x_0) \in \mathcal{L}(\mathbb{R}^n).$$

This construction extends the notion of linearization given in §1.8 of Chapter 1. We expect that

$$(4.3.40) \quad \Phi_F^t(x_0 + y) \approx x_0 + e^{tL}y,$$

for  $\|y\|$  small. Cf. Exercise 6 of §4.2 (but mind the change in notation). Going further, we expect some important qualitative features of the flow  $\Phi_F^t$  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 (4.3.31), with critical points at  $p_k = (k\pi, 0)$ , we have

$$(4.3.41) \quad L_0 = DF(0, 0) = \begin{pmatrix} 0 & 1 \\ -\frac{g}{\ell} & 0 \end{pmatrix}.$$

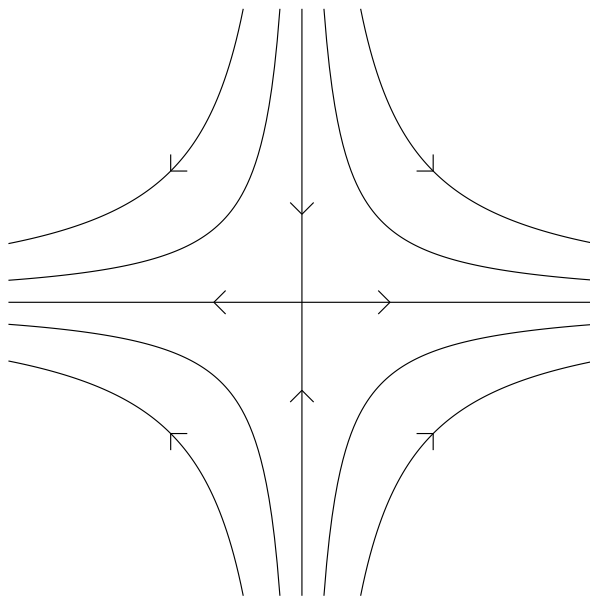


Figure 4.3.3. Saddle

The eigenvalues of this matrix are  $\pm i\sqrt{g/\ell}$ , and the orbits of  $e^{tL_0}$  are ellipses, with qualitative features like Figure 4.3.2, a center. Meanwhile,

$$(4.3.42) \quad L_1 = DF(\pm\pi, 0) = \begin{pmatrix} 0 & 1 \\ \frac{g}{\ell} & 0 \end{pmatrix}.$$

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 Figure 4.3.3, 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

$$(4.3.43) \quad F(x) = Jx - \|x\|^2 x, \quad x \in \mathbb{R}^2,$$

with  $J$  as in (4.3.36). 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

$$(4.3.44) \quad \begin{aligned} \frac{d}{dt}\|x(t)\|^2 &= 2x \cdot x' \\ &= 2x \cdot (Jx - \|x\|^2 x) \\ &= -2\|x\|^4, \end{aligned}$$



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

$$(4.3.45) \quad \frac{d\rho}{dt} = -2\rho^2.$$

This is separable and we have

$$(4.3.46) \quad \rho(0) = \rho_0 \implies \rho(t) = \frac{\rho_0}{1 + 2t\rho_0} \rightarrow 0 \text{ as } t \nearrow +\infty,$$

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 (4.3.43), the fact that  $(0, 0)$  is a center for  $F$ , given by (4.3.31), is no accident, but rather a consequence of the fact that  $F$  has the form (4.3.35),

$$(4.3.47) \quad F(x) = -J\nabla\mathcal{E}(x),$$

so that, as derived in (4.3.34)–(4.3.37), 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 (4.3.47), (nondegenerate) critical points of  $F$  and (nondegenerate) critical points of  $\mathcal{E}$  coincide. If  $x_0 \in \Omega$  is such a point

$$(4.3.48) \quad DF(x_0) = -JD^2\mathcal{E}(x_0),$$

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

$$(4.3.49) \quad D^2\mathcal{E} = \begin{pmatrix} \partial^2\mathcal{E}/\partial\theta^2 & \partial^2\mathcal{E}/\partial\psi\partial\theta \\ \partial^2\mathcal{E}/\partial\theta\partial\psi & \partial^2\mathcal{E}/\partial\psi^2 \end{pmatrix}.$$

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

$$(4.3.50) \quad \begin{aligned} D^2\mathcal{E}(x_0) \text{ positive definite} &\Leftrightarrow \mathcal{E} \text{ has a local minimum at } x_0, \\ D^2\mathcal{E}(x_0) \text{ negative definite} &\Leftrightarrow \mathcal{E} \text{ has a local maximum at } x_0, \\ D^2\mathcal{E}(x_0) \text{ indefinite} &\Leftrightarrow \mathcal{E} \text{ has a saddle at } x_0, \end{aligned}$$

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

$$(4.3.51) \quad \begin{aligned} A \text{ positive definite} &\Leftrightarrow \det A > 0 \text{ and } \operatorname{Tr} A > 0, \\ A \text{ negative definite} &\Leftrightarrow \det A > 0 \text{ and } \operatorname{Tr} A < 0, \\ A \text{ indefinite} &\Leftrightarrow \det A < 0. \end{aligned}$$

Furthermore, if  $A$  is such a matrix and

$$(4.3.52) \quad B = -JA,$$

then

$$(4.3.53) \quad \det B = \det A,$$

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

$$(4.3.54) \quad \begin{aligned} B \text{ has 2 real eigenvalues of opposite signs} &\Leftrightarrow \det B < 0, \\ B \text{ has 2 purely imaginary eigenvalues} &\Leftrightarrow \det B > 0 \text{ and } \operatorname{Tr} B = 0, \end{aligned}$$

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

**Proposition 4.3.3.** *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 (4.3.47). Then*

$$(4.3.55) \quad \begin{aligned} DF(x_0) \text{ has 2 purely imaginary eigenvalues} \\ \Leftrightarrow \mathcal{E} \text{ has a local max or local min at } x_0, \end{aligned}$$

and

$$(4.3.56) \quad \begin{aligned} DF(x_0) \text{ has 2 real eigenvalues of opposite sign} \\ \Leftrightarrow \mathcal{E} \text{ has a saddle at } x_0. \end{aligned}$$

We move on to the  $2 \times 2$  system

$$(4.3.57) \quad \begin{aligned} \frac{d\theta}{dt} &= \psi, \\ \frac{d\psi}{dt} &= -\frac{\alpha}{m}\psi - \frac{g}{\ell}\sin\theta, \end{aligned}$$

which arises from the damped pendulum equation (cf. Chapter 1, (1.7.6)),

$$(4.3.58) \quad \frac{d^2\theta}{dt^2} + \frac{\alpha}{m}\frac{d\theta}{dt} + \frac{g}{\ell}\sin\theta = 0,$$

by adding the variable  $\psi = d\theta/dt$ . Here  $g, \ell, \alpha, m > 0$ . The system (4.3.57) has the form (4.3.2) with  $x = (\theta, \psi)$  and

$$(4.3.59) \quad F(\theta, \psi) = \begin{pmatrix} \psi \\ -\frac{\alpha}{m}\psi - \frac{g}{\ell}\sin\theta \end{pmatrix}.$$

The phase portrait for this system is illustrated in Figure 4.3.4. We compare and contrast this portrait with that depicted in Figure 4.3.1.

To start, the vector field (4.3.59) has the same critical points as the field given by (4.3.31), 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. Figure 4.3.4 depicts orbits spiraling into these critical points, as opposed to the picture in Figure 4.3.1 of closed orbits circling these critical points. Let us consider the linearizations about these critical points. For  $F$  as in (4.3.59), we have

$$(4.3.60) \quad L = DF(0, 0) = \begin{pmatrix} 0 & 1 \\ -\frac{g}{\ell} & -\frac{\alpha}{m} \end{pmatrix},$$

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

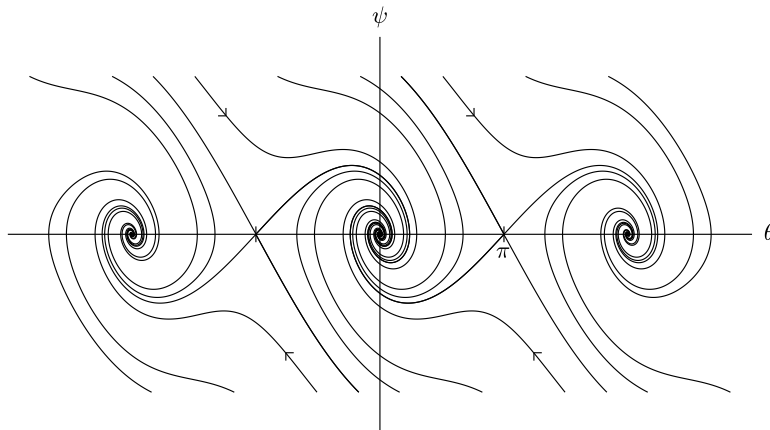
$$(4.3.61) \quad \lambda_{\pm} = -\frac{\alpha}{2m} \pm \sqrt{\frac{\alpha^2}{m^2} - \frac{4g}{\ell}}.$$

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  $\lambda_+$  and  $\lambda_-$  are distinct real numbers, each negative.



**Figure 4.3.4.** Damped pendulum phase plane

In all three cases, we have  $e^{tL}v \rightarrow 0$  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 Figure 4.3.5(a). Figure 4.3.4 depicts such behavior. In Case III, the orbits look like those in Figure 4.3.5(c). In Case II, the orbits look like a cross between Figure 4.3.5(b) and Figure 4.3.5(c). These critical points are all called *sinks*. (Reverse the sign on  $F$ , and the associated orbits are called *sources*; cf. Figure 4.3.6.) The three cases described above correspond to damped oscillatory, critically damped, and overdamped motion, as discussed in §1.9 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 (4.3.32), i.e.,

$$(4.3.62) \quad \mathcal{E}(\theta, \psi) = \frac{\psi^2}{2} - \frac{g}{\ell} \cos \theta.$$

This time, instead of (4.3.33), we have

$$(4.3.63) \quad \begin{aligned} \frac{d}{dt} \mathcal{E}(\theta, \psi) &= \psi \psi' + \frac{g}{\ell} (\sin \theta) \theta' \\ &= -\psi \left( \frac{\alpha}{m} \psi + \frac{g}{\ell} \sin \theta \right) + \frac{g}{\ell} (\sin \theta) \psi \\ &= -\frac{\alpha}{m} \psi^2. \end{aligned}$$

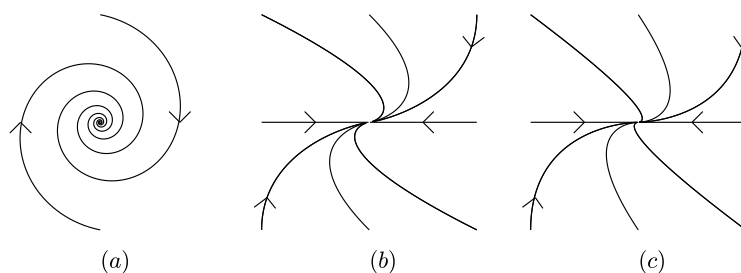


Figure 4.3.5. Sinks

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

**Proposition 4.3.4.** *Let  $F$  be a smooth vector field on  $\Omega \subset \mathbb{R}^n$ , with a critical point at  $x_0 \in \Omega$ . Assume*

$$(4.3.64) \quad \text{all the eigenvalues of } DF(x_0) \text{ have negative real part.}$$

*Then there exists  $\delta > 0$  such that*

$$(4.3.65) \quad \|x - x_0\| \leq \delta \implies \lim_{t \rightarrow +\infty} \Phi_F^t x = x_0.$$

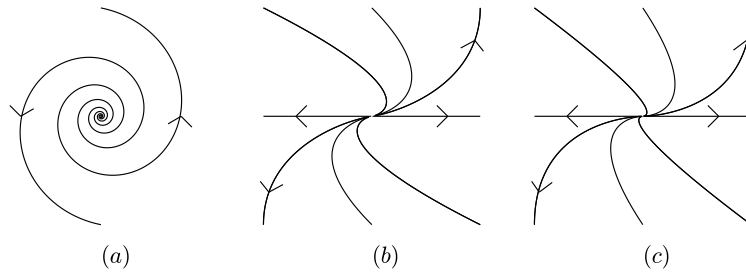
To prove this, we bring in the following linear algebra result.

**Lemma 4.3.5.** *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  $\langle \cdot, \cdot \rangle$  on  $\mathbb{R}^n$  and a positive constant  $K$  such that*

$$(4.3.66) \quad \langle Lv, v \rangle \leq -K \langle v, v \rangle, \quad \forall v \in \mathbb{R}^n.$$

We show how Lemma 4.3.5 allows us to prove Proposition 4.3.4. Apply the lemma to  $L = DF(x_0)$ . Note that there exist  $a, b \in (0, \infty)$  such that

$$(4.3.67) \quad a\|v\|^2 \leq \langle v, v \rangle \leq b\|v\|^2, \quad \forall v \in \mathbb{R}^n,$$



**Figure 4.3.6.** Sources

where  $\langle v, v \rangle$  is as in (4.3.66) and, as usual,  $\|v\|^2 = v \cdot v$ . Since  $F$  is smooth,

$$(4.3.68) \quad F(x_0 + y) = Ly + R(y),$$

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

$$(4.3.69) \quad \|R(y)\| \leq C\|y\|^2 \leq C'\langle y, y \rangle.$$

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

$$(4.3.70) \quad \begin{aligned} \frac{d}{dt} \langle y(t), y(t) \rangle &= 2 \langle y'(t), y(t) \rangle \\ &= 2 \langle F(x_0 + y), y \rangle \\ &= 2 \langle Ly, y \rangle + 2 \langle R(y), y \rangle. \end{aligned}$$

Now (4.3.66) applies to the first term in the last line of (4.3.70), while Cauchy's inequality plus (4.3.69) yields

$$(4.3.71) \quad \begin{aligned} |\langle R(y), y \rangle| &\leq \langle R(y), R(y) \rangle^{1/2} \langle y, y \rangle^{1/2} \\ &\leq C \langle y, y \rangle^{3/2}. \end{aligned}$$

Hence

$$(4.3.72) \quad \begin{aligned} \frac{d}{dt} \langle y, y \rangle &\leq -2K \langle y, y \rangle + C \langle y, y \rangle^{3/2} \\ &\leq -K \langle y, y \rangle, \end{aligned}$$

the last inequality holding provided  $\langle y, y \rangle^{1/2} \leq K/C$ . As long as  $\delta$  in (4.3.65) is small enough that  $\{x \in \mathbb{R}^n : \|x - x_0\| \leq \delta\}$  is contained in  $\Omega$  and  $\|v\| \leq \delta \Rightarrow \langle v, v \rangle^{1/2} \leq K/C$ , if  $x = x_0 + y_0$  and  $\|y_0\| \leq \delta$ , then (4.3.72) holds for  $y(t) = \Phi_F^t(x_0 + y_0) - x_0$  for all  $t \geq 0$ , and yields

$$(4.3.73) \quad \langle y(t), y(t) \rangle \leq e^{-Kt} \langle y_0, y_0 \rangle,$$

which in turn gives (4.3.65).

We now prove Lemma 4.3.5. As shown in §2.8 of Chapter 2,  $\mathbb{C}^n$  has a basis  $\{v_1, \dots, v_n\}$  with respect to which  $L$  is upper triangular, i.e.,

$$(4.3.74) \quad Lv_j = \lambda_j v_j + \sum_{k < j} a_{jk} v_k.$$

Alternatively, Appendix 2.B of Chapter 2 shows that  $\mathbb{C}^n$  has an orthonormal basis  $\{v_j\}$  for which (4.3.74) holds. The eigenvalues of  $L$  are  $\lambda_j$ , so by hypothesis there exists  $K_1 \in (0, \infty)$  such that  $\operatorname{Re} \lambda_j \leq -K_1$  for all  $j$ . Now if we take  $\varepsilon > 0$  and set  $w_j = \varepsilon^j v_j$ , we get

$$(4.3.75) \quad Lw_j = \lambda_j w_j + \sum_{k < j} \varepsilon^{j-k} a_{jk} w_k.$$

Then setting

$$(4.3.76) \quad \left\langle \sum a_j w_j, \sum b_k w_k \right\rangle = \operatorname{Re} \sum a_j \bar{b}_j$$

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 (4.3.66) follows, with  $K = K_1/2$ , from (4.3.75).

See the exercises for another proof of Lemma 4.3.5.

Having discussed the critical points of the vector field (4.3.59) at  $(0, 0)$  and related issues, we now consider the critical points at  $(\pm\pi, 0)$ . We have

$$(4.3.77) \quad DF(\pm\pi, 0) = \begin{pmatrix} 0 & 1 \\ \frac{g}{\ell} & -\frac{\alpha}{m} \end{pmatrix}.$$

This matrix has eigenvalues

$$(4.3.78) \quad \lambda_{\pm} = -\frac{\alpha}{2m} \pm \sqrt{\frac{\alpha^2}{m^2} + \frac{4g}{\ell}},$$

one positive and one negative. These critical points are saddles. The orbits near these critical points have a behavior such as described in (4.3.38). Unlike the case of  $F$  given by (4.3.31), where the orbits are level curves of  $\mathcal{E}$ , the proof of this is more subtle in the present situation. See Appendix 4.C for a proof.

Having studied the various critical points depicted in Figures 4.3.1 and 4.3.4, 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  $\Phi_F^t$  satisfying

$$(4.3.79) \quad \lim_{t \rightarrow -\infty} x(t) = p_1, \quad \lim_{t \rightarrow +\infty} x(t) = p_2$$

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 Figure 4.3.1, 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 Figure 4.3.1 portrays the motion of an idealized pendulum, without friction. The motion of a pendulum with friction is portrayed in Figure 4.3.4.

In Figure 4.3.4, 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  $\lim_{t \rightarrow -\infty} x(t) = (-\pi, 0)$  and  $\psi(t) > 0$  for large negative  $t$ , the fact that  $\lim_{t \rightarrow +\infty} x(t) = (0, 0)$  can be deduced from (4.3.63), i.e.,

$$(4.3.80) \quad \frac{d}{dt} \mathcal{E}(\theta, \psi) = -\frac{\alpha}{m} \psi^2.$$

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

$$(4.3.81) \quad F(\theta, \psi) = \begin{pmatrix} \frac{g}{\ell} \sin \theta \\ \psi \end{pmatrix}.$$

See Figure 4.3.7. In this case,

$$(4.3.82) \quad F = \nabla \mathcal{E} = \begin{pmatrix} \partial \mathcal{E} / \partial \theta \\ \partial \mathcal{E} / \partial \psi \end{pmatrix},$$

with  $\mathcal{E}$  given by (4.3.32), i.e.,

$$(4.3.83) \quad \mathcal{E}(\theta, \psi) = \frac{\psi^2}{2} - \frac{g}{\ell} \cos \theta.$$

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

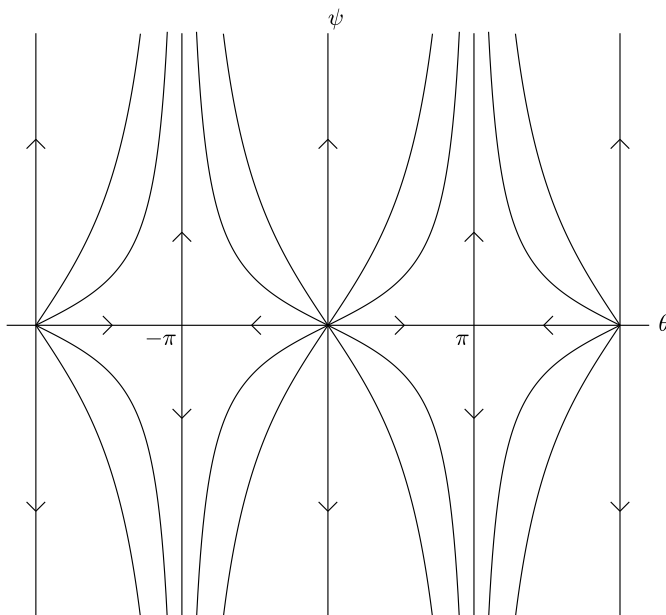
$$(4.3.84) \quad \frac{d}{dt} \mathcal{E}(x(t)) = \nabla \mathcal{E}(x(t)) \cdot \nabla \mathcal{E}(x(t)) = \|\nabla \mathcal{E}(x(t))\|^2.$$

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

$$(4.3.85) \quad DF(0, 0) = \begin{pmatrix} \frac{g}{\ell} & 0 \\ 0 & 1 \end{pmatrix},$$

which is positive definite. The origin is a (non-spiraling) *source*; cf. Figure 4.3.6. In particular, if  $x(t) = (\theta(t), \psi(t))$  is an orbit and  $x(0)$  is close to  $(0, 0)$ , then

$$(4.3.86) \quad \lim_{t \rightarrow -\infty} x(t) = (0, 0).$$



**Figure 4.3.7.** Gradient vector field,  $\nabla\mathcal{E}(\theta, \psi)$

This can be deduced from Proposition 4.3.4 by reversing time. It also follows directly from (4.3.84). For  $k$  odd, we have saddles:

$$(4.3.87) \quad DF(\pm\pi, 0) = \begin{pmatrix} -\frac{g}{\ell} & 0 \\ 0 & 1 \end{pmatrix}.$$

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

Note that an orbit for  $F = \nabla\mathcal{E}$  satisfies

$$(4.3.88) \quad \theta' = \frac{g}{\ell} \sin \theta, \quad \psi' = \psi,$$

so  $\psi(t) = Ae^t$ , and, for  $\theta$  away from  $\{k\pi : k \in \mathbb{Z}\}$ ,

$$(4.3.89) \quad \frac{g}{\ell}t + B = \int \frac{d\theta}{\sin \theta} = \int \frac{d\theta}{\cos(\theta - \pi/2)},$$

an integral that can be evaluated via results of Exercise 14 in §1.1.



---

**Exercises**

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

$$(4.3.90) \quad D\Phi_F^t(x)F(x) = F(\Phi_F^t(x))$$

and

$$(4.3.91) \quad Dv^t(x) = Dv(\Phi_F^t(x))D\Phi_F^t(x).$$

Relate these identities to the simultaneous validity of (4.3.10) and (4.3.12).

*Hint.* To get (4.3.90), use

$$(4.3.92) \quad \frac{d}{dt}\Phi_F^t(x) = \frac{d}{ds}\Phi_F^t \circ \Phi_F^s(x) \Big|_{s=0} = D\Phi_F^t(x)F(x),$$

and compare (4.3.8).

2. Extend Proposition 4.3.2 as follows. Replace hypothesis (4.3.26) by

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

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

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

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

then the flow generated by  $F$  is complete.

*(Hint.* Recall Exercise 12 of §4.1.)

Show that the flow is forward complete if

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

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

$$(4.3.93) \quad n(x) \cdot F(x) < 0, \quad \forall x \in \partial U.$$

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

5. In the setting of Exercise 4, relax the hypothesis (4.3.93) to

$$(4.3.94) \quad n(x) \cdot F(x) \leq 0, \quad \forall x \in \partial U.$$

Show that  $\Phi_F^t(x) \in \bar{U}$  if  $x \in \bar{U}$  and  $t \geq 0$ , and deduce that  $\Phi_F^t$  is forward complete on  $\bar{U}$ .

*Hint.* Take a  $C^1$  vector field  $X$  on  $\Omega$  such that  $X \cdot n < 0$  on  $\partial U$ . Set  $F_\tau = F + \tau X$  to produce a smooth family  $F_\tau$  of  $C^1$  vector fields on  $\Omega$  such that  $F_0 = F$  and, for  $0 < \tau < 1$ ,  $F_\tau$  has the property given in (4.3.93). Then make use of Exercise 4 and of results of §4.2.

6. In the setting of Exercise 5, replace the hypothesis (4.3.94) by

$$(4.3.95) \quad n(x) \cdot F(x) = 0, \quad \forall x \in \partial U.$$

Show that  $\Phi_F^t$  is complete on  $\bar{U}$ , and that  $\Phi_F^t(x) \in \partial U$  whenever  $x \in \partial U$  and  $t \in \mathbb{R}$ .

7. Show that if  $F$  is given by (4.3.31), then

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

while if  $F$  is given by (4.3.59), then

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

and if  $F$  is given by (4.3.81), then

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

8. Let  $A \in M(2, \mathbb{R})$ , and take  $J$  as in (4.3.36). Show that if  $A$  is positive definite then  $A = P^2$  with  $P$  positive definite. Show that

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

and deduce that

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

Relate this to Proposition 4.3.3.

9. Consider the system

$$(4.3.96) \quad \frac{dx}{dt} = y, \quad \frac{dy}{dt} = 1 - x^2.$$

Take  $E(x, y) = y^2/2 + x^3/3 - x$ . Show that if  $(x(t), y(t))$  solves (4.3.96), 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.

10. Find all the critical points of each of the following vector fields, and specify whether each one is a source, sink, saddle, or center.

$$F(x, y) = \begin{pmatrix} \sin x \cos y \\ \cos x \sin y \end{pmatrix}, \quad G(x, y) = \begin{pmatrix} \sin x \cos y \\ -\cos x \sin y \end{pmatrix}.$$

11. Returning to the context of Exercise 1, show that (4.2.2) gives

$$(4.3.97) \quad \frac{d}{dt} D\Phi_F^t(x) = DF(\Phi_F^t(x))D\Phi_F^t(x), \quad D\Phi_F^0(x) = I.$$

Recall from (3.8.6)–(3.8.10) of Chapter 3 that, for an  $n \times n$  matrix function  $M(t)$ ,

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

Deduce that

$$(4.3.98) \quad \begin{aligned} \frac{d}{dt} \det D\Phi_F^t(x) &= \operatorname{Tr} DF(\Phi_F^t(x)) \det D\Phi_F^t(x) \\ &= \operatorname{div} F(\Phi_F^t(x)) \det D\Phi_F^t(x). \end{aligned}$$

Relate this to (4.3.13), using the change of variable formula

$$(4.3.99) \quad \int u(x) dx = \int u(\Phi_F^t(x)) \det D\Phi_F^t(x) dx.$$

12. Use (3.8.10) of Chapter 3 to conclude from (4.3.98) that

$$(4.3.100) \quad \det D\Phi_F^t(x) = \exp\left\{\int_0^t \operatorname{div} F(\Phi_F^s(x)) ds\right\}.$$

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

$$(4.3.101) \quad \int_U \operatorname{div} F(x) dx = \int_{\partial U} n(x) \cdot F(x) dS(x),$$

where  $n(x)$  is the outward pointing unit normal to  $\partial U$  and  $dS(x)$  is  $(n-1)$ -dimensional surface area on  $\partial U$  (arc length if  $n=2$ ). Given this identity, we see that, in the setting of Proposition 4.3.1, (4.3.24) is equivalent to

$$(4.3.102) \quad \frac{d}{dt} \operatorname{Vol} \Phi_F^t(B) = \int_{\partial \Phi_F^t(B)} n(x) \cdot F(x) dS(x).$$

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

$$(4.3.103) \quad \frac{d}{dt} \operatorname{Vol} \Phi_F^t(U) \Big|_{t=0} = \int_{\partial U} n(x) \cdot F(x) dS(x).$$

Try to provide a direct demonstration of (4.3.103) (at least for  $n=2$ ).

Exercises 14–16 lead to another proof of Lemma 4.3.5.

14. Take  $L \in M(n, \mathbb{R})$ . Recall from §2.7 of Chapter 2 that  $\mathbb{C}^n$  has a basis of generalized eigenvectors for  $L$  and if  $v \in \mathcal{GE}(L, \lambda)$ , then  $e^{tL}v$  has the form

$$e^{tL}v = e^{t\lambda} \sum_{k=0}^{\ell} t^k v_k, \quad v_k \in \mathbb{C}^n.$$

Use these facts to show that if  $\operatorname{Re} \lambda < 0$  for each eigenvalue  $\lambda$  of  $L$ , then there exist  $C, K \in (0, \infty)$  such that

$$(4.3.104) \quad \|e^{tL}\| \leq Ce^{-Kt}, \quad \forall t \geq 0.$$

15. Assume  $L \in M(n, \mathbb{R})$  and  $\operatorname{Re} \lambda < 0$  for each eigenvalue  $\lambda$  of  $L$ . Let  $(v, w) = v \cdot w$

denote the standard Euclidean inner product on  $\mathbb{R}^n$ . Show that  $\langle v, w \rangle$ , given by

$$\langle v, w \rangle = \int_0^\infty (e^{tL}v, e^{tL}w) dt$$

is a well-defined, symmetric, positive-definite inner product on  $\mathbb{R}^n$ .

*Hint.* Use (4.3.104) to show that the integral is absolutely convergent.

16. In the setting of Exercise 15, show that, for  $v \in \mathbb{R}^n$ ,

$$\langle e^{sL}v, e^{sL}v \rangle = \int_s^\infty (e^{tL}v, e^{tL}v) dt,$$

and deduce that

$$\frac{d}{dt} \langle e^{sL}v, e^{sL}v \rangle \Big|_{s=0} = -(v, v).$$

On the other hand, show that also

$$\frac{d}{ds} \langle e^{sL}v, e^{sL}v \rangle \Big|_{s=0} = \langle Lv, v \rangle + \langle v, Lv \rangle = 2\langle Lv, v \rangle,$$

and obtain another proof of Lemma 4.3.5.

#### 4.4. Gradient vector fields

As mentioned in §4.3, 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

$$(4.4.1) \quad F = \nabla u,$$

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 (4.4.1) holds. Then

$$(4.4.2) \quad \frac{\partial F_k}{\partial x_j} = \frac{\partial}{\partial x_j} \frac{\partial u}{\partial x_k},$$

and

$$(4.4.3) \quad \frac{\partial}{\partial x_j} \frac{\partial u}{\partial x_k} = \frac{\partial}{\partial x_k} \frac{\partial u}{\partial x_j},$$

so if (4.4.1) holds then

$$(4.4.4) \quad \frac{\partial F_k}{\partial x_j} = \frac{\partial F_j}{\partial x_k}, \quad \forall j, k \in \{1, \dots, n\}.$$

We will establish the following converse.

**Proposition 4.4.1.** *Assume  $\Omega \subset \mathbb{R}^n$  is a connected open set satisfying the condition (4.4.13) given below. Let  $F$  be a  $C^1$  vector field on  $\Omega$ . If (4.4.4) holds on  $\Omega$ , then there exists  $u \in C^2(\Omega)$  such that (4.4.1) holds.*

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

$$(4.4.5) \quad \gamma : [0, 1] \longrightarrow \Omega, \quad \gamma(0) = p, \quad \gamma(1) = x.$$

We propose that, under the hypotheses of Proposition 4.4.1, we can take

$$(4.4.6) \quad u(x) = \int_{\gamma} F(y) \cdot dy.$$

Here the line integral is defined by

$$(4.4.7) \quad \int_{\gamma} F(y) \cdot dy = \int_0^1 F(\gamma(t)) \cdot \gamma'(t) dt.$$

For this to work, we need to know that (4.4.6) 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  $\gamma_s$  from  $p$  to  $x$ :

$$(4.4.8) \quad \begin{aligned} \gamma_s(t) &= \gamma(s, t), \quad \gamma : [0, 1] \times [0, 1] \longrightarrow \Omega, \\ \gamma(s, 0) &= p, \quad \gamma(s, 1) = x. \end{aligned}$$

**Lemma 4.4.2.** *If  $F$  is a  $C^1$  vector field satisfying (4.4.4) and  $\gamma_s$  is a smooth family satisfying (4.4.8), then*

$$(4.4.9) \quad \int_{\gamma_s} F(y) \cdot dy \text{ is independent of } s \in [0, 1].$$

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

$$(4.4.10) \quad \begin{aligned} & \int_0^1 F(\gamma(s, t)) \cdot \frac{\partial \gamma}{\partial t}(s, t) dt \\ &= \int_0^1 \sum_j F_j(\gamma(s, t)) \frac{\partial \gamma_j}{\partial t}(s, t) dt. \end{aligned}$$

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

$$(4.4.11) \quad \begin{aligned} \frac{d}{ds} \int_{\gamma_s} F(y) \cdot dy &= \int_0^1 \sum_{j,k} \frac{\partial F_j}{\partial x_k}(\gamma(s, t)) \frac{\partial}{\partial s} \gamma_k(s, t) \frac{\partial}{\partial t} \gamma_j(s, t) dt \\ &+ \int_0^1 \sum_j F_j(\gamma(s, t)) \frac{\partial}{\partial s} \frac{\partial}{\partial t} \gamma_j(s, t) dt. \end{aligned}$$

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 (4.4.11) 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 (4.4.11) becomes

$$(4.4.12) \quad - \int_0^1 \sum_{j,k} \frac{\partial F_j}{\partial x_k}(\gamma(s, t)) \frac{\partial}{\partial t} \gamma_k(s, t) \frac{\partial}{\partial s} \gamma_j(s, t) dt.$$

Now if we interchange the roles of  $j$  and  $k$  in (4.4.12), we cancel the first integral on the right side of (4.4.11), provided (4.4.4) holds. This proves the lemma.  $\square$

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

$$(4.4.13) \quad \begin{aligned} & \text{Given } p, x \in \Omega, \text{ if } \gamma_0 \text{ and } \gamma_1 \text{ are smooth paths from } p \text{ to } x, \\ & \text{they are connected by a smooth family } \gamma_s \text{ of paths from } p \text{ to } x. \end{aligned}$$

Here is a class of such domains.

**Lemma 4.4.3.** *If  $\Omega \subset \mathbb{R}^n$  is an open convex domain, then  $\Omega$  is simply connected.*

**Proof.** If  $\Omega$  is convex, two paths  $\gamma_0$  and  $\gamma_1$  from  $p \in \Omega$  to  $x \in \Omega$  are connected by

$$(4.4.14) \quad \gamma_s(t) = (1-s)\gamma_0(t) + s\gamma_1(t), \quad 0 \leq s \leq 1.$$

$\square$

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

Now that we have Lemma 4.4.2, under the hypotheses of Proposition 4.4.1 we simply write

$$(4.4.15) \quad u(x) = \int_p^x F(y) \cdot dy.$$

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

$$(4.4.16) \quad u(x) = \int_p^q F(y) \cdot dy + \int_q^x F(y) \cdot dy.$$

Again using the path independence, we see we can independently choose paths from  $p$  to  $q$  and from  $q$  to  $x$  in (4.4.16); these paths need not match up smoothly at  $q$ .

We are now in a position to complete the proof of Proposition 4.4.1. Take  $\delta > 0$  so that  $\{y \in \mathbb{R}^n : \|x - y\| \leq \delta\} \subset \Omega$ . Take  $k \in \{1, \dots, n\}$ , fix  $q_k$  such that  $|q_k - x_k| < \delta$ , and write

$$(4.4.17) \quad u(x) = \int_p^{(x_1, \dots, q_k, \dots, x_n)} F(y) \cdot dy + \int_{(x_1, \dots, q_k, \dots, x_n)}^x F(y) \cdot dy.$$

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 (4.4.17) is independent of  $x_k$ , so

$$(4.4.18) \quad \begin{aligned} \frac{\partial u}{\partial x_k}(x) &= \frac{\partial}{\partial x_k} \int_{(x_1, \dots, q_k, \dots, x_n)}^x F(y) \cdot dy \\ &= \frac{\partial}{\partial x_k} \int_{q_k}^{x_k} F_k(x_1, \dots, x_{k-1}, s, x_{k+1}, \dots, x_n) ds \\ &= F_k(x), \end{aligned}$$

the last identity by the fundamental theorem of calculus. This proves Proposition 4.4.1.

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

$$(4.4.19) \quad F(x) = \frac{Jx}{\|x\|^2}, \quad J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

with components

$$(4.4.20) \quad F_1(x) = \frac{-x_2}{x_1^2 + x_2^2}, \quad F_2(x) = \frac{x_1}{x_1^2 + x_2^2}.$$

We have

$$(4.4.21) \quad \frac{\partial F_1}{\partial x_2} = \frac{x_2^2 - x_1^2}{\|x\|^4} = \frac{\partial F_2}{\partial x_1}, \quad \text{on } \mathbb{R}^2 \setminus 0.$$

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 (4.4.1) is the angular coordinate  $\theta$ :

$$(4.4.22) \quad F(x) = \nabla\theta(x),$$

and this identity is true on any region  $\Omega$  formed by removing from  $\mathbb{R}^2$  a ray starting from the origin. However,  $\theta$  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  $2 \times 2$  system

$$(4.4.23) \quad \frac{dx}{dt} = f_1(x, y), \quad \frac{dy}{dt} = f_2(x, y).$$

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  $x$ :

$$(4.4.24) \quad \frac{dy}{dx} = \frac{f_2(x, y)}{f_1(x, y)},$$

which we rewrite as

$$(4.4.25) \quad \begin{aligned} g_1(x, y) dx + g_2(x, y) dy &= 0, \\ g_1(x, y) &= f_2(x, y), \quad g_2(x, y) = -f_1(x, y). \end{aligned}$$

The equation (4.4.25) is called exact if there exists  $u \in C^2(\Omega)$  such that

$$(4.4.26) \quad g_1 = \frac{\partial u}{\partial x}, \quad g_2 = \frac{\partial u}{\partial y}.$$

If there is such a  $u$ , solutions to (4.4.24) or (4.4.25) are given by

$$(4.4.27) \quad u(x, y) = C.$$

Now (4.4.26) is the condition that  $G = (g_1, g_2)^t$  be a gradient vector field on  $\Omega$ . Note that the relation between  $F = (f_1, f_2)^t$  and  $G = (g_1, g_2)^t$ , with components given by (4.4.25), is

$$(4.4.28) \quad G = -JF,$$

where

$$(4.4.29) \quad J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

As we have seen, when  $\Omega$  is simply connected, (4.4.26) holds for some  $u$  if and only if

$$(4.4.30) \quad \frac{\partial g_1}{\partial y} = \frac{\partial g_2}{\partial x}.$$

Note that this is equivalent to

$$(4.4.31) \quad \operatorname{div} F = 0.$$

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

$$(4.4.32) \quad \begin{aligned} \operatorname{curl} F &= \nabla \times F \\ &= \det \begin{pmatrix} i & j & k \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ F_1 & F_2 & F_3 \end{pmatrix} \\ &= \left( \frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z} \right) i + \left( \frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x} \right) j + \left( \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) k. \end{aligned}$$

We see that

$$(4.4.33) \quad (4.4.4) \text{ holds} \iff \operatorname{curl} F = 0.$$

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

$$(4.4.34) \quad \frac{\partial u}{\partial x_j}(x) = F_j(x), \quad 1 \leq j \leq n,$$



given the compatibility conditions (4.4.4), without evaluating line integrals. We start with

$$(4.4.35) \quad u_n(x) = \int F_n(x) dx_n, \quad \text{so} \quad \frac{\partial u_n}{\partial x_n} = F_n(x).$$

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

$$(4.4.36) \quad u(x) = u_n(x) + v(x'), \quad x' = (x_1, \dots, x_{n-1}).$$

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

$$(4.4.37) \quad \frac{\partial v}{\partial x_j} = F_j(x) - \frac{\partial u_n}{\partial x_j}, \quad 1 \leq j \leq n-1.$$

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

$$(4.4.38) \quad \begin{aligned} \frac{\partial}{\partial x_n} \left( F_j(x) - \frac{\partial u_n}{\partial x_j} \right) &= \frac{\partial F_j}{\partial x_n} - \frac{\partial}{\partial x_n} \frac{\partial u_n}{\partial x_j} \\ &= \frac{\partial F_n}{\partial x_j} - \frac{\partial}{\partial x_j} \frac{\partial u_n}{\partial x_n} \\ &= 0, \end{aligned}$$

the second identity by (4.4.4) (and (4.4.3)). Thus (4.4.37) takes the form

$$(4.4.39) \quad \frac{\partial v}{\partial x_j} = G_j(x'), \quad 1 \leq j \leq n-1,$$

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

$$(4.4.40) \quad \begin{aligned} \frac{\partial G_j}{\partial x_k} &= \frac{\partial F_j}{\partial x_k} - \frac{\partial}{\partial x_k} \frac{\partial u_n}{\partial x_j} \\ &= \frac{\partial F_k}{\partial x_j} - \frac{\partial}{\partial x_j} \frac{\partial u_n}{\partial x_k} \\ &= \frac{\partial G_k}{\partial x_j}, \end{aligned}$$

so the task of solving (4.4.39) is just like that in (4.4.34), but with one fewer variable. An iteration yields the solution to (4.4.34).

EXAMPLE. Take

$$(4.4.41) \quad F(x, y, z) = (y, x + z^2, 2yz)^t.$$

One readily verifies (4.4.4), or equivalently that  $\text{curl } F = 0$ . Here (4.4.35) gives

$$(4.4.42) \quad u_3(x, y, z) = \int 2yz dz = yz^2,$$

so

$$(4.4.43) \quad u(x, y, z) = yz^2 + v(x, y).$$

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

$$(4.4.44) \quad \frac{\partial v}{\partial y} = x,$$

so

$$(4.4.45) \quad v(x, y) = xy + w(x).$$

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

$$(4.4.46) \quad u(x, y, z) = yz^2 + xy,$$

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 (4.4.35)–(4.4.40) into an alternative proof of Proposition 4.4.1, at least if  $\Omega$  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 (4.4.19).

---

## Exercises

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

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

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

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

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

For (5)–(8), identify which equations are exact. If the equation is exact, write down the solution, in implicit form (4.4.27).

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

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

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

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

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.

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

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

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

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

13. Establish the following variant of Lemma 4.4.2:

**Lemma 4.4.2A.** *If  $F$  is a  $C^1$  vector field on  $\Omega$  satisfying (4.4.4) and  $\gamma_s$  is a smooth*

family satisfying

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

then

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

### 4.5. 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

$$(4.5.1) \quad m_k \frac{d^2 x_k}{dt^2} = \sum_{\{j:j \neq k\}} F_{jk}(x_k - x_j), \quad 1 \leq k \leq m.$$

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

$$(4.5.2) \quad F_{jk}(x_k - x_j) = -F_{kj}(x_j - x_k).$$

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

$$(4.5.3) \quad F_{jk}(x_k - x_j) = f_{jk}(\|x_k - x_j\|)(x_k - x_j).$$

In such a case, (4.5.2) is equivalent to

$$(4.5.4) \quad f_{jk}(r) = f_{kj}(r).$$

A force field of the form (4.5.3) is a gradient vector field:

$$(4.5.5) \quad \begin{aligned} f_{jk}(\|u\|)u &= -\nabla V_{jk}(u), \\ V_{jk}(u) &= v_{jk}(\|u\|), \quad v'_{jk}(r) = -rf_{jk}(r). \end{aligned}$$

If (4.5.4) holds,

$$(4.5.6) \quad V_{jk}(u) = V_{kj}(u).$$

The total energy of this system of interacting particles is

$$(4.5.7) \quad E = \frac{1}{2} \sum_k m_k \left\| \frac{dx_k}{dt} \right\|^2 + \frac{1}{2} \sum_{j \neq k} V_{jk}(x_k - x_j).$$

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,

$$(4.5.8) \quad \frac{dE}{dt} = \sum_k m_k \frac{d^2 x_k}{dt^2} \cdot \frac{dx_k}{dt} + \frac{1}{2} \sum_{j \neq k} \nabla V_{jk}(x_k - x_j) \cdot \left( \frac{dx_k}{dt} - \frac{dx_j}{dt} \right).$$

Next, (4.5.1) implies that the first sum on the right side of (4.5.8) is equal to

$$(4.5.9) \quad \sum_{j \neq k} F_{jk}(x_k - x_j) \cdot \frac{dx_k}{dt},$$

and (4.5.3)–(4.5.5) imply that the second sum on the right side of (4.5.8) is equal to

$$(4.5.10) \quad -\frac{1}{2} \sum_{j \neq k} F_{jk}(x_k - x_j) \cdot \left( \frac{dx_k}{dt} - \frac{dx_j}{dt} \right) = -\sum_{j \neq k} F_{jk}(x_k - x_j) \cdot \frac{dx_k}{dt}.$$

Comparing (4.5.9) and (4.5.10), we have energy conservation:

$$(4.5.11) \quad \frac{dE}{dt} = 0.$$

We can convert the second order system (4.5.1) 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*:

$$(4.5.12) \quad p_k = m_k \frac{dx_k}{dt}, \quad 1 \leq k \leq m.$$

We can express the energy  $E$  in (4.5.7) as a function of position  $x = (x_1, \dots, x_m)$  and momentum  $p = (p_1, \dots, p_m)$ :

$$(4.5.13) \quad E(x, p) = \sum_k \frac{1}{2m_k} \|p_k\|^2 + \frac{1}{2} \sum_{j \neq k} V_{jk}(x_k - x_j).$$

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

$$(4.5.14) \quad \frac{\partial E}{\partial p_{k\ell}} = \frac{1}{m_k} p_{k\ell},$$

and

$$(4.5.15) \quad \frac{\partial E}{\partial x_{k\ell}} = \sum_{\{j:j \neq k\}} \frac{\partial V_{jk}}{\partial u_\ell}(x_k - x_j),$$

invoking (4.5.6). Let us write (4.5.14)–(4.5.15) in vector form,

$$(4.5.16) \quad \frac{\partial E}{\partial p_k} = \frac{1}{m_k} p_k, \quad \frac{\partial E}{\partial x_k} = \sum_{\{j:j \neq k\}} \nabla V_{jk}(x_k - x_j),$$

where  $\partial E / \partial p_k = (\partial E / \partial p_{k1}, \dots, \partial E / \partial p_{kn})^t$ , etc. Now the system (4.5.1) yields the first order system

$$(4.5.17) \quad \frac{dx_k}{dt} = \frac{1}{m_k} p_k, \quad \frac{dp_k}{dt} = \sum_{\{j:j \neq k\}} F_{jk}(x_k - x_j),$$

which in turn, given (4.5.3)–(4.5.5), gives

$$(4.5.18) \quad \frac{dx_k}{dt} = \frac{\partial E}{\partial p_k}, \quad \frac{dp_k}{dt} = -\frac{\partial E}{\partial x_k}.$$

The system (4.5.18) 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

$$(4.5.19) \quad \begin{aligned} \frac{dx_k}{dt} &= \frac{\partial E}{\partial p_k}, \\ \frac{dp_k}{dt} &= -\frac{\partial E}{\partial x_k}, \end{aligned}$$

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

$$(4.5.20) \quad \frac{d}{dt} \begin{pmatrix} x \\ p \end{pmatrix} = X_E(x, p),$$

where  $X_E$  is a vector field on  $\Omega$ , called the Hamiltonian vector field associated to  $E$ . In this general setting,  $E$  is constant on each solution curve  $(x(t), p(t))$  of (4.5.19). Indeed, in such a case,

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

Returning to the setting (4.5.1)–(4.5.2), we next discuss the conservation of the total momentum

$$(4.5.22) \qquad P = \sum_k p_k = \sum_k m_k \frac{dx_k}{dt}.$$

Indeed,

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

the last identity by (4.5.2). Thus, for each solution  $x(t)$  to (4.5.1), there exist  $a, b \in \mathbb{R}^n$  such that

$$(4.5.24) \qquad \frac{1}{M} \sum_k m_k x_k(t) = a + bt, \quad M = \sum_k m_k.$$

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 (4.5.1):

$$(4.5.25) \qquad a = \frac{1}{M} \sum_k m_k x_k(0), \quad b = \frac{1}{M} \sum_k m_k x'_k(0).$$

Given this, we can obtain a system similar to (4.5.1) for the variables

$$(4.5.26) \qquad y_k(t) = x_k(t) - (a + bt).$$

We have  $y''_k = x''_k$  and  $y_k - y_j = x_k - x_j$ , so (4.5.1) gives

$$(4.5.27) \qquad m_k \frac{d^2y_k}{dt^2} = \sum_{\{j: j \neq k\}} F_{jk}(y_k - y_j), \quad 1 \leq k \leq m.$$

In this case we have the identity

$$(4.5.28) \qquad \sum_k m_k y_k(t) \equiv 0,$$

as a consequence of (4.5.24). We can use this to reduce the size of (4.5.27), from a system of  $mn$  equations to a system of  $(m-1)n$  equations, by substituting

$$(4.5.29) \qquad y_m = -\frac{1}{m_m} \sum_{\ell=1}^{m-1} m_\ell y_\ell$$

into (4.5.27), for  $1 \leq k \leq m-1$ . One calls  $(y_1, \dots, y_m)$  center of mass coordinates.

