Ordinary Differential Equations

Modeling, Structure, and Dynamics

I-Liang Chern

Department of Mathematics
National Taiwan University

Preface

Ordinary differential equations are a basic mathematical language for describing change. A population grows, a temperature relaxes, a mechanical system oscillates, an electrical circuit responds to an input, and a chemical system moves toward or away from equilibrium. In each case, the differential equation expresses a local law of change, while the central mathematical question is global: what behavior does this law produce?

This book develops ordinary differential equations through the interaction of modeling, calculation, geometry, analysis, and structure. The subject does not begin with an equation already given. One must first choose the state variables, identify the mechanisms to be retained, and translate the assumptions of the model into a mathematical law. Modeling therefore forms an essential part of the mathematics: it motivates the equations, while the resulting theory returns to explain the behavior of the models.

The title of the book reflects three complementary themes:

Modeling	: how assumptions about a phenomenon lead to an ODE,
Structure	: how algebraic, geometric, variational, and Hamiltonian ideas organize the equations,
Dynamics	: how solutions evolve, change stability, bifurcate, and approach their long-time behavior.

These themes remain connected throughout the book. Modeling involves the choice of state variables, the identification of mechanisms, the role of parameters, and the passage from assumptions to equations. Structure reveals relations among apparently different equations through modes, operators, symmetries, conservation laws, and geometric formulations. Dynamics returns to the behavior of solutions: equilibria, phase portraits, periodic orbits, bifurcations, attractors, and long-time behavior.

Exact solution methods are developed carefully when they are available, but formulas are only one part of the subject. Many differential equations cannot be solved explicitly, and even when they can, their most important features may be qualitative or structural. The guiding progression is therefore

$$\text{phenomenon} \longrightarrow \text{model} \longrightarrow \text{equation} \longrightarrow \text{structure} \longrightarrow \text{dynamics}.$$

These arrows also point backward. Dynamics may reveal that a model is incomplete or that a parameter has crossed a critical threshold, while structural properties may suggest better variables, conserved quantities, or numerical methods. Modeling and analysis therefore form a continuing cycle rather than a one-way procedure.

Organization of the book. Chapters 1–4 form an introductory course suitable for sophomore students. They begin with scalar models and classical solution methods, proceed through linear oscillation, and culminate in nonlinear planar systems and phase-plane geometry. Chapters 5–11

form an upper-division continuation, developing linear systems, operator methods, variational and Hamiltonian structures, analytic and numerical foundations, and long-time dynamics.

The chapters are also connected by three principal thematic threads:

$$\begin{aligned} \text{nonlinear dynamics}: &\quad \text{Chapters } 1 \longrightarrow 4 \longrightarrow 11, \\ \text{classical nonlinear theory and geometric mechanics}: &\quad \text{Chapters } 2 \longrightarrow 7 \longrightarrow 8, \\ \text{linear theory}: &\quad \text{Chapters } 3 \longrightarrow 5 \longrightarrow 6. \end{aligned}$$

Chapters 9 and 10 provide analytic and numerical foundations for these developments.

Appendix A complements these developments with a complete Python program for constructing a phase portrait of the Brusselator. It illustrates how to display a normalized direction field, nullclines, equilibria, and numerically computed trajectories using `NumPy`, `Matplotlib`, and `SciPy`. The program also provides a template that can be adapted to other planar systems studied in the book.

The first four chapters require mainly calculus, with multivariable calculus and elementary linear algebra entering progressively. Chapters 5–11 require greater mathematical maturity, especially in linear algebra and analysis. They develop matrix and spectral methods, variational and Hamiltonian structures, rigorous existence and stability theory, numerical analysis, and long-time dynamics. These chapters are intended primarily for junior and senior students in mathematics, science, and engineering, as well as readers from other disciplines seeking a more structural treatment of differential equations.

Throughout the book, examples recur from different viewpoints. Oscillators, coupled mechanical and electrical systems, population and reaction models, and other applications connect modeling with linear algebra, geometry, analysis, computation, and dynamical systems. Proofs are included when they clarify the mechanism of a result or prepare the reader for later theory.