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

$$(4.5.30) \quad y_2 = -\frac{m_1}{m_2}y_1,$$

and the system (4.5.27) reduces to

$$(4.5.31) \quad m_1 \frac{d^2 y_1}{dt^2} = F_{21} \left( \left( 1 + \frac{m_1}{m_2} \right) y_1 \right),$$

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

$$(4.5.32) \quad \frac{m_1 m_2}{m_1 + m_2} \frac{d^2 x}{dt^2} = F_{21}(x).$$

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

$$(4.5.33) \quad y_3 = -\frac{m_1}{m_3}y_1 - \frac{m_2}{m_3}y_2,$$

and the system (4.5.27) reduces to

$$(4.5.34) \quad \begin{aligned} m_1 y_1'' &= F_{21}(y_1 - y_2) + F_{31} \left( \left( 1 + \frac{m_1}{m_2} \right) y_1 + \frac{m_2}{m_1} y_2 \right), \\ m_2 y_2'' &= F_{12}(y_2 - y_1) + F_{32} \left( \frac{m_1}{m_3} y_1 + \left( 1 + \frac{m_2}{m_3} \right) y_2 \right). \end{aligned}$$

### Exercises

1. In (a)–(e), 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.

(a)  $F_{jk}(x) = x$

(b)  $F_{jk}(x) = -x$

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

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

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

In all cases, (4.5.2) must be enforced.

2. For an alternative derivation of (4.5.32), when  $m = 2$ , write (4.5.1) as

$$\frac{d^2 x_1}{dt^2} = \frac{1}{m_1} F_{21}(x_1 - x_2), \quad \frac{d^2 x_2}{dt^2} = \frac{1}{m_2} F_{12}(x_2 - x_1),$$

and subtract, using (4.5.2).

### 4.6. Central force problems and two-body planetary motion

As seen in §4.5, one can transform the  $m$ -body problem (4.5.1) to center of mass coordinates, under the hypothesis (4.5.2), and obtain a smaller system, which for  $m = 2$  is given by (4.5.32). Changing notation, we rewrite (4.5.32) as

$$(4.6.1) \quad m \frac{d^2x}{dt^2} = F(x).$$

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 (4.5.3)–(4.5.4) for the two body problem, we have

$$(4.6.2) \quad F(x) = f(\|x\|)x.$$

In such a case, (4.6.1) is called a *central force problem*. Parallel to (4.5.5), we have

$$(4.6.3) \quad \begin{aligned} F(x) &= -\nabla V(x), \\ V(x) &= v(\|x\|), \quad v'(r) = -rf(r). \end{aligned}$$

The total energy is given by

$$(4.6.4) \quad E = \frac{1}{2}m \left\| \frac{dx}{dt} \right\|^2 + V(x),$$

and if  $x(t)$  solves (4.6.1), then

$$(4.6.5) \quad \frac{dE}{dt} = m \frac{d^2x}{dt^2} \cdot \frac{dx}{dt} + \nabla V(x) \cdot \frac{dx}{dt} = 0,$$

yielding conservation of energy.

There are further conservation laws, starting with the following.

**Proposition 4.6.1.** *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 (4.6.1) for  $t \in I$  and (4.6.2) holds, we have*

$$(4.6.6) \quad x(t) \in W, \quad \forall t \in I.$$

**Proof.** One way to see this is to note that (4.6.1) is a well posed system for  $x(t)$  taking values in  $W$ . Then uniqueness of solutions yields (4.6.6). Here is another demonstration.

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

$$(4.6.7) \quad \begin{aligned} Av &= v, & \forall v \in W, \\ Av &= -v, & \forall v \in W^\perp. \end{aligned}$$

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

$$(4.6.8) \quad y(0) = x(0), \quad y'(0) = x'(0).$$

Also, given  $F(x)$  of the form (4.6.2), we have  $AF(x) = F(y)$ , so  $y(t)$  solves (4.6.1). The basic uniqueness result proven in §4.1 implies  $y \equiv x$  on  $I$ , which in turn gives (4.6.6).  $\square$

A third proof of Proposition 4.6.1, valid for  $n = 3$ , can be obtained from conservation of angular momentum, established in (4.6.11) below.



Proposition 4.6.1 guarantees that each path  $x(t)$  solving (4.6.1) 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 (4.6.1) and  $x(t)$  is a path in  $\mathbb{R}^3$ . We define the *angular momentum*

$$(4.6.9) \quad \alpha(t) = mx(t) \times x'(t).$$

We then have, under hypothesis (4.6.2),

$$(4.6.10) \quad \begin{aligned} \alpha'(t) &= mx(t) \times x''(t) \\ &= x(t) \times F(x) \\ &= f(\|x\|) x(t) \times x(t) \\ &= 0. \end{aligned}$$

This yields conservation of angular momentum:

$$(4.6.11) \quad x(t) \times x'(t) \equiv L,$$

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

$$(4.6.12) \quad x(t) \times x'(t) = (0, 0, x_1(t)x_2'(t) - x_1'(t)x_2(t)),$$

so the conservation law (4.6.11) gives

$$(4.6.13) \quad x_1(t)x_2'(t) - x_1'(t)x_2(t) \equiv L_3.$$

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

$$(4.6.14) \quad x(t) = x_1(t) + ix_2(t) = r(t)e^{i\theta(t)}.$$

A computation gives

$$(4.6.15) \quad \begin{aligned} x' &= (r' + ir\theta')e^{i\theta}, \\ x'' &= [r'' - r(\theta')^2 + i(2r'\theta' + r\theta'')]e^{i\theta}, \end{aligned}$$

so (4.6.1)–(4.6.2) becomes

$$(4.6.16) \quad m[r'' - r(\theta')^2 + i(2r'\theta' + r\theta'')] = f(r)r.$$

Equating real and imaginary parts separately, we get

$$(4.6.17) \quad \begin{aligned} r'' - r(\theta')^2 &= \frac{f(r)r}{m}, \\ 2r'\theta' + r\theta'' &= 0. \end{aligned}$$

Note that

$$(4.6.18) \quad \frac{d}{dt}(r^2\theta') = r(2r'\theta' + r\theta''),$$

so the second equation in (4.6.17) says  $r^2\theta'$  is independent of  $t$ . This is actually equivalent to the conservation of angular momentum, (4.6.13). In fact, we have  $x_1 = r \cos \theta$ ,  $x_2 = r \sin \theta$ , hence

$$(4.6.19) \quad x_1' = r' \cos \theta - r\theta' \sin \theta, \quad x_2' = r' \sin \theta + r\theta' \cos \theta,$$

and hence

$$(4.6.20) \quad x_1x_2' - x_1'x_2 = r^2\theta'.$$

Thus we have in two ways derived the identity

$$(4.6.21) \quad r^2\theta' = L.$$

(For notational simplicity, we drop the subscript 3 from (4.6.13).)

There is the following geometrical interpretation of (4.6.21). 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

$$(4.6.22) \quad A(t) = \frac{1}{2} \int_{\theta(t_0)}^{\theta(t)} r^2 d\theta = \frac{1}{2} \int_{t_0}^t r(s)^2 \theta'(s) ds,$$

so

$$(4.6.23) \quad A'(t) = \frac{1}{2} r^2 \theta' = \frac{L}{2}.$$

This says

$$(4.6.24) \quad \text{Equal areas are swept out in equal times,}$$

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

Next, we can plug  $\theta' = L/r^2$  into the first equation of (4.6.17), obtaining

$$(4.6.25) \quad \frac{d^2 r}{dt^2} = \frac{f(r)r}{m} + \frac{L^2}{r^3}.$$

This has the form

$$(4.6.26) \quad \frac{d^2 r}{dt^2} = g(r),$$

treated in Chapter 1, §1.5. We recall that treatment. Take  $w(r)$  such that  $g(r) = -w'(r)$ , so (4.6.26) becomes

$$(4.6.27) \quad \frac{d^2 r}{dt^2} = -w'(r).$$

Then form the "energy"

$$(4.6.28) \quad E = \frac{1}{2} \left( \frac{dr}{dt} \right)^2 + w(r),$$

and compute that if  $r(t)$  solves (4.6.27) then

$$(4.6.29) \quad \frac{dE}{dt} = \frac{d^2 r}{dt^2} \frac{dr}{dt} + w'(r) \frac{dr}{dt} = 0,$$

so for each solution to (4.6.27), there is a constant  $E$  such that

$$(4.6.30) \quad \frac{dr}{dt} = \pm \sqrt{2E - 2w(r)}.$$

Separation of variables gives

$$(4.6.31) \quad \int \frac{dr}{\sqrt{2E - 2w(r)}} = \pm t + C.$$

This integral can be quite messy.

Note that dividing (4.6.30) by (4.6.21) yields a differential equation for  $r$  as a function of  $\theta$ :

$$(4.6.32) \quad \frac{dr}{d\theta} = \pm \frac{r^2}{L} \sqrt{2E - 2w(r)},$$

which separates to

$$(4.6.33) \quad L \int \frac{dr}{r^2 \sqrt{2E - 2w(r)}} = \pm \theta + C.$$

Let us recall that

$$(4.6.34) \quad w'(r) = -\frac{f(r)r}{m} - \frac{L^2}{r^3}.$$

Typically the integral in (4.6.33) is as messy as the one in (4.6.31). 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 (4.5.1), this means specifying the vector field  $F_{21}$  on  $\mathbb{R}^3$ . The formula is

$$(4.6.35) \quad F_{21}(x) = -Gm_1m_2 \frac{x}{\|x\|^3}.$$

Here  $G$  is the universal gravitational constant. If we go to center of mass coordinates, the motion of the planet is governed by (4.5.32), yielding (4.6.1) with

$$(4.6.36) \quad F(x) = -Km \frac{x}{\|x\|^3}, \quad K = G(m + m_2).$$

Here  $m = m_1$  is the mass of the planet and  $m_2$  is the mass of the sun. Consequently we have (4.6.17) with

$$(4.6.37) \quad \frac{f(r)r}{m} = -\frac{K}{r^2},$$

and (4.6.25) becomes

$$(4.6.38) \quad \frac{d^2r}{dt^2} = -\frac{K}{r^2} + \frac{L^2}{r^3}.$$

Thus  $w(r)$  in (4.6.27)–(4.6.34) is given by

$$(4.6.39) \quad w(r) = -\frac{K}{r} + \frac{L^2}{2r^2}.$$

Thus the integral in (4.6.31) is

$$(4.6.40) \quad \int \frac{r \, dr}{\sqrt{2Er^2 + 2Kr - L^2}},$$

and the integral in (4.6.33) is

$$(4.6.41) \quad \int \frac{dr}{r\sqrt{2Er^2 + 2Kr - L^2}}.$$

The integral (4.6.40) can be evaluated by completing the square for  $2Er^2 + 2Kr - L^2$ . The integral (4.6.41) can also be evaluated, but rather than tackling this directly, we instead produce a differential equation for  $u$ , defined by

$$(4.6.42) \quad u = \frac{1}{r}.$$

By the chain rule,

$$(4.6.43) \quad \frac{dr}{dt} = -r^2 \frac{du}{dt} = -r^2 \frac{du}{d\theta} \frac{d\theta}{dt} = -L \frac{du}{d\theta},$$

the last identity by (4.6.21). Taking another  $t$ -derivative gives

$$(4.6.44) \quad \frac{d^2r}{dt^2} = -L \frac{d}{dt} \frac{du}{d\theta} = -L \frac{d^2u}{d\theta^2} \frac{d\theta}{dt} = -L^2 u^2 \frac{d^2u}{d\theta^2},$$

again using (4.6.21). Comparing this with (4.6.38), we get

$$(4.6.45) \quad -L^2 u^2 \frac{d^2u}{d\theta^2} = L^2 u^3 - K u^2,$$

or equivalently

$$(4.6.46) \quad \frac{d^2u}{d\theta^2} + u = \frac{K}{L^2}.$$

Miraculously, we have obtained a linear equation! The general solution to (4.6.46) is

$$(4.6.47) \quad u(\theta) = A \cos(\theta - \theta_0) + \frac{K}{L^2},$$

which by (4.6.42) gives

$$(4.6.48) \quad r \left[ A \cos(\theta - \theta_0) + \frac{K}{L^2} \right] = 1.$$

This is equivalent to

$$(4.6.49) \quad r [1 + e \cos(\theta - \theta_0)] = p, \quad p = \frac{L^2}{K}, \quad e = A \frac{L^2}{K}.$$

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 (4.6.24). To establish the third law, recall from

(4.6.23) that  $L/2$  is the rate at which such area is swept out, so the period  $T$  of the orbit satisfies

$$(4.6.50) \quad \begin{aligned} \frac{L}{2}T &= \text{area enclosed by the ellipse} \\ &= \pi ab, \end{aligned}$$

where  $a$  is the semi-major axis and  $b$  the semi-minor axis. For an ellipse given by (4.6.49), we have

$$(4.6.51) \quad a = \frac{p}{1 - e^2}, \quad b = \frac{p}{\sqrt{1 - e^2}} = p^{1/2}a^{1/2}$$

(cf. (4.6.58)–(4.6.59) below), which yields

$$(4.6.52) \quad T = \frac{2\pi ab}{L} = 2\pi \frac{\sqrt{p}}{L} a^{3/2} = \frac{2\pi}{\sqrt{K}} a^{3/2}.$$

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 [1] and [15], 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 [3] and also in Chapter 18 of [45]. While a derivation is well outside the scope of this book, we mention that the relativistic treatment leads to the following variant of (4.6.46):

$$(4.6.53) \quad \frac{d^2u}{d\theta^2} + u = A + \varepsilon u^2,$$

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

$$(4.6.54) \quad \frac{du}{d\theta} = v, \quad \frac{dv}{d\theta} = -u + A + \varepsilon u^2.$$

In analogy with (4.6.26)–(4.6.29), we can form

$$(4.6.55) \quad F(u, v) = \frac{1}{2}v^2 + \frac{1}{2}u^2 - Au - \frac{\varepsilon}{3}u^3,$$

and check that if  $(u(\theta), v(\theta))$  solves (4.6.54), then

$$(4.6.56) \quad \frac{d}{d\theta} F(u, v) = 0,$$

so the orbits for (4.6.54) 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 (4.6.54) has some solutions periodic in  $\theta$ . However, the period is generally not equal to  $2\pi$ . (See Appendix 4.E 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 (4.6.46), all the solutions in (4.6.47) are periodic in  $\theta$  of period  $2\pi$ .

### 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 (4.6.42), and generalize the calculations (4.6.43)–(4.6.46) to obtain a differential equation for  $u$  as a function of  $\theta$ , 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

$$(4.6.57) \quad r(1 + e \cos \theta) = p$$

is the equation in polar coordinates of an ellipse.

(a) Show that (4.6.57) describes a closed, bounded curve, since  $1 + e \cos \theta > 0$  for all  $\theta$  if  $0 < e < 1$ , and  $\cos \theta$  is periodic in  $\theta$  of period  $2\pi$ . 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

$$(4.6.58) \quad a = \frac{p}{1 - e^2}.$$

(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

$$(4.6.59) \quad \frac{(x + ea)^2}{a^2} + \frac{y^2}{b^2} = 1, \quad b^2 = (1 - e^2)a^2.$$

*Hint.* Use (4.6.57) and its square to write

$$(4.6.60) \quad \begin{aligned} r + ex &= p, \quad \text{so } r = p - ex, \\ x^2 + y^2 + 2erx + e^2x^2 &= p^2, \end{aligned}$$

hence  $x^2 + y^2 + 2e(p - ex)x + e^2x^2 = p^2$ , or equivalently

$$(4.6.61) \quad (1 - e^2)x^2 + 2epx + y^2 = p^2,$$

and proceed to derive (4.6.59), taking into account (4.6.58).

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

$$(4.6.62) \quad a = 1.496 \times 10^{11} \text{ m},$$

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

$$(4.6.63) \quad T = 31.536 \times 10^6 \text{ sec}.$$

The gravitational constant  $G$  has been measured as

$$(4.6.64) \quad G = 6.674 \times 10^{-11} \text{ m}^3/(\text{kg sec}^2).$$

With this information, use (4.6.36) and (4.6.52) to calculate the mass  $m_2$  of the sun. Assume the mass of the earth is negligible compared to  $m_2$ . You should get

$$(4.6.65) \quad m_2 = \alpha \times 10^{30} \text{ kg},$$

with  $\alpha$  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 [17] 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

$$(4.6.66) \quad -Gmm_e r^{-2},$$

where  $G$  is given in Exercise 5,

$$(4.6.67) \quad r = 6.38 \times 10^6 \text{ m}$$

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  $9.8 \text{ m/sec}^2$ , so we have

$$(4.6.68) \quad 9.8 \text{ m/sec}^2 = Gm_e r^{-2}.$$

Use this to compute  $m_e$ . You should get

$$(4.6.69) \quad m_e = \beta \times 10^{24} \text{ kg},$$

with  $\beta$  between 1 and 10.

REMARK. See Appendix 4.F for more on (4.6.66).

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  $1.07 \times 10^9$  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

$$m_J \approx 318 m_e.$$



### 4.7. 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  $u : [a, b] \rightarrow \Omega$ , consider

$$(4.7.1) \quad I(u) = \int_a^b L(u(t), u'(t)) dt.$$

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

$$(4.7.2) \quad \left. \frac{d}{ds} I(u_s) \right|_{s=0} = 0,$$

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

$$(4.7.3) \quad \left. \frac{\partial}{\partial s} u_s(t) \right|_{s=0} = w(t),$$

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

$$(4.7.4) \quad L_{x_k} = \frac{\partial L}{\partial x_k}, \quad L_{v_k} = \frac{\partial L}{\partial v_k}.$$

Then

$$(4.7.5) \quad \begin{aligned} \left. \frac{d}{ds} I(u_s) \right|_{s=0} &= \int_a^b \sum_k L_{x_k}(u(t), u'(t)) w_k(t) dt \\ &+ \int_a^b \sum_k L_{v_k}(u(t), u'(t)) w'_k(t) dt. \end{aligned}$$

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

$$(4.7.6) \quad \left. \frac{d}{ds} I(u_s) \right|_{s=0} = \int_a^b \sum_k \left[ L_{x_k}(u(t), u'(t)) - \frac{d}{dt} L_{v_k}(u(t), u'(t)) \right] w_k(t) dt.$$

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

$$(4.7.7) \quad \frac{d}{dt} L_{v_k}(u(t), u'(t)) - L_{x_k}(u(t), u'(t)) = 0, \quad \forall k.$$

This system is called the Lagrange equation for stationarity of (4.7.1). Applying the chain rule to the first sum, we can expand this out as

$$(4.7.8) \quad \begin{aligned} \sum_{\ell} L_{v_k v_{\ell}}(u(t), u'(t)) u'_{\ell}(t) + \sum_{\ell} L_{v_k x_{\ell}}(u(t), u'(t)) u'_{\ell}(t) \\ - L_{x_k}(u(t), u'(t)) = 0, \quad \forall k. \end{aligned}$$

This can be converted to a first order system for  $(u(t), u'(t))$ , to which the results of §4.1 apply, provided the  $n \times n$  matrix

$$(4.7.9) \quad \left( L_{v_k v_{\ell}}(x, v) \right)$$

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

$$(4.7.10) \quad T = \frac{1}{2}m\|v\|^2, \quad \text{and} \quad V = V(x),$$

respectively. The Lagrangian  $L(x, v)$  is given by the *difference*:

$$(4.7.11) \quad L(x, v) = T - V = \frac{1}{2}m\|v\|^2 - V(x).$$

In such a case,

$$(4.7.12) \quad L_{v_k}(x, v) = mv_k, \quad L_{x_k}(x, v) = -\frac{\partial V}{\partial x_k},$$

and the Lagrange system (4.7.7) becomes the standard Newtonian system

$$(4.7.13) \quad m\frac{d^2u}{dt^2} = -\nabla V(u).$$

In this setting, the integral (4.7.1) is called the *action*. The assertion that the laws of motion are given by the stationary condition for (4.7.1) where  $L$  is the Lagrangian (4.7.11) 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 §1.6 of Chapter 1. As there, we have a rigid rod, of length  $\ell$ , suspended at one end. We assume the rod has negligible mass, except for an object of mass  $m$  at the other end. See Figure 4.7.1. The rod makes an angle  $\theta$  with the downward vertical. We seek a differential equation for  $\theta$  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

$$(4.7.14) \quad z(t) = \ell e^{i\theta(t)}.$$

The velocity is

$$(4.7.15) \quad v(t) = z'(t) = i\ell\theta'(t)e^{i\theta(t)},$$

so the kinetic energy is

$$(4.7.16) \quad T = \frac{1}{2}m\|v(t)\|^2 = \frac{m\ell^2}{2}\theta'(t)^2.$$

Meanwhile the potential energy, due to the force of gravity, is

$$(4.7.17) \quad V = -mg\ell \cos \theta.$$

Taking  $\psi = \theta'$ , we have the Lagrangian

$$(4.7.18) \quad L(\theta, \psi) = \frac{m\ell^2}{2}\psi^2 + mg\ell \cos \theta,$$

$$L_\psi(\theta, \psi) = m\ell^2\psi, \quad L_\theta(\theta, \psi) = -mg\ell \sin \theta,$$

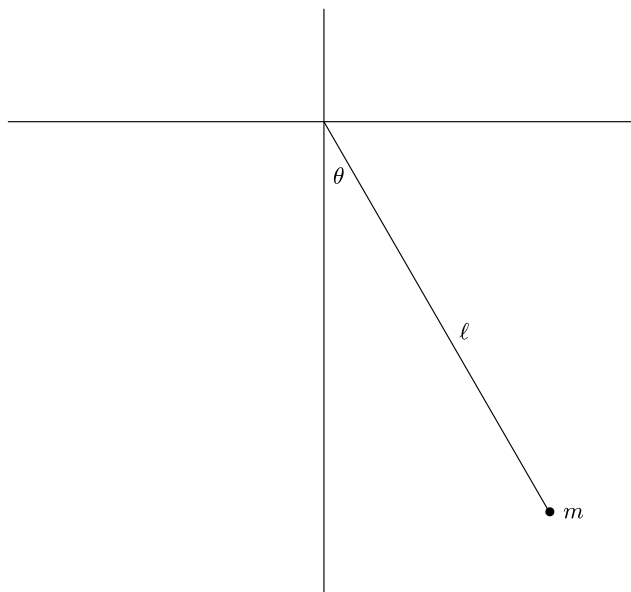


Figure 4.7.1. Pendulum

and Lagrange's equation

$$(4.7.19) \quad \frac{d}{dt} L_{\psi}(\theta(t), \theta'(t)) - L_{\theta}(\theta(t), \theta'(t)) = 0$$

yields the pendulum equation

$$(4.7.20) \quad \frac{d^2\theta}{dt^2} + \frac{g}{\ell} \sin \theta = 0,$$

in agreement with (4.6.6) of Chapter 1.

The approach above avoided a computation of the force acting on the pendulum (cf. (1.6.6) 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 §4.9.

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 (4.7.1). We look for equations for

$$(4.7.21) \quad u : [a, b] \longrightarrow M,$$

satisfying the stationary condition (4.7.2), 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

$$(4.7.22) \quad u_s : [a, b] \longrightarrow M.$$

Again we take  $w(t)$  as in (4.7.3), and this time we obtain an arbitrary smooth function  $w : [a, b] \rightarrow \mathbb{R}^n$ , satisfying  $w(a) = w(b) = 0$ , and the additional constraint

$$(4.7.23) \quad w(t) \cdot n(u(t)) \equiv 0.$$

The calculations (4.7.4)–(4.7.6) still apply, but from here we get a conclusion different from (4.7.7). Since (4.7.6) holds for all  $w(t)$  described as just above, the conclusion is

$$(4.7.24) \quad \frac{d}{dt} L_v(u(t), u'(t)) - L_x(u(t), u'(t)) \text{ is parallel to } n(u(t)),$$

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 (4.7.24) is

$$(4.7.25) \quad \left[ \frac{d}{dt} L_v(u(t), u'(t)) - L_x(u(t), u'(t)) \right] \times n(u(t)) = 0.$$

Let's specialize this constrained variational problem to the case

$$(4.7.26) \quad L(x, v) = \frac{1}{2} \|v\|^2.$$

The associated integral

$$(4.7.27) \quad E(u) = \frac{1}{2} \int_a^b \|u'(t)\|^2 dt$$

is called the *energy* of  $u : [a, b] \rightarrow M$ . In this case,  $L_v = v$  and  $L_x = 0$ , so (4.7.24) becomes

$$(4.7.28) \quad u''(t) \text{ is parallel to } n(u(t)).$$

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 (4.7.28) yields

$$(4.7.29) \quad u''(t) = n(u(t)) \cdot u''(t)n(u(t)).$$

An equation with a better form can be obtained by differentiating

$$(4.7.30) \quad u'(t) \cdot n(u(t)) \equiv 0,$$

to get

$$(4.7.31) \quad u'' \cdot n(u(t)) = -u'(t) \cdot \frac{d}{dt} n(u(t)).$$

Plugging this into the right side of (4.7.29) gives the differential equation

$$(4.7.32) \quad u''(t) + u'(t) \cdot \left( \frac{d}{dt} n(u(t)) \right) n(u(t)) = 0.$$

Note by (4.7.28) that  $u''$  is orthogonal to  $u'(t)$ , so

$$(4.7.33) \quad \frac{d}{dt} \|u'(t)\|^2 = 2u'(t) \cdot u''(t) \equiv 0.$$

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:

$$(4.7.34) \quad \ell(u) = \int_a^b \|u'(t)\| dt.$$

We will say a bit more about geodesics in Appendix 4.H. Further material can be found in Chapter 6 of [50], and also in texts on elementary differential geometry, such as [11].

We next present another approach to finding equations for stationary paths of (4.7.27). 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

$$(4.7.35) \quad u(t) = (x(t), \varphi(x(t))),$$

and

$$(4.7.36) \quad u'(t) = (x'(t), \nabla\varphi(x(t)) \cdot x'(t)),$$

so

$$(4.7.37) \quad \begin{aligned} \|u'(t)\|^2 &= \|x'(t)\|^2 + (\nabla\varphi(x(t)) \cdot x'(t))^2 \\ &= x'(t) \cdot G(x(t))x'(t), \end{aligned}$$

where

$$(4.7.38) \quad G(x) = \begin{pmatrix} 1 + \varphi_1(x)^2 & \varphi_1(x)\varphi_2(x) \\ \varphi_1(x)\varphi_2(x) & 1 + \varphi_2(x)^2 \end{pmatrix}, \quad \varphi_j(x) = \frac{\partial\varphi}{\partial x_j}.$$

Thus the problem of finding a constrained stationary path  $u(t)$  for the energy (4.7.27) is equivalent to the problem of finding an unconstrained stationary path  $x(t)$  for

$$(4.7.39) \quad \mathcal{E}(x) = \frac{1}{2} \int_a^b x'(t) \cdot G(x(t))x'(t) dt.$$

In this case,

$$(4.7.40) \quad \begin{aligned} L(x, v) &= \frac{1}{2}v \cdot G(x)v, \\ L_v(x, v) &= G(x)v, \quad \text{and,} \\ L_x(x, v) &= \frac{1}{2}v \cdot \nabla G(x)v, \end{aligned}$$

where the last identity means

$$(4.7.41) \quad L_{x_k}(x, v) = \frac{1}{2}v \cdot \frac{\partial G}{\partial x_k}v.$$

In this setting, the Lagrange equation (4.7.7) becomes

$$(4.7.42) \quad \frac{d}{dt} \left[ G(x(t))x'(t) \right] - \frac{1}{2}x'(t) \cdot \nabla G(x(t))x'(t) = 0,$$

i.e.,

$$(4.7.43) \quad \frac{d}{dt} \sum_j G_{kj}(x(t))x'_j(t) - \frac{1}{2} \sum_{i,j} x'_i(t) \frac{\partial G_{ij}}{\partial x_k} x'_j(t) = 0, \quad \forall k.$$

---

**Exercises**

1. Given a Lagrangian  $L(x, v)$ , we define the “energy”

$$(4.7.44) \quad \begin{aligned} E(x, v) &= L_v(x, v) \cdot v - L(x, v) \\ &= \sum_k L_{v_k}(x, v)v_k - L(x, v). \end{aligned}$$

Show that if  $u(t)$  solves the Lagrange equation (4.7.7), then

$$(4.7.45) \quad \frac{d}{dt}E(u(t), u'(t)) \equiv 0.$$

This is energy conservation, in this setting.

2. Suppose

$$(4.7.46) \quad L(x, v) = \frac{m}{2}v \cdot G(x)v - V(x),$$

Assume  $G(x) \in M(n, \mathbb{R})$  is symmetric and invertible, and define  $E(x, v)$  as in (4.7.44). Show that

$$(4.7.47) \quad E(x, v) = \frac{m}{2}v \cdot G(x)v + V(x).$$

3. Let  $L(x, v)$  be given by (4.7.46). Show that the Lagrange equation (4.7.7) is

$$(4.7.48) \quad m \frac{d}{dt} [G(u(t))u'(t)] - \frac{m}{2}u'(t) \cdot \nabla G(u(t))u'(t) = -\nabla V(u(t)),$$

where the second term is evaluated as in (4.7.42)–(4.7.43). Show in turn that this yields the first order system

$$\begin{aligned} \frac{du_k}{dt} &= v_k \\ m \sum_j G_{kj}(u(t)) \frac{dv_j}{dt} + m \sum_{i,j} v_i(t) \left[ \frac{\partial G_{kj}}{\partial x_i} - \frac{1}{2} \frac{\partial G_{ij}}{\partial x_k} \right] v_j(t) &= -\frac{\partial V}{\partial x_k}(u(t)). \end{aligned}$$

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 (4.7.21)–(4.7.24), and consider the following generalization of (4.7.26):

$$(4.7.49) \quad L(x, v) = \frac{m}{2}\|v\|^2 - V(x).$$

Establish the following replacement for (4.7.32):

$$(4.7.50) \quad mu''(t) + mu'(t) \cdot \left( \frac{d}{dt}n(u(t)) \right) n(u(t)) = -P_M(u(t))\nabla V(u(t)),$$

where, for  $x \in M$ ,  $w \in \mathbb{R}^n$ ,

$$(4.7.51) \quad P_M(x)w = w - (n(x) \cdot w)n(x).$$

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

$$(4.7.52) \quad M = \{x \in \mathbb{R}^3 : \|x\| = \ell\},$$

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, (4.7.50) produces, for

$$(4.7.53) \quad u(t) = \ell\omega(t),$$

the system

$$(4.7.54) \quad \omega''(t) + \|\omega'(t)\|^2\omega(t) = -\frac{g}{\ell}k + \frac{g}{\ell}(\omega(t) \cdot k)\omega(t).$$

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

$$(4.7.55) \quad \omega(t) = (\sin \theta(t), -\cos \theta(t))^t, \quad k = (0, 1)^t,$$

the equation (4.7.54) leads to the (planar) pendulum equation

$$(4.7.56) \quad \theta''(t) + \frac{g}{\ell} \sin \theta(t) = 0.$$

7. Let us return to the setting of Exercise 2, and set

$$(4.7.57) \quad p = L_v(x, v) = mG(x)v.$$

Also set

$$(4.7.58) \quad \mathcal{E}(x, p) = E(x, v) = E(x, G(x)^{-1}p/m).$$

Show that

$$(4.7.59) \quad \mathcal{E}(x, p) = \frac{1}{2m}p \cdot G(x)^{-1}p + V(x).$$

Show that the Lagrange equation (4.7.48) for  $u(t) = x(t)$  is equivalent to the following Hamiltonian system:

$$(4.7.60) \quad \frac{dx_k}{dt} = \frac{\partial \mathcal{E}}{\partial p_k}, \quad \frac{dp_k}{dt} = -\frac{\partial \mathcal{E}}{\partial x_k}.$$

*Hint.* To get started on (4.7.60), note that if (4.7.59) holds, then

$$(4.7.61) \quad \frac{\partial \mathcal{E}}{\partial p} = \frac{1}{m}G(x)^{-1}p = v,$$

and that the Lagrange equation implies

$$(4.7.62) \quad \frac{dp_k}{dt} = L_{x_k}(x, v) = \frac{m}{2}v \cdot \frac{\partial G}{\partial x_k}(x)v - \frac{\partial V}{\partial x_k}(x).$$

Furthermore, as in (3.8.13) of Chapter 3,

$$(4.7.63) \quad \frac{\partial}{\partial x_k}G(x)^{-1} = -G(x)^{-1}\frac{\partial G}{\partial x_k}(x)G(x)^{-1}.$$

*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 [1], [5], and Chapter 1 of [45].

Exercises 8–11 study surfaces of revolution that are surfaces of “least area.” To set this up, let  $u : [0, 1] \rightarrow (0, \infty)$  be smooth, and rotate the graph of  $y = u(x)$  about the  $x$ -axis in  $(x, y, z)$ -space. Elementary calculus gives the formula

$$(4.7.64) \quad A(u) = 2\pi \int_0^1 u(t) \sqrt{1 + u'(t)^2} dt$$

for the area of the resulting surface of revolution. The problem is to find  $u$  for which the area is minimal, given constraints

$$(4.7.65) \quad u(0) = \alpha, \quad u(1) = \beta, \quad \alpha, \beta > 0.$$

8. In (4.7.64),  $L(x, v) = x\sqrt{1 + v^2}$ . Show that the “energy”  $E(x, v)$  in (4.7.44) is given by

$$(4.7.66) \quad E(x, v) = -\frac{x}{\sqrt{1 + v^2}}.$$

9. Using (4.7.45), show that if  $u(t)$  solves the Lagrange equation (4.7.7) in this setting, then there is a constant  $a$  such that

$$(4.7.67) \quad \frac{u(t)}{\sqrt{1 + u'(t)^2}} = a,$$

hence

$$(4.7.68) \quad \frac{du}{dt} = \pm \sqrt{b^2 u^2 - 1}, \quad b = \frac{1}{a}.$$

10. Separate variables in (4.7.68) and use the substitution  $bu = \cosh v$  to evaluate the  $u$ -integral and conclude that

$$(4.7.69) \quad u(t) = \frac{1}{b} \cosh(bt + c),$$

for some constant  $c$ . Equation (4.7.69) is the equation of a catenary, seen before in (1.3.24) of Chapter 1, for the hanging cable.

11. Consider the problem of finding  $b$  and  $c$  in (4.7.69) such that the constraints (4.7.65) 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

$$(4.7.70) \quad V(u) = \int_{-A}^A u(t) \sqrt{1 + u'(t)^2} dt,$$



subject to the boundary conditions

$$(4.7.71) \quad u(-A) = u(A) = 0,$$

and the constraint that the curve  $y = u(x)$ ,  $-A \leq x \leq A$ , have length  $L$ ,

$$(4.7.72) \quad \ell(u) = \int_{-A}^A \sqrt{1 + u'(t)^2} dt = L.$$

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 (4.7.72), we bring in the Lagrange multiplier method. That is, we set

$$(4.7.73) \quad I_\lambda(u) = V(u) + \lambda \ell(u),$$

find the stationary path for (4.7.73) (subject to (4.7.71)) as a function of  $\lambda$ , and then find for which  $\lambda$  the constraint (4.7.72) holds. Note that  $I_\lambda(u)$  has the form (4.7.1) with

$$(4.7.74) \quad L_\lambda(x, v) = (x + \lambda)\sqrt{1 + v^2}.$$

12. Show that the “energy”  $E_\lambda(x, v)$  in (4.7.44) is given by

$$(4.7.75) \quad E_\lambda(x, v) = \frac{x + \lambda}{\sqrt{1 + v^2}}.$$

13. Using (4.7.45), show that if  $u(t)$  solves the Lagrange equation (4.7.7) in this setting, then there exists a constant  $a$  (maybe depending on  $\lambda$ ) such that

$$(4.7.76) \quad \frac{u(t) + \lambda}{\sqrt{1 + u'(t)^2}} = a,$$

hence

$$(4.7.77) \quad \frac{du}{dt} = \pm \sqrt{b^2(u + \lambda)^2 - 1}, \quad b = \frac{1}{a}.$$

14. Separate variables in (4.7.77) 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 (4.7.71) forces  $c = 0$ , so

$$(4.7.78) \quad u(t) = -\lambda + \frac{1}{b} \cosh bt.$$

15. Calculate the length of the curve  $y = u(x)$ ,  $-A \leq x \leq A$ , when  $u$  is given by (4.7.78), and show that the constraints (4.7.71)–(4.7.72) yield the equations

$$(4.7.79) \quad \sinh bA = \frac{bL}{2}, \quad \lambda = \frac{1}{b} \cosh bA.$$

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 Figure 4.7.1. Instead of

assuming all the mass is at the end of the rod, assume the rod has a mass distribution  $m(s) ds$ ,  $0 \leq s \leq \ell$ , so the total mass is  $m = \int_0^\ell m(s) ds$ . Show that for the potential energy  $V$  you replace (4.7.17) by

$$(4.7.80) \quad V = -m_a g \ell \cos \theta, \quad m_a = \int_0^\ell m(s) \frac{s}{\ell} ds,$$

and for the kinetic energy  $T$ , you replace (4.7.16) by

$$(4.7.81) \quad T = \frac{m_b \ell^2}{2} \theta'(t)^2, \quad m_b = \int_0^\ell m(s) \left(\frac{s}{\ell}\right)^2 ds.$$

Write down the replacement for the pendulum equation (4.7.20) in this setting. Specialize the calculation to the case

$$(4.7.82) \quad m(s) = \frac{m}{\ell}, \quad 0 \leq s \leq \ell,$$

which represents a rod with uniform mass distribution.

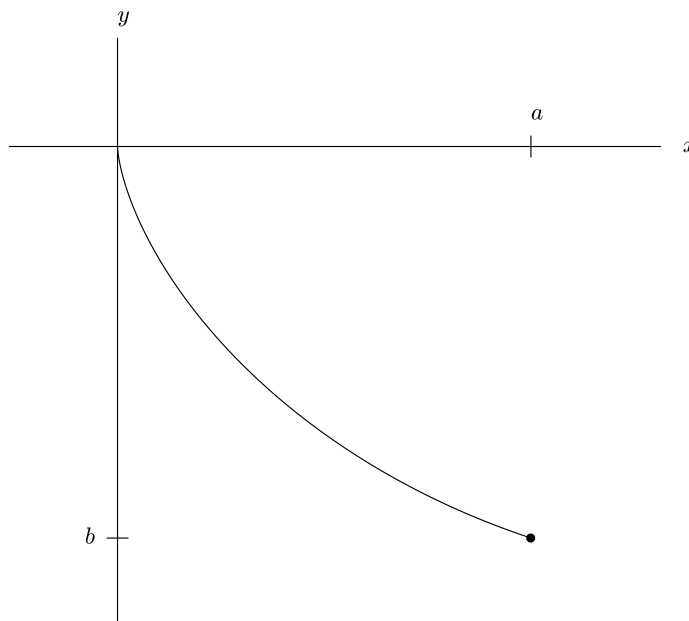


Figure 4.8.1. Brachistochrone problem

#### 4.8. The brachistochrone problem

The early masters of calculus enjoyed posing challenging problems to each other. The most famous of these is called the *brachistochrone 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 Figure 4.8.1. 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 §4.7. (In fact, this problem helped spark the *creation* of the variational method.) First, let  $\varphi : [0, a] \rightarrow \mathbb{R}$  with  $\varphi(0) = 0$ ,  $\varphi(a) = b$ , and consider the constrained motion of a particle,

$$(4.8.1) \quad u : [0, t_0] \longrightarrow M = \{(x, \varphi(x)) : 0 \leq x \leq a\},$$

under the force of gravity. Thus, in place of (4.7.27), we look for stationary paths for

$$(4.8.2) \quad I(u) = \int_0^a \left[ \frac{m}{2} \|u'(t)\|^2 - V(u(t)) \right] dt,$$

subject to the constraint (4.8.1), and with

$$(4.8.3) \quad V(x, y) = mgy.$$

We can convert this to an unconstrained variational problem as was done in (4.7.35)–(4.7.42), now with a nonzero  $V$ , and with lower dimension. We have

$$(4.8.4) \quad u(t) = (x(t), \varphi(x(t))),$$

and

$$(4.8.5) \quad \|u'(t)\|^2 = (1 + \varphi'(x(t))^2)x'(t)^2,$$

so the problem of finding a constrained stationary path  $u(t)$  for (4.8.2) is equivalent to the problem of finding an unconstrained stationary path  $x(t)$  for

$$(4.8.6) \quad J(x) = \int_0^a L(x(t), x'(t)) dt,$$

with

$$(4.8.7) \quad L(x, v) = \frac{m}{2}(1 + \varphi'(x)^2)v^2 - mg\varphi(x).$$

The path  $x(t)$  is governed by the differential equation

$$(4.8.8) \quad \frac{d}{dt}L_v(x(t), x'(t)) - L_x(x(t), x'(t)) = 0.$$

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:

$$(4.8.9) \quad E(x, v) = \frac{m}{2}(1 + \varphi'(x)^2)v^2 + mg\varphi(x),$$

that is, for a solution to (4.8.8),

$$(4.8.10) \quad \frac{m}{2}(1 + \varphi'(x(t))^2)x'(t)^2 + mg\varphi(x(t)) = E$$

is constant. In the current set-up,  $x(0) = 0$  and  $x'(0) = 0$ , so  $E = 0$ . We get

$$(4.8.11) \quad \frac{dx}{dt} = \pm \sqrt{\frac{-2g\varphi(x)}{1 + \varphi'(x)^2}},$$

which separates to

$$(4.8.12) \quad \frac{1}{\sqrt{2g}} \int_0^a \sqrt{\frac{1 + \varphi'(x)^2}{-\varphi(x)}} dx = \int_0^{t_0} dt.$$

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 (4.8.12).

Hence the brachistochrone problem is reduced to the problem of finding  $\varphi : [0, a] \rightarrow \mathbb{R}$ , minimizing

$$(4.8.13) \quad K(\varphi) = \int_0^a \mathcal{L}(\varphi(x), \varphi'(x)) dx,$$

subject to the condition

$$(4.8.14) \quad \varphi(0) = 0, \quad \varphi(a) = b,$$

where

$$(4.8.15) \quad \mathcal{L}(\varphi, \psi) = \sqrt{\frac{1 + \psi^2}{-\varphi}}.$$

Stationary paths for (4.8.13) satisfy the Lagrange equation

$$(4.8.16) \quad \frac{d}{dt} \mathcal{L}_\psi(\varphi(x), \varphi'(x)) - \mathcal{L}_\varphi(\varphi(x), \varphi'(x)) = 0.$$

Note that

$$(4.8.17) \quad \mathcal{L}_\psi(\varphi, \psi) = \frac{\psi}{\sqrt{-\varphi(1 + \psi^2)}}, \quad \mathcal{L}_\varphi(\varphi, \psi) = -\frac{1}{2} \frac{\sqrt{-\varphi(1 + \psi^2)}}{\varphi^2}.$$

Solutions to (4.8.16) have the property that

$$(4.8.18) \quad \mathcal{E}(\varphi(t), \varphi'(t)) = E$$

is constant, where (parallel to (4.7.44))

$$(4.8.19) \quad \mathcal{E}(\varphi, \psi) = \mathcal{L}_\psi(\varphi, \psi) - \mathcal{L}(\varphi, \psi).$$

Using (4.8.15) and (4.8.17), we have

$$(4.8.20) \quad \begin{aligned} \mathcal{E}(\varphi, \psi) &= \frac{\psi^2}{\sqrt{-\varphi(1 + \psi^2)}} - \sqrt{\frac{1 + \psi^2}{-\varphi}} \\ &= -\frac{1}{\sqrt{-\varphi(1 + \psi^2)}}. \end{aligned}$$

Thus, if  $\varphi(x)$  satisfies (4.8.16), then

$$(4.8.21) \quad \varphi(x)(1 + \varphi'(x)^2) = -k^2, \quad \text{const.},$$

where we have written the constant as  $-k^2$  to enforce the condition that  $\varphi(x) < 0$  for  $0 < x \leq a$ . For notational convenience, we make the change of variable

$$(4.8.22) \quad y(x) = -\varphi(x),$$

so (4.8.21) becomes

$$(4.8.23) \quad y(x)(1 + y'(x)^2) = k^2,$$

giving

$$(4.8.24) \quad \frac{dy}{dx} = \sqrt{\frac{k^2}{y} - 1}.$$

The equation (4.8.24) separates to

$$(4.8.25) \quad \int \frac{dy}{\sqrt{\frac{k^2}{y} - 1}} = \int dx.$$

The left integral has the form of (1.5.15) in Chapter 1, with  $E_0 = -1$ ,  $Km = k^2$ . Rather than recall the formulas (1.5.16)–(1.5.22) of Chapter 1, we implement the method previewed in Exercise 3 of that section. We use the change of variable

$$(4.8.26) \quad y = k^2 \sin^2 \tau, \quad 2\tau = \theta.$$

Then

$$(4.8.27) \quad dy = 2k^2 \sin \tau \cos \tau d\tau, \quad \sqrt{\frac{k^2}{y} - 1} = \frac{\cos \tau}{\sin \tau},$$

so

$$(4.8.28) \quad \begin{aligned} \int \frac{dy}{\sqrt{\frac{k^2}{y} - 1}} &= 2k^2 \int \sin^2 \tau d\tau \\ &= \frac{k^2}{2} \int (1 - \cos \theta) d\theta \\ &= \frac{k^2}{2} (\theta - \sin \theta), \end{aligned}$$

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

$$(4.8.29) \quad \begin{aligned} x = x(\theta) &= \frac{k^2}{2} (\theta - \sin \theta), \\ y = y(\theta) &= \frac{k^2}{2} (1 - \cos \theta). \end{aligned}$$

The choice of  $k^2 > 0$  is dictated by the implication

$$(4.8.30) \quad 0 < \theta < \pi k^2, \quad \frac{k^2}{2} (\theta - \sin \theta) = a \implies \frac{k^2}{2} (1 - \cos \theta) = |b|.$$

This solves the brachistochrone problem. The curve defined by (4.8.29) is known as a *cycloid*. See Figure 4.8.2. 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 (4.8.29), for some  $\theta \in (0, \pi k^2)$ .

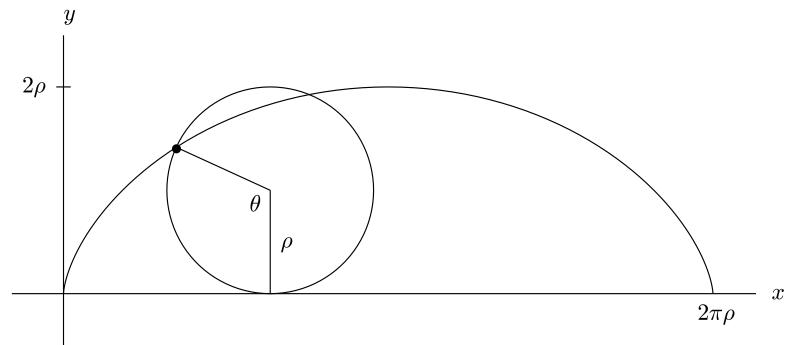
*Hint.* Consult Figure 4.8.2.

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 (4.8.29), set  $\varphi(\theta) = -y(\theta)$ . Let

$$(4.8.31) \quad \theta_1 = \frac{k^2}{2} \pi, \quad \theta_0 \in [0, \theta_1].$$

Show that the time it takes a particle starting at rest at  $(x(\theta_0), \varphi(\theta_0))$  to slide down the curve  $(x(\theta), \varphi(\theta))$ ,  $\theta_0 \leq \theta \leq \theta_1$ , to the point  $(x(\theta_1), \varphi(\theta_1))$  (the bottom of



**Figure 4.8.2.** Cycloid

the cycloid) is independent of  $\theta_0$ . One says the cycloid also solves the *tautochrone problem*.

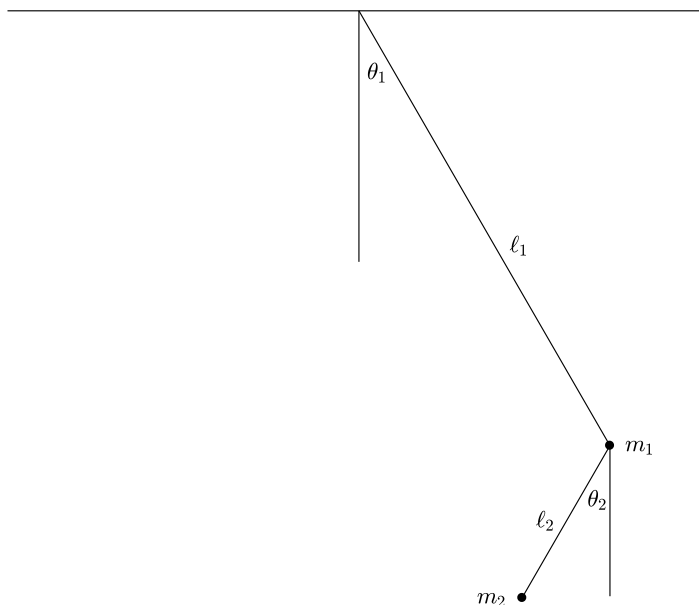


Figure 4.9.1. Double pendulum

#### 4.9. The double pendulum

Here we study the motion of a double pendulum, such as illustrated in Figure 4.9.1. We have a pair of rigid rods, of lengths  $\ell_1$  and  $\ell_2$ , 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

$$(4.9.1) \quad z_1(t) = \ell_1 e^{i\theta_1(t)},$$

and the position of mass 2 is

$$(4.9.2) \quad z_2(t) = z_1(t) + \ell_2 e^{i\theta_2(t)}.$$

Their velocities are

$$(4.9.3) \quad \begin{aligned} z_1' &= i\ell_1 \theta_1' e^{i\theta_1}, \\ z_2' &= i\ell_1 \theta_1' e^{i\theta_1} + i\ell_2 \theta_2' e^{i\theta_2}, \end{aligned}$$



with square norms

$$(4.9.4) \quad \begin{aligned} |z'_1|^2 &= \ell_1^2 (\theta'_1)^2, \\ |z'_2|^2 &= (\ell_1 \theta'_1 e^{i\theta_1} + \ell_2 \theta'_2 e^{i\theta_2})(\ell_1 \theta'_1 e^{-i\theta_1} + \ell_2 \theta'_2 e^{-i\theta_2}) \\ &= \ell_1^2 (\theta'_1)^2 + \ell_2^2 (\theta'_2)^2 + 2\ell_1 \ell_2 \theta'_1 \theta'_2 \cos(\theta_1 - \theta_2). \end{aligned}$$

The potential energy of this system is given by

$$(4.9.5) \quad \begin{aligned} V &= -m_1 g \operatorname{Re} z_1(t) - m_2 g \operatorname{Re} z_2(t) \\ &= -m_1 g \ell_1 \cos \theta_1 - m_2 g (\ell_1 \cos \theta_1 + \ell_2 \cos \theta_2), \end{aligned}$$

and the kinetic energy by

$$(4.9.6) \quad T = \frac{m_1}{2} |z'_1(t)|^2 + \frac{m_2}{2} |z'_2(t)|^2.$$

If we write

$$(4.9.7) \quad \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \quad \psi = \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix} = \begin{pmatrix} \theta'_1 \\ \theta'_2 \end{pmatrix},$$

then (4.9.4) gives

$$(4.9.8) \quad T = \frac{1}{2} \psi \cdot G(\theta) \psi,$$

with

$$(4.9.9) \quad G(\theta) = \begin{pmatrix} (m_1 + m_2)\ell_1^2 & m_2 \ell_1 \ell_2 \cos(\theta_1 - \theta_2) \\ m_2 \ell_1 \ell_2 \cos(\theta_1 - \theta_2) & m_2 \ell_2^2 \end{pmatrix}.$$

Thus the Lagrangian  $L = T - V$  is given by

$$(4.9.10) \quad L(\theta, \psi) = \frac{1}{2} \psi \cdot G(\theta) \psi - V(\theta),$$

with  $V(\theta)$  as in (4.9.5), and the equation of motion for the double pendulum is

$$(4.9.11) \quad \frac{d}{dt} L_{\psi}(\theta, \theta') - L_{\theta}(\theta, \theta') = 0.$$

As in (4.7.48), this expands out to the 2 by 2 system

$$(4.9.12) \quad \frac{d}{dt} \sum_j G_{kj}(\theta(t)) \theta'_j(t) - \frac{1}{2} \sum_{i,j} \theta'_i(t) \frac{\partial G_{ij}}{\partial \theta_k} \theta'_j(t) = -\frac{\partial V}{\partial \theta_k}(\theta(t)),$$

for  $k = 1, 2$ . Making explicit use of (4.9.5) and (4.9.9), we have

$$(4.9.13) \quad \begin{aligned} L_{\psi_1}(\theta, \psi) &= (m_1 + m_2)\ell_1^2 \psi_1 + m_2 \ell_1 \ell_2 \psi_2 \cos(\theta_1 - \theta_2), \\ L_{\psi_2}(\theta, \psi) &= m_2 \ell_2^2 \psi_2 + m_2 \ell_1 \ell_2 \psi_1 \cos(\theta_1 - \theta_2), \end{aligned}$$

and

$$(4.9.14) \quad \begin{aligned} L_{\theta_1}(\theta, \psi) &= -m_2 \ell_1 \ell_2 \psi_1 \psi_2 \sin(\theta_1 - \theta_2) - (m_1 + m_2)g \ell_1 \sin \theta_1, \\ L_{\theta_2}(\theta, \psi) &= m_2 \ell_1 \ell_2 \psi_1 \psi_2 \sin(\theta_1 - \theta_2) - m_2 g \ell_2 \sin \theta_2. \end{aligned}$$

Thus the explicit version of (4.9.11)–(4.9.12) is the pair of equations

$$(4.9.15) \quad \begin{aligned} (m_1 + m_2)\ell_1^2 \theta''_1 + m_2 \ell_1 \ell_2 \frac{d}{dt} [\theta'_2 \cos(\theta_1 - \theta_2)] \\ = -m_2 \ell_1 \ell_2 \theta'_1 \theta'_2 \sin(\theta_1 - \theta_2) - (m_1 + m_2)g \ell_1 \sin \theta_1, \end{aligned}$$

and

$$(4.9.16) \quad \begin{aligned} \ell_2^2 \theta_2'' + \ell_1 \ell_2 \frac{d}{dt} [\theta_1' \cos(\theta_1 - \theta_2)] \\ = \ell_1 \ell_2 \theta_1' \theta_2' \sin(\theta_1 - \theta_2) - g \ell_2 \sin \theta_2. \end{aligned}$$

Note that the masses  $m_1$  and  $m_2$  do not appear in (4.9.16);  $m_1$  does not appear in either term of  $(d/dt)L_{\psi_2} - L_{\theta_2}$ , and  $m_2$  factors out.

As in (4.7.44)–(4.7.47), we have the energy

$$(4.9.17) \quad E(\theta, \psi) = \frac{1}{2} \psi \cdot G(\theta) \psi + V(\theta),$$

and if  $\theta(t)$  solves (4.9.11), or equivalently (4.9.15)–(4.9.16), then

$$(4.9.18) \quad \frac{d}{dt} E(\theta(t), \theta'(t)) = 0.$$

By (4.9.5) and (4.9.9), the explicit form of the energy is

$$(4.9.19) \quad \begin{aligned} E(\theta, \psi) = \frac{1}{2} (m_1 + m_2) \ell_1^2 \psi_1^2 + m_2 \ell_1 \ell_2 \psi_1 \psi_2 \cos(\theta_1 - \theta_2) \\ + \frac{1}{2} m_2 \ell_2^2 \psi_2^2 - m_1 g \ell_1 \cos \theta_1 - m_2 g (\ell_1 \cos \theta_1 + \ell_2 \cos \theta_2). \end{aligned}$$

As in (4.7.57)–(4.7.60), we can convert the equations of motion to Hamiltonian form, by setting

$$(4.9.20) \quad p = G(\theta) \psi.$$

The energy (4.9.17) becomes

$$(4.9.21) \quad \begin{aligned} \mathcal{E}(\theta, p) = E(\theta, G(\theta)^{-1} p) \\ = \frac{1}{2} p \cdot G(\theta)^{-1} p + V(\theta), \end{aligned}$$

and (4.9.11) is equivalent to

$$(4.9.22) \quad \frac{d\theta_k}{dt} = \frac{\partial \mathcal{E}}{\partial p_k}, \quad \frac{dp_k}{dt} = -\frac{\partial \mathcal{E}}{\partial \theta_k}.$$

Note that, for  $G(\theta)$  given by (4.9.9),

$$(4.9.23) \quad G(\theta)^{-1} = \frac{1}{\det G(\theta)} \begin{pmatrix} m_2 \ell_2^2 & -m_2 \ell_1 \ell_2 \cos(\theta_1 - \theta_2) \\ -m_2 \ell_1 \ell_2 \cos(\theta_1 - \theta_2) & (m_1 + m_2) \ell_1^2 \end{pmatrix},$$

and

$$(4.9.24) \quad \det G(\theta) = m_1 m_2 \ell_1^2 \ell_2^2 + m_2^2 \ell_1^2 \ell_2^2 \sin^2(\theta_1 - \theta_2).$$

For notational simplicity we write

$$(4.9.25) \quad \mathcal{E}(\theta, p) = \frac{1}{2} p \cdot H(\theta) p + V(\theta), \quad H(\theta) = G(\theta)^{-1}.$$

Solutions to (4.9.22) are orbits of the flow generated by the Hamiltonian vector field

$$\begin{aligned}
 X_{\mathcal{E}}(\theta, p) &= -J\nabla_{\theta, p}\mathcal{E}(\theta, p) \\
 &= \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \begin{pmatrix} \nabla_{\theta}\mathcal{E} \\ \nabla_p\mathcal{E} \end{pmatrix} \\
 &= \begin{pmatrix} \nabla_p\mathcal{E} \\ -\nabla_{\theta}\mathcal{E} \end{pmatrix}.
 \end{aligned}
 \tag{4.9.26}$$

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 (4.9.26). From this formula we see that the critical points of  $X_{\mathcal{E}}$  coincide with the critical points of  $\mathcal{E}$ . Note that

$$\nabla_p\mathcal{E}(\theta, p) = H(\theta)p, \tag{4.9.27}$$

and  $H(\theta)$  is invertible for all  $\theta$ , so if  $\mathcal{E}$  has a critical point at  $(\theta, p)$ ,  $p = 0$ . Now

$$\nabla_{\theta}\mathcal{E}(\theta, 0) = \nabla V(\theta), \tag{4.9.28}$$

so we deduce that  $(\theta, p)$  is a critical point of  $X_{\mathcal{E}}$  if and only if  $p = 0$  and  $\nabla V(\theta) = 0$ . Rewriting (4.9.5) as

$$V(\theta) = -(m_1 + m_2)g\ell_1 \cos \theta_1 - m_2g\ell_2 \cos \theta_2, \tag{4.9.29}$$

we see that

$$\nabla V(\theta) = \begin{pmatrix} (m_1 + m_2)g\ell_1 \sin \theta_1 \\ m_2g\ell_2 \sin \theta_2 \end{pmatrix}, \tag{4.9.30}$$

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

$$(\theta_1, \theta_2, p_1, p_2) = (j\pi, k\pi, 0, 0), \quad j, k \in \mathbb{Z}. \tag{4.9.31}$$

Towards the goal of understanding the behavior of  $X_{\mathcal{E}}$  near these critical points, we examine its derivative. We have

$$DX_{\mathcal{E}}(\theta, 0) = \begin{pmatrix} 0 & H(\theta) \\ -D^2V(\theta) & 0 \end{pmatrix}. \tag{4.9.32}$$

The matrix  $H(\theta)$  is positive definite for all  $\theta$ , and in particular, since  $\sin j\pi = 0$  and  $\cos j\pi = (-1)^j$ ,

$$H(j\pi, k\pi) = \frac{1}{m_1m_2\ell_1^2\ell_2^2} \begin{pmatrix} m_2\ell_2^2 & (-1)^{j-k+1}m_2\ell_1\ell_2 \\ (-1)^{j-k+1}m_2\ell_1\ell_2 & (m_1 + m_2)\ell_1^2 \end{pmatrix}. \tag{4.9.33}$$

Also,

$$D^2V(j\pi, k\pi) = \begin{pmatrix} (-1)^j(m_1 + m_2)g\ell_1 & 0 \\ 0 & (-1)^k m_2g\ell_2 \end{pmatrix}. \tag{4.9.34}$$

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  $m_2 \rightarrow 0$  in the double pendulum system (4.9.15)–(4.9.16) and derive the limiting system

$$(4.9.35) \quad \begin{aligned} \theta_1'' + \frac{g}{\ell_1} \sin \theta_1 &= 0, \\ \theta_2'' + \frac{\ell_1}{\ell_2} \frac{d}{dt} [\theta_1' \cos(\theta_1 - \theta_2)] &= \frac{\ell_1}{\ell_2} \theta_1' \theta_2' \sin(\theta_1 - \theta_2) - \frac{g}{\ell_2} \sin \theta_2. \end{aligned}$$

2. Recall the spherical pendulum, introduced in Exercise 5 of §4.7. 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  $m_j(s) ds$ ,  $0 \leq s \leq \ell_j$ , 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 (4.9.5) and (4.9.6), and then obtain equations of motion, replacing (4.9.15)–(4.9.16).

*Note.* See Exercise 16 in §4.7 to get started.

### 4.10. Momentum-quadratic Hamiltonian systems

Most of the Lagrangians arising in the last three sections have been of the form

$$(4.10.1) \quad L(x, v) = \frac{1}{2}v \cdot G(x)v - V(x),$$

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 §4.7 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,

$$(4.10.2) \quad \frac{dx_k}{dt} = \frac{\partial \mathcal{E}}{\partial p_k}, \quad \frac{dp_k}{dt} = -\frac{\partial \mathcal{E}}{\partial x_k},$$

where

$$(4.10.3) \quad \mathcal{E}(x, p) = \frac{1}{2}p \cdot H(x)p + V(x), \quad H(x) = G(x)^{-1}.$$

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 (4.10.2) are orbits of the flow generated by the Hamiltonian vector field

$$(4.10.4) \quad \begin{aligned} X_{\mathcal{E}}(x, p) &= -J\nabla_{x,p}\mathcal{E}(x, p) \\ &= \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \begin{pmatrix} \nabla_x \mathcal{E} \\ \nabla_p \mathcal{E} \end{pmatrix} \\ &= \begin{pmatrix} \nabla_p \mathcal{E} \\ -\nabla_x \mathcal{E} \end{pmatrix}. \end{aligned}$$

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 (4.10.4).

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

$$(4.10.5) \quad \nabla_p \mathcal{E}(x, p) = H(x)p,$$

so, since  $H(x)$  is invertible, we see that if  $\mathcal{E}$  has a critical point at  $(x, p)$ , then  $p = 0$ . Now

$$(4.10.6) \quad \nabla_x \mathcal{E}(x, 0) = \nabla V(x),$$

so we deduce that the critical points of  $X_{\mathcal{E}}$  consist of

$$(4.10.7) \quad \{(x, 0) : \nabla V(x) = 0\}.$$

We next look at the linearization (cf. (4.3.39)) of  $X_{\mathcal{E}}$  at a critical point  $(x_0, 0)$ , given by

$$(4.10.8) \quad DX_{\mathcal{E}}(x_0, 0) = \begin{pmatrix} 0 & H(x_0) \\ -D^2V(x_0) & 0 \end{pmatrix}.$$

From here on, we assume  $H(x_0)$  is positive definite. For notational simplicity, we set

$$(4.10.9) \quad H = H(x_0), \quad W = D^2V(x_0), \quad L = \begin{pmatrix} 0 & H \\ -W & 0 \end{pmatrix}.$$

Then the linearization of (4.10.2) at  $(x_0, 0)$  is

$$(4.10.10) \quad \frac{dx}{dt} = Hp, \quad \frac{dp}{dt} = -Wx.$$

To analyze the structure of solutions to (4.10.10), it is convenient to directly tackle the second order system

$$(4.10.11) \quad \frac{d^2x}{dt^2} = -HWx,$$

and to do this we bring in the following.

**Lemma 4.10.1.** *Given that  $H \in M(n, \mathbb{R})$  is positive definite, there exists a positive definite  $A \in M(n, \mathbb{R})$  such that*

$$(4.10.12) \quad H = A^2.$$

**Proof.** From Chapter 2 we know that  $\mathbb{R}^n$  has an orthonormal basis  $\{v_j\}$  of eigenvectors of  $H$ , so  $Hv_j = \lambda_j v_j$ ,  $1 \leq j \leq n$ . Each  $\lambda_j$  is positive, so we can define  $A$  by  $Av_j = \sqrt{\lambda_j} v_j$ ,  $1 \leq j \leq n$ .  $\square$

If we make the change of variable

$$(4.10.13) \quad x = Ay,$$

then (4.10.11) is converted to

$$(4.10.14) \quad y'' + AW Ay = 0.$$

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 4.10.2.** *Let  $W \in M(n, \mathbb{R})$  be a symmetric matrix, and assume*

$$(4.10.15) \quad W \text{ has } k \text{ positive and } n - k \text{ negative eigenvalues.}$$

*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

$$(4.10.16) \quad v = Aw \in A\widetilde{\mathcal{W}}_+ \implies v \cdot Wv = w \cdot AWAw \geq 0 \implies v \in \mathcal{W}_+.$$

Thus

$$(4.10.17) \quad A : \widetilde{\mathcal{W}}_+ \longrightarrow \mathcal{W}_+, \text{ injectively,}$$

so

$$(4.10.18) \quad \dim \widetilde{\mathcal{W}}_+ \leq \dim \mathcal{W}_+.$$

A similar argument gives

$$(4.10.19) \quad \dim \widetilde{\mathcal{W}}_- \leq \dim \mathcal{W}_-,$$

and finishes the proof.  $\square$

To continue, under the hypotheses of Lemma 4.10.2, we have an orthonormal basis  $\{u_1, \dots, u_n\}$  of  $\mathbb{R}^n$  such that, with  $\mu_j \in (0, \infty)$ ,

$$(4.10.20) \quad \begin{aligned} AW Au_j &= \mu_j^2 u_j, & j \leq k, \\ AW Au_j &= -\mu_j^2 u_j, & j > k. \end{aligned}$$

in such a case, the general solution to (4.10.14) is

$$(4.10.21) \quad \begin{aligned} y(t) &= \sum_{j \leq k} (a_j \sin \mu_j t + b_j \cos \mu_j t) u_j \\ &\quad + \sum_{j > k} (a_j e^{\mu_j t} + b_j e^{-\mu_j t}) u_j. \end{aligned}$$

Such  $y(t)$  leads to

$$(4.10.22) \quad \begin{pmatrix} Ay(t) \\ A^{-1}y'(t) \end{pmatrix} = \begin{pmatrix} x(t) \\ p(t) \end{pmatrix} = e^{tL} \begin{pmatrix} v_0 \\ v_1 \end{pmatrix},$$

for general  $v_0, v_1 \in \mathbb{R}^n$ . As a result, we have the following.

**Proposition 4.10.3.** *Under the hypotheses of Lemma 10.2,  $L$ , given by (10.9), is diagonalizable, and its eigenvalues are*

$$(4.10.23) \quad \begin{aligned} \pm i\mu_j &\text{ for } j \leq k, \\ \pm\mu_j &\text{ for } j > k. \end{aligned}$$

**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  $t^\ell e^{t\lambda}$ ,  $\ell \geq 1$ , where  $\mu = \pm i\mu_j$  or  $\pm\mu_j$  in (4.10.23), depending on  $j$ .  $\square$

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 4.10.4.** *A critical point  $(x_0, 0)$  of  $X_\mathcal{E}$  is hyperbolic if and only if*

$$(4.10.24) \quad D^2V(x_0) \text{ is negative definite.}$$

If (4.10.24) 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 (4.3.78). See Appendix 4.C for further discussion.

The opposite extreme can also be read off from (4.10.23).

**Proposition 4.10.5.** *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*

$$(4.10.25) \quad D^2V(x_0) \text{ is positive definite.}$$

Recalling that  $\mathcal{E}(x, p)$  is given by (4.10.3), we see that (4.10.25) is equivalent to

$$(4.10.26) \quad D^2\mathcal{E}(x_0, 0) \in M(2n, \mathbb{R}) \text{ is positive definite,}$$

in which case  $\mathcal{E}$  has a local minimum at  $(x_0, 0)$ .

In case (4.10.25) holds, we can deduce from (4.10.21)–(4.10.22), 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 [1], Chapter 8, and [5], Appendices 7–8.

For  $n \geq 2$ , there can be cases intermediate between those covered by Proposition 4.10.4 and those covered by Proposition 4.10.5.

**Proposition 4.10.6.** *If  $(x_0, 0)$  is a critical point for  $X_{\mathcal{E}}$  and*

$$(4.10.27) \quad D^2V(x_0) \text{ has } k \text{ positive eigenvalues and } n - k \text{ negative eigenvalues,}$$

*then*

$$(4.10.28) \quad \begin{aligned} DX_{\mathcal{E}}(x_0, 0) &\text{ has } 2k \text{ imaginary eigenvalues, and} \\ &n - k \text{ positive, and } n - k \text{ negative eigenvalues.} \end{aligned}$$

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 §4.9. There  $V$  was given by (4.9.29), and the critical points by (4.9.31), i.e.,  $(j\pi, k\pi, 0, 0)$ , and  $D^2V(j\pi, k\pi)$  by (4.9.34). We have

$$(4.10.29) \quad \begin{aligned} j \text{ and } k \text{ even} &\implies D^2V(j\pi, k\pi) \text{ positive definite,} \\ j \text{ and } k \text{ odd} &\implies D^2V(j\pi, k\pi) \text{ negative definite,} \\ j \text{ and } k \text{ of opposite parity} &\implies D^2V(j\pi, k\pi) \text{ indefinite.} \end{aligned}$$

In the first case Proposition 4.10.5 applies, in the second case Proposition 4.10.4 applies, and in the third case Proposition 4.10.6 applies, with  $k = 1$  and  $n - k = 1$ .



---

**Exercises**

1. Establish analogues of Propositions 4.10.3, 4.10.5, and 4.10.6 in case  $H$  is allowed to be indefinite (nondegenerate), and we assume

$$(4.10.30) \quad D^2V(x_0) \text{ is either positive definite or negative definite.}$$

Exercises 2–6 deal with the  $2 \times 2$  system

$$(4.10.31) \quad \frac{d^2}{dt^2} \begin{pmatrix} x \\ y \end{pmatrix} = -\nabla_{x,y}V(x, y),$$

for various functions  $V$ . The associated energy function, as in (4.10.3), is

$$(4.10.32) \quad \mathcal{E}(x, y, p, q) = \frac{1}{2}(p^2 + q^2) + V(x, y).$$

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 4.10.6).

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 (4.10.32) replaced by

$$(4.10.33) \quad \mathcal{E}(x, y, p, q) = \frac{1}{2}(p^2 - q^2) + V(x, y).$$

Now Proposition 4.10.6 will not apply, but Exercise 1 might (or might not).

### 4.11. Numerical study – difference schemes

We describe some ways of numerically approximating the solution to a system of differential equations

$$(4.11.1) \quad \frac{dx}{dt} = F(x), \quad x(t_0) = x_0.$$

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 (4.11.1) at times  $t_0 + nh$ :

$$(4.11.2) \quad x_n \approx x(t_0 + nh).$$

Noting that a smooth solution to (4.11.1) satisfies

$$(4.11.3) \quad \begin{aligned} x(t+h) &= x(t) + hx'(t) + O(h^2) \\ &= x(t) + hF(x(t)) + O(h^2), \end{aligned}$$

we have the following crude difference scheme:

$$(4.11.4) \quad x_{n+1} = x_n + hF(x_n).$$

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 (4.11.1) as

$$(4.11.5) \quad x(t+h) = x(t) + \int_0^h F(x(t+s)) ds.$$

Consider methods of approximating

$$(4.11.6) \quad \int_0^h g(s) ds$$

better than  $hg(0) + O(h^2)$ , for smooth  $g$ . Two simple improvements are

$$(4.11.7) \quad \frac{h}{2} [g(0) + g(h)] + O(h^3),$$

the trapezoidal method, and

$$(4.11.8) \quad hg\left(\frac{h}{2}\right) + O(h^3),$$

the midpoint method. These lead respectively to

$$(4.11.9) \quad x(t+h) = x(t) + \frac{h}{2} [F(x(t)) + F(x(t+h))] + O(h^3)$$

and

$$(4.11.10) \quad x(t+h) = x(t) + hF\left(x\left(t + \frac{h}{2}\right)\right) + O(h^3).$$

Neither of them immediately converts to an explicit difference scheme, but in (4.11.9) we can substitute  $F(X(t+h)) = F(X(t) + hF(X(t))) + O(h^2)$  and in (4.11.10) we can substitute  $F(X(t+h/2)) = F(X(t) + (h/2)F(X(t))) + O(h^2)$ , to obtain the second order accurate difference schemes

$$(4.11.11) \quad x_{n+1} = x_n + \frac{h}{2} \left[ F(x_n) + F(x_n + hF(x_n)) \right]$$

and

$$(4.11.12) \quad x_{n+1} = x_n + hF\left(x_n + \frac{h}{2}F(x_n)\right).$$

Often (4.11.11) is called Heun's method and (4.11.12) a modified Euler method.

We now come to the heart of the matter for this section. The Runge-Kutta scheme for (4.11.1) is specified as follows. The approximation  $x_n$  to  $x(t_0 + nh)$  is given recursively by

$$(4.11.13) \quad x_{n+1} = x_n + \frac{h}{6} (K_{n1} + 2K_{n2} + 2K_{n3} + K_{n4}),$$

where

$$(4.11.14) \quad \begin{aligned} K_{n1} &= F(x_n), \\ K_{n2} &= F\left(x_n + \frac{1}{2}hK_{n1}\right), \\ K_{n3} &= F\left(x_n + \frac{1}{2}hK_{n2}\right), \\ K_{n4} &= F(x_n + hK_{n3}). \end{aligned}$$

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

$$(4.11.15) \quad \int_0^h g(s) ds = \frac{h}{6} \left( g(0) + 4g\left(\frac{h}{2}\right) + g(h) \right) + O(h^5).$$

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

$$(4.11.16) \quad \int_{-h}^h G(s) ds = \frac{h}{3} \left[ G(-h) + 4G(0) + G(h) \right] + O(h^5).$$

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  $\leq 3$ . Making a power series expansion  $G(s) = \sum_{j=0}^3 a_j s^j + O(s^4)$  then yields (4.11.16).

Now, write the equation (4.11.1) as the integral equation (4.11.5). By (4.11.15),

$$(4.11.17) \quad \int_0^h F(X(t+s)) ds = \frac{h}{6} \left[ F(X(t)) + 4F\left(X\left(t + \frac{h}{2}\right)\right) + F(X(t+h)) \right] + O(h^5).$$

We then have as an immediate consequence the following result on producing accurate difference schemes.

**Proposition 4.11.1.** *Suppose the approximation*

$$(4.11.18) \quad x(t+h) \approx x(t) + \Phi(x(t), h) = \mathcal{X}(x(t), h)$$

*produces a  $j$ th order accurate difference scheme for the solution to (4.11.1). If  $j \leq 3$ , then a difference scheme accurate of order  $j+1$  is given by*

$$(4.11.19) \quad x_{n+1} = x_n + \frac{h}{6} \left[ F(x_n) + 4F\left(\mathcal{X}\left(x_n, \frac{h}{2}\right)\right) + F(\mathcal{X}(x_n, h)) \right].$$

*Furthermore, if  $x(t+h) \approx \mathcal{X}_\ell(x(t), h)$  both work in (4.11.18),  $\ell = 0, 1$ , then you can use*

$$(4.11.20) \quad x_{n+1} = x_n + \frac{h}{6} \left[ F(x_n) + 4F\left(\mathcal{X}_0\left(x_n, \frac{h}{2}\right)\right) + F(\mathcal{X}_1(x_n, h)) \right].$$

We apply this to two second order methods derived before:

$$(4.11.21) \quad \mathcal{X}_0(x_n, h) = x_n + \frac{h}{2} \left[ F(x_n) + F(x_n + hF(x_n)) \right], \text{ Heun,}$$

and

$$(4.11.22) \quad \mathcal{X}_1(x_n, h) = x_n + hF\left(x_n + \frac{h}{2}F(x_n)\right), \text{ modified Euler.}$$

Thus a third order accurate scheme is produced. The last term in (4.11.19) becomes

$$(4.11.23) \quad \frac{h}{6} \left[ F(x_n) + 4F\left(x_n + \frac{h}{4} \left[ F(x_n) + F\left(x_n + \frac{h}{2}F\right) \right] \right) + F\left(x_n + hF\left(x_n + \frac{h}{2}F\right)\right) \right],$$

where  $F = F(x_n)$ . In terms of  $K_{n1}$ ,  $K_{n2}$  as defined in (4.11.14), we have

$$(4.11.24) \quad \frac{h}{6} \left[ K_{n1} + 4F\left(x_n + \frac{h}{4} [K_{n1} + K_{n2}] \right) + F(x_n + hK_{n2}) \right].$$

This could be used in a 3rd order accurate scheme, but some simplification of the middle term is desirable. Note that, for smooth  $H$ ,

$$(4.11.25) \quad H\left(x + \frac{1}{2}\eta\right) = \frac{1}{2}H(x) + \frac{1}{2}H(x + \eta) + O(|\eta|^2).$$

Consequently, as  $|K_{n1} - K_{n2}| = O(h)$ , by (11.14),

$$(4.11.26) \quad F\left(x_n + \frac{h}{4} [K_{n1} + K_{n2}] \right) = \frac{1}{2}F\left(x_n + \frac{h}{2}K_{n1}\right) + \frac{1}{2}F\left(x_n + \frac{h}{2}K_{n2}\right) + O(h^4).$$

Therefore we have the following.

**Proposition 4.11.2.** *A third order accurate difference scheme for (4.11.1) is given by*

$$(4.11.27) \quad x_{n+1} = x_n + \frac{h}{6} [K_{n1} + 2K_{n2} + 2K_{n3} + L_{n4}]$$

*where  $K_{n1}$ ,  $K_{n2}$ ,  $K_{n3}$  are given by (4.11.14) and*

$$(4.11.28) \quad L_{n4} = F(x_n + hK_{n2}).$$

We can now produce a 4th order accurate difference scheme by applying Proposition 4.11.1 with  $\mathcal{X}(x_n, h)$  defined by (4.11.27). Thus we obtain the difference scheme.

$$(4.11.29) \quad x_{n+1} = x_n + \frac{h}{6} \left\{ K_{n1} + 4F\left(x_n + \frac{h}{12}[K_{n1} + 2k_{n2} + 2k_{n3} + \ell_{n4}]\right) + F\left(x_n + \frac{h}{6}[K_{n1} + 2K_{n2} + 2K_{n3} + L_{n4}]\right) \right\},$$

where  $K_{nj}$ ,  $L_{n4}$  are as above and

$$(4.11.30) \quad \begin{aligned} k_{n2} &= F\left(x_n + \frac{h}{4}K_{n1}\right), \\ k_{n3} &= F\left(x_n + \frac{h}{4}k_{n2}\right), \\ \ell_{n4} &= F\left(x_n + \frac{h}{2}k_{n2}\right). \end{aligned}$$

This formula is more complicated than the Runge-Kutta formula (4.11.13). We say no more about how to obtain (4.11.13), which represents a masterpiece of insight.

We have dealt specifically with autonomous systems in (4.11.1), but a non-autonomous system

$$(4.11.31) \quad \frac{dx}{dt} = G(t, x), \quad x(t_0) = x_0,$$

can be treated similarly, as one can see by writing its autonomous analogue

$$(4.11.32) \quad \frac{d}{dt} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} G(y, x) \\ 1 \end{pmatrix}, \quad \begin{pmatrix} x(t_0) \\ y(t_0) \end{pmatrix} = \begin{pmatrix} x_0 \\ t_0 \end{pmatrix},$$

and applying the formulas just derived to (4.11.32).

We move briefly to another class of difference schemes, based on power series. It derives from the expansion

$$(4.11.33) \quad x(t+h) = x(t) + hx'(t) + \frac{h^2}{2}x''(t) + \cdots + \frac{h^k}{k!}x^{(k)}(t) + O(h^{k+1}).$$

To begin, differentiate (4.11.1), producing

$$(4.11.34) \quad x''(t) = F_2(x, x'), \quad F_2(x, x') = DF(x)x'.$$

Continue differentiating, getting

$$(4.11.35) \quad x^{(j)}(t) = F_j(x, x', \dots, x^{(j-1)}), \quad j \leq k.$$

Then one obtains a difference scheme for an approximation  $x_n$  to  $x(t_0 + nh)$ , of the form

$$(4.11.36) \quad x_{n+1} = x_n + hx'_n + \frac{h^2}{2}x''_n + \cdots + \frac{h^k}{k!}x_n^{(k)},$$

where

$$(4.11.37) \quad x'_n = F(x_n), \quad x''_n = F_2(x_n, x'_n),$$

and, inductively,

$$(4.11.38) \quad x_n^{(j)} = F_j(x_n, x'_n, \dots, x_n^{(j-1)}).$$

This difference scheme is  $k$ th 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

$$(4.11.39) \quad \begin{aligned} x' &= v, \\ y' &= w, \\ v' &= -x(x^2 + y^2)^{-3/2}, \\ w' &= -y(x^2 + y^2)^{-3/2}. \end{aligned}$$

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

$$(4.11.40) \quad \theta' = \psi, \quad \psi' = -\sin \theta,$$

we have

$$(4.11.41) \quad \begin{aligned} \theta'' &= \psi', & \psi'' &= -\psi \cos \theta, \\ \theta^{(3)} &= \psi'', & \psi^{(3)} &= -\psi' \cos \theta + \psi^2 \sin \theta, \\ \theta^{(4)} &= \psi^{(3)}, & \psi^{(4)} &= -\psi'' \cos \theta + 3\psi' \psi \sin \theta + \psi^3 \cos \theta, \end{aligned}$$

from which one can get a workable fourth order difference scheme of the form (4.11.36)–(4.11.38).

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 [7] and [39].

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 (4.11.39) and (4.11.40). 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 [37] and [13].

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. In many cases, various stability issues arise. One such case is discussed in Appendix 4.H, concerning an instability for geodesic equations, arising in §4.7, and how to handle it.

For more on such matters, 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 (4.11.40).
2. Write a program to apply the power series method described in (4.11.33)–(4.11.38) to (4.11.40). Produce a fourth order accurate method.
3. Consider applying the Runge-Kutta scheme to the problem of motion in a planar force field,

$$(4.11.42) \quad x'' = f(x, y), \quad y'' = g(x, y),$$

which can be written as the first order system

$$(4.11.43) \quad \begin{aligned} x' &= v, & v' &= f(x, y), \\ y' &= w, & w' &= g(x, y). \end{aligned}$$

Show that (4.11.13)–(4.11.14) in this context become

$$(4.11.44) \quad \begin{aligned} x &\mapsto x + \frac{h}{6}(v + 2v_2 + 2v_3 + v_4), \\ y &\mapsto y + \frac{h}{6}(w + 2w_2 + 2w_3 + w_4), \\ v &\mapsto v + \frac{h}{6}(a_1 + 2a_2 + 2a_3 + a_4), \\ w &\mapsto w + \frac{h}{6}(b_1 + 2b_2 + 2b_3 + b_4), \end{aligned}$$

where  $a_j, b_j, v_j$ , and  $w_j$  are computed as follows. First,

$$(4.11.45) \quad a_1 = f(x, y), \quad b_1 = g(x, y);$$

then

$$(4.11.46) \quad \begin{aligned} x_2 &= x + \frac{h}{2}v, & y_2 &= y + \frac{h}{2}w, \\ v_2 &= v + \frac{h}{2}a_1, & w_2 &= w + \frac{h}{2}b_1, \end{aligned}$$

and

$$(4.11.47) \quad a_2 = f(x_2, y_2), \quad b_2 = g(x_2, y_2);$$

then

$$(4.11.48) \quad \begin{aligned} x_3 &= x + \frac{h}{2}v_2, & y_3 &= y + \frac{h}{2}w_2, \\ v_3 &= v + \frac{h}{2}a_2, & w_3 &= w + \frac{h}{2}b_2, \end{aligned}$$

and

$$(4.11.49) \quad a_3 = f(x_3, y_3), \quad b_3 = g(x_3, y_3);$$

then

$$(4.11.50) \quad \begin{aligned} x_4 &= x + hv_3, & y_4 &= y + hw_3, \\ v_4 &= v + ha_3, & w_4 &= w + hb_3, \end{aligned}$$

and finally,

$$(4.11.51) \quad a_4 = f(x_4, y_4), \quad b_4 = g(x_4, y_4).$$

Write a program to implement this difference scheme. Test it for various functions  $f(x, y)$  and  $g(x, y)$ . Consider particularly

$$(4.11.52) \quad f(x, y) = -\frac{x}{(x^2 + y^2)^{3/2}}, \quad g(x, y) = -\frac{y}{(x^2 + y^2)^{3/2}},$$

arising in the Kepler problem, (4.11.39).

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 (4.9.15)–(4.9.16).



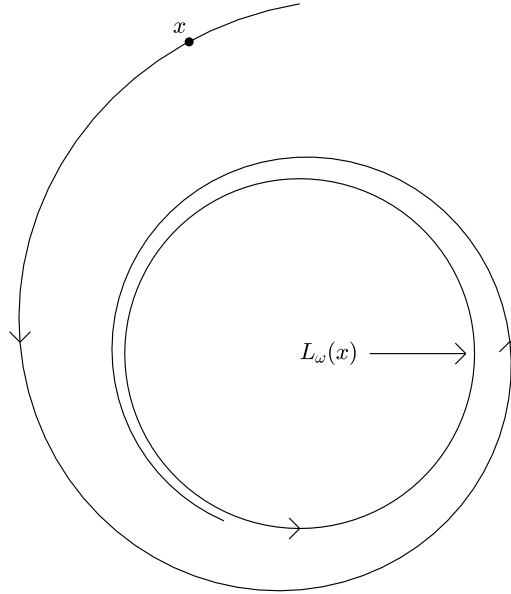


Figure 4.12.1. Limit cycle

#### 4.12. 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  $\Phi^t$ . Take  $x \in \mathcal{O}$ . If  $\Phi^t(x)$  is well defined for all  $t \geq 0$ , we define the  $\omega$ -limit set  $L_\omega(x)$  to consist of all points

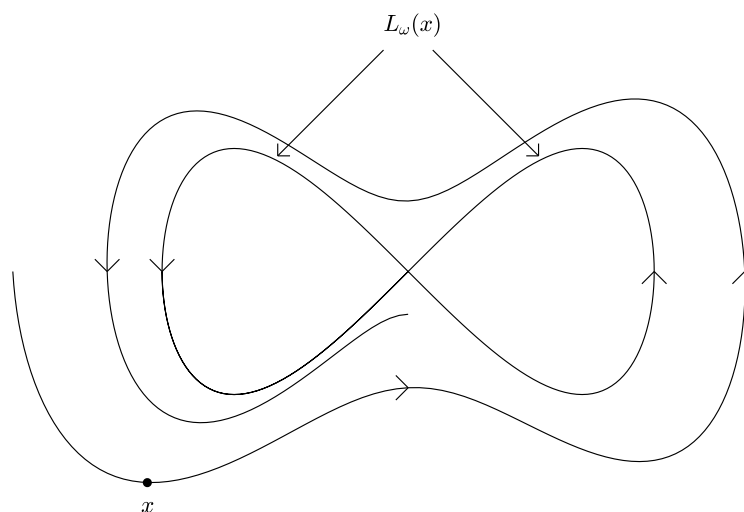
$$(4.12.1) \quad y \in \mathcal{O} \text{ such that there exist } t_k \nearrow +\infty \text{ with } \Phi^{t_k}(x) \rightarrow y.$$

Similarly, if  $\Phi^t(x)$  is well defined for all  $t \leq 0$ , we define the  $\alpha$ -limit set  $L_\alpha(x)$  to consist of all points  $y \in \mathcal{O}$  such that there exist  $t_k \searrow -\infty$  with  $\Phi^{t_k}(x) \rightarrow y$ . Sinks are  $\omega$ -limit sets for all nearby points. Other examples of  $\omega$ -limit sets are pictured in Figures 4.12.1–4.12.2. In Figure 4.12.1,  $L_\omega(x)$  is a periodic orbit, i.e., for some  $T \in (0, \infty)$ ,  $\Phi^T(y) = y$ . In Figure 4.12.2,  $L_\omega(x)$  is a figure eight, containing a hyperbolic critical point of the vector field. The reader might think up examples in which  $L_\omega(x)$  contains several hyperbolic critical points.

The next proposition records some general observations about limit sets that hold in any dimension. Take  $\mathcal{O} \subset \mathbb{R}^n$ ,  $F$ , and  $\Phi^t$  as above.

**Proposition 4.12.1.** *Assume that  $K \subset \mathcal{O}$  is a closed, bounded (hence compact) set in  $\mathbb{R}^n$  and that  $\Phi^t(K) \subset K$  for each  $t > 0$ . Take  $x \in K$ . Then*

$$(4.12.2) \quad L_\omega(x) \text{ is a nonempty, compact subset of } K,$$



**Figure 4.12.2.** Another limit set

given by

$$\begin{aligned}
 (4.12.3) \quad L_\omega(x) &= \bigcap_{s \in \mathbb{R}^+} \overline{\{\Phi^t(x) : t \geq s\}} \\
 &= \bigcap_{k \in \mathbb{N}} \overline{\{\Phi^t(x) : t \geq k\}}.
 \end{aligned}$$

We have

$$(4.12.4) \quad \Phi^t(L_\omega(x)) = L_\omega(x), \quad \forall t \geq 0,$$

and hence

$$(4.12.5) \quad \Phi^t : L_\omega(x) \longrightarrow L_\omega(x), \quad \forall t \in \mathbb{R}.$$

Furthermore,

$$(4.12.6) \quad y \in L_\omega(x) \implies L_\omega(y) \subset L_\omega(x).$$

**Proof.** The result (4.12.3) is a straightforward consequence of the definition of  $L_\omega(x)$ . The fact that this set is nonempty follows from Proposition 4.B.6, in Appendix 4.B. The results (4.12.4)–(4.12.6) are left as exercises.  $\square$

We now specialize to planar vector fields, where  $\omega$ -limit sets tend to have rather special properties. The following result, characterizing  $\omega$ -limit sets without critical

points in planar regions (under a few additional hypotheses), is called the Poincaré-Bendixson theorem.

**Theorem 4.12.2.** *Let  $\mathcal{O}$  be a planar domain, and let  $F$  generate a flow  $\Phi^t$  on  $\mathcal{O}$ . Assume there is a compact set  $K \subset \mathcal{O}$  that 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  $\Phi$ .*

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 [14], §18, and in [32]. We do mention that actually we will need this result only for piecewise smooth simple closed curves, where a simpler proof exists; see [42], pp. 34–40, [45], Chapter 1, §19, or [50], §5.3. 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 §4.15 illustrate how much more complex matters can be in higher dimension.