Exercises provide practice with individual ideas, homework problems develop more substantial arguments, and projects emphasize modeling, computation, or synthesis. Each chapter concludes with Historical Notes and Further Reading.

Suggested course arrangements. The book contains more material than can be covered in a single standard course. The following paths are guides rather than fixed syllabi.

One-semester introductory course. A first course may be based on Part I:

Chapter 1 $\longrightarrow$ selected topics from Chapter 2 $\longrightarrow$ Chapter 3 $\longrightarrow$ selected topics from Chapter 4.

Chapter 1 should form the starting point because it introduces differential equations through modeling, equilibrium, phase lines, stability, and bifurcation. Chapter 2 supplies the main scalar solution methods and the geometric interpretation of first-order equations. Chapter 3 develops linear oscillators, forcing, resonance, and planar linear systems. Chapter 4 introduces nonlinear models and phase-plane analysis.

For a standard course with limited time, the instructor may emphasize Sections 2.1 and 2.4 in Chapter 2 and the central portions of Chapter 4 on modeling, linearization, invariant regions, and phase portraits. The sections on Pfaffian equations, orthogonal trajectories, envelopes, and selected extended examples may be chosen according to the interests and preparation of the class.

A more computational first course may supplement Part I with selected material from Chapter 10. A more linear-algebraic first course may conclude with selected sections from Chapter 5.

Two-semester sequence. The first semester may follow Part I. The second semester should normally begin with Chapter 5, since linear systems and modal analysis provide a common language for much of the later material. After this common foundation, several coherent paths are possible.

An applied and computational path is

$$\text{Chapter 5} \longrightarrow \text{Chapter 6} \longrightarrow \text{Chapter 9} \longrightarrow \text{Chapter 10} \longrightarrow \text{selected topics from Chapter 11.}$$

This path emphasizes coupled models, modal analysis, Laplace transforms, Green's functions, well-posedness, numerical approximation, dissipative dynamics, Hopf bifurcation, and attractors.

A geometric and mechanical path is

$$\text{Chapter 5} \longrightarrow \text{Chapters 7–8} \longrightarrow \text{Chapter 9} \longrightarrow \text{selected topics from Chapter 11.}$$

This path emphasizes modes, variational principles, Hamiltonian systems, symmetry, conservation laws, analytic foundations, and long-time behavior.

A balanced second course may take Chapters 5 and 9 as a common core, add selected material from Chapter 6 and Chapters 7–8, and conclude with Chapter 10 or Chapter 11 according to whether the principal emphasis is numerical approximation or dynamical systems.

For students in engineering or applied science, examples from circuits, mechanics, reaction systems, populations, and numerical simulation may be used to organize the course. For mathematics students, the same examples may serve as concrete settings for linear algebra, geometry, analysis, and dynamical systems. The examples are therefore not tied to one particular course path; they form a common thread through the book.

No single route is intended to be canonical. The organization is designed so that instructors and independent readers may select a path suited to their background while retaining the main progression from modeling to equations, from equations to structure, and from structure to dynamics.

Learning in the era of artificial intelligence. Artificial intelligence is likely to change not only the speed of learning but also its pattern. Students can use AI tools to review missing background, obtain alternative explanations, check calculations, and explore examples at a pace adapted to their individual preparation. A textbook therefore need not be confined as rigidly as before to a single linear sequence of topics.

The breadth and modular organization of this book are particularly suited to this emerging mode of learning. Its applications are not presented as isolated examples; they are connected through recurring mathematical ideas such as equilibrium, stability, modes, conservation, dissipation, bifurcation, and long-time behavior. Students may enter through a model or application, use AI and computational tools to strengthen the necessary background, and then follow the mathematical connections to other parts of the subject. AI can assist with local explanations and calculations, while the book provides the structure, progression, and mathematical judgment needed to organize these pieces into a coherent understanding of ordinary differential equations.