To tackle Theorem 4.12.2, first note that the hypotheses imply  $L_\omega(x)$  is a nonempty subset of  $K$ . Let  $y \in L_\omega(x)$ , and say

$$(4.12.7) \quad y_k = \Phi^{t_k}(x), \quad t_k \nearrow +\infty, \quad y_k \rightarrow y.$$

We have  $F(y) \neq 0$ . Let  $\Gamma$  be a smooth curve segment in  $\mathcal{O}$ , containing  $y$ , such that the tangent to  $\Gamma$  at  $y$  is linearly independent of  $F(y)$ . Shrinking  $\Gamma$  if necessary, we can assume that for each  $z \in \Gamma$ , the tangent to  $\Gamma$  at  $z$  is linearly independent of  $F(z)$ . We say  $F$  is transverse to  $\Gamma$ ; cf. Figure 4.12.3.

With  $y_k$  as in (4.12.7), we can assume all  $y_k$  are sufficiently close to  $y$  to lie in orbits through  $\Gamma$ , and adjusting each  $t_k$  as needed, we can take

$$(4.12.8) \quad y_k \in \Gamma, \quad \forall k.$$

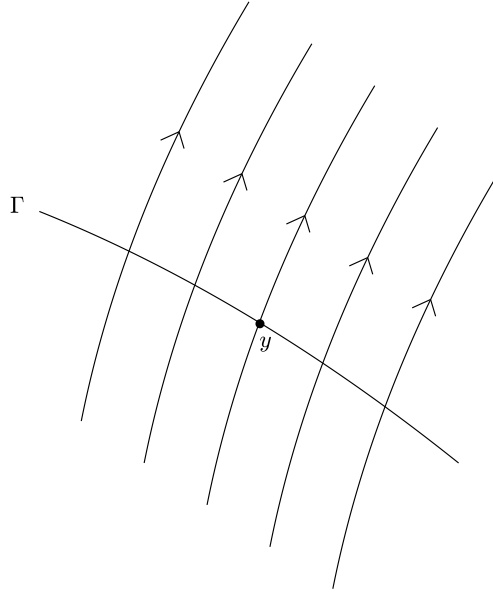
At this point, it is useful to revise the list  $\{t_k\}$  slightly. Let  $t_1 \in \mathbb{R}^+$ ,  $y_1 = \Phi^{t_1}(x)$  be as above. Now let  $t_k \nearrow +\infty$  denote all the successive times when  $\Phi^t(x)$  intersects  $\Gamma$ , so we may be adding times to the set denoted  $t_k$  in (4.12.7). Shortly we will show that (4.12.7) continues to hold for this expanded set of points  $y_k = \Phi^{t_k}(x)$ . First, we make the following useful observation.

**Lemma 4.12.3.** *With  $t_j < t_{j+1} < t_{j+2}$  as above,*

$$(4.12.9) \quad y_{j+1} \text{ lies between } y_j \text{ and } y_{j+2} \text{ on } \Gamma.$$

**Proof.** Consider the curve  $C_j$  starting at  $y_j$ , running to  $y_{j+1}$  along  $\Phi^t(x)$ ,  $t_j \leq t \leq t_{j+1}$ , and returning to  $y_j$  along  $\Gamma$ . Cf. Figure 4.12.4. This is a simple closed curve, and the Jordan curve theorem applies.

Now for  $s$  and  $\sigma$  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



**Figure 4.12.3.** Curve transverse to orbits of  $\Phi^t$

components of  $\mathbb{R}^2 \setminus C_j$ . Since  $\{\Phi^s(y_{j+1}) : s \geq 0\}$  cannot cross  $C_j$  at any point but a point in  $\Gamma$ , 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  $\Gamma$ .  $\square$

Having Lemma 4.12.3, we see that the expanded set of points  $\{y_k\} \subset \Gamma$  interlaces the original set, so (4.12.7) continues to hold. We see that the convergence of  $y_k$  to  $y$  is monotone on  $\Gamma$ . If by chance some  $y_j = y$ , then  $y_k = y$  for all  $k \geq j$ . 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 4.12.2 is that the orbit through  $y$  is periodic. The next result takes us closer to that goal.

**Lemma 4.12.4.** *Suppose  $s > 0$  and  $\Phi^s(y) \in \Gamma$ . Then  $\Phi^s(y) = y$ .*

**Proof.** We have

$$(4.12.10) \quad \sup_{0 \leq t \leq s+1} \|\Phi^t(y_k) - \Phi^t(y)\| = \varepsilon_k \rightarrow 0, \quad \text{as } k \rightarrow \infty.$$

It follows that there exist  $\delta_k \rightarrow 0$  such that  $\Phi^{s+\delta_k}(y_k) \in \Gamma$ , and hence

$$(4.12.11) \quad \Phi^{s+\delta_k}(y_k) = y_{k+\ell(k)}, \quad \text{for some } \ell(k) \in \{1, 2, 3, \dots\}.$$

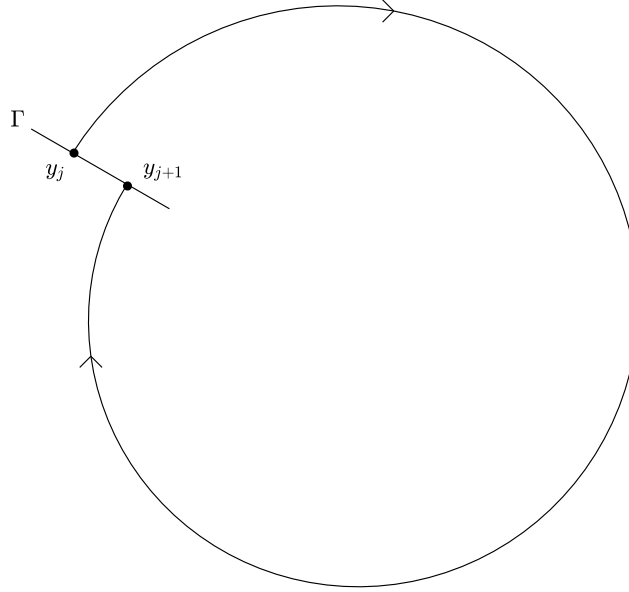


Figure 4.12.4. Orbit of  $\Phi^t$

Thus

$$(4.12.12) \quad \Phi^s(y) = \lim_{k \rightarrow \infty} \Phi^{s+\delta_k}(y_k) = \lim_{k \rightarrow \infty} y_{k+\ell(k)} = y,$$

as asserted.  $\square$

We are ready for the endgame in the proof of Theorem 4.12.2. 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

$$(4.12.13) \quad z_j = \Phi^{s_j}(y) \longrightarrow z \in K.$$

We have  $F(z) \neq 0$ , so there is a curve segment  $\tilde{\Gamma}$  through  $z$ , transverse to  $F$ . Adjusting  $s_j$ , we can arrange

$$(4.12.14) \quad z_j \in \tilde{\Gamma}_j.$$

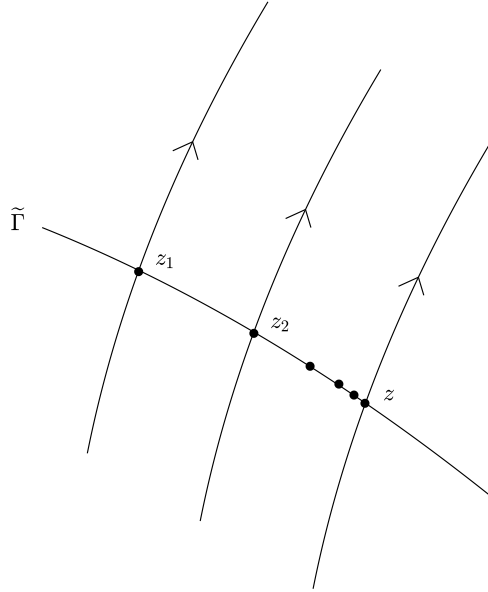
We need only two such points in such a curve  $\tilde{\Gamma}$ ; say, upon relabeling,

$$(4.12.15) \quad z_1 = \Phi^{s_1}(y), \quad z_2 = \Phi^{s_2}(y) = \Phi^{s_2-s_1}(z_1) \in \tilde{\Gamma}.$$

See Figure 4.12.5.

Note that

$$(4.12.16) \quad \Phi^{t_k+s_1}(x) \longrightarrow z_1,$$



**Figure 4.12.5.** March of  $z_j = \Phi^{s_j}(y) \rightarrow z$

so we can use the previous results, with  $t_k$  replaced by  $t_k + s_1$  and  $y$  by  $z_1$ , and  $\Gamma$  by  $\tilde{\Gamma}$ . In this case, the analogue of the hypothesis in Lemma 4.12.4 applies:

$$(4.12.17) \quad s_2 - s_1 > 0, \quad \Phi^{s_2 - s_1}(z_1) \in \tilde{\Gamma}.$$

The conclusion of Lemma 4.12.4 is

$$(4.12.18) \quad \Phi^{s_2 - s_1}(z_1) = z_1,$$

i.e., actually  $z_2 = z_1$ . (The same argument gives  $z_{j+1} = z_j$  for all  $j$ , so actually  $z_1 = z$ .)

Thus the orbit of  $\Phi$  through  $y$  is periodic, of period  $s = s_2 - s_1$ . Since  $y \in L_\omega(x)$ , it follows that this periodic orbit is contained in  $L_\omega(x)$ .

Note that if  $\Phi^s(y) = y$ , then we can apply (4.12.11) to deduce that the times  $t_k$  arising in  $\Phi^{t_k}(x) = y_k$  satisfy

$$(4.12.19) \quad \limsup_{k \rightarrow \infty} (t_{k+1} - t_k) \leq s.$$

The last point to cover in the proof of Theorem 4.12.2 is that the periodic orbit through  $y$  contains all of  $L_\omega(x)$ . Indeed, let  $\tilde{y}$  be another point in  $L_\omega(x)$ . We have

$$(4.12.20) \quad \Phi^{\tau_j}(x) \rightarrow \tilde{y}, \quad \tau_j \nearrow +\infty.$$

Using (4.12.19), we can write  $\tau_j = t_k + \sigma_k$ ,  $0 \leq \sigma_k \leq s + 1$ , and passing to a subsequence obtain

$$(4.12.21) \quad \Phi^{\tau_j}(x) = \Phi^{\sigma_k}(y_k) \longrightarrow \Phi^\sigma(y),$$

hence  $\tilde{y} = \Phi^\sigma(y)$ . This completes the proof of Theorem 4.12.2.  $\square$

The following equation, known as the van der Pol equation, illustrates the workings of Theorem 4.12.2. The equation is

$$(4.12.22) \quad x'' - \mu(1 - x^2)x' + x = 0.$$

Here  $\mu$  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 (4.12.22) converts to the first order system

$$(4.12.23) \quad x' = y, \quad y' = -x + \mu(1 - x^2)y.$$

Figure 4.12.6 is a phase portrait for the case  $\mu = 1$ . The vector field  $F$  associated with (4.12.23) has one critical point, at the origin. The linearization of (4.12.23) at the origin is

$$(4.12.24) \quad \frac{d}{dt} \begin{pmatrix} \xi \\ \eta \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & \mu \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix},$$

and the eigenvalues of this matrix are

$$(4.12.25) \quad \frac{\mu}{2} \pm \frac{1}{2} \sqrt{\mu^2 - 4}.$$

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 (4.12.23),

$$(4.12.26) \quad \frac{d}{dt}(x^2 + y^2) = 2\mu(1 - x^2)y^2,$$

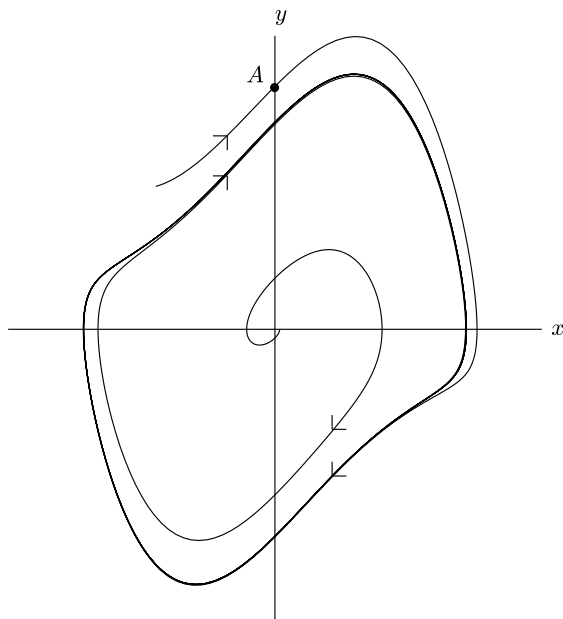
which is  $\geq 0$  for  $|x| \leq 1$ , and in particular is  $\geq 0$  near the origin.

An examination of Figure 4.12.6 indicates the presence of a periodic orbit, attracting all the other orbits. Let us see how this fits into the set-up of Theorem 4.12.2. 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  $\Phi^t$  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 Figure 4.12.6 and again in Figure 4.12.7.

A numerical integration of (4.12.23) (using the Runge-Kutta scheme) shows that

$$(4.12.27) \quad \begin{aligned} &\Phi^t(A) \text{ winds clockwise about the origin,} \\ &\text{and again hits the positive } y\text{-axis} \\ &\text{at the point } B, \text{ lying below } A. \end{aligned}$$

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 (4.12.23) that on this line



**Figure 4.12.6.** Van der Pol limit cycle

segment the vector field  $F$  points to the right. Thus the closed region  $\tilde{K}$  bounded by this curve has the invariance property

$$(4.12.28) \quad \Phi^t(\tilde{K}) \subset \tilde{K}, \quad \forall t \geq 0.$$

We then pick  $\varepsilon > 0$  small enough (in particular  $< 1$ ), and set

$$(4.12.29) \quad K = \tilde{K} \setminus \{(x, y) : x^2 + y^2 < \varepsilon^2\}.$$

The fact that

$$(4.12.30) \quad \Phi^t(K) \subset K, \quad \forall t \geq 0$$

follows from (4.12.28) and (4.12.26). We have removed the only critical point of  $F$ , so  $K$  contains no critical points, and Theorem 4.12.2 applies.

It must be said that the validity of the argument just given relies on the accuracy of the statement (4.12.27) about the orbit through  $A$ . Here we have relied on a numerical approximation to that orbit. We applied the Runge-Kutta scheme, described in §4.11, 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 4.12.2 is a special case of the following result.



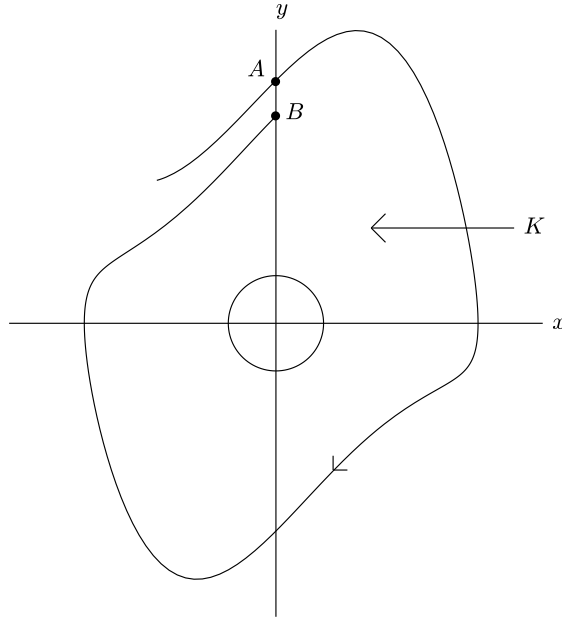


Figure 4.12.7. Van der Pol orbit

**Bendixson's Theorem.** Let  $F$  be a  $C^1$  vector field on  $\mathcal{O} \subset \mathbb{R}^2$ , generating a flow  $\Phi^t$ . 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 [10], Chapter 16, or in [29], Chapter 10. Note that alternative (c) is illustrated in Figure 4.12.2. We emphasize that both this result and Theorem 4.12.2 are results for planar vector fields. In higher dimension, matters are completely different, as we will discuss in §4.15.

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,

$$(4.12.31) \quad \frac{d}{dt}V(x(t)) = \nabla V(x(t)) \cdot F(x(t)).$$

If also  $V$  has the property

$$(4.12.32) \quad \nabla V(y) \cdot F(y) < 0, \quad \forall y \in \mathcal{O} \setminus p,$$

we say  $V$  is a strong Lyapunov function for  $F$ . In such a case,

$$(4.12.33) \quad \frac{d}{dt}V(x(t)) < 0, \quad \text{whenever } x(t) \neq p.$$

If we replace (4.12.32) by the weaker property

$$(4.12.34) \quad \nabla V(y) \cdot F(y) \leq 0, \quad \forall y \in \mathcal{O},$$

we say  $V$  is a Lyapunov function for  $F$ . In such a case,

$$(4.12.35) \quad \frac{d}{dt}V(x(t)) \leq 0, \quad \forall t \geq 0.$$

Thus, as  $t \nearrow +\infty$ ,  $V(x(t))$  monotonically approaches a limit,  $V_0$ , which must be  $\geq V(p)$ , and furthermore,

$$(4.12.36) \quad \lim_{t \rightarrow +\infty} \frac{d}{dt}V(x(t)) = 0.$$

This has the following immediate consequence.

**Proposition 4.12.5.** *Let  $F$  be a  $C^1$  vector field on  $\mathcal{O} \subset \mathbb{R}^n$ , generating a flow  $\Phi^t$ . 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*

$$(4.12.37) \quad L_\omega(x_0) \subset \{y \in \mathcal{O} : \nabla V(y) \cdot F(y) = 0\}.$$

If  $V$  is a strong Lyapunov function, then

$$(4.12.38) \quad L_\omega(x_0) = \{p\}.$$

## Exercises

1. Let  $\mathcal{O} \subset \mathbb{R}^n$  be open and  $\bar{\Omega} \subset \mathcal{O}$  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  $\Phi^t$ . Assume

$$(4.12.39) \quad F \cdot n \leq 0 \quad \text{on } \partial\Omega.$$

Show that

$$(4.12.40) \quad \Phi^t(\bar{\Omega}) \subset \bar{\Omega}, \quad \forall t \geq 0.$$

Compare Exercises 4–5 of §4.3.

2. In the setting of Exercise 1, show that

$$(4.12.41) \quad \Phi^t(\bar{\Omega}) \subset \Phi^s(\bar{\Omega}) \quad \text{for } 0 < s < t.$$

Set

$$(4.12.42) \quad \mathcal{B} = \bigcap_{t \in \mathbb{R}^+} \Phi^t(\bar{\Omega}) = \bigcap_{k \in \mathbb{Z}^+} \Phi^k(\bar{\Omega}).$$

Show that

$$(4.12.43) \quad \Phi^t(\mathcal{B}) = \mathcal{B}, \quad \forall t \geq 0.$$

REMARK. It can be shown from material in Appendix 4.B that  $\mathcal{B}$  is nonempty, closed, and bounded.

3. In the setting of Exercise 2, show that

$$(4.12.44) \quad \forall x \in \bar{\Omega}, \quad L_\omega(x) \subset \mathcal{B}.$$

4. In the setting of Exercise 2, show that

$$(4.12.45) \quad \operatorname{div} F < 0 \text{ on } \bar{\Omega} \implies \operatorname{Vol}(\mathcal{B}) = 0.$$

5. In the setting of Exercise 4, assume that  $n = 2$ , that  $\bar{\Omega} \subset \mathbb{R}^2$  is an annulus, and that  $F$  has no critical points in  $\bar{\Omega}$ , so by Theorem 4.12.2 there is a periodic orbit of  $\Phi$  in  $\bar{\Omega}$ . Show that, due to (4.12.45), there can be only *one* periodic orbit of  $\Phi$  in  $\bar{\Omega}$ .

*Hint.* Feel free to use the Jordan Curve Theorem.

Exercises 6–8 deal with a nonlinear RLC circuit, as pictured in Figure 4.12.8. The setup is as in §1.13 of Chapter 1 (see also Chapter 3, §3.5), except that Ohm's law is modified. The voltage drop across the "resistor" is given by

$$(4.12.46) \quad V = f(I),$$

where  $f$  can be nonlinear, and not necessarily monotonic. As an example, one could have

$$(4.12.47) \quad f(I) = \mu \left( \frac{1}{3} I^3 - I \right).$$

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

$$(4.12.48) \quad V = L \frac{dI}{dt}, \quad V = \frac{Q}{C}.$$

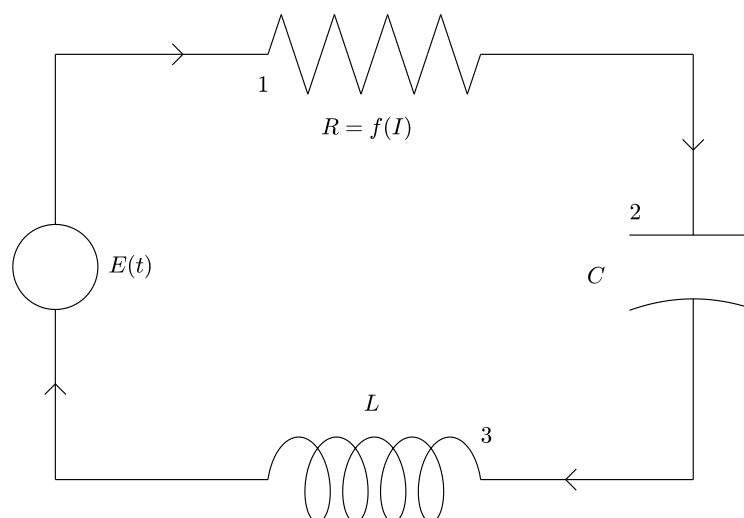
Units of current, etc., are as in §1.13 of Chapter 1.

6. Modify the computations done in (1.14.1)–(1.14.7) of Chapter 1 and show that the current  $I(t)$  satisfies the differential equation

$$(4.12.49) \quad \frac{d^2 I}{dt^2} + \frac{f'(I)}{L} \frac{dI}{dt} + \frac{1}{LC} I = \frac{E'(t)}{L}.$$

Show that rescaling  $I$  and  $t$  leads to (4.12.23), when  $f(I)$  is given by (4.12.47) and  $E \equiv 0$ . More generally, rescale (4.12.49) to

$$(4.12.50) \quad \frac{d^2 x}{dt^2} + f'(x) \frac{dx}{dt} + x = g(t).$$



**Figure 4.12.8.** RLC circuit with nonlinear resistance

7. Assume  $g \equiv 0$  in (4.12.50). Parallel to (4.12.23), 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

$$(4.12.51) \quad \begin{aligned} \frac{dx}{dt} &= y - f(x), \\ \frac{dy}{dt} &= -x. \end{aligned}$$

This is called a Lienard equation.

8. Show that if  $(x(t), y(t))$  solves (4.12.51), then

$$(4.12.52) \quad \frac{d}{dt}(x^2 + y^2) = -2xf(x).$$

### 4.13. 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:

$$(4.13.1) \quad \frac{dx}{dt} = ax.$$

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 (4.13.1) 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 (4.13.1) posits that the resources can support a population no greater than  $K$ . The following is called the logistic equation:

$$(4.13.2) \quad \frac{dx}{dt} = ax(1 - bx),$$

where  $b = 1/K$ . In this model, (4.13.1) 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 (4.13.2) can be solved by separation of variables:

$$(4.13.3) \quad \frac{dx}{x(1 - bx)} = a dt.$$

The reader can perform the integration as an exercise.

The function  $F(x) = ax(1 - bx)$  on the right side of (4.13.2) 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 Figure 4.13.1 for the “phase portrait.”

We turn to a class of  $2 \times 2$  systems called “predator-prey” equations. For this, we set

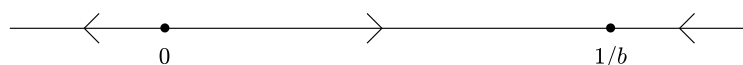
$$(4.13.4) \quad \begin{aligned} x(t) &= \text{population of predators,} \\ y(t) &= \text{population of prey,} \\ \zeta(y) &= \text{rate at which each predator consumes prey.} \end{aligned}$$

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:

$$(4.13.5) \quad \begin{aligned} \frac{dx}{dt} &= -ax + b\zeta x, \\ \frac{dy}{dt} &= ry - \zeta x, \end{aligned}$$

or

$$(4.13.6) \quad \begin{aligned} \frac{dx}{dt} &= -ax + b\zeta x, \\ \frac{dy}{dt} &= ry(1 - cy) - \zeta x. \end{aligned}$$



**Figure 4.13.1.** Phase portrait for logistic equation

Here,  $a, b, c$ , and  $r$  are positive constants. As for the rate of feeding  $z$ , we assume

$$(4.13.7) \quad \zeta = \zeta(y).$$

Clearly if  $y = 0$  then  $\zeta = 0$ . One possibility that is used is

$$(4.13.8) \quad \zeta(y) = \kappa y,$$

for some positive constant  $\kappa$ . 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 Figure 4.13.2, representing a feeding rate bounded by  $\beta$ .

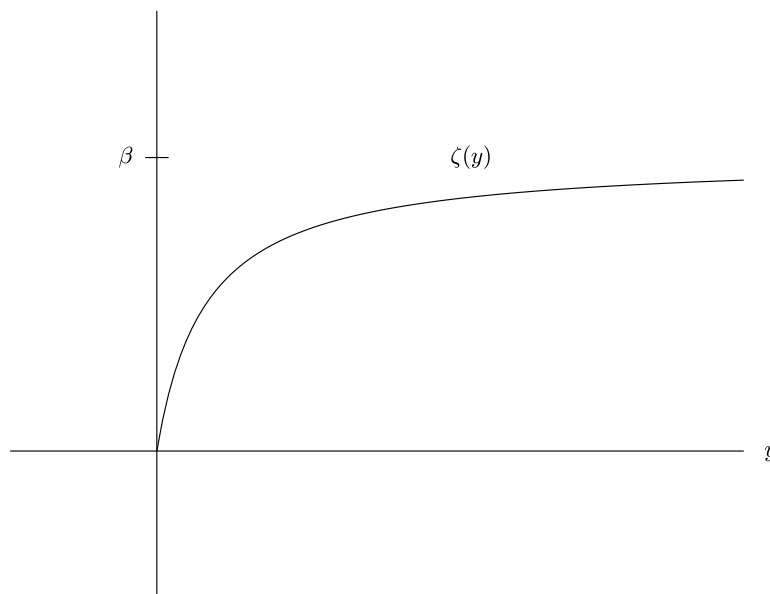
A class of functions of this sort is given by

$$(4.13.9) \quad \zeta(y) = \frac{\kappa y}{1 + \gamma y}, \quad \frac{\kappa}{\gamma} = \beta.$$

Another class is

$$(4.13.10) \quad \zeta(y) = \beta(1 - e^{-\gamma y}), \quad \beta\gamma = \kappa.$$

Let us examine various cases in more detail.



**Figure 4.13.2.** Rate of feeding curve for predator-prey system

### Volterra-Lotka equations

The case (4.13.5) with  $\zeta$  given by (4.13.8) produces systems called Volterra-Lotka equations:

$$(4.13.11) \quad \begin{aligned} \frac{dx}{dt} &= -ax + \sigma xy, & \sigma &= b\kappa, \\ \frac{dy}{dt} &= ry - \kappa xy. \end{aligned}$$

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

$$(4.13.12) \quad DV(0, 0) = \begin{pmatrix} -a & 0 \\ 0 & r \end{pmatrix},$$

so the origin is a saddle. The other critical point is

$$(4.13.13) \quad (x_0, y_0) = \left( \frac{r}{\kappa}, \frac{a}{\sigma} \right).$$

Note that

$$(4.13.14) \quad DV(x_0, y_0) = \begin{pmatrix} 0 & \sigma x_0 \\ -\kappa y_0 & 0 \end{pmatrix},$$

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 (4.13.11) we get

$$(4.13.15) \quad \frac{dy}{dx} = \frac{y(r - \kappa x)}{x(\sigma y - a)},$$

which separates to

$$(4.13.16) \quad \left(\sigma - \frac{a}{y}\right) dy = \left(\frac{r}{x} - \kappa\right) dx.$$

Integrating yields

$$(4.13.17) \quad \sigma y - a \log y = r \log x - \kappa x + C.$$

We deduce that the following smooth function on the region  $x, y > 0$ ,

$$(4.13.18) \quad H(x, y) = \sigma y - a \log y + \kappa x - r \log x,$$

is constant on orbits of (4.13.11), i.e., these orbits lie on level curves of  $H$ . Note that

$$(4.13.19) \quad \nabla H(x, y) = \begin{pmatrix} \kappa - \frac{r}{x} \\ \sigma - \frac{a}{y} \end{pmatrix}, \quad D^2 H(x, y) = \begin{pmatrix} \frac{r}{x^2} & 0 \\ 0 & \frac{a}{y^2} \end{pmatrix},$$

hence, with  $(x_0, y_0)$  as in (4.13.13),

$$(4.13.20) \quad \nabla H(x_0, y_0) = 0, \quad D^2 H(x_0, y_0) = \begin{pmatrix} \frac{r}{x_0^2} & 0 \\ 0 & \frac{a}{y_0^2} \end{pmatrix},$$

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 (4.13.11) is pictured in Figure 4.13.3.

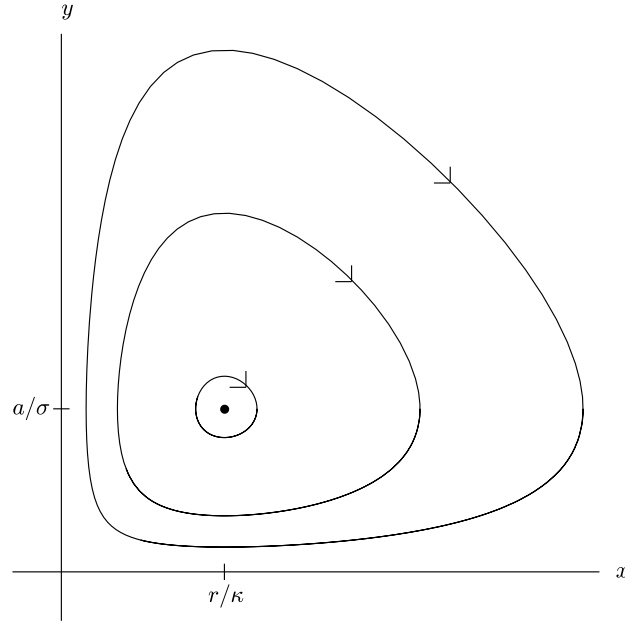
The system (4.13.11) was studied independently by Lotka and Volterra around 1925, by Lotka as a model of some chemical reactions and by Volterra as a predator-prey model, specifically for sharks preying on another species of fish. Volterra made the following further observation. Bring in another type 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 (4.13.11) gets revised to

$$(4.13.21) \quad \begin{aligned} \frac{dx}{dt} &= -ax + \sigma xy - ex, \\ \frac{dy}{dt} &= ry - \kappa xy - ey. \end{aligned}$$

Now (4.13.21) has the same form as (4.13.11), with  $a$  replaced by  $a + e$  and with  $r$  replaced by  $r - e$ , all these constants remaining positive as long as

$$(4.13.22) \quad 0 < e < r.$$





**Figure 4.13.3.** Phase portrait for Volterra-Lotka system

Then the previous analysis applies. The system (4.13.21) has a stable critical point at

$$(4.13.23) \quad (x_1, y_1) = \left( \frac{r-e}{\kappa}, \frac{a+e}{\sigma} \right).$$

Note that at this critical point there are fewer sharks and more prey, compared to (4.13.13). Of course, this depends on the hypothesis (4.13.22). If  $e > r$ , things are catastrophically different.

### First modification

We turn from Volterra-Lotka equations to predator-prey models given by (4.13.6), still keeping (4.13.8). Then we have the following system:

$$(4.13.24) \quad \begin{aligned} \frac{dx}{dt} &= -ax + \sigma xy, & \sigma &= b\kappa, \\ \frac{dy}{dt} &= ry(1-cy) - \kappa xy. \end{aligned}$$

As with (4.13.11), 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 ( $x \geq 0, y \geq 0$ ) is invariant under the flow. Note furthermore that, for

$$(4.13.25) \quad V(x, y) = (-ax + \sigma xy, ry(1-cy) - \kappa xy)^t,$$

we have

$$(4.13.26) \quad V\left(x, \frac{1}{c}\right) = \left(\left(\frac{\sigma}{c} - a\right)x, -\frac{\kappa}{c}x\right)^t,$$

which points downward for  $x > 0$ . It follows that

$$(4.13.27) \quad \mathcal{R} = \left\{(x, y) : x \geq 0, 0 \leq y \leq \frac{1}{c}\right\}$$

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 (4.13.25). Two of these are

$$(4.13.28) \quad (0, 0) \quad \text{and} \quad \left(0, \frac{1}{c}\right).$$

$DV(0, 0)$  is again given by (4.13.12), so  $(0, 0)$  is a saddle. Also,

$$(4.13.29) \quad DV\left(0, \frac{1}{c}\right) = \begin{pmatrix} -a + \frac{\sigma}{c} & 0 \\ -\frac{\kappa}{c} & -r \end{pmatrix}.$$

$V$  has a third critical point, at

$$(4.13.30) \quad y_0 = \frac{a}{\sigma}, \quad x_0 = \frac{r}{\kappa} \left(1 - \frac{ca}{\sigma}\right) = \frac{rc}{\kappa\sigma} \left(\frac{\sigma}{c} - a\right).$$

Note how this point is shifted to the left from the point (4.13.13). There are three cases to consider.

CASE I.  $\sigma/c - a < 0$ .

In this case, the critical point (4.13.30) is not in the first quadrant, so  $V$  has only the critical points (4.13.28) in  $\mathcal{R}$ . In this case (4.13.29) has two negative eigenvalues, so the critical point  $(0, 1/c)$  is a sink. Note that the  $x$ -component of  $V(x, y)$  is

$$(4.13.31) \quad x(\sigma y - a) \leq x\left(\frac{\sigma}{c} - a\right), \quad \text{for } x \geq 0, y \leq \frac{1}{c},$$

so  $V$  points to the left everywhere in  $\mathcal{R}$  except the left edge. Consequently, the population of predators is driven to extinction as  $t \rightarrow +\infty$ , 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 (4.13.29) 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

$$(4.13.32) \quad \begin{aligned} DV(x_0, y_0) &= \begin{pmatrix} -a + \sigma y_0 & \sigma x_0 \\ -\kappa y_0 & r(1 - 2cy_0) - \kappa x_0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & \frac{rc}{\kappa} \left(\frac{\sigma}{c} - a\right) \\ -\frac{\kappa a}{\sigma} & -\frac{rca}{\sigma} \end{pmatrix}. \end{aligned}$$

Note that

$$(4.13.33) \quad \begin{aligned} \det DV(x_0, y_0) &= \frac{rca}{\sigma} \left( \frac{\sigma}{c} - a \right) > 0, \\ \text{Tr } DV(x_0, y_0) &= -\frac{rca}{\sigma} < 0. \end{aligned}$$

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  $t \rightarrow +\infty$ . To see this, we construct a Liapunov function. We do this by modifying  $H(x, y)$  in (4.13.18), which has a minimum at the point (4.13.13), to one that has a minimum at the point (4.13.30). We take

$$(4.13.34) \quad \tilde{H}(x, y) = \sigma y - a \log y + \kappa x - r \left( 1 - \frac{ca}{\sigma} \right) \log x.$$

If  $(x(t), y(t))$  solves (4.13.24), a computation gives

$$(4.13.35) \quad \frac{d}{dt} \tilde{H}(x, y) = -\frac{rc}{\sigma} (\sigma y - a)^2.$$

By Proposition 4.12.5, if we take any point  $p \in \mathcal{R}$ , with positive  $x$  and  $y$ -coordinates (so it is in the domain of  $\tilde{H}$ ), the  $\omega$ -limit set of  $p$  satisfies

$$(4.13.36) \quad L_\omega(p) \subset \left\{ (x, y) \in \mathcal{R} : y = \frac{a}{\sigma} \right\}.$$

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 Figure 4.13.4 for a phase portrait treating Case II.

CASE III.  $\sigma/c - a = 0$ .

In this case  $(x_0, y_0) = (0, 1/c)$ . In (4.13.29) the eigenvalues are 0 and  $-r$ , so  $(0, 1/c)$  is a degenerate critical point. In place of (4.13.31) we have that the  $x$ -component of  $V(x, y)$  is

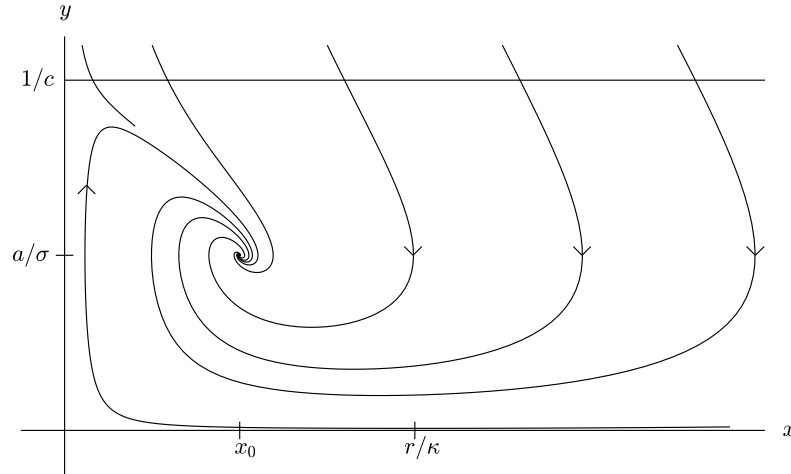
$$(4.13.37) \quad x(\sigma y - a) \leq 0, \quad \text{for } x \geq 0, y \leq \frac{1}{c},$$

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  $t \rightarrow +\infty$ .

### Second modification

We now move to the next level of sophistication, using the system (4.13.6) with  $\zeta = \zeta(y)$ , described as in Figure 4.13.2. Thus, we look at systems of the form

$$(4.13.38) \quad \begin{aligned} \frac{dx}{dt} &= -ax + bx\zeta(y), \\ \frac{dy}{dt} &= ry(1 - cy) - x\zeta(y). \end{aligned}$$



**Figure 4.13.4.** First modification of Volterra-Lotka system (Case II)

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 Figure 4.13.2, we make the following hypotheses:

- (4.13.39)      (a)  $\zeta : [0, \infty) \rightarrow [0, \infty)$  is smooth,  
                   (b)  $\zeta(0) = 0$ ,  
                   (c)  $\zeta'(y) > 0, \quad \forall y \geq 0$ ,  
                   (d)  $\sup \zeta(y) = \beta < \infty$ ,  
                   (e)  $\zeta''(y) \leq 0$ .

All these conditions are satisfied by the examples (4.13.9) and (4.13.10). Hypothesis (c) implies  $\zeta$  is strictly monotone increasing, and hypothesis (e) implies  $\zeta$  is concave.

In this case, the vector field is

$$(4.13.40) \quad V(x, y) = (x(b\zeta(y) - a), ry(1 - cy) - x\zeta(y))^t.$$

Parallel to (4.13.26),

$$(4.13.41) \quad V\left(x, \frac{1}{c}\right) = \left(\left(b\zeta\left(\frac{1}{c}\right) - a\right)x, -\zeta\left(\frac{1}{c}\right)x\right)^t,$$

which points downward for  $x > 0$ , and again it follows that the region  $\mathcal{R}$ , given by (4.13.27), is invariant under the flow  $\Phi^t$  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

$$(0, 0) \text{ and } \left(0, \frac{1}{c}\right),$$

and again  $DV(0, 0)$  is given by (4.13.12), so  $(0, 0)$  is a saddle. This time,

$$(4.13.42) \quad DV\left(0, \frac{1}{c}\right) = \begin{pmatrix} b\zeta\left(\frac{1}{c}\right) - a & 0 \\ -\zeta\left(\frac{1}{c}\right) & -r \end{pmatrix}.$$

Also, a critical point would occur at  $(x_0, y_0)$  if these coordinates satisfy

$$(4.13.43) \quad \zeta(y_0) = \frac{a}{b}, \quad x_0 = \frac{b}{a}ry_0(1 - cy_0).$$

Under the hypotheses (4.13.39), the first equation in (4.13.43) has a (unique) solution if and only if

$$(4.13.44) \quad \frac{a}{b} < \beta.$$

From here on we will assume (4.13.44) holds, and leave it to the reader to consider the behavior of the flow when (4.13.44) fails. Given (4.13.44),  $x_0$  and  $y_0$  are well defined by (4.13.43). Parallel to the study of (4.13.30), again we have three cases.

CASE I.  $1 - cy_0 < 0$ ,

CASE II.  $1 - cy_0 > 0$ ,

CASE III.  $1 - cy_0 = 0$ .

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

$$(4.13.45) \quad (x_0, y_0) \in \mathcal{R}.$$

Given  $\zeta(y_0) = a/b$  and the hypotheses (4.13.39) on  $\zeta$ , we have

$$(4.13.46) \quad \zeta\left(\frac{1}{c}\right) > \frac{a}{b} \iff \frac{1}{c} > y_0 \iff 1 - cy_0 > 0,$$

and hence in Case II,  $DV(0, 1/c)$  has one positive eigenvalue and one negative eigenvalue, so

$$(4.13.47) \quad \left(0, \frac{1}{c}\right) \text{ is a saddle.}$$

(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 (4.13.32):

$$(4.13.48) \quad \begin{aligned} DV(x_0, y_0) &= \begin{pmatrix} b\zeta(y_0) - a & b\zeta'(y_0)x_0 \\ -\zeta(y_0) & r(1 - 2cy_0) - x_0\zeta'(y_0) \end{pmatrix} \\ &= \begin{pmatrix} 0 & b\zeta'(y_0)x_0 \\ -\frac{a}{b} & r(1 - 2cy_0) - x_0\zeta'(y_0) \end{pmatrix}, \end{aligned}$$

and parallel to (4.13.33) we have

$$\begin{aligned}
 \det DV(x_0, y_0) &= ax_0\zeta'(y_0) > 0, \\
 \text{Tr } DV(x_0, y_0) &= r(1 - 2cy_0) - x_0\zeta'(y_0) \\
 &= r \left[ -cy_0 + (1 - cy_0) \left\{ 1 - \frac{\zeta'(y_0)y_0}{\zeta(y_0)} \right\} \right].
 \end{aligned}
 \tag{4.13.49}$$

Let us set

$$Z_0 = 1 - \frac{\zeta'(y_0)y_0}{\zeta(y_0)}.$$

Given  $\zeta$ , this is a function of  $a/b$ , but it is independent of  $c$  and  $r$ . Note that, since  $\zeta(0) = 0$ ,

$$\frac{\zeta(y_0)}{y_0} = \zeta'(\tilde{y}), \quad \text{for some } \tilde{y} \in (0, y_0),$$

by the mean value theorem, so the hypotheses on  $\zeta$  in (4.13.39) imply

$$0 < Z_0 < 1.$$

(Note that in the context of the previous model, with  $\zeta(y)$  given by (13.8),  $Z_0 = 0$ .)

We have

$$\text{Tr } DV(x_0, y_0) = r[Z_0(1 - cy_0) - cy_0].$$

This gives rise to three cases.

CASE IIA.  $Z_0 < cy_0/(1 - cy_0)$ .

Then  $\text{Tr } DV(x_0, y_0) < 0$ , so, by (4.13.49),

$$(x_0, y_0) \text{ is a sink.}$$

CASE IIB.  $Z_0 > cy_0/(1 - cy_0)$ .

Then  $\text{Tr } DV(x_0, y_0) > 0$ , so, by (4.13.49),

$$(x_0, y_0) \text{ is a source.}$$

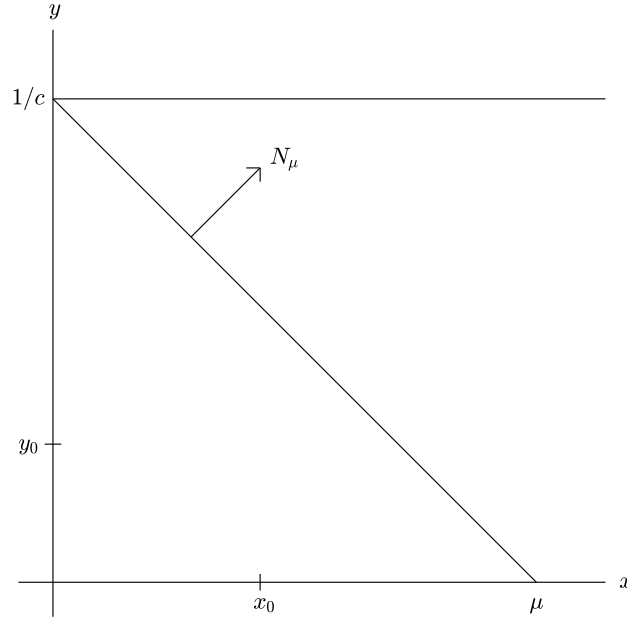
CASE IIC.  $Z_0 = cy_0/(1 - cy_0)$ .

Then  $\text{Tr } DV(x_0, y_0) = 0$ , so, by (4.13.49), 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  $\Phi^t$  for  $t \geq 0$ . Namely, consider the triangle  $\mathcal{T}_\mu$  with vertices at  $(0, 1/c)$ ,  $(0, 0)$ , and  $(\mu, 0)$ , as pictured in Figure 4.13.5.

**Claim.** If  $\mu > 0$  is large enough, the triangle  $\mathcal{T}_\mu$  is invariant under  $\Phi^t$ , for  $t \geq 0$ .

**Proof.** Note that  $V$  is vertical on the left edge of  $\mathcal{T}_\mu$ , with critical points at the



**Figure 4.13.5.** Invariant domain

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  $\mu$  is sufficiently large. This line segment is given by

$$(4.13.56) \quad x = \mu(1 - cy), \quad 0 \leq y \leq \frac{1}{c},$$

and the vector

$$(4.13.57) \quad N_\mu = \begin{pmatrix} 1 \\ \mu c \end{pmatrix}$$

is normal to this segment, and points away from  $\mathcal{T}_\mu$ . We want to show that  $V \cdot N_\mu \leq 0$  along this line segment, for  $\mu$  large. Indeed, from (4.13.40),

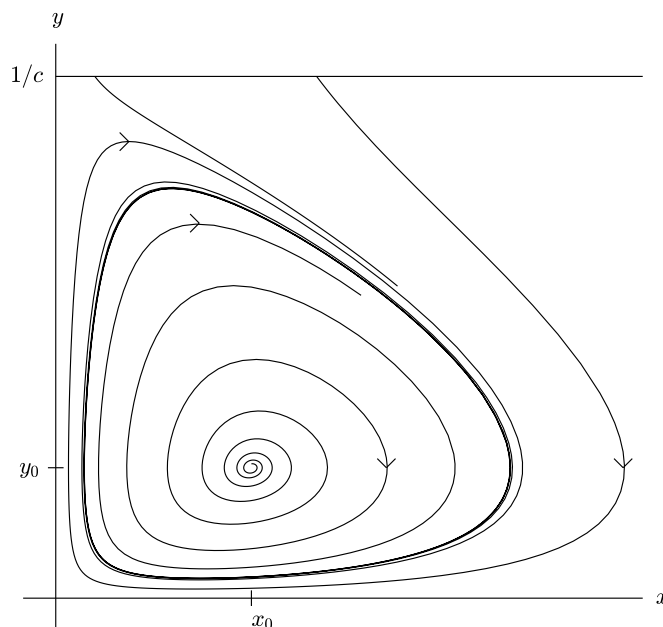
$$(4.13.58) \quad \begin{aligned} V(\mu(1 - cy), y) \cdot N_\mu &= (1 - cy)[\mu(b\zeta(y) - a) + \mu cry - \mu^2 c\zeta(y)] \\ &= \mu(1 - cy)[-a + cry - (\mu c - b)\zeta(y)], \end{aligned}$$

and under the hypotheses (4.13.39) on  $\zeta$ , this is

$$(4.13.59) \quad \leq 0, \quad \forall y \in \left[0, \frac{1}{c}\right],$$

if  $\mu$  is sufficiently large, say  $\mu \geq \mu_0$ . □

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  $t$ .



**Figure 4.13.6.** Second modification of the Volterra-Lotka system (Case IIB)

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

$$(4.13.60) \quad \text{in Case IIA, } \Phi^t(p) \longrightarrow (x_0, y_0), \quad \text{as } t \rightarrow +\infty,$$

for all  $p$  in the interior of  $\mathcal{R}$ , so the phase portrait has qualitative features similar to Figure 4.13.4. 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

$$(4.13.61) \quad \mathcal{T}_{\mu_0} \setminus U \text{ is invariant under } \Phi^t, \text{ for } t \geq 0.$$

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 4.12.2, shows that

$$(4.13.62) \quad \text{in Case IIB, } L_\omega(p) \text{ is a periodic orbit,}$$

for all  $p \neq (x_0, y_0)$  in the interior of  $\mathcal{R}$ . The phase portrait is depicted in Figure 4.13.6.



---

**Exercises**

Exercises 1–5 deal with the system (4.13.37), i.e.,

$$(4.13.63) \quad \begin{aligned} x' &= -ax + bx\zeta(y), \\ y' &= ry(1 - cy) - x\zeta(y), \end{aligned}$$

where  $\zeta(y)$  is given by (4.13.9), i.e.,

$$(4.13.64) \quad \zeta(y) = \frac{\kappa y}{1 + \gamma y}, \quad \frac{\kappa}{\gamma} = \beta.$$

As usual,  $a, b, c, \kappa, \gamma, r \in (0, \infty)$ . The exercises deal with when Cases I–III, specified below (4.13.44), hold. Recall these cases apply if and only if there is a critical point  $(x_0, y_0)$  given by (4.13.43), i.e., of and only if

$$(4.13.65) \quad \frac{a}{b} < \beta = \frac{\kappa}{\gamma}.$$

We will assume this holds.

1. Show that the critical point  $(x_0, y_0)$  is given by

$$(4.13.66) \quad y_0 = \frac{a}{b\kappa - a\gamma}, \quad x_0 = \frac{b}{a}ry_0(1 - cy_0).$$

2. Show that

$$\begin{aligned} \text{Case I} &\iff ac > b\kappa - a\gamma, \\ \text{Case II} &\iff ac < b\kappa - a\gamma, \\ \text{Case III} &\iff ac = b\kappa - a\gamma. \end{aligned}$$

3. Let  $Z_0$  be given by (4.13.50), i.e.,

$$(4.13.67) \quad Z_0 = 1 - \frac{\zeta'(y_0)y_0}{\zeta(y_0)}.$$

Show that

$$(4.13.68) \quad Z_0 = \frac{a\gamma}{b\kappa}.$$

4. In Case II, recall Cases IIA–IIC, specified below (4.13.53). Show that

$$\begin{aligned} \text{Case IIA} &\iff \frac{\gamma}{\kappa} < \frac{c}{b\kappa - a\gamma - ac}, \\ \text{Case IIB} &\iff \frac{\gamma}{\kappa} > \frac{c}{b\kappa - a\gamma - ac}, \\ \text{Case IIC} &\iff \frac{\gamma}{\kappa} = \frac{c}{b\kappa - a\gamma - ac}, \end{aligned}$$

5. Let us take

$$(4.13.69) \quad a = 1, \quad b = 2, \quad \kappa = 1, \quad \gamma = 1.$$

Note that (4.13.65) holds. Show that

$$\text{Case I} \iff c > 1,$$

$$\text{Case II} \iff c < 1,$$

$$\text{Case III} \iff c = 1.$$

In Case II, show that

$$\text{Case IIA} \iff c > \frac{1}{3},$$

$$\text{Case IIB} \iff c < \frac{1}{3},$$

$$\text{Case IIC} \iff c = \frac{1}{3}.$$

Figure 4.13.6 was produced using the parameters in (4.13.69), together with  $c = 1/4$ ,  $r = 1$ .

Exercises 6–10 deal with the system (4.13.63), where  $\zeta(y)$  is given by (4.13.10), i.e.,

$$(4.13.70) \quad \zeta(y) = \beta(1 - e^{-\gamma y}), \quad \beta\gamma = \kappa.$$

Again there is a critical point  $(x_0, y_0)$ , given by (4.13.43), if and only if (4.13.65) holds. We assume this holds, so  $b\beta > a$ .

6. Show that the critical point  $(x_0, y_0)$  is given by

$$(4.13.71) \quad y_0 = \frac{1}{\gamma} \log \frac{b\beta}{b\beta - a}, \quad x_0 = \frac{b}{a} r y_0 (1 - c y_0).$$

7. For  $Z_0$ , defined by (4.13.67), show that

$$(4.13.72) \quad Z_0 = 1 - \frac{b\beta - a}{a} \log \frac{b\beta}{b\beta - a}.$$

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  $\gamma$  as in (4.13.69). Work out a parallel to Exercise 5.

For Exercises 11–12, consider the following system, for  $x$  predators and  $y$  prey, presented in [44], p. 376:

$$(4.13.73) \quad \begin{aligned} x' &= ax \left( b - \frac{x}{y} \right), \\ y' &= ry(1 - cy) - x\zeta(y). \end{aligned}$$

Here the equation for  $y$  is as in (4.13.63), 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 (4.13.73) as parallel as possible to the analysis done in this section for (4.13.63).

12. Take  $\zeta(y)$  as in (4.13.64) and work out results parallel to those of Exercises 1–5.

Exercises 13–15 are for readers who can use numerical software, with graphics capabilities.

13. The following system is known as the basic model of virus dynamics (cf. [34], p. 100, [52], p. 26):

$$(4.13.74) \quad \begin{aligned} \frac{dx}{dt} &= \lambda - dx - \beta xv, \\ \frac{dy}{dt} &= \beta xv - ay, \\ \frac{dv}{dt} &= ky - uv. \end{aligned}$$

Here,  $x$  represents the uninfected cell population,  $y$  the infected cell population, and  $v$  the virus population. The positive parameters  $\lambda, d, \beta, a, k$ , and  $u$  are taken to be constant. The ratio

$$(4.13.75) \quad R_0 = \frac{\lambda\beta k}{adu}$$

is called the basic reproductive ratio. Graph solution curves for (4.13.74), 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

$$(4.13.76) \quad \begin{aligned} \frac{dx}{dt} &= \lambda - dx - bxy, \\ \frac{dy}{dt} &= -ay + bxy. \end{aligned}$$

Study this system, with an eye to comparison with the Volterra-Lotka system (4.13.11). Here, replace (4.13.75) by

$$(4.13.77) \quad R_0 = \frac{b\lambda}{ad}.$$

15. The following system modifies (4.13.76) by introducing  $z(t)$ , the population of

“killer T cells,” which kill off infected cells, thereby negatively affecting  $y$ :

$$(4.13.78) \quad \begin{aligned} \frac{dx}{dt} &= \lambda - dx - bxy, \\ \frac{dy}{dt} &= bxy - ay - pyz, \\ \frac{dz}{dt} &= cyz - bz, \end{aligned}$$

now with positive parameters  $\lambda, d, b, a, p$ , and  $c$ . Continue to define  $R_0$  by (4.13.77). Consider particularly cases where

$$(4.13.79) \quad R_0 > 1, \quad c\left(\frac{\lambda}{a} - \frac{d}{b}\right) > b.$$

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 [34] and [52], and references therein.

#### 4.14. Competing species equations

The following system models the populations  $x(t)$  and  $y(t)$  of two competing species:

$$(4.14.1) \quad \begin{aligned} \frac{dx}{dt} &= ax(1 - bx) - cxy, \\ \frac{dy}{dt} &= \alpha y(1 - \beta y) - \gamma xy. \end{aligned}$$

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 (4.14.1), 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

$$(4.14.2) \quad \begin{aligned} \frac{dx}{dt} &= ax(1 - x) - cxy, \\ \frac{dy}{dt} &= \alpha y(1 - y) - \gamma xy. \end{aligned}$$

which we will consider henceforth. We call this system CSE. We take  $a, c, \alpha, \gamma \in (0, \infty)$ . Associated to this system is the vector field

$$(4.14.3) \quad V = \begin{pmatrix} ax(1 - x) - cxy \\ \alpha y(1 - y) - \gamma xy \end{pmatrix}.$$

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  $\Phi^t$  generated by  $V$ . Hence the quadrant  $\{x \geq 0, y \geq 0\}$ , which is the region of biological significance, is invariant under  $\Phi^t$ . Note also that

$$(4.14.4) \quad V(x, 1) = \begin{pmatrix} ax(1 - x) - cx \\ -\gamma x \end{pmatrix}, \quad V(1, y) = \begin{pmatrix} -cy \\ \alpha y(1 - y) - \gamma y \end{pmatrix},$$

so  $\Phi^t$  leaves invariant the region

$$(4.14.5) \quad \mathcal{B} = \{(x, y) : 0 \leq x, y \leq 1\},$$

for  $t \geq 0$ .

The vector field  $V$  has the following critical points,

$$(4.14.6) \quad (0, 0), \quad (0, 1), \quad (1, 0),$$

and a fourth critical point  $(x_0, y_0)$ , satisfying

$$(4.14.7) \quad cy_0 = a(1 - x_0), \quad \gamma x_0 = \alpha(1 - y_0).$$

A calculation gives

$$(4.14.8) \quad x_0 = \alpha \frac{a - c}{a\alpha - c\gamma}, \quad y_0 = a \frac{\alpha - \gamma}{a\alpha - c\gamma}.$$

The point  $(x_0, y_0)$  may or may not lie in the first quadrant. We investigate this further below.

We have

$$(4.14.9) \quad DV(0, 0) = \begin{pmatrix} a & 0 \\ 0 & \alpha \end{pmatrix},$$

so  $(0, 0)$  is a source. Also,

$$(4.14.10) \quad DV(0, 1) = \begin{pmatrix} a - c & 0 \\ -\gamma & -\alpha \end{pmatrix}, \quad DV(1, 0) = \begin{pmatrix} -a & -c \\ 0 & \alpha - \gamma \end{pmatrix},$$

and each of these might be a saddle or a sink, depending on the signs of  $a - c$  and  $\alpha - \gamma$ . Next,

$$(4.14.11) \quad \begin{aligned} DV(x_0, y_0) &= \begin{pmatrix} a(1 - 2x_0) - cy_0 & -cx_0 \\ -\gamma y_0 & \alpha(1 - 2y_0) - \gamma x_0 \end{pmatrix} \\ &= \begin{pmatrix} -ax_0 & -cx_0 \\ -\gamma y_0 & -\alpha y_0 \end{pmatrix}, \end{aligned}$$

the second identity by (4.14.7). Hence

$$(4.14.12) \quad \begin{aligned} \det DV(x_0, y_0) &= (a\alpha - c\gamma)x_0y_0, \\ \text{Tr } DV(x_0, y_0) &= -ax_0 - \alpha y_0. \end{aligned}$$

At this point, it is natural to consider the following cases of CSE:

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 (4.14.10) that

$$(4.14.13) \quad (0, 1) \text{ and } (1, 0) \text{ are saddles.}$$

In this case,  $a\alpha > c\gamma$ , so, by (4.14.8),

$$(4.14.14) \quad x_0 > 0, \quad y_0 > 0,$$

and the critical point  $(x_0, y_0)$  is in the first quadrant. Then we see from (4.14.12) that

$$(4.14.15) \quad \det DV(x_0, y_0) > 0, \quad \text{Tr } DV(x_0, y_0) < 0,$$

so

$$(4.14.16) \quad (x_0, y_0) \text{ is a sink.}$$

We have

$$(4.14.17) \quad \Phi^t(x, y) \longrightarrow (x_0, y_0) \text{ as } t \rightarrow +\infty,$$

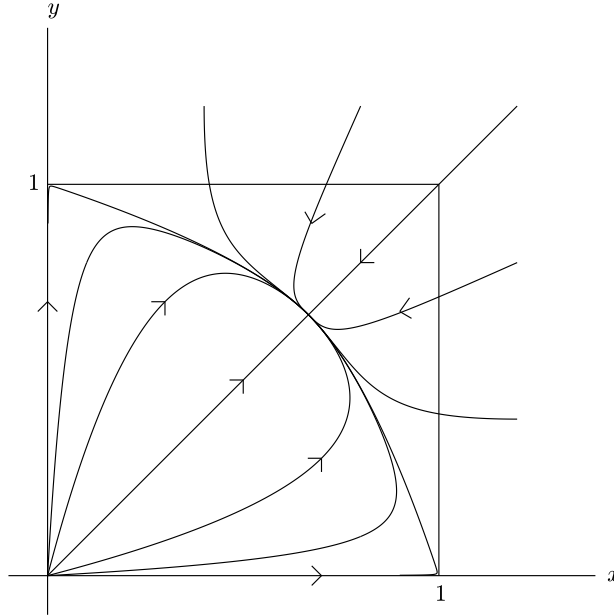
whenever  $x > 0$  and  $y > 0$ . The two competing species tend to an equilibrium of coexistence. The phase portrait for this case, with  $a = 2, \alpha = 2, c = 1, \gamma = 1$ , is illustrated in Figure 4.14.1.

In Case II, we see from (4.14.10) that

$$(4.14.18) \quad (0, 1) \text{ and } (1, 0) \text{ are sinks.}$$

In this case,  $a\alpha < c\gamma$ , so, by (4.14.8), again (4.14.14) holds, and the critical point  $(x_0, y_0)$  is in the first quadrant. We see from (4.14.12) that

$$(4.14.19) \quad \det DV(x_0, y_0) < 0,$$



**Figure 4.14.1.** Case I of CSE

so

$$(4.14.20) \quad (x_0, y_0) \text{ is a saddle.}$$

The phase portrait for this case, with  $a = 1, \alpha = 1, c = 2, \gamma = 2$ , is illustrated in Figure 4.14.2. For almost all initial data  $(x, y)$  in the first quadrant,  $\Phi^t(x, y)$  tends to either  $(0, 1)$  or  $(1, 0)$  as  $t \rightarrow +\infty$ . One species or the other tends toward extinction, depending on the initial conditions.

In Case III, we see from (4.14.10) that

$$(4.14.21) \quad (0, 1) \text{ is a saddle and } (1, 0) \text{ is a sink.}$$

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 (4.14.8),

$$(4.14.22) \quad x_0 > 0, \quad y_0 < 0,$$

so the critical point  $(x_0, y_0)$  is not in the first quadrant. We see from (4.14.12) that

$$(4.14.23) \quad \det DV(x_0, y_0) < 0,$$

so

$$(4.14.24) \quad (x_0, y_0) \text{ is a saddle.}$$

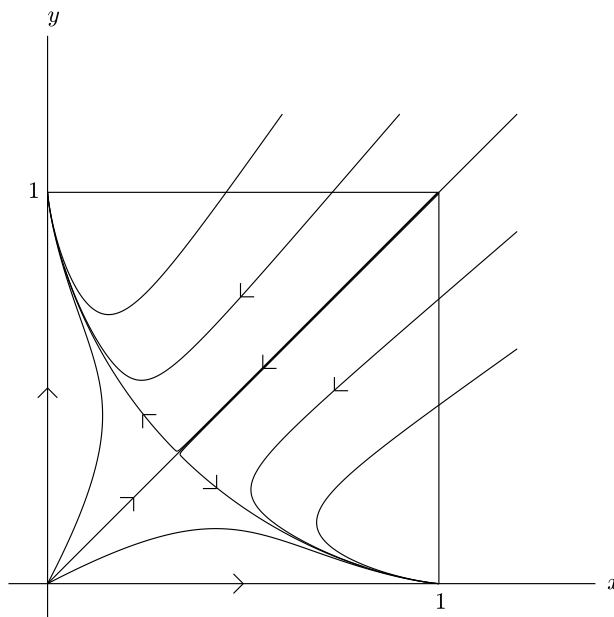


Figure 4.14.2. Case II of CSE

The phase portrait for this case, with  $a = 2, \alpha = 1, c = 1/4, \gamma = 2$ , is illustrated in Figure 4.14.3. We have

$$(4.14.25) \quad \Phi^t(x, y) \longrightarrow (1, 0) \text{ as } t \rightarrow +\infty,$$

whenever  $x > 0$  and  $y > 0$ . Species  $y$  tends to extinction.

CASE IIIB.  $a\alpha < c\gamma$ .

This time, by (4.14.8),

$$(4.14.26) \quad x_0 < 0, \quad y_0 > 0,$$

and again the critical point  $(x_0, y_0)$  is not in the first quadrant. We see from (4.14.22) that

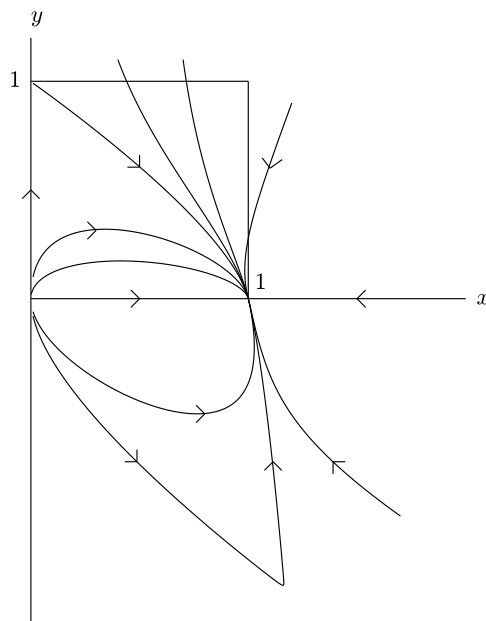
$$(4.14.27) \quad \det DV(x_0, y_0) > 0.$$

Thus

$$(4.14.28) \quad (x_0, y_0) \text{ is a source or a sink,}$$

depending on the sign of  $\text{Tr } DV(x_0, y_0)$ . The phase portrait for this case, with  $a = 2, \alpha = 1/2, c = 1, \gamma = 2$ , is illustrated in Figure 4.14.4. (In this example,  $(x_0, y_0)$  is a sink.) Again (4.14.25) holds whenever  $x > 0$  and  $y > 0$ .





**Figure 4.14.3.** Case IIIA of CSE

To summarize Case III, the flows in the first quadrant have the same qualitative features in the two sub-cases; (4.14.25) 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 (4.14.2), 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

$$x, y \geq 0, x^2 + y^2 \geq R^2 \implies \frac{d}{dt}(x + y) \leq 0.$$

Deduce global existence of solutions to (4.14.2), for  $t \geq 0$ , given  $(x(0), y(0))$  in the first quadrant.

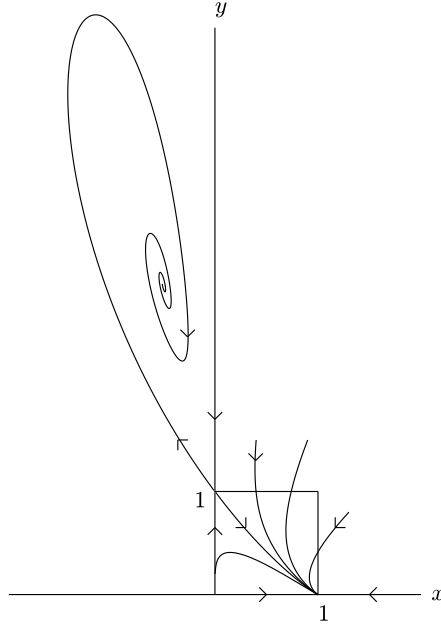


Figure 4.14.4. Case IIIB of CSE

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 (4.14.5), 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

$$\frac{dx}{dt} = ax(1-x) - xy,$$

$$\frac{dy}{dt} = 2y(1-y) - xy,$$

with  $a \in (0, \infty)$ . Specify when Cases I–IV hold. Record the possible outcomes, as regards coexistence/extinction.

### 4.15. 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 §4.12. 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  $2 \times 2$  system of differential equations, which is converted to a  $4 \times 4$  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 §§4.5–4.6. 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 §4.9, 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 devote this section to a discussion of two special cases of  $3 \times 3$  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:

$$(4.15.1) \quad \begin{aligned} x' &= \sigma(y - x), \\ y' &= rx - y - xz, \\ z' &= xy - bz. \end{aligned}$$

An alternative presentation is

$$(4.15.2) \quad \frac{d}{dt} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -\sigma & \sigma & 0 \\ r & -1 & 0 \\ 0 & 0 & -b \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -x \\ 0 & x & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

Denoting the right side of (4.15.2) by  $V(x, y, z)$ , we see that the first matrix on the right side is  $DV(0, 0, 0)$ . One assumes the parameters  $\sigma, b$ , and  $r$  are all positive.

Lorenz took

$$(4.15.3) \quad \sigma = 10, \quad b = \frac{8}{3},$$

and considered various values of  $r$ , with emphasis on

$$(4.15.4) \quad r = 28.$$

Phase portraits of some orbits for (4.15.1), with  $\sigma$  and  $b$  given by (4.15.3) and with various values of  $r$  are given in Figure 4.15.1. Each of the four portraits depicts the forward orbits through the point

$$(4.15.5) \quad x = \frac{1}{100}, \quad y = 0, \quad z = 5.$$

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 (4.15.1) can be established with the help of the remarkable function

$$(4.15.6) \quad f(x, y, z) = rx^2 + \sigma y^2 + \sigma(z - 2r)^2.$$

A calculation shows that if  $(x(t), y(t), z(t))$  solves (4.15.1), then

$$(4.15.7) \quad \frac{d}{dt}f(x, y, z) = -2\sigma(rx^2 + y^2 + bz^2 - 2brz).$$

Clearly there exists  $K \in (0, \infty)$  such that

$$(4.15.8) \quad B = \{(x, y, z) \in \mathbb{R}^3 : f(x, y, z) \leq K\}$$

is a closed, bounded subset of  $\mathbb{R}^3$  and the right side of (4.15.7) is  $< 0$  on the complement of  $B$ . Hence

$$(4.15.9) \quad \Phi^t(B) \subset B, \quad \forall t > 0,$$

where  $\Phi^t$  is the flow generated by  $V(x, y, z)$ . Moreover, for each  $(x, y, z) \in \mathbb{R}^3$ ,

$$(4.15.10) \quad \Phi^t(x, y, z) \in B, \quad \text{for all sufficiently large } t > 0.$$

Note that (4.15.9) plus the identity  $\Phi^t = \Phi^s \circ \Phi^{t-s}$  implies

$$(4.15.11) \quad \Phi^t(B) \subset \Phi^s(B) \quad \text{for } 0 < s < t,$$

so  $B(t) = \Phi^t(B)$  is a family of closed, bounded sets that is decreasing as  $t \nearrow +\infty$ . Now set

$$(4.15.12) \quad \mathcal{B} = \bigcap_{t \in \mathbb{R}^+} B(t) = \bigcap_{k \in \mathbb{Z}^+} B(k).$$

The set  $\mathcal{B}$  is called the *attractor* for (4.15.1). We have

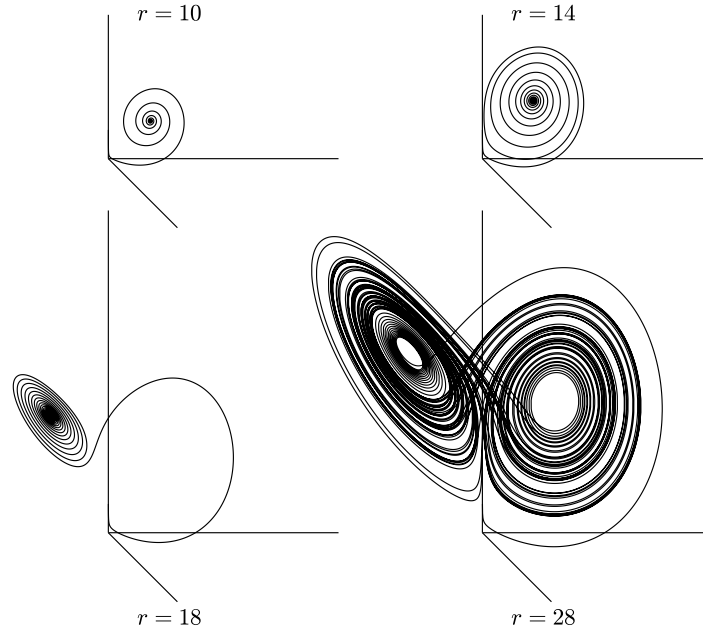
$$(4.15.13) \quad \Phi^t(\mathcal{B}) = \mathcal{B}, \quad \forall t \geq 0.$$

Note that

$$(4.15.14) \quad \operatorname{div} V = -\sigma - 1 - b < 0,$$

so results of §3 imply

$$(4.15.15) \quad \operatorname{Vol} \mathcal{B} = 0.$$



**Figure 4.15.1.** Lorenz system,  $r = 10, 14, 18, 28$

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 (4.15.2), we see its eigenvalues are

$$(4.15.16) \quad \lambda_{\pm} = -\frac{\sigma+1}{2} \pm \frac{1}{2}\sqrt{(\sigma+1)^2 + 4\sigma(r-1)}, \quad \lambda_3 = -b,$$

with eigenvectors

$$(4.15.17) \quad v_{\pm} = \begin{pmatrix} \sigma \\ \lambda_{\pm} + \sigma \\ 0 \end{pmatrix}, \quad v_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

It follows from (4.15.16) that

$$(4.15.18) \quad \begin{aligned} 0 < r < 1 &\implies DV(0) \text{ has 3 negative eigenvalues,} \\ r > 1 &\implies DV(0) \text{ has 2 negative and one positive eigenvalue.} \end{aligned}$$

For  $r > 1$ , the positive eigenvalue is  $\lambda_+$  and its associated eigenvector is  $v_+$ . There is a parallel to the results in (4.3.38) describing saddles. It is shown in [19] that there is a smooth 2-dimensional surface through the origin consisting of points  $p$  such that  $\Phi^t(p) \rightarrow 0$  as  $t \rightarrow +\infty$  and a smooth 1-dimensional curve through the origin consisting of points  $p$  such that  $\Phi^t(p) \rightarrow 0$  as  $t \rightarrow -\infty$ . 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 Appendix 4.C for further discussion.

For  $r > 1$ ,  $V$  has two additional critical points, satisfying

$$(4.15.19) \quad x = y, \quad (r - 1 - z)x = 0, \quad bz = x^2,$$

i.e.,

$$(4.15.20) \quad C_{\pm} = (\pm\sqrt{b(r-1)}, \pm\sqrt{b(r-1)}, r-1).$$

We have

$$(4.15.21) \quad DV(C_{\pm}) = \begin{pmatrix} -\sigma & \sigma & 0 \\ 1 & -1 & \pm\xi \\ \pm\xi & \pm\xi & -b \end{pmatrix}, \quad \xi = \sqrt{b(r-1)}.$$

Note that  $DV(C_+)$  and  $DV(C_-)$  are conjugate by the action of

$$(4.15.22) \quad \begin{pmatrix} -1 & & \\ & -1 & \\ & & 1 \end{pmatrix},$$

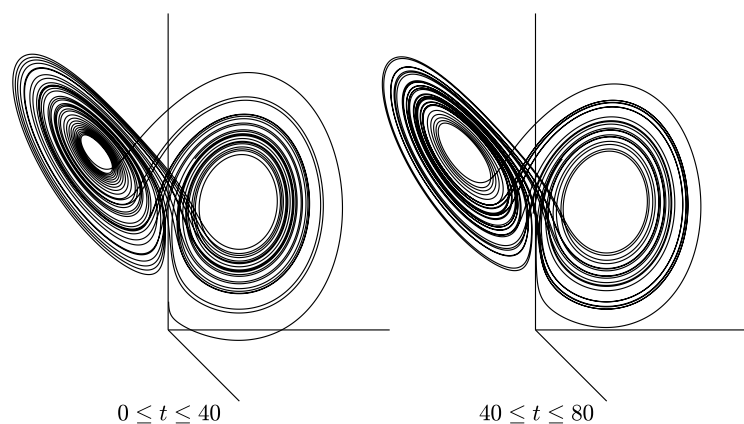
so they have the same eigenvalues. This mirrors the fact that (4.15.1) is invariant under the transformation  $(x, y, z) \mapsto (-x, -y, z)$ . Further calculations give the following results, when  $\sigma$  and  $b$  are given by (4.15.3):

$$(4.15.23) \quad \begin{array}{l} DV(C_{\pm}) \text{ has} \\ 3 \text{ negative eigenvalues for } 1 < r < 1.346 \dots \\ 1 \text{ negative and 2 with negative real part for } 1.346 \dots < r < 24.74 \dots \\ 1 \text{ negative and 2 with positive real part for } r > 24.74 \dots \end{array}$$

In the first two cases in (4.15.23), Proposition 4.3.4 applies, and for all points  $p$  sufficiently close to  $C_+$ ,  $\Phi^t(p) \rightarrow C_+$  as  $t \rightarrow +\infty$ , and similarly for  $C_-$ . The third case in (4.15.23) is like the second case in (4.15.18), 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 (4.15.18).

With these calculations in hand, let's take a closer look at the four phase portraits depicted in Figure 4.15.1, orbits with initial data given by (4.15.5). We see from (4.15.1) 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 point  $(0.01, 0, 5)$  is 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  $v_+$ , where  $v_+$  is given by (4.15.17). The orbit from  $(+0.01, 0, 5)$  spirals into the critical point  $C_+$ , in the first two portraits, where  $r = 10$  and  $14$ . Similarly, the orbit from  $(-0.01, 0, 5)$  spirals into  $C_-$ ,

Around  $r \approx 14$ , something new happens. These orbits pass close to the origin. At a certain critical value  $r_h \approx 14$ , the unstable manifold is actually a pair of homoclinic orbits, approaching the origin both as  $t \rightarrow -\infty$  and as  $t \rightarrow +\infty$ . For larger values of  $r$ , the orbit from  $(+0.01, 0, 5)$  crosses over and spirals into  $C_-$ , as



**Figure 4.15.2.** Lorenz system,  $r = 28$ , for two ranges of  $t$

depicted in the third portrait in Figure 4.15.1, for  $r = 18$ . Similarly, the orbit from  $(-0.01, 0, 5)$  spirals into  $C_+$ .

This spiraling into  $C_{\pm}$  does not endure as  $r$  increases. As stated in (4.15.23), there is a critical  $r_c \approx 24.74$  past which  $DV(C_{\pm})$  has two eigenvalues with positive rather than negative real part. In the fourth phase portrait of Figure 4.15.1, we have  $r = 28 > r_c$ . The orbit starting from  $(0.01, 0, 5)$  approaches the unstable manifold of  $C_-$  and then spirals out from this critical point. After some spiraling out from  $C_-$ , this orbit 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 two phase portraits in Figure 4.15.2 show

$$(4.15.24) \quad \Phi^t(0.08, 0, 5), \quad 40j < t < 40(j+1), \quad j = 0, 1.$$

The portraits differ in fine detail from each other, but they are fairly similar, and seem to reveal what is called a strange attractor.

We make one further comment about Figures 4.15.1–4.15.2. 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:

$$(4.15.25) \quad \frac{d^2x}{dt^2} = f(x) + r \cos t.$$

Here  $r$  is a parameter. When converted to a first order system and put in autonomous form, this becomes

$$(4.15.26) \quad \begin{aligned} \frac{dx}{dt} &= y, \\ \frac{dy}{dt} &= f(x) + r \cos z, \\ \frac{dz}{dt} &= 1. \end{aligned}$$

We take

$$(4.15.27) \quad f(x) = x - x^3.$$

The equation (4.15.25) is called a periodically forced Duffing equation if  $r \neq 0$ .

For  $r = 0$ , (4.15.25) is called Duffing's equation, and it reduces to a  $2 \times 2$  system, whose phase portrait is given in Figure 4.15.3. We take orbits through the points

$$(4.15.28) \quad x = \sqrt{2} + \frac{3k}{10}, \quad k = -3, -2, 0, 1, \quad y = 0,$$

and their mirror images about the  $y$ -axis. There are two homoclinic orbits, each tending to the origin as  $t \rightarrow \pm\infty$ . All the other orbits are closed, and lie on level curves of

$$(4.15.29) \quad E(x, y) = \frac{y^2}{2} - \frac{x^2}{2} + \frac{x^4}{4}.$$

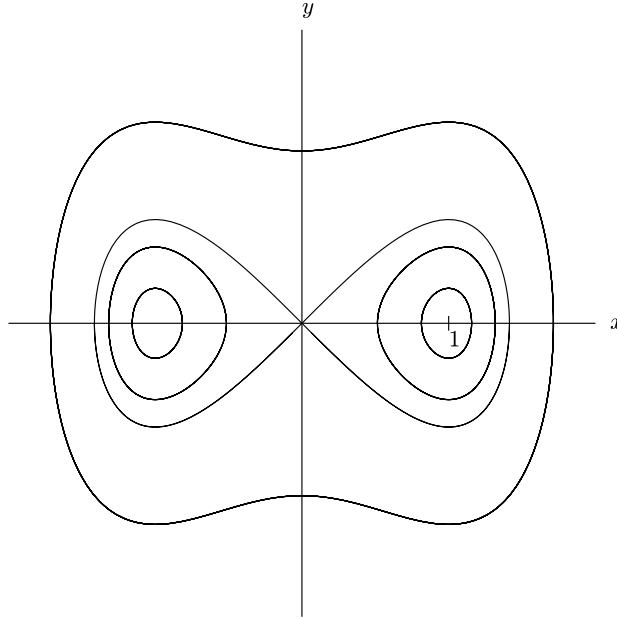
For  $r \neq 0$ , matters are more complicated, since  $z$  is coupled to  $(x, y)$  in (4.15.26). We need a different way to portray the orbits  $(x(t), y(t), z(t))$ . In this case, unlike for the Lorenz system, a linear projection of  $(x, y, z)$  space onto  $(u, v)$  space is not the best way to proceed. Taking into account the periodicity of the right side of (4.15.26) 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 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 Figures 4.15.4–4.15.6, we draw curves  $(u(t), v(t))$ , with

$$(4.15.30) \quad u(t) = (x(t) + 2) \left( \sin t - \frac{\cos t}{2} \right), \quad v(t) = y(t) - (x(t) + 2) \frac{\cos t}{2}.$$

For initial data, we take  $x$  and  $y$  as in (4.15.28) and  $z = 0$ . We use a fourth order Runge-Kutta scheme.



**Figure 4.15.3.** Phase portrait for Duffing's equation

Figure 4.15.4 draws such curves when  $(x, y, z)$  solve (4.15.26) with  $r = 0$ . In all but the third portrait, the orbits lie on smooth donut-shaped surfaces (called tori). The third 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 4.15.5 gives this representation of orbits of (4.15.26), with

$$(4.15.31) \quad r = 0.1.$$

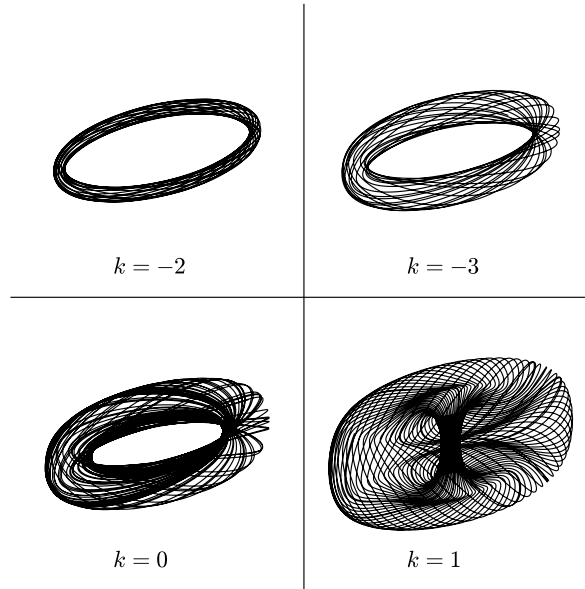
One of the four orbits seems to lie on a smooth torus (somewhat deformed). The other three are all apparently a mess, and also, apparently, about the same mess. In Figure 4.15.6 we present an enlarged version of one such orbit, with initial data

$$(4.15.32) \quad x = \sqrt{2}, \quad y = 0,$$

An alternative to depicting orbits of the system (4.15.26)–(4.15.27) is to depict orbits of the associated *Poincaré map*, characterized as follows. Take an initial point  $p = (x_0, y_0, 0)$ . Solve (4.15.26) 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

$$(4.15.33) \quad (x(0), y(0)) \mapsto (x(2\pi), y(2\pi)).$$

This is the Poincaré map associated to the system (4.15.26).



**Figure 4.15.4.** 3D orbits for unforced Duffing's equation,  $x = \sqrt{2} + 3k/10$

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

$$(4.15.34) \quad P: \mathcal{O} \longrightarrow S,$$

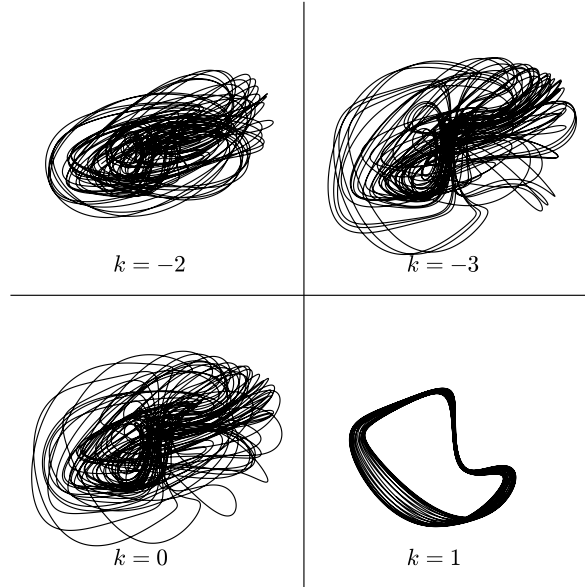
defined on an open subset  $\mathcal{O} \subset S$ , where  $p \in \mathcal{O}$  and  $P(p) = q$  is the point  $\Phi_X^t(p)$  with smallest  $t > 0$  such that  $\Phi_X^t(p) \in S$ . See Figure 4.15.7.

In the setting of (4.15.26), (4.15.33), the orbits for Poincaré map  $(x(0), y(0)) = (x, 0)$ , with

$$(4.15.35) \quad x = -1.1, 1.3, \sqrt{2}, 1.85, \text{ and } 2.1,$$

are presented in Figure 4.15.8. These five initial data give rise to five orbits for the Poincaré map. Of these, four seem to lie along smooth curves. The orbit through  $(x, y) = (\sqrt{2}, 0)$  populates the fuzzy grey area, formed from 20,000 points in the orbit of the Poincaré map (or rather an approximation via a Runge-Kutta difference scheme). Note that  $(x, y) = (\sqrt{2}, 0)$  is the initial datum leading to Figure 4.15.6.

The appearance of smooth closed curves  $\gamma_j$ , invariant under the Poincaré map, suggests the following.



**Figure 4.15.5.** Orbits for forced Duffing's equation,  $x = \sqrt{2} + 3k/10$

**Assertion.** Each  $\gamma_j$  bounds a region  $\bar{\Omega}_j \subset \mathbb{R}^2$ , smoothly equivalent to the disk

$$(4.15.36) \quad \bar{D} = \{x \in \mathbb{R}^2 : \|x\| \leq 1\},$$

that is to say, there are smooth one-to-one maps  $\varphi_j : \bar{\Omega}_j \rightarrow \bar{D}$  with smooth inverses  $\varphi_j^{-1} : \bar{D} \rightarrow \bar{\Omega}_j$ , and the Poincaré map takes  $\bar{\Omega}_j$  into itself, i.e.,

$$(4.15.37) \quad P : \bar{\Omega}_j \longrightarrow \bar{\Omega}_j.$$

Granted this, we can make use of the following result, known as Brouwer's fixed-point theorem.

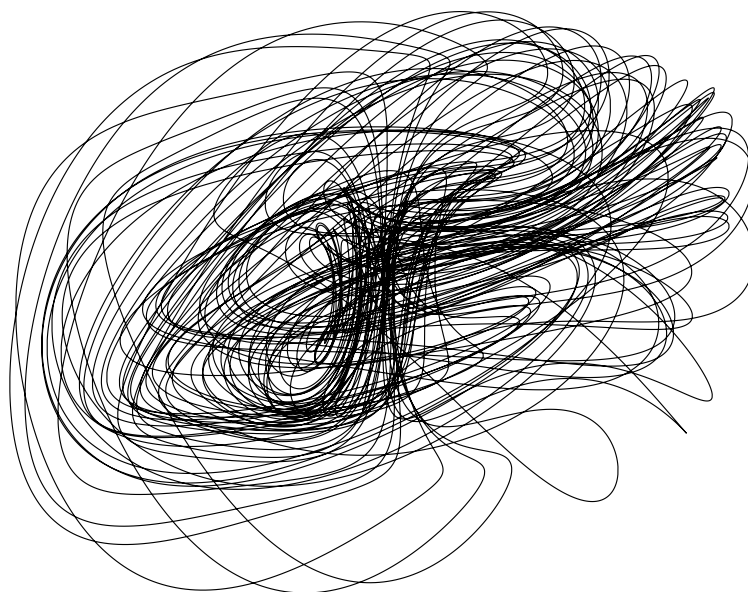
**Theorem.** Each smooth map

$$(4.15.38) \quad \psi : \bar{D} \longrightarrow \bar{D}$$

has a fixed point, i.e., there exists  $p \in \bar{D}$  such that  $\psi(p) = p$ .

See Appendix 4.G for a proof of this result. Given the assertion above, we can take  $\psi = \varphi_j \circ P \circ \varphi_j^{-1}$  and conclude that

$$(4.15.39) \quad P(q) = q, \quad q = \varphi_j^{-1}(p).$$



**Figure 4.15.6.** Orbit for forced Duffing's equation,  $x = \sqrt{2}$

Such fixed points of the Poincaré map give rise to periodic solutions to the associated systems of differential equations (in this case, (4.15.25)). 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 [16], which treats four examples, including the Lorenz system and the forced Duffing system. Other material on chaotic systems can be found in [2], [6], [18], [23], [25], and [30]. A detailed study of the Lorenz system is given in [41].

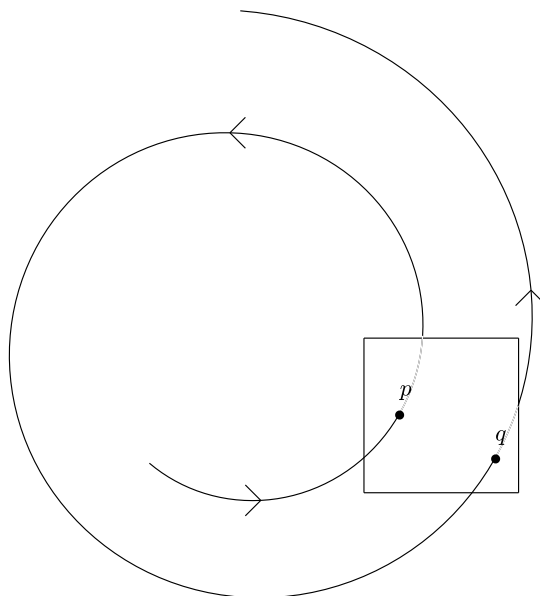


Figure 4.15.7. Poincaré map

## Exercises

1. Consider the double pendulum system, in the limit  $m_2 = 0$ , given by (4.9.35). Substitute

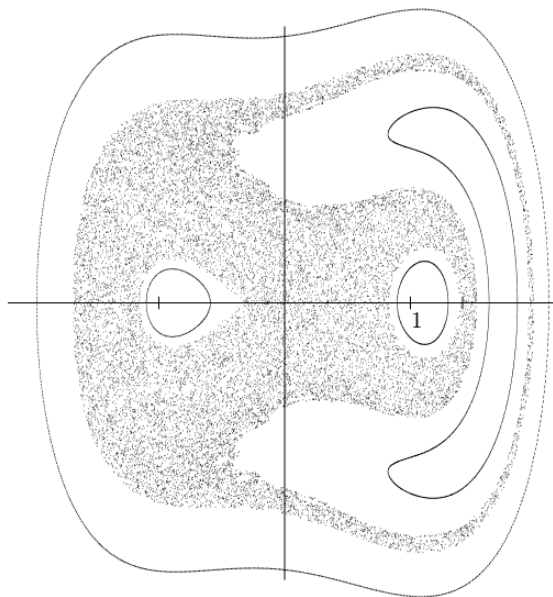
$$(4.15.40) \quad \theta_1(t) \approx r \cos \omega t, \quad \omega = \sqrt{\frac{g}{\ell_1}}$$

into the second equation in (4.9.35), expand in powers of  $r$ , and throw away terms containing second and higher powers of  $r$ . Show that you get

$$(4.15.41) \quad \theta_2''(t) + \frac{g}{\ell_2} \sin \theta_2(t) = r \omega^2 \frac{\ell_1}{\ell_2} \cos \theta_2(t) \cos \omega t.$$

Exercises 2–10 are for readers who can use numerical software, with graphics capabilities.

2. Write a program to exhibit solution curves of (4.15.41), in a fashion analogous to the treatment of (4.15.25), involving an analogue of (4.15.26). Try various values of  $r, g/\ell_2$ , etc., and see when the behavior is more chaotic or less chaotic.



**Figure 4.15.8.** Poincaré map for forced Duffing,  $x = -1.1, 1.3, \sqrt{2}, 1.85, 2.1$

3. In place of (4.15.41), consider the periodically forced pendulum equation

$$(4.15.42) \quad \theta''(t) + \frac{g}{\ell} \sin \theta(t) = \rho \cos \omega t,$$

and write a program to exhibit solution curves of this equation, in the spirit of Exercise 2.