Acknowledgments. I am especially grateful to Professor Yu-Chen Shu for developing the companion website and its interactive tools, which enable students to explore figures and models dynamically.

I am grateful for the assistance provided by ChatGPT, developed by OpenAI, during the revision of this book. It served as an editorial aid in reviewing the exposition, checking the consistency of notation, and suggesting improvements in organization and language. It also assisted with computational examples and with generating MATLAB code used to produce some of the figures. I reviewed all mathematical arguments and made the final editorial decisions, and I take full responsibility for the content and for any remaining errors.

I-Liang Chern
National Taiwan University
Taipei, Taiwan
September 2026

Contents

Preface **iii**

Orientation to Part I **1**

1 Modeling and Qualitative Behavior of First-Order ODEs **3**
1.1 What is Mathematical Modeling? . . . 4
1.2 Relaxation Processes and Equilibria . . . 4
1.2.1 Linear relaxation . . . 4
1.2.2 Examples from mechanics, heat transfer, chemistry, and economics . . . 5
1.3 Phase Lines and Stability . . . 10
1.3.1 The phase-line method . . . 10
1.3.2 Local stability and linearization . . . 12
1.4 Population Dynamics . . . 15
1.4.1 Malthusian growth . . . 15
1.4.2 Logistic growth . . . 16
1.4.3 The Allee effect and population thresholds . . . 19
1.5 Elementary Bifurcations in One Dimension . . . 21
1.5.1 Saddle-node bifurcation . . . 22
1.5.2 Transcritical bifurcation . . . 26
1.5.3 Bistability and hysteresis . . . 28
1.5.4 Pitchfork bifurcation . . . 33
1.5.5 Interpretation in applications . . . 38

2 Scalar Differential Equations: Methods and Geometry **43**
2.1 Classical Solution Methods . . . 43
2.1.1 Separation of variables . . . 43
2.1.2 Linear first-order equations . . . 47
2.1.3 Reducible nonlinear equations . . . 52
2.2 Geometric Viewpoint of First-Order Equations . . . 55
2.2.1 Direction fields, vector fields, and Pfaffian equations . . . 55
2.2.2 Families of integral curves and first integrals . . . 59
2.2.3 Exact differential forms and integrating factors . . . 62
2.3 Families of Plane Curves . . . 67
2.3.1 From curve families to differential equations . . . 67
2.3.2 Orthogonal trajectories . . . 71

2.3.3 Envelopes . . . 73
2.4 Existence and Uniqueness: A Preview . . . 79
2.4.1 The local existence and uniqueness theorem . . . 80
2.4.2 Maximal intervals of existence . . . 80
2.4.3 Non-crossing and barriers . . . 82
2.4.4 Nonuniqueness and envelopes . . . 83

3 Second-Order Linear Equations and Linear Oscillators 89
3.1 Physical Origins of Second-Order Linear Equations . . . 90
3.1.1 The spring–mass–dashpot system . . . 90
3.1.2 The RLC circuit . . . 91
3.1.3 The standard oscillator equation . . . 93
3.2 Homogeneous Motion and Characteristic Roots . . . 94
3.2.1 Exponential modes and the characteristic equation . . . 94
3.2.2 Linear independence and the Wronskian . . . 95
3.2.3 Classification of characteristic roots . . . 97
3.3 Linear Oscillators . . . 99
3.3.1 Undamped oscillation . . . 100
3.3.2 Damped oscillation . . . 101
3.4 Forced Motion and Duhamel's Principle . . . 103
3.4.1 General solution of the inhomogeneous equation . . . 103
3.4.2 Variation of constants . . . 104
3.4.3 Duhamel's formula . . . 105
3.4.4 Impulse Response and Green's Function . . . 105
3.5 Sinusoidal Forcing, Beating, Resonance, and Synchronization . . . 109
3.5.1 Undamped forced oscillation . . . 109
3.5.2 Damped forced oscillation . . . 112
3.6 Phase Portraits of Planar Linear Systems . . . 114
3.6.1 From a second-order equation to a first-order system . . . 115
3.6.2 From a planar linear system to a second-order equation . . . 116
3.6.3 Eigenvalues and invariant directions . . . 117
3.6.4 Basic phase portraits . . . 118
3.6.5 Trace-determinant classification . . . 120

4 Nonlinear Systems in the Plane 127
4.1 Planar Autonomous Systems . . . 127
4.1.1 Vector fields and phase portraits . . . 128
4.1.2 Equilibria and local behavior . . . 129
4.1.3 Global phase-plane geometry . . . 136
4.2 Interacting Population and Reaction Models . . . 139
4.2.1 Predator–prey dynamics . . . 140
4.2.2 Competition models . . . 144
4.2.3 Epidemic models . . . 150
4.2.4 Chemical reaction systems–the Brusselator . . . 153
4.3 Conservative, Dissipative, and Oscillatory Systems . . . 159
4.3.1 The simple pendulum as a conservative system . . . 159

4.3.2 The damped simple pendulum as a dissipative system . . . 163
4.3.3 The Duffing oscillator . . . 166
4.3.4 The Van der Pol oscillator . . . 170
4.3.5 The FitzHugh–Nagumo system . . . 172
4.4 From Phase Portraits to Time-Dependent Solutions . . . 176
4.4.1 Lotka–Volterra: closed orbits and periodic solutions . . . 176
4.4.2 SIR: phase-plane orbits and epidemic time evolution . . . 177
4.4.3 FitzHugh–Nagumo: excitation pulses and relaxation oscillations . . . 179
4.4.4 What has been solved? . . . 180

Orientation to Part II **187**

5 Linear Systems and Modal Analysis **189**
5.1 Modeling Linear Dynamical Systems . . . 189
5.1.1 Coupled spring–mass systems . . . 190
5.1.2 Coupled RLC circuits . . . 192
5.1.3 Chemical reaction chains . . . 196
5.1.4 Diffusion on a graph . . . 198
5.1.5 State-space formulation . . . 201
5.2 Abstract Linear Theory . . . 203
5.2.1 Existence and uniqueness . . . 204
5.2.2 The solution space . . . 205
5.2.3 Fundamental solutions and fundamental matrices . . . 206
5.2.4 Representation of solutions . . . 208
5.2.5 Nonhomogeneous systems and Duhamel's principle . . . 209
5.3 Modal Representation . . . 211
5.3.1 Eigenvectors and ordinary modes . . . 211
5.3.2 Complex modes and real modal pairs . . . 212
5.3.3 Mode coalescence and generalized eigenvectors . . . 215
5.3.4 Jordan chains and generalized modes . . . 218
5.4 Decoupling and the Structure of a Matrix . . . 221
5.4.1 Invariant subspaces and decoupling . . . 221
5.4.2 Generalized eigenspaces and primary decomposition . . . 223
5.4.3 Kernel filtration and Jordan block sizes . . . 225
5.4.4 Jordan chains, Jordan bases, and Jordan form . . . 227
5.5 Complete Modal Representation and Linear Stability . . . 233
5.5.1 Jordan coordinates and decoupled subsystems . . . 234
5.5.2 Complete modal representation . . . 236
5.5.3 Linear stability . . . 237
5.5.4 Modal analysis of important coupled systems . . . 239
5.6 Matrix Exponential . . . 245
5.6.1 Definition and evolution operator . . . 246
5.6.2 Exponential of a Jordan block . . . 247
5.6.3 Computing the matrix exponential from Jordan form . . . 249
5.6.4 Basic properties of the matrix exponential . . . 250
5.6.5 Product approximation for nonautonomous systems . . . 252