4. Write a program to exhibit solutions to the full double pendulum system (4.9.15)–(4.9.16). Take, e.g.,  $m_1 = m_2 = 1, \ell_1 = \ell_2 = 1$ , and variants.

5. 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. Also consider other forcing periods, i.e., replace  $\cos t$  by  $\cos \omega t$ .

Exercises 6–9 deal with systems of the form

$$(4.15.43) \quad \frac{d^2}{dt^2} \begin{pmatrix} x \\ y \end{pmatrix} = -\nabla V(x, y).$$

These are  $2 \times 2$  second order systems, which convert to  $4 \times 4$  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.

6. Take

$$V(x, y) = x^2 + axy + y^4.$$

Try various  $a \in [0, 10]$ .

7. Take

$$V(x, y) = x^4 + axy + y^4, \quad a \in [-2, 2].$$

8. Take

$$V(x, y) = x^4 + ax^2y + y^4, \quad a \in [-1, 1].$$

9. Take

$$V(x, y) = \frac{1}{2}(x^2 + y^2) + a(x^4 - x^2y + y^4), \quad a \in [0, 1].$$

10. Taking off from models in §§4.13–4.14, see if you can construct models of interactions of 3 species that exhibit chaotic behavior.



#### 4.A. The derivative in several variables

Here we present basic definitions and results on multivariable differential calculus, useful for the material in Chapter 4. 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  $F : \mathcal{O} \rightarrow \mathbb{R}^m$  a continuous function. We say  $F$  is differentiable at a point  $x \in \mathcal{O}$ , with derivative  $L$ , if  $L : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is a linear transformation such that, for  $y \in \mathbb{R}^n$ , small,

$$(4.A.1) \quad F(x + y) = F(x) + Ly + R(x, y)$$

with  $\|R(x, y)\| = o(\|y\|)$ , i.e.,

$$(4.A.2) \quad \frac{\|R(x, y)\|}{\|y\|} \rightarrow 0 \quad \text{as } y \rightarrow 0.$$

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,

$$(4.A.3) \quad DF(x) = \left( \frac{\partial F_j}{\partial x_k} \right),$$

so that, if  $v = (v_1, \dots, v_n)^t$ , (regarded as a column vector) then

$$(4.A.4) \quad DF(x)v = \left( \sum_k \frac{\partial F_1}{\partial x_k} v_k, \dots, \sum_k \frac{\partial F_m}{\partial x_k} v_k \right)^t.$$

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  $\leq k$  exist and are continuous. If  $F$  is  $C^k$  for all  $k$ , we say  $F$  is  $C^\infty$ .

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  $F : \mathcal{O} \rightarrow \mathbb{R}^m$  be differentiable at  $x \in \mathcal{O}$ , as above, let  $U$  be a neighborhood of  $z = F(x)$  in  $\mathbb{R}^m$ , and let  $G : U \rightarrow \mathbb{R}^k$  be differentiable at  $z$ . Consider  $H = G \circ F$ . We have

$$(4.A.5) \quad \begin{aligned} H(x + y) &= G(F(x + y)) \\ &= G(F(x) + DF(x)y + R(x, y)) \\ &= G(z) + DG(z)(DF(x)y + R(x, y)) + R_1(x, y) \\ &= G(z) + DG(z)DF(x)y + R_2(x, y) \end{aligned}$$

with

$$\frac{\|R_2(x, y)\|}{\|y\|} \rightarrow 0 \quad \text{as } y \rightarrow 0.$$

Thus  $G \circ F$  is differentiable at  $x$ , and

$$(4.A.6) \quad D(G \circ F)(x) = DG(F(x)) \cdot DF(x).$$

In case  $k = 1$ , so  $G : U \rightarrow \mathbb{R}$ , we can rewrite (4.A.6) as

$$(4.A.7) \quad D(G \circ F)(x) = \nabla G(F(x))^t DF(x),$$

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

$$(4.A.8) \quad \frac{d}{dx} G(F(x)) = \nabla G(F(x)) \cdot F'(x).$$

This leads to such formulas as (4.3.10).

Another useful remark is that, by the Fundamental Theorem of Calculus, applied to  $\varphi(t) = F(x + ty)$ ,

$$(4.A.9) \quad F(x + y) = F(x) + \int_0^1 DF(x + ty)y \, dt,$$

provided  $DF$  is continuous. A closely related application of the Fundamental Theorem of Calculus is that, if we assume  $F : \mathcal{O} \rightarrow \mathbb{R}^m$  is differentiable in each variable separately, and that each  $\partial F/\partial x_j$  is continuous on  $\mathcal{O}$ , then

$$(4.A.10) \quad F(x + y) = F(x) + \sum_{j=1}^n [F(x + z_j) - F(x + z_{j-1})] = F(x) + \sum_{j=1}^n A_j(x, y)y_j,$$

$$A_j(x, y) = \int_0^1 \frac{\partial F}{\partial x_j}(x + z_{j-1} + ty_j e_j) \, dt,$$

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 (4.A.4). Thus we have established the following.

**Proposition 4.A.1.** *If  $\mathcal{O}$  is an open subset of  $\mathbb{R}^n$  and  $F : \mathcal{O} \rightarrow \mathbb{R}^m$  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 4.A.2.** *Assume  $F : \mathcal{O} \rightarrow \mathbb{R}^m$  is of class  $C^2$ , with  $\mathcal{O}$  open in  $\mathbb{R}^n$ . Then, for each  $x \in \mathcal{O}$ ,  $1 \leq j, k \leq n$ ,*

$$(4.A.11) \quad \frac{\partial}{\partial x_j} \frac{\partial F}{\partial x_k}(x) = \frac{\partial}{\partial x_k} \frac{\partial F}{\partial x_j}(x).$$

To prove Proposition 4.A.2, it suffices to treat real valued functions, so consider  $f : \mathcal{O} \rightarrow \mathbb{R}$ . For  $1 \leq j \leq n$ , we set

$$(4.A.12) \quad \Delta_{j,h} f(x) = \frac{1}{h}(f(x + he_j) - f(x)),$$

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,

$$(4.A.13) \quad \Delta_{j,h} f(x) = \partial_j f(x + \alpha_j h e_j),$$

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,

$$(4.A.14) \quad \begin{aligned} \Delta_{k,h} \Delta_{j,h} f(x) &= \partial_k(\Delta_{j,h} f)(x + \alpha_k h e_k) \\ &= \Delta_{j,h}(\partial_k f)(x + \alpha_k h e_k) \\ &= \partial_j \partial_k f(x + \alpha_k h e_k + \alpha_j h e_j), \end{aligned}$$

with  $\alpha_j, \alpha_k \in (0, 1)$ . Here we have used the elementary result

$$(4.A.15) \quad \partial_k \Delta_{j,h} f = \Delta_{j,h} (\partial_k f).$$

We deduce the following.

**Proposition 4.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*

$$(4.A.16) \quad \partial_j \partial_k f(x_0) = \lim_{h \rightarrow 0} \Delta_{k,h} \Delta_{j,h} f(x_0).$$

Clearly

$$(4.A.17) \quad \Delta_{k,h} \Delta_{j,h} f = \Delta_{j,h} \Delta_{k,h} f,$$

so we have the following, which easily implies Proposition 4.A.2.

**Corollary 4.A.4.** *In the setting of Proposition 4.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*

$$(4.A.18) \quad \partial_j \partial_k f(x_0) = \partial_k \partial_j f(x_0).$$

If  $U$  and  $V$  are open subsets of  $\mathbb{R}^n$  and  $F : U \rightarrow V$  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 4.A.5.** *Let  $F$  be a  $C^k$  map from an open neighborhood  $\Omega$  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  $F : U \rightarrow V$  is one-to-one and onto, and  $F^{-1} : V \rightarrow U$  is a  $C^k$  map. (So  $F : U \rightarrow V$  is a diffeomorphism.)*

Proofs of Theorem 4.A.5 can be found in a number of texts, including [31], Chapter 2 of [50], and Chapter 1 of [45].

## 4.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  $p_j \rightarrow p$ ) if and only if

$$(4.B.1) \quad \|p_j - p\| \longrightarrow 0.$$

Here  $\|\cdot\|$  is the norm on  $\mathbb{R}^n$  arising in §2.10 of Chapter 2, and the meaning of (4.B.1) is that for every  $\varepsilon > 0$  there exists  $N$  such that

$$(4.B.2) \quad j \geq N \implies \|p_j - p\| < \varepsilon.$$

A set  $S \subset \mathbb{R}^n$  is said to be *closed* if and only if

$$(4.B.3) \quad p_j \in S, p_j \rightarrow p \implies p \in S.$$

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

$$(4.B.4) \quad B_\varepsilon(q) = \{p \in \mathbb{R}^n : \|p - q\| < \varepsilon\},$$

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

$$(4.B.5) \quad \|p_j - p_k\| \longrightarrow 0, \quad \text{as } j, k \rightarrow \infty.$$

It is easy to see that if  $p_j \rightarrow p$  for some  $p \in \mathbb{R}^n$ , then (4.B.5) holds. The completeness property is the converse.

**Theorem 4.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  $p_j \rightarrow p$  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 [8], from an axiomatic viewpoint, and in [27] and [49], 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.

$$(4.B.6) \quad \begin{array}{l} \text{Each infinite sequence } (p_j) \text{ in } K \text{ has a subsequence} \\ \text{that converges to a point in } K. \end{array}$$

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 4.B.2.** *If  $K \subset \mathbb{R}^n$  is closed and bounded, then it is compact.*

We start with a special case.

**Proposition 4.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

$$(4.B.7) \quad |q_j - q_{j+k}| \leq 2^{-j}|b-a|, \quad \forall k \geq 0.$$

Thus  $(q_j)$  is a Cauchy sequence, so by the completeness of  $\mathbb{R}$ , it converges, to the desired limit  $p \in [a, b]$ .  $\square$

From Proposition 4.B.3 it is easy enough to show that any closed, bounded box

$$(4.B.8) \quad \mathcal{B} = \{(x_1, \dots, x_n) \in \mathbb{R}^n : a_j \leq x_j \leq b_j, \forall j\}$$

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 4.B.2.

We next discuss continuity. If  $S \subset \mathbb{R}^n$ , a function

$$(4.B.9) \quad f : S \rightarrow \mathbb{R}^m$$

is said to be continuous at  $p \in S$  provided

$$(4.B.10) \quad p_j \in S, p_j \rightarrow p \implies f(p_j) \rightarrow f(p).$$

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 4.B.4.** *If  $K \subset \mathbb{R}^n$  is compact and  $f : K \rightarrow \mathbb{R}^m$  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  $p_{k_\nu} \rightarrow p$  in  $K$ , and then  $q_{k_\nu} \rightarrow f(p)$  in  $\mathbb{R}^m$ .  $\square$

This leads to the second connection.

**Proposition 4.B.5.** *If  $K \subset \mathbb{R}^n$  is compact and  $f : K \rightarrow \mathbb{R}^m$  is continuous, then there exists  $p \in K$  such that*

$$(4.B.11) \quad \|f(p)\| = \max_{x \in K} \|f(x)\|,$$

and there exists  $q \in K$  such that

$$(4.B.12) \quad \|f(q)\| = \min_{x \in K} \|f(x)\|.$$

The meaning of (4.B.11) is that  $\|f(p)\| \geq \|f(x)\|$  for all  $x \in K$ , and the meaning of (4.B.12) is similar.

For the proof, consider

$$(4.B.13) \quad g : K \rightarrow \mathbb{R}, \quad g(p) = \|f(p)\|.$$

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 4.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 (4.B.11)–(4.B.12).

A variant of Proposition 4.B.5, with a very similar proof, is that if  $K \subset \mathbb{R}^n$  is compact and  $f : K \rightarrow \mathbb{R}$  is continuous, then there exist  $p, q \in K$  such that

$$(4.B.14) \quad f(p) = \max_{x \in K} f(x), \quad f(q) = \min_{x \in K} f(x).$$

We next define the *closure*  $\bar{S}$  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,  $p \in \bar{S}$  if and only if there exists an infinite sequence  $(p_j)$  of points in  $S$  such that  $p_j \rightarrow p$ .

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  $x \leq R$  for all  $x \in S$ . Hence  $x \leq R$  for all  $x \in \bar{S}$ . In such a case, there exists an interval  $[R - k, R]$  whose intersection with  $\bar{S}$  is nonempty, hence compact. We set

$$(4.B.15) \quad \sup S = \max_{\bar{S} \cap [R-k, R]} x,$$

the right side well defined by (4.B.14), with  $f(x) = x$ . There is a similar definition of

$$(4.B.16) \quad \inf S,$$

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 4.B.9 below.

**Proposition 4.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  $\bigcap_m X_m \neq \emptyset$ .*

**Proof.** Pick  $x_m \in X_m$ . If  $K$  is compact,  $(x_m)$  has a convergent subsequence,  $x_{m_k} \rightarrow y$ . Since  $\{x_{m_k} : k \geq \ell\} \subset X_{m_\ell}$ , which is closed, we have  $y \in \bigcap_m X_m$ .  $\square$

**Corollary 4.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  $\bigcup_m U_m \supset K$ , then  $U_M \supset K$  for some  $M$ .*

**Proof.** Consider  $X_m = K \setminus U_m$ .  $\square$

Before getting to Proposition 4.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

$$(4.B.17) \quad \mathcal{R} = \{B_{r_j}(q_j) : q_j \in \mathbb{Q}^n, r_j \in \mathbb{Q} \cap (0, \infty)\}.$$

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 4.B.8.** *Let  $\Omega \subset \mathbb{R}^n$  be a nonempty open set. Then*

$$(4.B.18) \quad \Omega = \bigcup \{B : B \in \mathcal{R}, B \subset \Omega\}.$$

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 4.B.9.** *If  $K \subset \mathbb{R}^n$  is compact, then it has the following property.*

$$(4.B.19) \quad \text{Every open cover } \{U_\alpha : \alpha \in \mathcal{A}\} \text{ of } K \text{ has a finite subcover.}$$

**Proof.** By Lemma 4.B.8, it suffices to prove the following.

$$(4.B.20) \quad \begin{array}{l} \text{Every countable cover } \{B_j : j \in \mathbb{N}\} \text{ of } K \text{ by open balls} \\ \text{has a finite subcover.} \end{array}$$

For this, we set

$$(4.B.21) \quad U_m = B_1 \cup \cdots \cup B_m$$

and apply Corollary 4.B.7. □

### 4.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

$$(4.C.1) \quad \mathbb{C}^n = W_+ \oplus W_-,$$

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

$$(4.C.2) \quad \mathbb{R}^n = V_+ \oplus V_-, \quad V_{\pm} = W_{\pm} \cap \mathbb{R}^n.$$

We have

$$(4.C.3) \quad v \in W_{\pm} \implies e^{tL}v \rightarrow 0 \text{ as } t \rightarrow \mp\infty,$$

and a fortiori

$$(4.C.4) \quad v \in V_{\pm} \implies e^{tL}v \rightarrow 0 \text{ as } t \rightarrow \mp\infty.$$

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

$$(4.C.5) \quad V_- = \mathbb{R}^n \implies \Phi_F^t(x) \rightarrow p \text{ as } t \rightarrow +\infty,$$

for  $x$  sufficiently close to  $p$ , where  $\Phi_F^t$  is the flow generated by  $F$ , was proven in §4.3 (cf. Proposition 4.3.4), and similarly we have

$$(4.C.6) \quad V_+ = \mathbb{R}^n \implies \Phi_F^t(x) \rightarrow p \text{ as } t \rightarrow -\infty,$$

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

$$(4.C.7) \quad \{p\} = S_+ \cap S_-,$$

and

$$(4.C.8) \quad x \in S_{\pm} \implies \Phi_F^t(x) \rightarrow p \text{ as } t \rightarrow \mp\infty.$$

The surfaces  $S_-$  and  $S_+$  are called, respectively, the stable and unstable manifolds of  $F$  at  $p$ . They have the further property that if  $\gamma$  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  $\text{dist}(x, S_-) < \delta$ , then for some  $t_1 > 0$ ,  $\|\Phi_F^{t_1}(x) - p\| < \varepsilon$ , and for all  $t \geq t_1$ ,  $\text{dist}(\Phi_F^t(x), S_+) < \varepsilon$ , 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

$$(4.C.9) \quad \|F(x)\| \leq C\|x\|, \quad \forall x \in \mathbb{R}^n.$$



Hence, as seen in §4.3 (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

$$(4.C.10) \quad F(x) = Lx + \sum_{j,k} x_j x_k G_{jk}(x),$$

where

$$(4.C.11) \quad L = DF(0),$$

and  $G_{jk}$  are  $C^1$  vector fields, given by

$$(4.C.12) \quad G_{jk}(x) = \int_0^1 \int_0^1 \frac{\partial^2}{\partial x_k \partial x_j} F(stx) ds dt.$$

We define the family of vector fields  $F_\varepsilon$  by

$$(4.C.13) \quad F_\varepsilon(x) = \frac{1}{\varepsilon} F(\varepsilon x),$$

for  $\varepsilon > 0$ . By (4.C.10),

$$(4.C.14) \quad F_\varepsilon(x) = Lx + \varepsilon G_\varepsilon(x),$$

where

$$(4.C.15) \quad G_\varepsilon(x) = \sum_{j,k} x_j x_k G_{jk}(\varepsilon x).$$

Passing to the limit  $\varepsilon \rightarrow 0$  gives  $F_0(x) = Lx$ . Results of §4.2 yield the following.

**Lemma 4.C.1.** *Given  $\delta > 0$ ,  $T < \infty$ , there exists  $\varepsilon_0 = \varepsilon_0(\delta, T, F) > 0$  such that for all  $\varepsilon \in (0, \varepsilon_0]$ ,*

$$(4.C.16) \quad \begin{aligned} \|x\| \leq 2, \quad |t| \leq T, \quad \|\Phi_{F_\varepsilon}^s(x)\| \leq 2 \quad \forall s \in [0, t] \\ \implies \|\Phi_{F_\varepsilon}^t(x) - e^{tL}x\| \leq \delta. \end{aligned}$$

Specializing to  $n = 2$ , we can assume that

$$(4.C.17) \quad L = \begin{pmatrix} a & \\ & -b \end{pmatrix}, \quad a, b > 0.$$

We take the box

$$(4.C.18) \quad \mathcal{O} = \{(x_1, x_2) : |x_1|, |x_2| \leq 1\},$$

and set

$$(4.C.19) \quad \mathcal{O}_k = 2^{-k} \mathcal{O}.$$

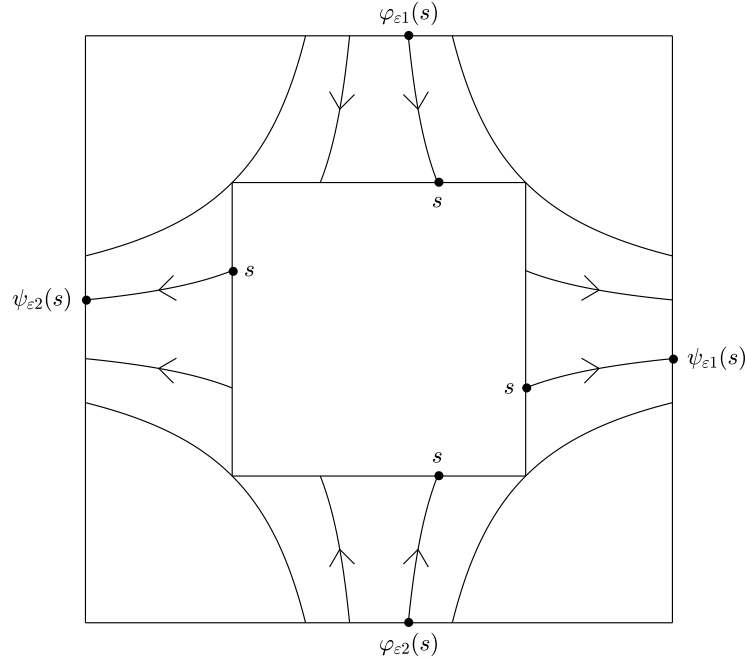
We define four families of maps

$$(4.C.20) \quad \varphi_{\varepsilon j}, \psi_{\varepsilon j} : \left[-\frac{1}{2}, \frac{1}{2}\right] \longrightarrow [-1, 1], \quad j = 1, 2, \quad 0 \leq \varepsilon \leq 1,$$

as follows. For  $j = 1$ , define  $t_\varepsilon(s)$  as the smallest positive number such that

$$\Phi_{F_\varepsilon}^{-t_\varepsilon(s)} \left( s, \frac{1}{2} \right) \in \{(\sigma, 1) : -1 \leq \sigma \leq 1\},$$

and then set  $\varphi_{\varepsilon 1}(s)$  to be the  $x_1$ -coordinate of  $\Phi_{F_\varepsilon}^{-t_\varepsilon(s)}(s, 1/2)$ . 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  $F_\varepsilon$ .



**Figure 4.C.1.** The maps  $\varphi_{\varepsilon j}$  and  $\psi_{\varepsilon j}$

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 Figure 4.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

$$(4.C.21) \quad \begin{aligned} |\varphi_{\varepsilon j}(s) - \varphi_{\varepsilon j}(t)| &\leq A|s - t|, \\ |\psi_{\varepsilon j}(s) - \psi_{\varepsilon j}(t)| &\leq A|s - t|, \end{aligned}$$

for all  $s, t \in [-1/2, 1/2]$ .

Results of §4.2 then establish the following.

**Lemma 4.C.2.** *There exists  $\varepsilon_1 = \varepsilon_1(F) > 0$  and  $A = A(F) < 1$  such that whenever  $0 \leq \varepsilon \leq \varepsilon_1$ , the maps  $\varphi_{\varepsilon j}$  and  $\psi_{\varepsilon j}$  in (4.C.20) are well defined on  $[-1/2, 1/2]$  and (4.C.21) holds for all  $s, t \in [-1/2, 1/2]$ .*

We make a further adjustment. Take  $\varepsilon_2 \leq \min(\varepsilon_1(F), \varepsilon_0(1/10, 10, F))$ . Further shrinking  $\varepsilon_2$  is necessary, arrange that, whenever  $\varepsilon \in (0, \varepsilon_2]$ ,

$$(4.C.22) \quad \|x\| \leq 2 \implies \|F_\varepsilon(x) - Lx\| \leq \frac{1}{2}\|Lx\|,$$

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  $F_{\varepsilon_2}$ , denoting

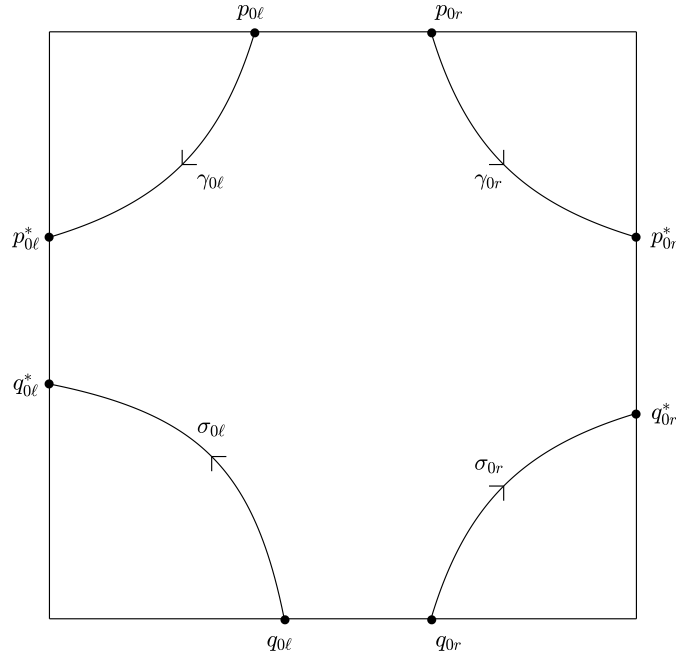


Figure 4.C.2. Step 0

this scaled vector field by  $F$ . Then (4.C.16) holds with  $T = 10$  and  $\delta = 1/10$  for all  $\varepsilon \in (0, 1]$  and (4.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

$$(4.C.23) \quad \Phi_k^t = \Phi_{F_\varepsilon}^t, \quad \varepsilon = 2^{-k}.$$

Note that dilation by the factor  $2^k$  takes the flow  $\Phi_F^t$  on  $\mathcal{O}_k$  to the flow  $\Phi_k^t$  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 sides of  $\mathcal{O}_k$ . Then the maps (4.C.20) can by slight abuse of terminology be labeled

$$(4.C.24) \quad \begin{aligned} \varphi_{\varepsilon 1} : \mathcal{T}_1 &\rightarrow \mathcal{T}, & \varphi_{\varepsilon 2} : \mathcal{B}_1 &\rightarrow \mathcal{B}, \\ \psi_{\varepsilon 1} : \mathcal{R}_1 &\rightarrow \mathcal{R}, & \psi_{\varepsilon 2} : \mathcal{L}_1 &\rightarrow \mathcal{L}. \end{aligned}$$

Pick two points  $p_{0\ell}, p_{0r} \in \mathcal{T}$  such that for some  $t_{0\ell}, t_{0r} \in (0, 1)$ ,

$$(4.C.25) \quad p_{0\ell}^* = \Phi_F^{t_{0\ell}}(p_{0\ell}) \in \mathcal{L}, \quad p_{0r}^* = \Phi_F^{t_{0r}}(p_{0r}) \in \mathcal{R},$$

and pick two points  $q_{0\ell}, q_{0r} \in \mathcal{B}$  such that for some  $s_{0\ell}, s_{0r} \in (0, 1)$ ,

$$(4.C.26) \quad q_{0\ell}^* = \Phi_F^{s_{0\ell}}(q_{0\ell}) \in \mathcal{L}, \quad q_{0r}^* = \Phi_F^{s_{0r}}(q_{0r}) \in \mathcal{R}.$$

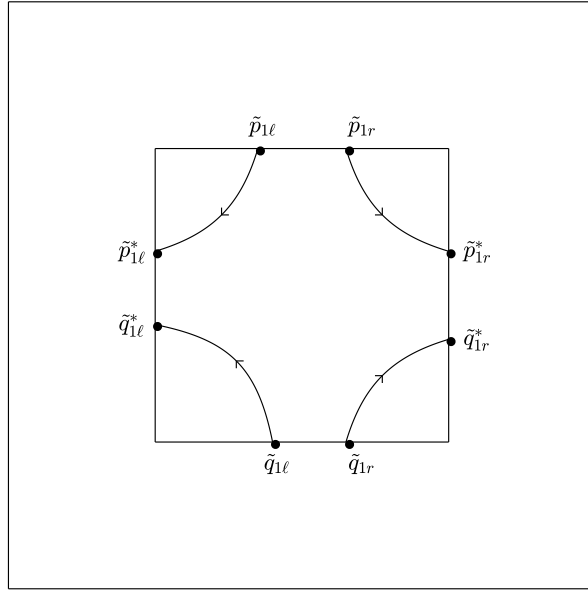


Figure 4.C.3. Beginning of Step 1

See Figure 4.C.2. The possibility to do this is guaranteed by Lemma 4.C.1. Denote the orbits of  $\Phi_0^t = \Phi_F^t$  through  $p_{0l}, p_{0r}$  by  $\gamma_{0l}, \gamma_{0r}$  and those through  $q_{0l}, q_{0r}$  by  $\sigma_{0l}, \sigma_{0r}$ .

Let us call the construction just described Step 0. To continue, at Step 1, pick  $\tilde{p}_{1l}, \tilde{p}_{1r} \in \mathcal{T}_1$  and  $\tilde{q}_{1l}, \tilde{q}_{1r} \in \mathcal{B}_1$  such that the following holds. Note that  $2\tilde{p}_{1l}, 2\tilde{p}_{1r} \in \mathcal{T}$  and  $2\tilde{q}_{1l}, 2\tilde{q}_{1r} \in \mathcal{B}$ . We require that, for some  $t_{1l}, t_{1r} \in (0, T_1)$ ,

$$(4.C.27) \quad \Phi_1^{t_{1l}}(2\tilde{p}_{1l}) \in \mathcal{L}, \quad \Phi_1^{t_{1r}}(2\tilde{p}_{1r}) \in \mathcal{R},$$

and for some  $s_{1l}, s_{1r} \in (0, T_1)$ ,

$$(4.C.28) \quad \Phi_1^{s_{1l}}(2\tilde{q}_{1l}) \in \mathcal{L}, \quad \Phi_1^{s_{1r}}(2\tilde{q}_{1r}) \in \mathcal{R}.$$

The conditions (4.C.27) and (4.C.28) are equivalent to

$$(4.C.29) \quad \tilde{p}_{1l}^* = \Phi_F^{t_{1l}}(\tilde{p}_{1l}) \in \mathcal{L}_1, \quad \tilde{p}_{1r}^* = \Phi_F^{t_{1r}}(\tilde{p}_{1r}) \in \mathcal{R}_1,$$

and

$$(4.C.30) \quad \tilde{q}_{1l}^* = \Phi_F^{s_{1l}}(\tilde{q}_{1l}) \in \mathcal{L}_1, \quad \tilde{q}_{1r}^* = \Phi_F^{s_{1r}}(\tilde{q}_{1r}) \in \mathcal{R}_1.$$

See Figure 4.C.3. Denote the orbits of  $F$  through  $\tilde{p}_{1l}, \tilde{p}_{1r}$  by  $\gamma_{1l}, \gamma_{1r}$  and those through  $\tilde{q}_{1l}, \tilde{q}_{1r}$  by  $\sigma_{1l}, \sigma_{1r}$ . When picking  $\tilde{p}_{1l}, \tilde{p}_{1r}, \tilde{q}_{1l}$ , and  $\tilde{q}_{1r}$ , one can and should enforce the following condition. If  $\gamma_{0l}$  intersects  $\mathcal{T}_1$ ,  $\tilde{p}_{1l}$  should be to the right of such an intersection, if  $\gamma_{0r}$  intersects  $\mathcal{T}_1$ ,  $\tilde{p}_{1r}$  should be to the left of such an

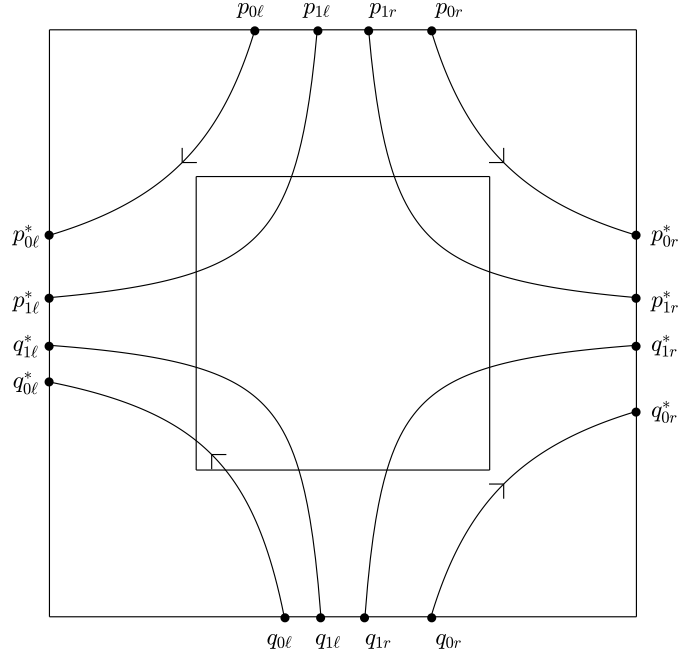


Figure 4.C.4. Completion of Step 1

intersection, and similarly for cases when  $\sigma_{0\ell}$  or  $\sigma_{0r}$  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  $\sigma_{1r}$  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 Figure 4.C.4. That such an intersection must occur is guaranteed by (4.C.22). This, together with the fact that orbits of  $F$  cannot intersect, guarantees that

$$(4.C.31) \quad p_{0\ell} < p_{1\ell} < p_{1r} < p_{0r},$$

in the sense that  $p < p'$  means  $p$  is to the left of  $p'$ . In a similar sense, made clear in Figure 4.C.4, we have

$$(4.C.32) \quad \begin{aligned} q_{0\ell} &< q_{1\ell} < q_{1r} < q_{0r}, \\ p_{0r}^* &< p_{1r}^* < q_{1r}^* < q_{0r}^*, \\ p_{0\ell}^* &< p_{1\ell}^* < q_{1\ell}^* < q_{0\ell}^*. \end{aligned}$$

Furthermore, as a consequence of (4.C.21), we have

$$(4.C.33) \quad \begin{aligned} |p_{1\ell} - p_{1r}| &\leq A|\tilde{p}_{1\ell} - \tilde{p}_{1r}| \leq A, \\ |q_{1\ell} - q_{1r}| &\leq A|\tilde{q}_{1\ell} - \tilde{q}_{1r}| \leq A, \\ |p_{1r}^* - q_{1r}^*| &\leq A|\tilde{p}_{1r}^* - \tilde{q}_{1r}^*| \leq A, \\ |p_{1\ell}^* - q_{1\ell}^*| &\leq A|\tilde{p}_{1\ell}^* - \tilde{q}_{1\ell}^*| \leq A. \end{aligned}$$

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  $t_{k\ell}, t_{kr} \in (0, T_k)$ ,

$$(4.C.34) \quad \Phi_k^{t_{k\ell}}(2^k \tilde{p}_{k\ell}) \in \mathcal{L}, \quad \Phi_k^{t_{kr}}(2^k \tilde{p}_{kr}) \in \mathcal{R},$$

and for some  $s_{k\ell}, s_{kr} \in (0, T_k)$ ,

$$(4.C.35) \quad \Phi_k^{s_{k\ell}}(2^k \tilde{q}_{k\ell}) \in \mathcal{L}, \quad \Phi_k^{s_{kr}}(2^k \tilde{q}_{kr}) \in \mathcal{R}.$$

The conditions (4.C.34) and (4.C.35) are equivalent to

$$(4.C.36) \quad \tilde{p}_{k\ell}^* = \Phi_F^{t_{k\ell}}(\tilde{p}_{k\ell}) \in \mathcal{L}_k, \quad \tilde{p}_{kr}^* = \Phi_F^{t_{kr}}(\tilde{p}_{kr}) \in \mathcal{R}_k,$$

and

$$(4.C.37) \quad \tilde{q}_{k\ell}^* = \Phi_F^{s_{k\ell}}(\tilde{q}_{k\ell}) \in \mathcal{L}_k, \quad \tilde{q}_{kr}^* = \Phi_F^{s_{kr}}(\tilde{q}_{kr}) \in \mathcal{R}_k.$$

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 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 4.C.1, we can take

$$(4.C.38) \quad T_k \rightarrow \infty \text{ as } k \rightarrow \infty,$$

and hence take (with  $z = \ell$  or  $r$ )

$$(4.C.39) \quad \|2^k \tilde{p}_{kz} - (0, 1)\| \leq \eta_k, \quad \|2^k \tilde{q}_{kz} - (0, 1)\| \leq \eta_k, \quad \eta_k \rightarrow 0 \text{ as } k \rightarrow \infty.$$

It then follows that (again with  $z = \ell$  or  $r$ )

$$(4.C.40) \quad \|\tilde{p}_{kz}^* - (2^{-k}, 0)\| \leq 2^{-k} \tilde{\eta}_k, \quad \|\tilde{q}_{kz}^* - (2^{-k}, 0)\| \leq 2^{-k} \tilde{\eta}_k, \quad \tilde{\eta}_k \rightarrow 0 \text{ as } k \rightarrow \infty.$$

Now we continue the orbits  $\gamma_{k\ell}, \gamma_{kr}, \sigma_{k\ell}$ , and  $\sigma_{kr}$  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 Figure 4.C.5. That such intersections must occur is guaranteed by (4.C.22). As before, the fact that orbits of  $\Phi_F^t$  cannot intersect guarantees that

$$(4.C.41) \quad p_{0\ell} < \cdots < p_{k\ell} < p_{kr} < \cdots < p_{0r},$$

in the sense specified in (4.C.31), and, as in (4.C.32),

$$(4.C.42) \quad \begin{aligned} q_{0\ell} &< \cdots < q_{k\ell} < q_{kr} < \cdots < q_{0r}, \\ p_{0r}^* &< \cdots < p_{kr}^* < q_{kr}^* < \cdots < q_{0r}^*, \\ p_{0\ell}^* &< \cdots < p_{k\ell}^* < q_{k\ell}^* < \cdots < q_{0\ell}^*. \end{aligned}$$

Furthermore, as a consequence of (4.C.21), we have

$$(4.C.43) \quad \begin{aligned} |p_{k\ell} - p_{kr}| &\leq A^k |\tilde{p}_{k\ell} - \tilde{p}_{kr}| \leq A^k 2^{-k} \eta_k, \\ |q_{k\ell} - q_{kr}| &\leq A^k |\tilde{q}_{k\ell} - \tilde{q}_{kr}| \leq A^k 2^{-k} \eta_k, \\ |p_{k\ell}^* - p_{kr}^*| &\leq A^k |\tilde{p}_{k\ell}^* - \tilde{p}_{kr}^*| \leq A^k 2^{-k} \tilde{\eta}_k, \\ |q_{k\ell}^* - q_{kr}^*| &\leq A^k |\tilde{q}_{k\ell}^* - \tilde{q}_{kr}^*| \leq A^k 2^{-k} \tilde{\eta}_k. \end{aligned}$$

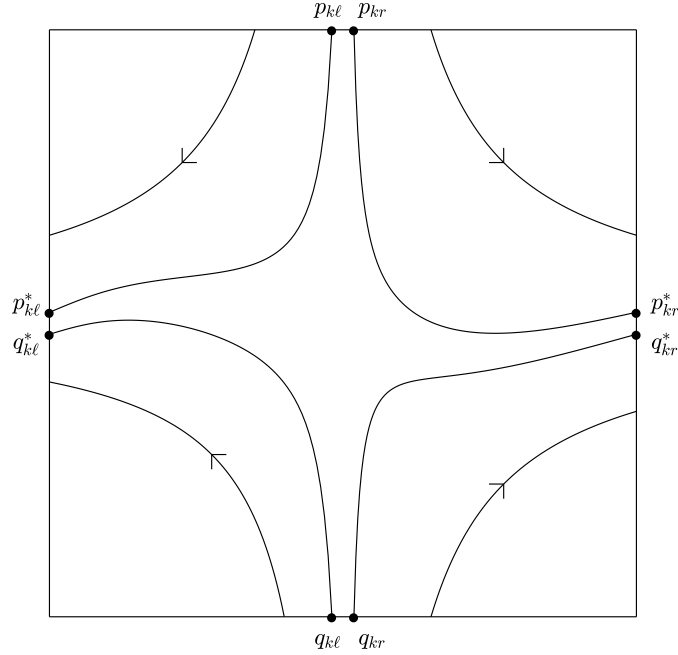


Figure 4.C.5. Step  $k$

In particular, these distances are converging to 0 quite rapidly. We obtain limits

$$(4.C.44) \quad \begin{aligned} p_{k\ell}, p_{kr} &\rightarrow p_t \in \mathcal{T}, & q_{k\ell}, q_{kr} &\rightarrow p_b \in \mathcal{B}, \\ p_{k\ell}^*, q_{kr}^* &\rightarrow p_r^* \in \mathcal{R}, & p_{k\ell}^*, q_{k\ell}^* &\rightarrow p_\ell^* \in \mathcal{L}. \end{aligned}$$

See Figure 4.C.6. We have

$$(4.C.45) \quad \Phi_F^t(p_t), \Phi_F^t(p_b) \rightarrow 0 \text{ as } t \rightarrow +\infty,$$

and

$$(4.C.46) \quad \Phi_F^t(p_\ell^*), \Phi_F^t(p_r^*) \rightarrow 0 \text{ as } t \rightarrow -\infty,$$

since the paths in (4.C.45) meet each  $\mathcal{O}_k$  for large positive  $t$  and those in (4.C.46) meet each  $\mathcal{O}_k$  for large negative  $t$ . Furthermore, by (4.C.39)–(4.C.40), plus the fact that all these paths solve  $dx/dt = F(x)$ , the curves in (4.C.45) fit together to form a  $C^1$  curve tangent to the  $x_2$ -axis at  $p = 0$ , and those in (4.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 (4.C.17), we can take

$$(4.C.47) \quad L = \begin{pmatrix} A & \\ & -b \end{pmatrix}, \quad A \in M(2, \mathbb{R}), \quad b > 0,$$

and, via Lemma 4.3.5, arrange that

$$(4.C.48) \quad Av \cdot v \geq a\|v\|^2, \quad a > 0, \quad \forall v \in \mathbb{R}^2.$$

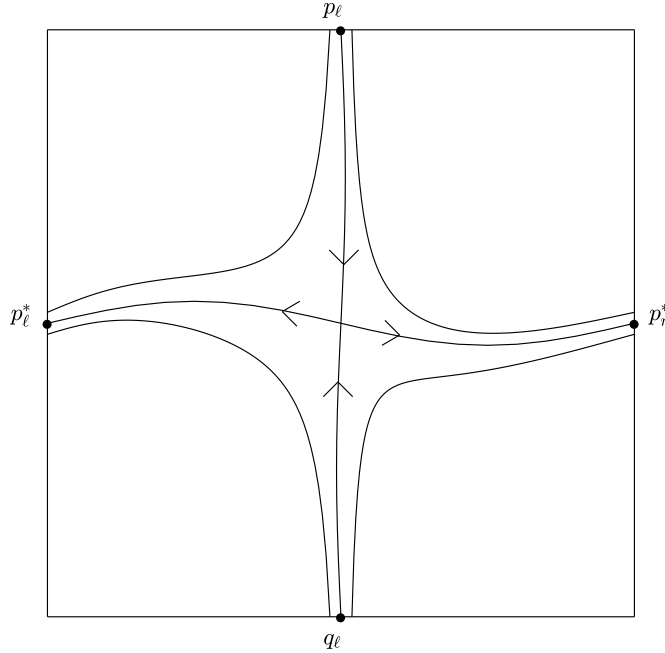


Figure 4.C.6. Limiting configuration as  $k \rightarrow \infty$

In place of (4.C.18), we use the cylinder

$$(4.C.49) \quad \mathcal{O} = \{(x_1, x_2, x_3) : x_1^2 + x_2^2 \leq 1, |x_3| \leq 1\}.$$

with boundary

$$(4.C.50) \quad \partial\mathcal{O} = \mathcal{T} \cup \mathcal{B} \cup \mathcal{L},$$

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 (4.C.24), we have (at least for small  $\varepsilon$ ) maps

$$(4.C.51) \quad \begin{aligned} \varphi_{\varepsilon 1} : \mathcal{T}_1 &\rightarrow \mathcal{T}, & \varphi_{\varepsilon 2} : \mathcal{B}_1 &\rightarrow \mathcal{B}, \\ \psi_\varepsilon : \mathcal{L}_1 &\rightarrow \mathcal{L}, \end{aligned}$$

with  $\varphi_{\varepsilon j}$  defined by backward flow of  $\Phi_{F_\varepsilon}^t$  and  $\psi_\varepsilon$  defined by forward flow. Again the maps  $\varphi_{\varepsilon j}$  are contractions for small  $\varepsilon$ . The maps  $\psi_\varepsilon$  are not contractions, but composing them on the left with the projection  $S^1 \times [-1, 1] \rightarrow [-1, 1]$  produces a contraction, for small  $\varepsilon$ , 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  $\Phi_{F_\varepsilon}^t$  yields a pair of flared tubes, as pictured in Figure 4.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 [10] and [19] for other approaches to this result.



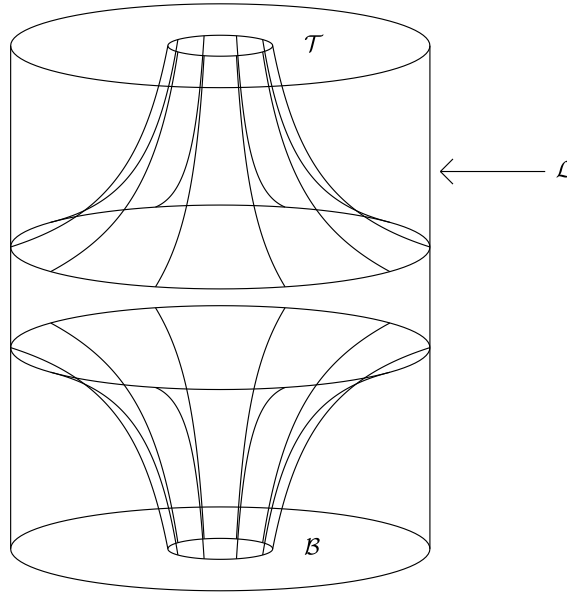


Figure 4.C.7. Set-up for a saddle in 3D

#### 4.D. Blown up phase portrait at a critical point

Let  $\Omega \subset \mathbb{R}^n$  be open,  $0 \in \Omega$ , and  $F : \Omega \rightarrow \mathbb{R}^n$  a smooth vector field. Assume  $F(0) = 0$ , and that 0 is a non-degenerate critical point, i.e.,

$$(4.D.1) \quad A = DF(0) \in M(n, \mathbb{R}) \text{ is invertible.}$$

We want to “blow up” the portrait of solutions to

$$(4.D.2) \quad x' = F(x),$$

by using spherical polar coordinates,

$$(4.D.3) \quad x = r\omega, \quad r = r(t) \in [0, \infty), \quad \omega = \omega(t) \in S^{n-1},$$

where  $S^{n-1}$  is the unit sphere in  $\mathbb{R}^n$ . Note that

$$(4.D.4) \quad x' = r'\omega + r\omega',$$

and  $\omega(t) \cdot \omega(t) \equiv 1 \Rightarrow \omega' \perp \omega$ , so the two vectors on the right side of (4.D.4) are mutually orthogonal. We obtain

$$(4.D.5) \quad \begin{aligned} r' &= x' \cdot \omega = \omega \cdot F(r\omega), \\ \omega' &= r^{-1}P_\omega x' = r^{-1}P_\omega F(r\omega), \end{aligned}$$

where  $P_\omega$  is the orthogonal projection of  $\mathbb{R}^n$  onto the orthogonal complement of  $\text{Span } \omega$ , i.e.,

$$(4.D.6) \quad P_\omega v = v - (v \cdot \omega)\omega.$$

To proceed, write

$$(4.D.7) \quad F(x) = Ax + R(x),$$

where

$$(4.D.8) \quad \begin{aligned} R(x) &= \sum_{j,k} \left( \int_0^1 (1-s) \partial_j \partial_k F(sx) ds \right) x_j x_k \\ &= r^2 \sum_{j,k} \left( \int_0^1 (1-s) \partial_j \partial_k F(sr\omega) ds \right) \omega_j \omega_k \\ &= r^2 G(r, \omega). \end{aligned}$$

Here, if  $B_R(0) \subset \Omega$ , we have

$$(4.D.9) \quad G : (-R, R) \times S^{n-1} \longrightarrow \mathbb{R}^n, \text{ smooth.}$$

Thus we can rewrite the system (4.D.5) as

$$(4.D.10) \quad \begin{aligned} r' &= (\omega \cdot A\omega)r + \omega \cdot G(r, \omega)r^2, \\ \omega' &= P_\omega A\omega + rP_\omega G(r, \omega). \end{aligned}$$

This is a smooth system of ODE on  $(-R, R) \times S^{n-1}$ .

Now the null space  $\mathcal{N}(P_\omega)$  is  $\text{Span } \omega$ , and

$$(4.D.11) \quad P_\omega A\omega = A\omega - (\omega \cdot A\omega)\omega,$$

so

$$(4.D.12) \quad \begin{aligned} P_\omega A\omega = 0 &\iff A\omega \parallel \omega \\ &\iff \omega \text{ is an eigenvector of } A, \end{aligned}$$

in which case the associated eigenvalue  $\lambda$  is necessarily equal to  $\omega \cdot A\omega$  (and this eigenvalue is nonzero if  $A$  is invertible). Consequently, the right side of (4.D.10) vanishes on

$$(4.D.13) \quad \{(r, \omega) : r = 0, \omega \in S^{n-1}, A\omega = (\omega \cdot A\omega)\omega\},$$

and we have the following.

**Proposition 4.D.1.** *If (4.D.1) holds, there exists a  $a > 0$  such that, on*

$$(4.D.14) \quad \mathcal{O}_a = (-a, a) \times S^{n-1},$$

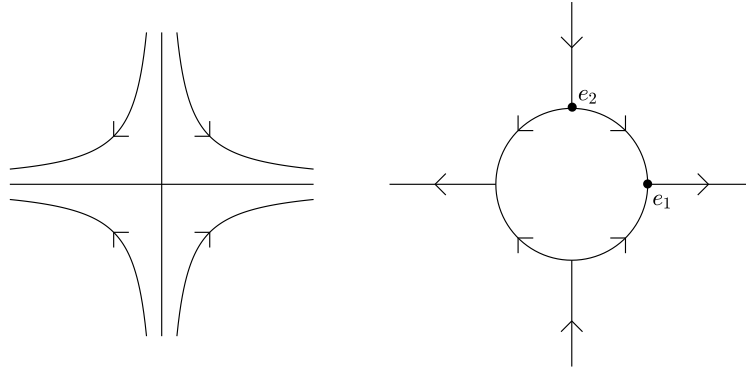
*the right side of (4.D.10) vanishes only on the set (4.D.13). If each real eigenspace of  $A$  is one dimensional, then each critical point of (4.D.10) is isolated.*

Solutions to (4.D.10) define a flow on  $\mathcal{O}_a$ , and  $\{(r, \omega) \in \mathcal{O}_a : r = 0\}$  is invariant under this flow. The flow restricted to this set is the flow on  $S^{n-1}$  defined by

$$(4.D.15) \quad \omega' = P_\omega A\omega,$$

which is

$$(4.D.16) \quad \omega(t) = \|e^{tA}\omega_0\|^{-1} e^{tA}\omega_0, \quad \omega_0 = \omega(0).$$



**Figure 4.D.1.** Saddle critical point and its blowup

One convenient way to visualize the flow on  $\mathcal{O}_a$  is to use the diffeomorphism

$$(4.D.17) \quad \Phi : \mathcal{O}_a \longrightarrow U_a \subset \mathbb{R}^n,$$

given by

$$(4.D.18) \quad \Phi(r, \omega) = e^r \omega = y,$$

where

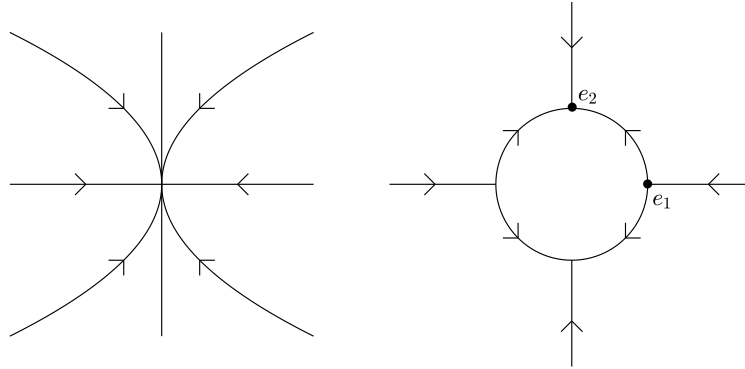
$$(4.D.19) \quad U_a = \{y \in \mathbb{R}^n : e^{-a} < |y| < e^a\}.$$

The resulting flow on  $U_a$ , acting on  $y = e^r \omega$ , is defined by

$$(4.D.20) \quad \begin{aligned} y' &= e^r r' \omega + e^r \omega' \\ &= e^r \left\{ [(\omega \cdot A\omega)r + \omega \cdot G(r, \omega)r^2] \omega \right. \\ &\quad \left. + P_\omega [A\omega + rG(r, \omega)] \right\} \\ &= X(y), \end{aligned}$$

where  $G(r, \omega)$  is given by (4.D.8). Note that

$$(4.D.21) \quad G(0, \omega) = \frac{1}{2} \sum_{j,k} \partial_j \partial_k F(0) \omega_j \omega_k.$$



**Figure 4.D.2.** A generic sink and its blowup

We see that  $X$  is a smooth vector field on  $U_a$  whose critical points lie on  $S^{n-1} \subset U_a$ , and  $\omega_0 \in S^{n-1}$  is a critical point of  $X$  if and only if

$$(4.D.22) \quad A\omega_0 = \lambda\omega_0,$$

for some  $\lambda \in \mathbb{R}$ , necessarily

$$(4.D.23) \quad \lambda = \omega_0 \cdot A\omega_0,$$

and  $\lambda \neq 0$ , given (4.D.1).

It is of interest to specify when such a critical point  $\omega_0$  of  $X$  is nondegenerate, i.e., when  $DX(\omega_0) \in M(n, \mathbb{R})$  is invertible. To begin this analysis, we see from (4.D.20) that

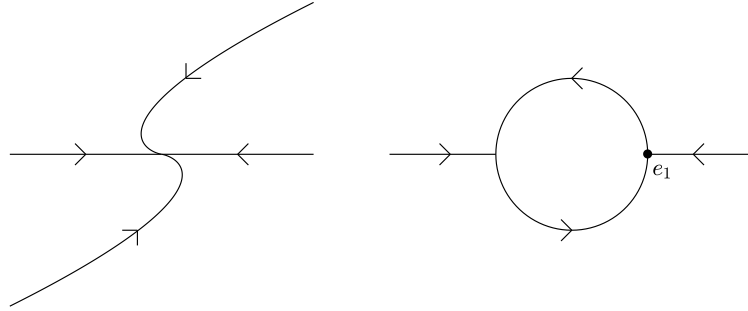
$$(4.D.24) \quad DX(\omega_0)\omega_0 = \lambda\omega_0.$$

We next evaluate  $DX(\omega_0)\xi$  when  $\omega_0$  is a critical point and  $\xi \perp \omega_0$ . We have from (4.D.20) that

$$(4.D.25) \quad DX(\omega_0)\xi = \frac{d}{ds} P_{\omega(s)} A\omega(s) \Big|_{s=0},$$

where

$$(4.D.26) \quad \omega : (-\varepsilon, \varepsilon) \rightarrow S^{n-1}, \quad \omega(0) = \omega_0, \quad \omega'(0) = \xi.$$



**Figure 4.D.3.** A non-generic sink and its blowup

A computation gives

$$(4.D.27) \quad \frac{d}{ds} P_{\omega(s)} A \omega(s) \Big|_{s=0} = P_{\omega_0} A \xi - \lambda \xi,$$

provided (4.D.22) holds, and  $\xi \perp \omega_0$ . Hence

$$(4.D.28) \quad DX(\omega_0) \xi = P_{\omega_0} A \xi - \lambda \xi,$$

provided  $\omega_0$  is a critical point of  $X$  and  $\xi \perp \omega_0$ . It follows that, if  $\xi \perp \omega_0$ ,

$$(4.D.29) \quad DX(\omega_0)(\alpha \omega_0 + \xi) = \alpha \lambda \omega_0 + P_{\omega_0} A \xi - \lambda \xi,$$

so  $\omega_0 \in S^{n-1}$  is a nondegenerate critical point of  $X$  if and only if

$$(4.D.30) \quad \begin{aligned} \alpha \lambda \omega_0 + P_{\omega_0} A \xi - \lambda \xi &= 0, \quad \xi \perp \omega_0 \\ \implies \alpha &= 0 \quad \text{and} \quad \xi = 0. \end{aligned}$$

Since  $\lambda \neq 0$ , we have the conclusion  $\alpha = 0$ , so the criterion boils down to

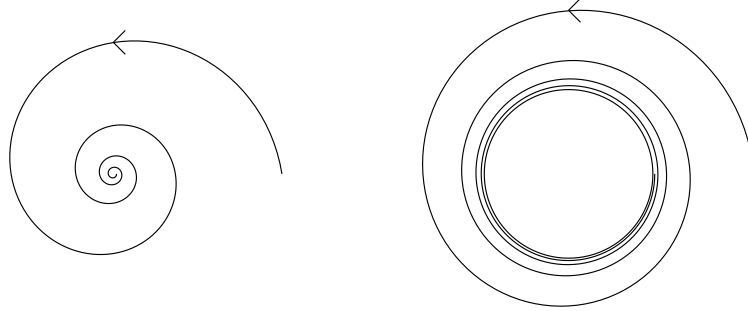
$$(4.D.31) \quad \xi \perp \omega_0, \quad P_{\omega_0} A \xi - \lambda \xi = 0 \implies \xi = 0.$$

Now, for  $\xi \perp \omega_0$ ,

$$(4.D.32) \quad P_{\omega_0} A \xi - \lambda \xi = 0 \iff (A - \lambda I) \xi \in \text{Span } \omega_0.$$

If  $\dim \mathcal{E}(A, \lambda) \geq 2$ , one has nonzero  $\xi \in \mathcal{E}(A, \lambda)$  orthogonal to  $\omega_0$ , so nondegeneracy requires

$$(4.D.33) \quad \dim \mathcal{E}(A, \lambda) = 1.$$



**Figure 4.D.4.** A spiral sink and its blowup

Furthermore, if  $\dim \mathcal{N}((A - \lambda I)^2) \geq 2$ , this space contains a nonzero element  $\xi$  orthogonal to  $\omega_0$ , and, by (4.D.33),  $(A - \lambda I)\xi \in \text{Span } \omega_0$ , so nondegeneracy requires

$$(4.D.34) \quad \dim \mathcal{N}((A - \lambda I)^2) = 1, \quad \text{hence } \mathcal{N}((A - \lambda I)^2) = \mathcal{E}(A, \lambda).$$

We have the following:

**Proposition 4.D.2.** *Maintain the hypothesis (4.D.1). Let  $\omega_0 \in S^{n-1}$  be a critical point of  $X$ , so (4.D.22) holds, with  $\lambda \in \mathbb{R}$ ,  $\lambda \neq 0$ . Then  $\omega_0$  is a nondegenerate critical point of  $X$  if and only if*

$$(4.D.35) \quad \mathcal{GE}(A, \lambda) = \mathcal{E}(A, \lambda), \quad \text{and } \dim \mathcal{E}(A, \lambda) = 1.$$

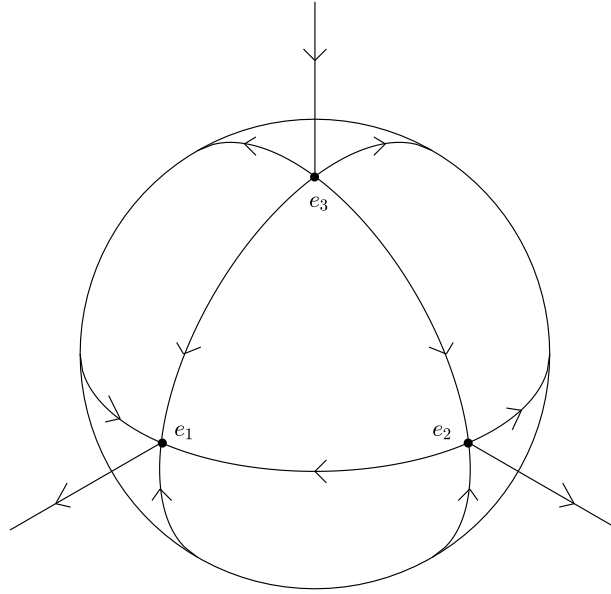
Figures 4.D.1–4.D.4 depict the blowups of critical points of four planar vector fields. These have the form (4.D.7) with  $A \in M(2, \mathbb{R})$  given, respectively, by

$$(4.D.36) \quad \begin{pmatrix} 1 & \\ & -1 \end{pmatrix}, \quad \begin{pmatrix} -2 & \\ & -1 \end{pmatrix}, \quad \begin{pmatrix} -1 & -1 \\ 0 & -1 \end{pmatrix}, \quad \begin{pmatrix} -1 & -1 \\ 1 & -1 \end{pmatrix}.$$

The critical points are, respectively, a

$$(4.D.37) \quad \text{saddle, generic sink, nongeneric sink, spiral sink.}$$

We provide a rather sketchy depiction of the phase portraits of the blow ups. We sketch orbits on  $S^1 = \{|y| = 1\}$ . Other orbits sketched are confined to  $\{|y| > 1\}$ , corresponding to  $\{r > 0\}$ , and we only sketch orbits that lead to (or from) critical points of  $X$  on  $S^1$ , except for the spiral sink, where  $X$  has no critical points.



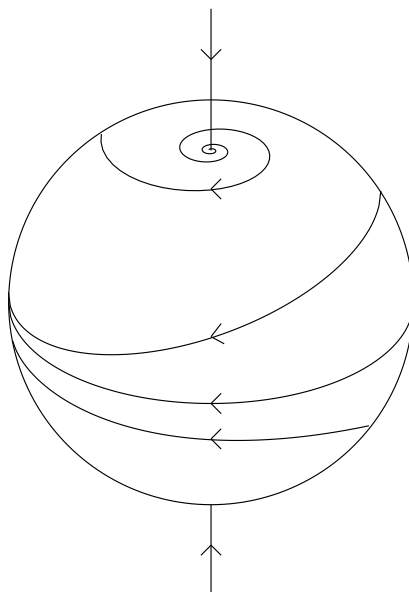
**Figure 4.D.5.** Blowup of a 3D saddle

Figure 4.D.1, depicting the blowup of a saddle, has  $\pm e_1$  and  $\pm e_2$  as critical points of  $X$ , each one in turn being a saddle. Figure 4.D.2, depicting the blowup of a generic sink, also has  $\pm e_1$  and  $\pm e_2$  as critical points of  $X$ . In this case,  $\pm e_1$  are saddles and  $\pm e_2$  are sinks. In Figure 4.D.3, depicting a non-generic sink, the only eigenvectors of  $A$  in  $S^1$  are  $\pm e_1$ , and these are the only critical points of  $X$ . Consistent with Proposition 4.D.2, these critical points are degenerate, and the orbits pictured here illustrate this. Figure 4.D.4 deals with a spiral sink. In this case,  $A$  has no real eigenvectors, so, as noted above,  $X$  has no critical points on  $S^1$ . This figure illustrates how orbits of  $X$  spiral into  $S^1$ .

In Figure 4.D.5 we depict the blowup of a 3D saddle. In this case,  $F(x)$  has the form (4.D.7), with

$$(4.D.38) \quad A = \begin{pmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \lambda_3 \end{pmatrix}, \quad \lambda_3 < 0 < \lambda_2 < \lambda_1.$$

In this situation, the critical points of  $X$  on  $S^2 = \{|y| = 1\}$  are  $e_j$ ,  $1 \leq j \leq 3$ , each of which is a saddle, though the saddles are of three different types. When one restricts the flow generated by  $X$  to  $S^2$ , one sees that  $\pm e_1$  are sinks,  $\pm e_2$  are saddles, and  $\pm e_3$  are sources. Note the heteroclinic orbits connecting these various critical points.



**Figure 4.D.6.** Blown up of a 3D spiral sink

In Figure 4.D.6 we depict the behavior of the blowup of a 3D spiral sink. In this case,  $F(x)$  has the form (4.D.7), with

$$(4.D.39) \quad A = \begin{pmatrix} B & \\ & \lambda_3 \end{pmatrix}, \quad B = \begin{pmatrix} -1 & -\mu \\ \mu & -1 \end{pmatrix}, \quad \mu, \lambda_3 \in (-\infty, 0).$$

The eigenvalues of  $B$  are  $-1 \pm \mu i$ . There are two critical points of  $X$  on  $S^2$ , at  $\pm e_3$ . Figure 4.D.6 depicts the case  $\lambda_3 = -2$ . In this case,  $\pm e_3$  are spiral saddles for  $X$ . When one restricts the flow generated by  $X$  to  $S^2$ , the points  $\pm e_3$  are spiral sources. This is somewhat analogous to the saddle behavior of  $\pm e_1$  in Figure 4.D.2. In addition, the equator of  $S^2$  (in the  $(e_1 e_2)$ -plane) is an attracting cycle for the flow generated by  $X$ .

The reader is invited to consider the behavior of blowups in case (4.D.39) when one takes  $\lambda_3 = -1/2$ , or  $-1$ .



#### 4.E. Periodic solutions of $x'' + x = \varepsilon\psi(x)$

Equations of the form

$$(4.E.1) \quad x'' + x = \varepsilon\psi(x)$$

with “small”  $\varepsilon$  arise in a number of cases, and it is of interest to analyze various features of these solutions. For example, as mentioned in §4.6, the relativistic correction for planetary motion gives rise to the equation (4.6.53), which takes the form (4.E.1) for  $x = u - A$ , with

$$(4.E.2) \quad \psi(x) = (x + A)^2.$$

Another example,

$$(4.E.3) \quad \psi(x) = -x^3,$$

yields a special case of Duffing’s equation. As we mentioned in §4.6, solutions to (4.6.53) tend not to be periodic of period  $2\pi$ , and this leads to the phenomenon of precession of perihelia. It is of general interest to compute the period of a solution to (4.E.1), and we discuss this problem here. We assume  $\psi$  is smooth.

We rewrite (4.E.1) as a first order system and also explicitly record the dependence on  $\varepsilon$ :

$$(4.E.4) \quad \begin{aligned} x'_\varepsilon(t) &= y_\varepsilon(t), \\ y'_\varepsilon(t) &= -x_\varepsilon(t) + \varepsilon\psi(x_\varepsilon(t)). \end{aligned}$$

We pick  $a \in (0, \infty)$  and impose the initial conditions

$$(4.E.5) \quad x_\varepsilon(0) = a, \quad y_\varepsilon(0) = 0.$$

Note that if we take

$$(4.E.6) \quad F_\varepsilon(x, y) = \frac{y^2}{2} + \frac{x^2}{2} - \varepsilon\Psi(x),$$

where  $\Psi'(x) = \psi(x)$ , then  $(d/dt)F_\varepsilon(x_\varepsilon(t), y_\varepsilon(t)) = 0$  for solutions to (4.E.4), so orbits of (4.E.4) lie on level curves of  $F_\varepsilon$ . For  $\varepsilon$  sufficiently small with respect to  $a$ , the level curves of  $F_\varepsilon$  on  $\{(x, y) : x^2 + y^2 \leq 2a^2\}$  will be close to those of  $F_0$ , that is to say, such level curves of  $F_\varepsilon$  will be closed curves, close to circles, and the associated solutions to (4.E.4)–(4.E.5) will be periodic. The period  $T(\varepsilon)$  will have the following two properties, at least for small  $\varepsilon$ :

$$(4.E.7) \quad T(\varepsilon) = 2\pi + O(\varepsilon), \quad y_\varepsilon(T(\varepsilon)) = 0.$$

We will calculate a more precise approximation to  $T(\varepsilon)$ , accurate for small  $\varepsilon$ .

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 §4.2 that  $x_\varepsilon(t)$  and  $y_\varepsilon(t)$  are smooth functions of  $\varepsilon$ . Hence, for each  $N \in \mathbb{N}$ , we can write

$$(4.E.8) \quad \begin{aligned} x_\varepsilon(t) &= a \cos t + \sum_{k=1}^N X_k(t)\varepsilon^k + R_{1N}(t, \varepsilon), \\ y_\varepsilon(t) &= -a \sin t + \sum_{k=1}^N Y_k(t)\varepsilon^k + R_{2N}(t, \varepsilon), \end{aligned}$$

where

$$(4.E.9) \quad |R_{jN}(t, \varepsilon)| \leq C_{KN}\varepsilon^{N+1}, \quad \forall |t| \leq K.$$

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 (4.E.8)

$$(4.E.10) \quad x''_\varepsilon(t) + x_\varepsilon(t) = \sum_{k=1}^N [X''_k(t) + X_k(t)]\varepsilon^k + O(\varepsilon^{N+1}),$$

while

$$(4.E.11) \quad \begin{aligned} \varepsilon\psi(x_\varepsilon(t)) &= \varepsilon\psi\left(a \cos t + \sum_{k=1}^N X_k(t)\varepsilon^k\right) + O(\varepsilon^{N+1}) \\ &= \varepsilon\left[\psi(a \cos t) + \sum_{j=1}^N \frac{1}{j!}\psi^{(j)}(a \cos t)\left(\sum_{k=1}^N X_k(t)\varepsilon^k\right)^j\right] + O(\varepsilon^{N+1}). \end{aligned}$$

We match up the coefficients of  $\varepsilon^k$  in (4.E.10) and (4.E.11) to obtain equations for  $X_k(t)$ . The case  $k = 1$  gives

$$(4.E.12) \quad X''_1(t) + X_1(t) = \psi(a \cos t),$$

and from (4.E.5) the initial conditions are seen to be

$$(4.E.13) \quad X_1(0) = 0, \quad X'_1(0) = 0.$$

The solution to (4.E.12)–(4.E.13) is given by Duhamel's formula, cf. (3.4.9) of Chapter 3:

$$(4.E.14) \quad X_1(t) = \int_0^t \sin(t-s)\psi(a \cos s) ds.$$

It is convenient to expand  $\sin(t-s)$  and rewrite (4.E.14) as

$$(4.E.15) \quad X_1(t) = (\sin t) \int_0^t \cos s \psi(a \cos s) ds - (\cos t) \int_0^t \sin s \psi(a \cos s) ds.$$

Regarding  $Y_k(t)$ , we have

$$(4.E.16) \quad Y_k(t) = X'_k(t),$$

for all  $k$ , and in particular

$$(4.E.17) \quad Y_1(t) = (\cos t) \int_0^t \cos s \psi(a \cos s) ds + (\sin t) \int_0^t \sin s \psi(a \cos s) ds.$$

In case  $\psi(x)$  is given by (4.E.2), we have

$$(4.E.18) \quad \psi(a \cos s) = \frac{a^2}{2} \cos 2s + 2aA \cos s + \left(A^2 + \frac{a^2}{2}\right),$$

hence

$$(4.E.19) \quad \begin{aligned} &\int_0^t \cos s \psi(a \cos s) ds \\ &= \left(A^2 + \frac{3a^2}{4}\right) \sin t + \frac{aA}{2} \sin 2t + \frac{a^2}{12} \sin 3t + aAt, \end{aligned}$$

and

$$(4.E.20) \quad \int_0^t \sin s \psi(a \cos s) ds \\ = \left( A^2 + \frac{aA}{2} + \frac{a^2}{2} \right) - \left( A^2 + \frac{5a^2}{12} \right) \cos t - \frac{aA}{2} \cos 2t - \frac{a^2}{12} \cos 3t.$$

One can compute higher terms in (4.E.8). For example, matching up coefficients of  $\varepsilon^2$  in (4.E.10) and (4.E.11) yields

$$(4.E.21) \quad X_2''(t) + X_2(t) = \psi'(a \cos t) X_1(t).$$

Again  $X_2(0) = X_2'(0) = 0$ , and, parallel to (4.E.14), we have

$$(4.E.22) \quad X_2(t) = \int_0^t \sin(t-s) \psi'(a \cos s) X_1(s) ds.$$

$Y_2(t)$  is given by (4.E.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 (4.E.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

$$(4.E.23) \quad T_0(\varepsilon) = 2\pi, \quad T_{n+1}(\varepsilon) = T_n(\varepsilon) - \frac{y_\varepsilon(T_n(\varepsilon))}{y'_\varepsilon(T_n(\varepsilon))}.$$

(A general treatment of Newton's method is given in Chapter 5 of [49].) This sequence converges fast:

$$(4.E.24) \quad T(\varepsilon) = T_n(\varepsilon) + O(\varepsilon^{2^n}),$$

provided one has  $y_\varepsilon(t)$  evaluated exactly. Given an approximation to  $y_\varepsilon(t)$ ,

$$(4.E.25) \quad y_\varepsilon(t) = \tilde{y}_\varepsilon(t) + O(\varepsilon^N), \quad y'_\varepsilon(t) = \tilde{y}'_\varepsilon(t) + O(\varepsilon^N),$$

we can work with  $\tilde{T}_n(\varepsilon)$ , given by

$$(4.E.26) \quad \tilde{T}_0(\varepsilon) = 2\pi, \quad \tilde{T}_{n+1}(\varepsilon) = \tilde{T}_n(\varepsilon) - \frac{\tilde{y}_\varepsilon(\tilde{T}_n(\varepsilon))}{\tilde{y}'_\varepsilon(\tilde{T}_n(\varepsilon))},$$

and we get

$$(4.E.27) \quad T(\varepsilon) = \tilde{T}_n(\varepsilon) + O(\varepsilon^N), \quad \text{provided } 2^n \geq N.$$

In particular, taking

$$(4.E.28) \quad y_\varepsilon(t) = a \sin t + Y_1(t)\varepsilon + O(\varepsilon^2),$$

we have

$$(4.E.29) \quad T(\varepsilon) = \tilde{T}_1(\varepsilon) + O(\varepsilon^2),$$

with

$$(4.E.30) \quad \tilde{T}_1(\varepsilon) = 2\pi - \frac{\tilde{y}_\varepsilon(2\pi)}{\tilde{y}'_\varepsilon(2\pi)} \\ = 2\pi + \frac{1}{a} Y_1(2\pi)\varepsilon,$$

hence, by (4.E.17),

$$(4.E.31) \quad T(\varepsilon) = 2\pi + \frac{\varepsilon}{a} \int_0^{2\pi} \cos s \psi(a \cos s) ds + O(\varepsilon^2).$$

In case  $\psi(x)$  is given by (4.E.2), we have from (4.E.19) that

$$(4.E.32) \quad \int_0^{2\pi} \cos s \psi(a \cos s) ds = 2\pi a A,$$

so in this case

$$(4.E.33) \quad T(\varepsilon) = 2\pi(1 + A\varepsilon) + O(\varepsilon^2).$$

Given an approximation  $\tilde{y}_\varepsilon(t)$  satisfying (4.E.25) with  $N = 3$  or  $4$ , we can iterate (4.E.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 (4.E.4), and address a limitation of the approximations of the form (4.E.8). As follows from (4.E.15)–(4.E.20), the first order approximation has the form

$$(4.E.34) \quad \begin{aligned} x_\varepsilon(t) &= a \cos t + X_1(t)\varepsilon + O(\varepsilon^2), \\ y_\varepsilon(t) &= -a \sin t + Y_1(t)\varepsilon + O(\varepsilon^2), \end{aligned}$$

and, in the case that  $\psi(x)$  is given by (4.E.2),

$$(4.E.35) \quad \begin{aligned} X_1(t) &= X_1^b(t) + aAt \sin t, \\ Y_1(t) &= Y_1^b(t) - aAt \cos t, \end{aligned}$$

where  $X_1^b(t)$  and  $Y_1^b(t)$  are periodic in  $t$ , of period  $2\pi$ , being sums of products of  $\sin kt$  and  $\cos kt$  ( $0 \leq k \leq 3$ ). In (4.E.34), 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 (4.E.35) 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 (4.E.34) are concerned,

$$(4.E.36) \quad \begin{aligned} &a \cos t + X_1^b(t)\varepsilon \quad \text{and} \\ &-a \sin t + Y_1^b(t)\varepsilon \end{aligned}$$

are uniformly bounded, of period  $2\pi$ , but

$$(4.E.37) \quad aA\varepsilon t \sin t \quad \text{and} \quad -aA\varepsilon t \cos t$$

are unbounded as  $|t| \rightarrow \infty$ . These terms are called *secular terms*, and it is desirable to have a replacement for (4.E.8), in which such secular terms do not appear. To get this, we proceed as follows.

The functions

$$(4.E.38) \quad x_\varepsilon^\#(t) = x_\varepsilon\left(\frac{T(\varepsilon)t}{2\pi}\right), \quad y_\varepsilon^\#(t) = y_\varepsilon\left(\frac{T(\varepsilon)t}{2\pi}\right)$$

are periodic of period  $2\pi$  in  $t$  and smooth in  $\varepsilon$ . Hence we have expansions

$$(4.E.39) \quad \begin{aligned} x_\varepsilon^\#(t) &= a \cos t + \sum_{k=1}^N X_k^\#(t) \varepsilon^k + O(\varepsilon^{N+1}), \\ y_\varepsilon^\#(t) &= -a \sin t + \sum_{k=1}^N Y_k^\#(t) \varepsilon^k + O(\varepsilon^{N+1}). \end{aligned}$$

Note that

$$(4.E.40) \quad \frac{d}{dt} x_\varepsilon^\#(t) = \frac{T(\varepsilon)}{2\pi} y_\varepsilon^\#(t),$$

which leads to a variant of (4.E.16). We have the following.

**Proposition 4.E.1.** *The solution to (4.E.4)–(4.E.5) has the expansion*

$$(4.E.41) \quad \begin{aligned} x_\varepsilon(t) &= a \cos \frac{2\pi t}{T(\varepsilon)} + \sum_{k=1}^N X_k^\# \left( \frac{2\pi t}{T(\varepsilon)} \right) \varepsilon^k + O(\varepsilon^{N+1}), \\ y_\varepsilon(t) &= -a \sin \frac{2\pi t}{T(\varepsilon)} + \sum_{k=1}^N Y_k^\# \left( \frac{2\pi t}{T(\varepsilon)} \right) \varepsilon^k + O(\varepsilon^{N+1}). \end{aligned}$$

*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

$$(4.E.42) \quad X_0(t) = X_0^\#(t) = a \cos t, \quad Y_0(t) = Y_0^\#(t) = -a \sin t.$$

It remains to compute  $X_k^\#(t)$  and  $Y_k^\#(t)$  for  $k \geq 1$ . To this end, set

$$(4.E.43) \quad \frac{T(\varepsilon)}{2\pi} = 1 + \gamma(\varepsilon), \quad \gamma(\varepsilon) = \varepsilon \sum_{\ell \geq 0} \gamma_\ell \varepsilon^\ell.$$

If we compare the expressions for  $x_\varepsilon(t)$  in (4.E.8) and (4.E.41) and make the substitution  $s = 2\pi t/T(\varepsilon)$ , we obtain

$$(4.E.44) \quad \begin{aligned} \sum_{k \geq 0} X_k^\#(s) \varepsilon^k &= \sum_{i \geq 0} X_i(s + \gamma(\varepsilon)s) \varepsilon^i \\ &= \sum_{i \geq 0} \sum_{j \geq 0} \frac{1}{j!} X_i^{(j)}(s) s^j \gamma(\varepsilon)^j \varepsilon^i \\ &= \sum_{i \geq 0} \sum_{j \geq 0} \frac{1}{j!} X_i^{(j)}(s) s^j \left( \sum_{\ell \geq 0} \gamma_\ell \varepsilon^\ell \right)^j \varepsilon^{i+j}. \end{aligned}$$

We conclude that  $X_k^\#(s)$  is equal to the coefficient of  $\varepsilon^k$  in the last power series. For  $k = 0$ , we get

$$(4.E.45) \quad X_0^\#(s) = X_0(s) = a \cos s,$$

as already noted in (4.E.42). For  $k = 1$ , we get

$$(4.E.46) \quad \begin{aligned} X_1^\#(s) &= X_1(s) + \gamma_0 s X_0'(s) \\ &= X_1(s) - \gamma_0 a s \sin s. \end{aligned}$$

When  $\psi(x)$  is given by (4.E.2), we have from (4.E.35) that this is

$$(4.E.47) \quad \begin{aligned} &= X_1^b(s) + aAs \sin s - \gamma_0 a s \sin s \\ &= X_1^b(s), \end{aligned}$$

the last identity by (4.E.43) and (4.E.33), 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 (4.E.39). We forego the details.

We end this appendix with an indication of how to extend the scope of (4.E.1). We treat the pendulum equation

$$(4.E.48) \quad u'' + \sin u = 0,$$

and seek information on small oscillations, solving (4.E.48) with initial data

$$(4.E.49) \quad u(0) = a\sqrt{\varepsilon}, \quad u'(0) = 0.$$

Thus we set

$$(4.E.50) \quad x(t) = \frac{1}{\sqrt{\varepsilon}} u(t),$$

which solves

$$(4.E.51) \quad x'' + \frac{\sin \sqrt{\varepsilon} x}{\sqrt{\varepsilon}} = 0, \quad x(0) = a, \quad x'(0) = 0.$$

If we set

$$(4.E.52) \quad \frac{\sin \tau}{\tau} = 1 - \tau^2 F(\tau), \quad F(\tau) = \frac{1}{3!} - \frac{\tau^2}{5!} + \cdots,$$

we get

$$(4.E.53) \quad \begin{aligned} x'' + x &= \varepsilon x^3 F(\sqrt{\varepsilon} x) \\ &= \varepsilon \frac{x^3}{3!} - \varepsilon^2 \frac{x^5}{5!} + \cdots. \end{aligned}$$

This has a form similar to (4.E.1), but generalized to

$$(4.E.54) \quad x'' + x = \varepsilon\psi(\varepsilon, x),$$

with  $\psi$  smooth in  $(\varepsilon, x)$ . Treatments of the solutions to (4.E.1) and their periods  $T(\varepsilon)$  extend to the case (4.E.54). The reader is invited to work out details.

#### 4.F. 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

$$(4.F.1) \quad F(x) = Gm_1m_2 \frac{p-x}{\|p-x\|^3}.$$

Here  $G$  is the gravitational constant, given by (4.6.64). As indicated in Exercise 6 of §4.6, 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  $\rho$ , and

$$(4.F.2) \quad \rho(p+Ry) = \rho(p+y), \quad \forall R \in O(3), \quad y \in \mathbb{R}^3,$$

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

$$(4.F.3) \quad \|y\| > a \implies \rho(p+y) = 0.$$

The planet's mass is

$$(4.F.4) \quad m_1 = \int \rho(y) \, dy.$$

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

$$(4.F.5) \quad G(x) = Gm_2 \int \frac{y-x}{\|y-x\|^3} \rho(y) \, dy.$$

We will show that if (4.F.2)–(4.F.4) hold and  $\|p-x\| > a$ , then  $F(x) = G(x)$ .

For notational simplicity, we may as well take

$$(4.F.6) \quad p = 0,$$

so

$$(4.F.7) \quad F(x) = -Gm_1m_2 \frac{x}{\|x\|^3}.$$

Note that

$$(4.F.8) \quad F(x) = -\nabla V(x), \quad G(x) = -\nabla W(x),$$

with

$$(4.F.9) \quad V(x) = -\frac{Gm_1m_2}{\|x\|}, \quad W(x) = -Gm_2 \int_{\|y\| \leq a} \frac{1}{\|x-y\|} \rho(y) \, dy,$$

so it suffices to prove that these potential energies coincide for  $\|x\| > a$ , i.e.,

$$(4.F.10) \quad \|x\| > a \implies V(x) = W(x).$$

As a first step toward proving (4.F.10), note that clearly, for all  $R \in O(3)$ ,

$$(4.F.11) \quad V(Rx) = V(x),$$

and furthermore

$$\begin{aligned}
 (4.F.12) \quad W(Rx) &= -Gm_1 \int \frac{1}{\|Rx - y\|} \rho(y) dy \\
 &= -Gm_1 \int \frac{1}{\|Rx - Rz\|} \rho(Rz) dz \\
 &= -Gm_1 \int \frac{1}{\|x - z\|} \rho(z) dz \\
 &= W(x),
 \end{aligned}$$

the second identity by change of variable and the third by (4.F.2). Consequently, we have

$$(4.F.13) \quad V(x) = v(r), \quad W(x) = w(r), \quad r = \|x\|,$$

and it remains to show that

$$(4.F.14) \quad r > a \implies v(r) = w(r).$$

As another step toward showing this, we note that, given  $a \in (0, \infty)$ , there exists  $C < \infty$  such that

$$(4.F.15) \quad \|y\| \leq a, \quad \|x\| \geq a + 1 \implies \left| \frac{1}{\|x\|} - \frac{1}{\|x - y\|} \right| \leq \frac{C}{\|x\|^2},$$

and hence, by (4.F.4), (4.F.9), and (4.F.13), there exists  $C_2 < \infty$  such that

$$\begin{aligned}
 (4.F.16) \quad r = \|x\| \geq a + 1 &\implies |V(x) - W(x)| \leq \frac{C_2}{\|x\|^2} \\
 &\implies |v(r) - w(r)| \leq \frac{C_2}{r^2}.
 \end{aligned}$$

The next step toward establishing (4.F.14) involves the following harmonicity,

$$(4.F.17) \quad \Delta V(x) = 0, \quad \forall x \in \mathbb{R}^3 \setminus 0,$$

where  $\Delta$  is the Laplace operator,

$$(4.F.18) \quad \Delta f(x) = \frac{\partial^2 f}{\partial x_1^2} + \frac{\partial^2 f}{\partial x_2^2} + \frac{\partial^2 f}{\partial x_3^2}.$$

To see this, recall from (1.A.11) of Chapter 1 that (on  $\mathbb{R}^3$ )

$$(4.F.19) \quad f(x) = g(r) \implies \Delta f(x) = g''(r) + \frac{2}{r}g'(r),$$

and by results on Euler equations from §1.15 of Chapter 1,

$$(4.F.20) \quad g''(r) + \frac{2}{r}g'(r) = 0 \iff g(r) = \frac{c_1}{r} + c_2,$$

Since

$$(4.F.21) \quad V(x) = v(r) = -\frac{Gm_1 m_2}{r},$$

we have (4.F.17), and hence we also have

$$(4.F.22) \quad \Delta \left( \frac{1}{\|x - y\|} \right) = 0 \quad \text{for } x \neq y,$$



so a direct consequence of the integral formula (4.F.9) for  $W(x)$  is

$$(4.F.23) \quad \Delta W(x) = 0 \quad \text{for } \|x\| > a.$$

Hence, by (4.F.13), (4.F.19), and (4.F.20),

$$(4.F.24) \quad \begin{aligned} r > a &\implies w''(r) + \frac{2}{r}w'(r) = 0 \\ &\implies w(r) = \frac{c_1}{r} + c_2, \end{aligned}$$

for some constants  $c_1$  and  $c_2$ . This identity together with (4.F.21) and (4.F.16) proves (4.F.14). Hence we have (4.F.10), so indeed, under the hypotheses (4.F.2)–(4.F.4) (and with  $p = 0$ ),

$$(4.F.25) \quad \|x\| > a \implies F(x) = G(x).$$

We mention the following refinement of (4.F.23),

$$(4.F.26) \quad \Delta W = 4\pi Gm_2\rho.$$

This is not needed to establish (4.F.14), so we will not prove it here. A proof can be found in [45], Chapter 3, §4. Further exploration of the relation between the Laplace operator and the “potential” function  $W$ , through (4.F.9), leads to the subject of potential theory, addressed in Chapters 3–5 of [45] 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. [20].

### 4.G. Brouwer's fixed-point theorem

Here we prove the following fixed-point theorem of L. Brouwer, which arose in §4.15. Take

$$(4.G.1) \quad \bar{D} = \{x \in \mathbb{R}^2 : \|x\| \leq 1\}.$$

**Theorem 4.G.1.** *Each smooth map  $F : \bar{D} \rightarrow \bar{D}$  has a fixed point.*

The proof proceeds by contradiction. We are claiming that  $F(x) = x$  for some  $x \in \bar{D}$ . If not, then for each  $x \in \bar{D}$  define  $\varphi(x)$  to be the endpoint of the ray from  $F(x)$  to  $x$ , continued until it hits

$$(4.G.2) \quad \partial D = \{x \in \mathbb{R}^2 : \|x\| = 1\}.$$

An explicit formula is

$$(4.G.3) \quad \begin{aligned} \varphi(x) &= x + t(x - F(x)), \quad t = \frac{\sqrt{b^2 + 4ac} - b}{2a}, \\ a &= \|x - F(x)\|^2, \quad b = 2x \cdot (x - F(x)), \quad c = 1 - \|x\|^2. \end{aligned}$$

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  $\varphi$  would have the following properties:

$$(4.G.4) \quad \varphi : \bar{D} \rightarrow \partial D \text{ smoothly, } x \in \partial D \Rightarrow \varphi(x) = x.$$

Such a map is called a smooth retraction. The contradiction that proves Theorem 4.G.1 is provided by the following result, called Brouwer's no-retraction theorem.

**Theorem 4.G.2.** *There is no smooth retraction  $\varphi : \bar{D} \rightarrow \partial D$  of  $\bar{D}$  onto its boundary.*

**Proof.** This proof, also by contradiction, brings in material developed in §4. Suppose we had such a retraction  $\varphi$ . Consider the closed curve

$$(4.G.5) \quad \gamma : [0, 2\pi] \rightarrow \partial D, \quad \gamma(t) = (\cos t, \sin t),$$

and form

$$(4.G.6) \quad \gamma_s(t) = \varphi(s\gamma(t)), \quad 0 \leq s \leq 1.$$

This would be a smooth family of maps

$$(4.G.7) \quad \gamma_s : [0, 2\pi] \rightarrow \partial D, \quad \gamma_s(0) = \gamma_s(2\pi),$$

such that  $\gamma_1 = \gamma$  and  $\gamma_0(t) = \varphi(0)$  for all  $t$ . The variant of Lemma 4.4.2 given in Exercise 13 of §4.4 implies

$$(4.G.8) \quad \int_{\gamma_s} F(y) \cdot dy \text{ is independent of } s \in [0, 1],$$

for each  $C^1$  vector field  $F$  defined on a neighborhood of  $\partial D$  and satisfying (4.4.4). Clearly the line integral (4.G.7) is 0 for  $s = 0$ , so we deduce that

$$(4.G.9) \quad \int_{\gamma} F(y) \cdot dy = 0$$

for each such vector field. In particular, this would apply to the vector field given by (4.4.19)–(4.4.20), i.e.,

$$(4.G.10) \quad F(x) = \frac{1}{\|x\|^2} \begin{pmatrix} -x_2 \\ x_1 \end{pmatrix},$$

which is smooth on  $\mathbb{R}^2 \setminus 0$  and satisfies (4.4.4) (cf. (4.4.21)). On the other hand, we compute

$$(4.G.11) \quad \int_{\gamma} F(y) \cdot dy = \int_0^{2\pi} (-\sin t, \cos t) \cdot (-\sin t, \cos t) dt \\ = 2\pi,$$

contradicting (4.G.9) and hence contradicting the existence of such a retraction.  $\square$

The fixed-point theorem is valid for all continuous  $F : \bar{D} \rightarrow \bar{D}$ . 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  $\tilde{F} : \bar{D} \rightarrow \bar{D}$  that would also have no fixed point.

Furthermore, Theorem 4.G.1 holds in  $n$  dimensions, i.e., when

$$(4.G.12) \quad \bar{D} = \{x \in \mathbb{R}^n : \|x\| \leq 1\}.$$

The reduction to Theorem 4.G.2, in the setting of (4.G.12), is the same as above, but the proof of Theorem 4.G.2 in the  $n$ -dimensional setting requires a further argument. Proofs using topology can be found in [14] and [32]. Proofs using differential forms can be found in [26], [50], Chapter 5, [45], Chapter 1, and [46], Appendix G. We have no space to introduce differential forms here, but as shown in [45], and also in [1] and [5], they give rise to many important results in the study of differential equations, at the next level.

#### 4.H. Geodesic equations on surfaces

The notion of a geodesic on a surface in Euclidean space was introduced in §4.7. Here we say a bit more about this. Let  $M \subset \mathbb{R}^k$  be a smooth,  $m$ -dimensional surface (classically,  $k = 3, m = 2$ ), and let  $u : [a, b] \rightarrow M$  be a smooth curve. The length of this curve is

$$(4.H.1) \quad L(u) = \int_a^b \|u'(t)\| dt,$$

where  $\|u'(t)\|^2 = u'(t) \cdot u'(t)$ , the dot product taken in  $\mathbb{R}^k$ . We consider smooth curves that are length minimizing, among curves with the same endpoints. Such curves are called geodesics. They have the following property. Let  $u_s$  be a smooth family of curves satisfying

$$(4.H.2) \quad u_s : [a, b] \rightarrow M, \quad u_s(a) \equiv p, \quad u_s(b) \equiv q,$$

with  $u_0 = u$ . Then  $L(u_s) \geq L(u_0)$  for all  $s$ , so

$$(4.H.3) \quad \frac{d}{ds} L(u_s) \Big|_{s=0} = 0.$$

In other words,  $u_0$  is a critical point of the length functional. We define the term “geodesics” to include all such critical paths.

The quantity  $L(u)$  is unchanged under reparametrization. We will reparametrize  $u_0$  so that  $\|u'_0(t)\| \equiv c_0$  is constant. Then

$$(4.H.4) \quad \begin{aligned} \frac{d}{ds} L(u_s) \Big|_{s=0} &= \frac{d}{ds} \int_a^b (u'_s(t) \cdot u'_s(t))^{1/2} dt \Big|_{s=0} \\ &= \frac{1}{2c_0} \int_a^b \frac{\partial}{\partial s} u'_s(t) \cdot u'_s(t) dt \Big|_{s=0}. \end{aligned}$$

Equivalently,

$$(4.H.5) \quad \frac{d}{ds} L(u_s) \Big|_{s=0} = \frac{1}{c_0} \frac{d}{ds} E(u_s) \Big|_{s=0},$$

where

$$(4.H.6) \quad E(u_s) = \frac{1}{2} \int_a^b \|u'_s(t)\|^2 dt$$

is the *energy* of the curve  $u_s : [a, b] \rightarrow M$ . This is exactly the energy functional introduced in (4.7.27), and the analysis in (4.7.28)–(4.7.33) applies.

In particular, if  $k = m + 1$  and  $n(x)$  is the unit normal to  $M$  at  $x$ , then the condition that  $u = u_0$  is a critical path for  $E$  is given by (4.7.32), i.e.,

$$(4.H.7) \quad u''(t) = -u'(t) \cdot \frac{d}{dt} n(u(t)) n(u(t)).$$

Now suppose that  $M$  is a level set of a smooth function  $f : \mathcal{O} \rightarrow \mathbb{R}$ , defined on an open set  $\mathcal{O} \subset \mathbb{R}^k$ , i.e., for some  $c \in \mathbb{R}$ ,

$$(4.H.8) \quad M = \{x \in \mathcal{O} : f(x) = c\}, \quad \nabla f \neq 0 \text{ on } M.$$

Then, for  $x \in M$ ,

$$(4.H.9) \quad n(x) = \frac{\nabla f(x)}{\|\nabla f(x)\|},$$

and the ODE (4.H.7) can be written

$$(4.H.10) \quad u''(t) = -\frac{u'(t) \cdot D^2 f(u(t)) u'(t)}{\|\nabla f(u(t))\|^2} \nabla f(u(t)),$$

where  $D^2 f(x)$  is the  $k \times k$  matrix of second order partial derivatives of  $f$  at  $x$ . Passing from (4.H.9) to (4.H.10) uses the fact that a path  $u(t)$  on  $M$  satisfies  $u'(t) \cdot n(u(t)) = 0$ . The equation (4.H.10) can be written as a first order system:

$$(4.H.11) \quad \begin{aligned} u' &= v, \\ v' &= -\frac{v \cdot D^2 f(u) v}{\|\nabla f(u)\|^2} \nabla f(u). \end{aligned}$$

Solutions to the system (4.H.11) define a flow

$$(4.H.12) \quad \mathcal{F}^t : T\mathcal{O} \longrightarrow T\mathcal{O}, \quad T\mathcal{O} = \mathcal{O} \times \mathbb{R}^k.$$

We have

$$(4.H.13) \quad \mathcal{F}^t : TM \longrightarrow TM, \quad \mathcal{F}^t|_{TM} = \mathcal{G}^t,$$

where

$$(4.H.14) \quad TM = \{(x, v) \in \mathcal{O} \times \mathbb{R}^k : x \in M, v \in T_x M\},$$

where  $T_x M$  denotes the set of vectors  $v \in \mathbb{R}^k$  tangent to  $M$  at  $x$ , i.e.,

$$v \cdot \nabla f(x) = 0,$$

and

$$(4.H.15) \quad \mathcal{G}^t : TM \longrightarrow TM$$

is the geodesic flow, i.e., for  $v \in T_x M$ ,

$$(4.H.16) \quad \mathcal{G}^t(x, v) = (\gamma(t), \gamma'(t)),$$

where  $\gamma(t)$  is the constant speed geodesic on  $M$  satisfying

$$(4.H.17) \quad \gamma(0) = x, \quad \gamma'(0) = v.$$

However, as we illustrate below, it is often the case that

$$(4.H.18) \quad TM \text{ is an unstable invariant surface for the flow } \mathcal{F}^t.$$

This has consequences for the numerical treatment of geodesic curves on  $M$ . Indeed, one has the important task of stabilizing a numerical approximation to the flow  $\mathcal{F}^t$ , so it reliably acts on  $TM$ .

### Numerical considerations

To be specific, suppose we apply a 4th order Runge-Kutta scheme to the first-order system (4.H.11), with step size  $h$ . We start at a point  $(x, v) \in TM$ . At time  $h$  we obtain an approximation

$$(4.H.19) \quad (\tilde{x}(h), \tilde{y}(h)) \text{ to } \mathcal{F}^h(x, v).$$

In fact,

$$(4.H.20) \quad (\tilde{x}(h), \tilde{y}(h)) = \mathcal{F}^h(x, v) + O(h^5).$$

In particular,

$$(4.H.21) \quad \text{dist}((\hat{x}(h), \hat{v}(h)), TM) = O(h^5).$$

We can construct a *retraction* of a neighborhood of  $TM$  in  $T\mathbb{R}^k$  onto  $TM$ , and apply this to get an approximation

$$(4.H.22) \quad \begin{aligned} (\hat{x}(h), \hat{v}(h)) &= \mathcal{G}^h(x, v) + O(h^5), \\ (\hat{x}(h), \hat{v}(h)) &\in TM. \end{aligned}$$

This provides a useful modified Runge-Kutta approximation to the geodesic flow  $\mathcal{G}^t$  on  $TM$ . It remains 4th order accurate.

We next examine how the instability advertised in (4.H.18) arises in case  $M$  is a  $(k-1)$ -dimensional ellipsoid in  $\mathbb{R}^k$ .

### Geodesic equations on ellipsoids

Here we look at the ellipsoid  $M_c \subset \mathbb{R}^k$ , given by

$$(4.H.23) \quad M_c = \{x \in \mathbb{R}^k : f(x) = c\},$$

where

$$(4.H.24) \quad f(x) = x \cdot Ax, \quad A = A^t \in M(k, \mathbb{R}), \text{ positive definite.}$$

We pick  $c \in (0, \infty)$ . In such a case, we have

$$(4.H.25) \quad \nabla f(x) = 2Ax, \quad D^2 f(x) = 2A,$$

and the geodesic equation (4.H.10) becomes (with slightly different notation)

$$(4.H.26) \quad x''(t) = -\frac{x'(t) \cdot Ax'(t)}{Ax(t) \cdot Ax(t)} Ax(t).$$

We write this as a first-order system:

$$(4.H.27) \quad \begin{aligned} x' &= v, \\ v' &= -\varphi(x, v)Ax, \end{aligned}$$

with

$$(4.H.28) \quad \varphi(x, v) = \frac{v \cdot Av}{\|Ax\|^2}.$$

Note that, for  $x \in M_c$ ,

$$(4.H.29) \quad v \in T_x M_c \iff v \cdot Ax = 0.$$

The system (4.H.27) generates a flow  $\mathcal{F}^t$  on  $T\mathbb{R}^k = \mathbb{R}^{2k}$ , specializing to the geodesic flow  $\mathcal{G}^t$  on  $TM_c \subset \mathbb{R}^{2k}$ . Under  $\mathcal{G}^t$ , acting on  $(x, v) \in TM_c$ , the quantity  $\|v\|$  is constant on each orbit. This need not be the case on other orbits of  $\mathcal{F}^t$ , as we will soon see.

In fact, we have the following computations for solutions  $(x, v)$  to (4.H.27):

$$(4.H.30) \quad \frac{d}{dt} x \cdot Ax = 2v \cdot Ax,$$

$$(4.H.31) \quad \frac{d}{dt} \|v\|^2 = 2v \cdot v' = -2\varphi(x, v)v \cdot Ax,$$

and

$$\begin{aligned}
 (4.H.32) \quad \frac{d}{dt} v \cdot Ax &= v \cdot Ax' + v' \cdot Ax \\
 &= v \cdot Av - \varphi(x, v) \|Ax\|^2 \\
 &= 0.
 \end{aligned}$$

In particular, on each orbit  $\gamma$  of  $\mathcal{F}^t$ , the quantity  $v \cdot Ax$  is constant, say

$$(4.H.33) \quad v \cdot Ax = \kappa \text{ on } \gamma,$$

and then (4.H.30)–(4.H.31) yield

$$\begin{aligned}
 (4.H.34) \quad \frac{d}{dt} x \cdot Ax &= 2\kappa, \\
 \frac{d}{dt} \|v\|^2 &= -2\varphi(x, v)\kappa.
 \end{aligned}$$

We see that, on such an orbit,

$$(4.H.35) \quad x(t) \cdot Ax(t) = c + 2\kappa t.$$

If  $\kappa \neq 0$ , then, as  $c + 2\kappa t \searrow 0$ , one has  $x(t) \rightarrow 0$  and, by (4.H.34),  $\|v(t)\| \nearrow \infty$ . Hence the solution to (4.H.27) ceases to exist at  $c + 2\kappa t = 0$ , and we avoid the absurd conclusion that  $x(t) \cdot Ax(t) < 0$  for  $c + 2\kappa t < 0$ .

#### 4.I. Rigid body motion in $\mathbb{R}^n$ and geodesics on $SO(n)$

Suppose there is a rigid body in  $\mathbb{R}^n$ , with a mass distribution at  $t = 0$  given by a function  $\rho(x)$ , which we will assume is piecewise continuous and has compact support. We also assume  $\rho \geq 0$  and it is not identically zero. Suppose the body moves, subject to no external forces, only the constraint of being rigid. We want to describe the motion of such a body. According to the Lagrangian approach to mechanics, we seek a critical path of the integrated kinetic energy, subject to this constraint.

If  $\xi(t, x)$  is the position at time  $t$  of the point on the body whose position at time 0 is  $x$ , then we can write the Lagrangian as

$$(4.I.1) \quad I(\xi) = \frac{1}{2} \int_{t_0}^{t_1} \int_{\mathbb{R}^n} \rho(x) |\dot{\xi}(t, x)|^2 dx dt.$$

Here  $\dot{\xi}(t, x) = \partial \xi / \partial t$ .

Using center of mass coordinates, we will assume that the center of mass of the body is at the origin, and its total linear momentum is zero, so

$$(4.I.2) \quad \xi(t, x) = W(t)x, \quad W(t) \in SO(n),$$

where  $SO(n)$  is the set of rotations of  $\mathbb{R}^n$ , introduced in §2.12. Thus, describing the motion of the body becomes the problem of specifying the curve  $W(t)$  in  $SO(n)$ . We can write (4.I.1) as

$$(4.I.3) \quad I(\xi) = J(W) = \frac{1}{2} \int_{t_0}^{t_1} \int_{\mathbb{R}^n} \rho(x) |W'(t)x|^2 dx dt.$$

We look for an extremum, or other critical point, where we vary the family of paths  $W : [t_0, t_1] \rightarrow SO(n)$  (keeping the endpoints fixed).

We want to reduce the formula (4.I.3) for  $J(W)$  to a single integral, over  $t$ . To do this, we bring in the following.

**Lemma 4.I.1.** *If  $A, B \in M(n, \mathbb{R})$ , then*

$$(4.I.4) \quad \int_{\mathbb{R}^n} \rho(x) (Ax, Bx) dx = \text{Tr}(A\mathcal{I}_\rho B^t),$$

where  $\mathcal{I}_\rho \in M(n, \mathbb{R})$  is defined by

$$(4.I.5) \quad \mathcal{I}_\rho = \int_{\mathbb{R}^n} \rho(x) xx^t dx.$$

**Proof.** It suffices to note that

$$(4.I.6) \quad (Ax, Bx) = \text{Tr}(Axx^t B^t),$$

as a consequence of the identity  $(x, y) = \text{Tr} xy^t$ , for  $x, y \in \mathbb{R}^n$ , regarded as column vectors.  $\square$



Note that  $\mathcal{I}_\rho$  is a symmetric, positive-definite  $n \times n$  matrix. Now, using (4.I.4), we can write the Lagrangian (4.I.3) as

$$(4.I.7) \quad \begin{aligned} J(W) &= \frac{1}{2} \int_{t_0}^{t_1} \operatorname{Tr}(W'(t)\mathcal{I}_\rho W'(t)^t) dt \\ &= \frac{1}{2} \int_{t_0}^{t_1} Q_\rho(W'(t), W'(t)) dt, \end{aligned}$$

where  $Q_\rho$  is the inner product on  $M(n, \mathbb{R})$  defined by

$$(4.I.8) \quad Q_\rho(A, B) = \operatorname{Tr}(A\mathcal{I}_\rho B^t).$$

Note that this inner product is invariant under left multiplication by elements of  $SO(n)$ , i.e.,

$$(4.I.9) \quad \begin{aligned} W \in SO(n) \Rightarrow Q_\rho(WA, WB) &= \operatorname{Tr}(W\mathcal{I}_\rho B^t W^{-1}) \\ &= Q_\rho(A, B). \end{aligned}$$

On the other hand, for  $W \in SO(n)$ ,

$$(4.I.10) \quad Q_\rho(AW, BW) = \operatorname{Tr}(AW\mathcal{I}_\rho W^{-1}B^t),$$

which is equal to  $Q_\rho(A, B)$  for all  $A, B \in M(n, \mathbb{R})$  if and only if  $W\mathcal{I}_\rho = \mathcal{I}_\rho W$ . In turn, this holds for all  $W \in SO(n)$  if and only if  $\mathcal{I}_\rho$  is a scalar multiple of the identity matrix  $I$ .

Finding a critical path  $W : I \rightarrow SO(n)$  for (4.I.7) is a constrained variational problem, similar to those described in (4.7.21)–(4.7.27). Parallel to (4.7.28), the condition for a path  $W$  to be critical is

$$(4.I.11) \quad W''(t) \perp_{T_{W(t)}SO(n)}, \quad \forall t \in I,$$

orthogonality being with respect to the inner product  $Q_\rho$ , i.e.,

$$(4.I.12) \quad A \in T_{W(t)}SO(n) \implies Q_\rho(W''(t), A) = 0.$$

Given  $V \in SO(n)$ , we can define the vector space  $T_V SO(n)$  as the space of all matrices  $W'(0)$ , for smooth curves  $W : (-\varepsilon, \varepsilon) \rightarrow SO(n)$  satisfying  $W(0) = V$ . For example,

$$(4.I.13) \quad T_I SO(n) = \operatorname{Skew}(n) = \{X \in M(n, \mathbb{R}) : X^t = -X\},$$

and, for  $V \in SO(n)$ ,

$$(4.I.14) \quad \begin{aligned} T_V SO(n) &= \{VX : X \in \operatorname{Skew}(n)\} \\ &= \{YV : Y \in \operatorname{Skew}(n)\}. \end{aligned}$$

Comparison with (4.H.1)–(4.H.6) shows that these critical paths are *geodesics* on  $SO(n)$ , where the length of a curve  $W : [t_0, t_1] \rightarrow SO(n)$  is given by

$$(4.I.15) \quad L_\rho(W) = \int_{t_0}^{t_1} Q_\rho(W'(t), W'(t))^{1/2} dt.$$

To proceed, we see from (4.I.12)–(4.I.14) that the condition for  $W : I \rightarrow SO(n)$  to be a critical path for (4.I.7) is

$$(4.I.16) \quad \operatorname{Tr}(W(t)^{-1}W''(t)\mathcal{I}_\rho X) = 0, \quad \forall X \in \operatorname{Skew}(n),$$

upon setting  $A = W(t)X$  in (4.I.12). It is convenient to bring in

$$(4.I.17) \quad Z(t) = W(t)^{-1}W'(t),$$

and derive an equation for  $Z(t)$  from (4.I.16). First, note the following (which echoes part of (4.I.14)).

**Lemma 4.I.2.** *If  $W : I \rightarrow SO(n)$  is a smooth curve, then*

$$(4.I.18) \quad Z(t) \in \text{Skew}(n), \quad \forall t \in I.$$

**Proof.** Differentiating  $W(t)^t W(t) = I$  gives

$$W'(t)^t W(t) = -W(t)^t W'(t),$$

hence

$$Z(t)^t = W'(t)^t W(t) = -Z(t).$$

□

To recast (4.I.16) in terms of  $Z(t)$ , note that (4.I.17) yields

$$(4.I.19) \quad \begin{aligned} Z'(t) &= W(t)^{-1}W''(t) - W(t)^{-1}W'(t)W(t)^{-1}W'(t) \\ &= W(t)^{-1}W''(t) - Z(t)^2. \end{aligned}$$

Now, given  $B \in M(n, \mathbb{R})$ ,

$$(4.I.20) \quad \text{Tr}(BX) = 0 \quad \forall X \in \text{Skew}(n) \iff B = B^t.$$

Hence the condition (4.I.16) is equivalent to the statement that

$$(4.I.21) \quad [Z'(t) + Z(t)^2]\mathcal{I}_\rho \text{ is symmetric.}$$

If we denote the matrix in (4.I.21) by  $B$  and compute  $B - B^t$ , we arrive at the following result.

**Proposition 4.I.3.** *If we define  $Z(t)$  by (4.I.17), the condition that  $W : I \rightarrow SO(n)$  be a critical path for (4.I.7) is equivalent to*

$$(4.I.22) \quad Z'(t)\mathcal{I}_\rho + \mathcal{I}_\rho Z'(t) + Z(t)^2\mathcal{I}_\rho - \mathcal{I}_\rho Z(t)^2 = 0.$$

To work on (4.I.22), let us define

$$(4.I.23) \quad \mathcal{L}_\rho : \text{Skew}(n) \rightarrow \text{Skew}(n), \quad \mathcal{L}_\rho X = \frac{1}{2}(X\mathcal{I}_\rho + \mathcal{I}_\rho X).$$

Then (4.I.22) can be written

$$(4.I.24) \quad 2\mathcal{L}_\rho Z'(t) - [\mathcal{I}_\rho, Z(t)^2] = 0,$$

where, generally,  $[A, B] = AB - BA$ . In turn, if we set

$$(4.I.25) \quad M(t) = \mathcal{L}_\rho Z(t) = \frac{1}{2}(Z(t)\mathcal{I}_\rho + \mathcal{I}_\rho Z(t)),$$

and note that

$$(4.I.26) \quad [\mathcal{I}_\rho, Z^2] = 2[M, Z],$$

we can recast (4.I.24) as

$$(4.I.27) \quad M'(t) = [M(t), Z(t)],$$

or equivalently

$$(4.I.28) \quad M'(t) = [M(t), \mathcal{L}_\rho^{-1}M(t)],$$

a system of ODE with a quadratic nonlinearity. The following result leads to valuable information about  $M(t)$ .

**Proposition 4.I.4.** *Suppose (4.I.27) holds for  $t \in I$ , that  $t_0 \in I$ , and  $M_0 = M(t_0)$ . Then there exists  $U : I \rightarrow SO(n)$  such that*

$$(4.I.29) \quad M(t) = U(t)M_0U(t)^{-1}, \quad t \in I.$$

**Proof.** We produce a linear ODE for  $U(t)$ . Differentiating (4.I.29) gives

$$(4.I.30) \quad \begin{aligned} M'(t) &= U'(t)M_0U(t)^{-1} - U(t)M_0U(t)^{-1}U'(t)U(t)^{-1} \\ &= U'(t)U(t)^{-1}M(t) - M(t)U'(t)U(t)^{-1} \\ &= [M(t), Z(t)], \end{aligned}$$

provided  $Z = -U'U^{-1}$ , i.e.,

$$(4.I.31) \quad U'(t) = U(t)Z(t).$$

To obtain (4.I.29), take  $U$  to solve (4.I.31), with  $U(t_0) = I$ , and verify that  $Z(t) \in \text{Skew}(n) \Rightarrow U(t) \in SO(n)$ .  $\square$

Note that (4.I.29) implies

$$(4.I.32) \quad \|M(t)\| = \|M_0\|, \quad \forall t \in I.$$

Hence, by Proposition 4.1.2, we have:

**Proposition 4.I.5.** *Given  $t_0 \in \mathbb{R}$  and initial data  $M(t_0) = M_0 \in \text{Skew}(n)$ , the system (4.I.28) has a unique solution for all  $t \in \mathbb{R}$ ,  $M : \mathbb{R} \rightarrow \text{Skew}(n)$ .*

Having a solution to (4.I.28), we can retrace our steps, obtaining  $Z(t) = \mathcal{L}_\rho^{-1}M(t)$ , satisfying (4.I.22), and then solve the linear system

$$(4.I.33) \quad W'(t) = W(t)Z(t), \quad W(t_0) = W_0 \in SO(n),$$

to obtain a critical path for (4.I.7).

The identity (4.I.32) says the operator norm  $\|M(t)\|$  is a conserved quantity for solutions to (4.I.28). We record some other conserved quantities.

**Proposition 4.I.6.** *For each solution  $M : \mathbb{R} \rightarrow \text{Skew}(n)$  to (4.I.28) and each  $k \in \mathbb{N}$ , the quantities*

$$(4.I.34) \quad \text{Tr } M(t)^{2k}$$

are independent of  $t$ . So is

$$(4.I.35) \quad Q_\rho(Z(t), Z(t)),$$

with  $Z(t) = \mathcal{L}_\rho^{-1}M(t)$ .

**Proof.** From (4.I.29) we have

$$(4.I.36) \quad M(t)^{2k} = U(t)M_0^{2k}U(t)^{-1},$$

and taking traces yields (4.I.34). To get (4.I.35), note that, when  $W : \mathbb{R} \rightarrow SO(n)$  is a critical path for (4.I.7), then

$$(4.I.37) \quad \frac{d}{dt}Q_\rho(W'(t), W'(t)) = 2Q_\rho(W''(t), W'(t)) = 0,$$

the last identity by (4.I.11)–(4.I.12). Since  $Z(t) = W(t)^{-1}W'(t)$ , (4.I.9) gives

$$(4.I.38) \quad Q_\rho(Z(t), Z(t)) = Q_\rho(W'(t), W'(t)),$$

and we have (4.I.35).  $\square$

NOTE. The conserved quantities listed in Proposition 4.I.6 include two quadratic forms in  $Z$ , namely

$$(4.I.39) \quad \begin{aligned} Q_\rho(Z, Z) &= -\operatorname{Tr} Z\mathcal{I}_\rho Z, \\ Q_m(Z, Z) &= \frac{1}{4}\operatorname{Tr}(Z\mathcal{I}_\rho + \mathcal{I}_\rho Z)^2. \end{aligned}$$

Let us specialize to  $n = 3$ . Assume the standard basis  $\{e_1, e_2, e_3\}$  of  $\mathbb{R}^3$  diagonalizes  $\mathcal{I}_\rho$ , and set

$$(4.I.40) \quad Z = \kappa(x) = \begin{pmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{pmatrix}, \quad \mathcal{I}_\rho = \begin{pmatrix} a_1 & & \\ & a_2 & \\ & & a_3 \end{pmatrix}.$$

Here the isomorphism  $\kappa : \mathbb{R}^3 \rightarrow \operatorname{Skew}(3)$  is chosen to satisfy

$$(4.I.41) \quad x \times y = \kappa(x)y, \quad x, y \in \mathbb{R}^3,$$

where  $x \times y$  is the cross product on  $\mathbb{R}^3$ . Compare §2.12, Exercise 9. A calculation gives

$$(4.I.42) \quad \begin{aligned} -\operatorname{Tr} Z\mathcal{I}_\rho Z &= (a_2 + a_3)x_1^2 + (a_1 + a_3)x_2^2 + (a_1 + a_2)x_3^2 \\ &= x \cdot \mathcal{J}_\rho x, \end{aligned}$$

where

$$(4.I.43) \quad \begin{aligned} \mathcal{J}_\rho &= (\operatorname{Tr} \mathcal{I}_\rho)I - \mathcal{I}_\rho \\ &= \begin{pmatrix} a_2 + a_3 & & \\ & a_1 + a_3 & \\ & & a_1 + a_2 \end{pmatrix} \\ &= \begin{pmatrix} \alpha_1 & & \\ & \alpha_2 & \\ & & \alpha_3 \end{pmatrix}. \end{aligned}$$

Next, we have

$$(4.I.44) \quad \begin{aligned} M &= Z\mathcal{I}_\rho + \mathcal{I}_\rho Z \\ &= \frac{1}{2} \begin{pmatrix} 0 & -\alpha_3 x_3 & \alpha_2 x_2 \\ \alpha_3 x_3 & 0 & -\alpha_1 x_1 \\ -\alpha_2 x_2 & \alpha_1 x_1 & 0 \end{pmatrix}, \end{aligned}$$

hence

$$(4.1.45) \quad \begin{aligned} -\operatorname{Tr} M^2 &= \|M\|_{HS}^2 = \frac{1}{2}(\alpha_1^2 x_1^2 + \alpha_2^2 x_2^2 + \alpha_3^2 x_3^2) \\ &= \frac{1}{2}x \cdot \mathcal{J}_\rho^2 x. \end{aligned}$$

Note also that

$$(4.1.46) \quad M = \kappa(\mathcal{J}_\rho x).$$

We want to rewrite the equations (4.1.27)–(4.1.28) as equations for

$$(4.1.47) \quad x(t) = \kappa^{-1}Z(t), \quad y(t) = \kappa^{-1}M(t) = \mathcal{J}_\rho x(t),$$

using the cross product. Complementing (4.1.41), we have

$$(4.1.48) \quad \kappa(x \times y) = [\kappa(x), \kappa(y)];$$

see again §2.12, Exercise 9. Given this, we read off from (4.1.27) that, if  $x$  and  $y$  are given by (4.1.47), then

$$(4.1.49) \quad \frac{dy}{dt} = -x \times y.$$

In this setting,  $x(t)$  is called the *angular velocity* of the rotating body,  $y(t)$  its *angular momentum*, and  $\mathcal{J}_\rho$  the *inertia tensor*. The equation (4.1.49) is the standard form of Euler's equation for the free motion of a rigid body in  $\mathbb{R}^3$ .

We can rederive the conservation laws for  $x \cdot \mathcal{J}_\rho x$  and  $x \cdot \mathcal{J}_\rho^2 x$  directly from (4.1.49), upon noting that  $x \times y$  is orthogonal to  $x$  and to  $y = \mathcal{J}_\rho x$ , hence

$$(4.1.50) \quad \begin{aligned} 0 &= x \cdot \mathcal{J}_\rho x' = \frac{1}{2} \frac{d}{dt} x \cdot \mathcal{J}_\rho x, \\ 0 &= \mathcal{J}_\rho x \cdot \mathcal{J}_\rho x' = \frac{1}{2} \frac{d}{dt} x \cdot \mathcal{J}_\rho^2 x. \end{aligned}$$

Explicitly, the conservation laws we get are

$$(4.1.51) \quad \begin{aligned} \alpha_1 x_1^2 + \alpha_2 x_2^2 + \alpha_3 x_3^2 &= C_1, \\ \alpha_1^2 x_1^2 + \alpha_2^2 x_2^2 + \alpha_3^2 x_3^2 &= C_2, \end{aligned}$$

since we have chosen coordinates on  $\mathbb{R}^3$  so that  $\mathcal{J}_\rho$  is given by (4.1.43), and  $\mathcal{I}_\rho$  by (4.1.40). Note also that

$$(4.1.52) \quad a_1 > a_2 > a_3 > 0 \implies 0 < \alpha_1 < \alpha_2 < \alpha_3,$$

and more generally  $a_1 \geq a_2 \geq a_3 > 0 \implies 0 < \alpha_1 \leq \alpha_2 \leq \alpha_3$ .

Given that  $y = \mathcal{J}_\rho x$ , we have

$$(4.1.53) \quad x \times y = \begin{pmatrix} (\alpha_3 - \alpha_2)x_2x_3 \\ (\alpha_1 - \alpha_3)x_1x_3 \\ (\alpha_2 - \alpha_1)x_1x_2 \end{pmatrix},$$

and (4.1.49) becomes

$$(4.1.54) \quad \begin{aligned} \alpha_1 x_1' + (\alpha_3 - \alpha_2)x_2x_3 &= 0, \\ \alpha_2 x_2' + (\alpha_1 - \alpha_3)x_1x_3 &= 0, \\ \alpha_3 x_3' + (\alpha_2 - \alpha_1)x_1x_2 &= 0. \end{aligned}$$

If any of the quantities  $\alpha_\ell$  coincide, the system (4.I.54) simplifies. For example, if  $\alpha_1 = \alpha_2$ , we get  $x'_3 = 0$ , hence  $x_3 = \xi_3 = \text{const.}$ , and

$$(4.I.55) \quad \begin{aligned} x'_1 &= -\gamma\xi_3x_2, \\ x'_2 &= \gamma\xi_3x_1 \end{aligned}$$

(with  $\gamma = (\alpha_3 - \alpha_1)/\alpha_1$ ), a constant coefficient linear system. If  $\alpha_1, \alpha_2, \alpha_3$  are all distinct, as in (4.I.52), then we can deduce from (4.I.51) identities of the form

$$(4.I.56) \quad \begin{aligned} x_1^2 + \gamma_3x_2^2 &= c_3, \\ x_1^2 + \gamma_2x_3^2 &= c_2, \\ x_2^2 + \gamma_1x_3^2 &= c_1, \end{aligned}$$

and then (4.I.54) transforms into equations such as

$$(4.I.57) \quad x'_1 = A_1(c_3 - x_1^2)^{1/2}(c_2 - x_1^2)^{1/2},$$

etc. The equation (4.I.57) and its analogues for  $x_2$  and  $x_3$  are separable. One gets

$$(4.I.58) \quad \int \frac{dx_1}{\sqrt{(c_3 - x_1^2)(c_2 - x_1^2)}} = A_1 \int dt,$$

the left side being an *elliptic integral*, which one can read about in Chapter 6 of [47].

An alternative presentation of (4.I.51) is

$$(4.I.59) \quad \begin{aligned} y_1^2 + y_2^2 + y_3^2 &= C_2, \\ \beta_1y_1^2 + \beta_2y_2^2 + \beta_3y_3^2 &= C_1, \end{aligned}$$

with  $y_j = \alpha_jx_j$ ,  $\beta_j = 1/\alpha_j$ . One obtains variants of (4.I.54)–(4.I.58), with  $y_j$  in place of  $x_j$ . Note that

$$(4.I.60) \quad 0 < \alpha_1 < \alpha_2 < \alpha_3 \implies 0 < \beta_3 < \beta_2 < \beta_1.$$

One variant of (4.I.56), following from (4.I.59), is

$$(4.I.61) \quad (\beta_1 - \beta_2)y_1^2 - (\beta_2 - \beta_3)y_3^2 = C_1 - \beta_2C_2.$$

Orbits  $y(t)$ , solving (4.I.49) with  $x = \mathcal{J}_\rho^{-1}y$ , lie on curves in the intersection of a sphere  $|y|^2 = C_2$  with a surface given by (4.I.61). See Figure 4.I.1 for an illustration. (In this illustration, the observer is looking down the  $y_2$ -axis, and  $\beta_1 - \beta_2 = \beta_2 - \beta_3$ .)

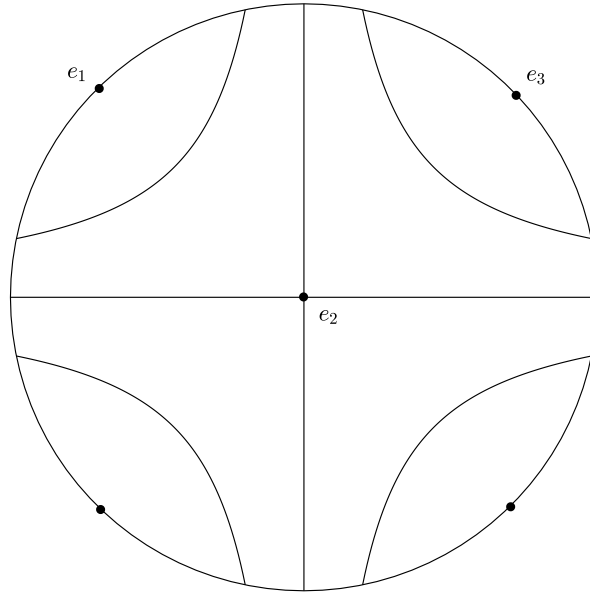
Let us write the system (4.I.49) as

$$(4.I.62) \quad \frac{dy}{dt} = F(y), \quad F(y) = y \times \mathcal{J}_\rho^{-1}y.$$

Then  $F$  is a vector field on  $\mathbb{R}^3$  that is tangent to each sphere  $S_C = \{|y|^2 = C\}$ , and we can regard the solution to (4.I.62) as defining a flow on each such sphere, and  $F|_{S_C}$  as a vector field on  $S_C$ . It has six critical points. In case  $C = 1$ , the critical points are

$$(4.I.63) \quad \begin{aligned} e_1, -e_1, e_3, -e_3, & \text{ centers,} \\ e_2, -e_2, & \text{ saddles.} \end{aligned}$$

The result (4.I.63) has the following significance. Suppose  $\mathcal{B} \subset \mathbb{R}^3$  is a rigid body with inertia tensor  $\mathcal{J}_\rho$  given by (4.I.43), whose diagonal entries satisfy the



**Figure 4.I.1.** Orbits of  $y' = y \times x$  on  $|y|^2 = 1$

hypotheses in (4.I.60). Suppose  $\mathcal{B}$  is set rotating, with angular momentum  $y_0 = y(0)$  at time  $t = 0$ . If  $y_0$  is parallel to any of the six vectors in (4.I.63), then  $\mathcal{B}$  will rotate steadily, with constant angular momentum  $y_0$  (hence constant angular velocity  $x_0 = \mathcal{J}_\rho^{-1}y_0$ ). Furthermore, if  $y_0/|y_0|$  is close to one of the four centers  $\pm e_1, \pm e_3$ , then  $y(t)/|y(t)|$  remains close to such a center for all  $t$ . On the other hand, suppose  $y_0/|y_0|$  is close to but not exactly equal to  $\pm e_2$ . Then  $y(t)/|y(t)|$  travels along a path taking it close to  $-y_0/|y_0|$ , then back to  $y_0/|y_0|$ , infinitely often. Thus rotation of  $\mathcal{B}$  about the  $e_1$  and  $e_3$ -axes is stable, but rotation about the  $e_2$ -axis is unstable. (In this connection, note from (4.I.40) that  $\kappa(e_j) \in \text{Skew}(3)$  generates rotation about the  $e_j$ -axis.)

---

## Bibliography

- [1] R. Abraham and J. Marsden, *Foundations of Mechanics, 2nd Ed.*, Benjamin Cummins, Reading, Mass., 1978.
- [2] R. Abraham and C. Shaw, *Dynamics – The Geometry of Behavior, Vols. 1–3*, Aerial Press, Santa Cruz, 1984.
- [3] R. Adler, M. Bazin, and M. Schiffer, *Introduction to General Relativity*, McGraw-Hill, New York, 1975.
- [4] L. Ahlfors, *Complex Analysis*, McGraw-Hill, New York, 1966.
- [5] V. Arnol'd, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, New York, 1978.
- [6] D. Arrowsmith and C. Place, *An Introduction to Dynamical Systems*, Cambridge Univ. Press, Cambridge, 1990.
- [7] K. Atkinson, *An Introduction to Numerical Methods*, John Wiley and Sons, New York, 1978.
- [8] R. Bartle and D. Sherbert, *Introduction to Real Analysis*, J. Wiley and Sons, New York, 2000.
- [9] W. Boyce and R. DiPrima, *Elementary Differential Equations and Boundary Value Problems, 7th Ed.*, John Wiley, New York, 2001.
- [10] E. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.
- [11] M. DoCarmo, *Differential Geometry of Curves and Surfaces*, Prentice Hall, Englewood Cliffs, New Jersey, 1976.
- [12] F. Gantmacher, *Applications of the Theory of Matrices*, Dover, New York, 2005.
- [13] A. Gray, M. Mezzino, and M. Pinsky, *Introduction to Ordinary Differential Equations with Mathematica*, Springer-Verlag, New York, 1997.
- [14] M. Greenberg and J. Harper, *Algebraic Topology, a First Course*, Addison-Wesley, New York, 1981.
- [15] N. Grossman, *The Sheer Joy of Celestial Mechanics*, Birkhauser, Boston, 1996.
- [16] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.



- 
- [17] 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.
- [18] J. Hale and H. Kocak, *Dynamics and Bifurcations*, Springer-Verlag, New York, 1991.
- [19] P. Hartman, *Ordinary Differential Equations*, Baltimore, 1973.
- [20] W. Heiskanen and M. Moriz, *Physical Geodesy*, W. H. Freeman, San Francisco, 1967.
- [21] D. Henderson, *Differential Geometry – A Geometric Introduction*, Prentice Hall, Upper Saddle River, New Jersey, 1998.
- [22] M. Hirsch and S. Smale, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, New York, 1974.
- [23] M. Hirsch, S. Smale, and R. Devaney, *Differential Equations, Dynamical Systems, and an Introduction to Chaos*, Elsevier Academic Press, New York, 2004.
- [24] J. Hubbard and B. West, *Differential Equations: A Dynamical Systems Approach*, Springer-Verlag, New York, 1995.
- [25] E. Jackson, *Perspectives of Nonlinear Dynamics, Vols. 1–2*, Cambridge Univ. Press, Cambridge, 1991.
- [26] Y. Kannai, *An elementary proof of the no-retraction theorem*, Amer. Math. Monthly 88 (1981), 264–268.
- [27] S. Krantz, *The Elements of Advanced Mathematics*, CRC Press, Boca Raton, 1995.
- [28] N. Lebedev, *Special Functions and Their Applications*, Dover, New York, 1972.
- [29] S. Lefschetz, *Differential Equations: General Theory*, Wiley-Interscience, New York, 1963.
- [30] A. Lichtenberg and M. Lieberman, *Regular and Chaotic Dynamics*, Springer-Verlag, New York, 1982.
- [31] L. Loomis and S. Sternberg, *Advanced Calculus*, Addison-Wesley, New York, 1968.
- [32] J. Munkres, *Topology, a First Course*, Prentice Hall, Englewood Cliffs, New Jersey, 1974.
- [33] J. Murray, *Mathematical Biology*, Springer-Verlag, New York, 1989.
- [34] M. Nowak and R. May, *Virus dynamics - Mathematical Principles of Immunology and Virology*, Oxford Univ. Press, Oxford, 2000.
- [35] J. Oprea, *Differential Geometry and its Applications*, Prentice Hall, Upper Saddle River, New Jersey, 1997.
- [36] L. Perko, *Differential Equations and Dynamical Systems*, Springer-Verlag, New York, 1991.
- [37] J. Polking and D. Arnold, *Ordinary Differential Equations Using MATLAB*, Prentice Hall, Upper Saddle River, New Jersey, 2003.
- [38] J. Polking, A. Boggess, and D. Arnold, *Differential Equations with Boundary Problems*, Prentice Hall, Upper Saddle River, New Jersey, 2006.
- [39] L. Shampine, *Numerical Solution of Ordinary Differential Equations*, Chapman-Hall, New York, 1994.
- [40] G. Simmons, *Differential Equations, with Applications and Historical Notes*, McGraw-Hill, New York, 1982.
- [41] C. Sparrow, *The Lorenz Equation: Bifurcations, Chaos, and Strange Attractors*, Springer-Verlag, New York, 1982.
- [42] J. Stoker, *Differential Geometry*, Wiley-Interscience, New York, 1969.

- 
- [43] G. Strang, *Linear Algebra and its Applications*, Harcourt Brace Jovanovich, San Diego, 1988.
  - [44] C. Taubes, *Modeling Differential Equations in Biology*, Prentice Hall, Upper Saddle River, New Jersey, 2001.
  - [45] M. Taylor, *Partial Differential Equations, Vols. 1–3*, Springer-Verlag, New York, 1996 (2nd Ed., 2011).
  - [46] M. Taylor, *Measure Theory and Integration*, GSM #76, Amer. Math. Soc., Providence RI, 2006.
  - [47] M. Taylor, *Introduction to Complex Analysis*, GSM #202, Amer. Math. Soc., Providence RI, 2019.
  - [48] M. Taylor, *Linear Algebra*, Undergraduate texts #45, Amer. Math. Soc., Providence RI, 2020.
  - [49] M. Taylor, *Introduction to Analysis in One Variable*, Undergraduate texts #47, Amer. Math. Soc., Providence RI, 2020.
  - [50] M. Taylor, *Introduction to Analysis in Several Variables – Advanced Calculus*, Undergraduate texts #46, Amer. Math. Soc., Providence RI, 2020.
  - [51] S. Wiggins, *Introduction to Applied Nonlinear Dynamical Systems and Chaos*, Springer-Verlag, New York, 1990.
  - [52] D. Wodnarz, *Killer Cell Dynamics – Mathematical and Computational Approaches to Immunology*, Springer-Verlag, New York, 2007.



---

# Index

- acceleration, 25, 192
- action, 293
- adjoint, 140
- Airy's equation, 65
- amp, 59
- angular momentum, 284, 416
- angular velocity, 416
- arclength, 192
- attractor, 360
- autonomous system, 244
- autonomous systems, 253
  
- basis, 106
- Bendixson's theorem, 331
- Bernoulli equation, 17
- Bessel function, 70
- Bessel function of the second kind, 72
- Bessel functions, 86
- Bessel's equation, 65, 69, 217
- binormal, 192
- blowup of a critical point, 388
- brachistochrone problem, 302
- Brouwer's fixed-point theorem, 367, 405
- Brouwer's no-retraction theorem, 405
  
- capacitor, 59
- catenary, 21, 299
- Cauchy's inequality, 136
- Cayley-Hamilton theorem, 133, 175
- celestial mechanics, 288
- center, 257
- central force problem, 283
- chain rule, 15, 372
- chaos, 359
  
- characteristic equation, 75
- characteristic polynomial, 42, 123
- circuit, 59, 180
- column vectors, 100
- commute, 104
- compact, 375
- companion matrix, 174
- competing species equations, 352
- complete, 255
- complete elliptic integral, 36
- confluent hypergeometric equation, 227
- conservation laws, 283, 416
- conservation of energy, 279, 283, 297
- conservation of momentum, 281
- constrained variational problem, 294
- cos, 8
- cosh, 11
- coulomb, 60
- Cramer's formula, 119
- critical point, 252, 257, 379
- critically damped, 46, 262
- cross product, 151, 192
- curl, 275
- current, 59
- curvature, 192
- curve, 192
- cycloid, 305
  
- damped oscillatory, 45, 262
- damped pendulum, 38, 261
- derivative, 372
- determinant, 114
- diagonal, 124

- diagonalizable, 124
- difference schemes, 317
- dimension, 106
- divergence, 254
- dot product, 135
- double pendulum, 307, 370
- Duffing equation, 364
- Duhamel formula, 199
- Duhamel's formula, 177, 188, 207, 244, 397
- eigenvalue, 123, 163
- eigenvector, 123, 163
- electrical circuit, 59, 180
- elliptic integral, 33, 417
- energy conservation, 25
- Euler equations, 67, 217, 403
- Euler's constant, 81
- Euler's formula, 1, 8
- exact equation, 274
- existence of solutions, 240
- exponential function, 1, 3
- farad, 59
- first-order system, 161, 174
- Floquet representation, 205
- flow, 253, 379
- force, 25
- force of gravity, 29
- forced pendulum, 53
- Frenet frame, 192
- Frenet-Serret equations, 193, 194
- fundamental theorem of algebra, 75, 123, 157
- fundamental theorem of calculus, 9, 240, 274, 373
- fundamental theorem of linear algebra, 107
- gamma function, 70, 81, 88
- Gaussian elimination, 121
- general relativity, 288
- general solution, 44
- generalized eigenvector, 125, 165
- geodesic, 296, 321, 407, 412
- gradient vector field, 266, 272
- Gramm-Schmidt construction, 137
- Gronwall's inequality, 247
- Hamiltonian system, 280
- Hamiltonian vector field, 257, 281
- hanging cable, 19, 299
- Hankel function, 73
- harmonicity, 403
- henry, 59
- heteroclinic orbit, 266
- Hilbert-Schmidt norm, 141
- hyperbolic critical point, 258
- hypergeometric equation, 226
- inductor, 59
- inertia tensor, 416
- initial condition, 42
- initial value problem, 45
- injective, 104
- inner product, 135
- instability, 321
- integrating factor, 277
- inverse, 104
- inverse function theorem, 374
- isomorphism, 104
- Jordan canonical form, 154
- Jordan curve theorem, 326
- joule, 60
- Kepler problem, 286
- Kepler's laws, 286
- Kepler's second law, 285
- kinetic energy, 25, 279, 308
- Kirchhoff's laws, 59, 180
- Lagrange equation, 292
- Lagrange multiplier, 300
- Lagrangian, 293, 308, 411
- Laplace operator, 86, 403
- Laplace transform, 78, 231
- law of cosines, 151
- law of gravitation, 286
- law of gravity, 26
- law of sines, 151
- Lienard equation, 335
- limit set, 324
- line integral, 272
- linear subspace, 100
- linear transformation, 102
- linearization, 40, 258
- linearly dependent, 106
- linearly independent, 106
- Lipschitz, 240
- logarithm, 5, 229
- logarithm of a matrix, 229
- logistic equation, 336
- Lorenz equations, 359
- lower triangular, 131
- Lyapunov function, 333

- mass, 25
- matrix, 102
- matrix exponential, 161
- matrix multiplication, 103
- matrix representation, 111
- minimal polynomial, 126
- minor, 119
- momentum, 280
  
- Newton, 60
- Newton's law, 25, 279
- Newton's method, 398
- nilpotent, 131
- nonhomogeneous equation, 48
- nonlinear circuit, 334
- normal, 192
- null space, 103
  
- ohm, 59
- operator norm, 140
- orbit, 237, 253
- orthogonal, 137, 148
- orthogonal complement, 138
- orthonormal basis, 136
- overdamped, 46, 262
  
- pendulum, 29, 256, 293, 294, 401
- pendulum, periodically forced, 370
- period, 396
- periodic orbit, 324
- permutation, 116
- phase portrait, 237, 256
- Picard iteration, 240
- Poincaré map, 365
- Poincaré-Bendixson theorem, 326, 347
- positive definite, 146
- positive semidefinite, 146
- potential energy, 25, 279, 308
- potential theory, 402
- power series, 3, 65, 91, 209, 218
- predator-prey equations, 336
- product formula, 14
- Pythagorean theorem, 135
  
- range, 103
- ratio test, 92
- regular singular point, 217
- resistance, 38
- resistor, 59
- resonance, 55
- rigid body, 411
- RLC circuit, 59
- row operation, 120
  
- row vectors, 100
- Runge-Kutta scheme, 318, 408
  
- saddle, 258, 379
- sec, 10
- second order equations, 23
- second-order systems, 185
- secular terms, 399
- self-adjoint, 144
- separation of variables, 18, 87
- separatrices, 34
- simply connected, 273
- sin, 8
- sinh, 11
- sink, 262, 379
- Skew( $n$ ), 412
- skew-adjoint, 144
- SO( $n$ ), 148, 411
- source, 262, 379
- span, 106
- speed, 192
- spring, 57
- standard basis, 106
- stationary action principle, 293
- SU( $n$ ), 148
- surjective, 104
  
- tan, 10
- tangent vector, 192
- tautochrone problem, 306
- torsion, 192
- total energy, 25, 279
- trace, 140
- transposition, 116
- triangle inequality, 136
- trigonometric functions, 8, 171
- trigonometry, 7
  
- undetermined coefficients, 48
- uniqueness of solutions, 240
- unitary, 148
- upper triangular, 131
  
- van der Pol equation, 330
- variable coefficient systems, 198
- variation of parameters, 62, 206
- variational method, 302
- variational problems, 292
- vector, 99
- vector addition, 99
- vector field, 237, 253
- vector space, 99
- velocity, 192

virus dynamics, 350

volt, 59

voltage, 59

Volterra-Lotka equations, 338

watt, 60

Wronskian, 63, 66, 71, 198, 206