6 Laplace Transform Methods **259**
6.1 Laplace Transform . . . 260
6.1.1 Definition and basic examples . . . 261
6.1.2 Operational rules and operator algebra . . . 263
6.2 Scalar Equations and Polynomial Differential Operators . . . 267
6.2.1 From $P(D)y = f$ to $P(s)Y = I(s) + F(s)$ . . . 267
6.2.2 Partial fractions, poles, and scalar modes . . . 269
6.2.3 Examples from second-order equations . . . 271
6.3 Green's Functions, Convolution, and Impulses . . . 273
6.3.1 Green's functions as inverse operators . . . 273
6.3.2 Convolution and Duhamel's principle . . . 275
6.3.3 Discontinuous and impulsive forcing . . . 277
6.4 Linear Systems, Resolvents, and Modes . . . 280
6.4.1 Laplace transform of first-order systems . . . 281
6.4.2 Properties of the resolvent . . . 283
6.4.3 Resolvent poles, Jordan structure, and modal decomposition . . . 285

Orientation to Part III **293**

7 Variational Principles and Euler–Lagrange Equations **295**
7.1 Variational Problems and the Euler–Lagrange Equation . . . 295
7.1.1 Functionals and variations . . . 296
7.1.2 The Euler–Lagrange equation . . . 297
7.1.3 Constrained variations and Lagrange multipliers . . . 299
7.2 First Integrals and Quadrature . . . 302
7.2.1 Cyclic variables and momentum . . . 302
7.2.2 The Beltrami identity . . . 304
7.2.3 Periods of closed orbits . . . 307
7.2.4 Separatrices and connecting orbits . . . 310
7.3 Lagrangian Mechanics . . . 313
7.3.1 Hamilton's principle . . . 313
7.3.2 Generalized coordinates and the mass matrix . . . 314
7.3.3 Conservation of energy, conjugate momentum, and cyclic coordinates . . . 317
7.3.4 Symmetry and Noether's theorem . . . 319
7.3.5 The Kepler problem . . . 322

8 Hamiltonian Systems and Phase-Space Geometry **331**
8.1 From Lagrangian to Hamiltonian Form . . . 332
8.1.1 From velocity to momentum variables . . . 332
8.1.2 The Legendre transform . . . 333
8.1.3 Hamilton's equations . . . 335
8.1.4 Canonical form of Hamilton's equations . . . 336
8.1.5 Two coupled linear oscillators . . . 338
8.2 Canonical Structure of Hamiltonian Flows . . . 340
8.2.1 Canonical pairing and canonical transformations . . . 340
8.2.2 Invariant energy surfaces . . . 345

8.2.3 Equilibria and linear Hamiltonian systems . . . 345
8.2.4 Liouville's theorem . . . 347
8.2.5 Normal modes, invariant tori, and recurrence . . . 350
8.3 Poisson Brackets, Symmetry, and First Integrals . . . 352
8.3.1 Evolution of observables and the Poisson bracket . . . 352
8.3.2 Generators and the Hamiltonian form of Noether's theorem . . . 354
8.3.3 Commuting first integrals and integrability . . . 356
8.4 Point-Vortex Dynamics . . . 358
8.4.1 Passive tracers and stream functions . . . 358
8.4.2 Hamiltonian equations for point vortices . . . 359
8.4.3 Symmetries and conserved quantities . . . 361
8.4.4 Two-vortex motion . . . 362

Orientation to Part IV **369**

9 Existence, Uniqueness, and Well-Posedness **371**
9.1 Local Existence and Uniqueness . . . 371
9.1.1 The integral formulation . . . 371
9.1.2 Function spaces and the contraction principle . . . 372
9.1.3 The Picard operator and the Cauchy–Lipschitz theorem . . . 375
9.1.4 Existence and uniqueness hypotheses . . . 378
9.2 Gronwall Estimates and Continuous Dependence . . . 381
9.2.1 Gronwall's inequality . . . 381
9.2.2 Uniqueness and stability estimates . . . 383
9.2.3 Continuous dependence on initial data . . . 385
9.2.4 Continuous dependence on parameters . . . 387
9.3 Flows and Differentiable Dependence . . . 390
9.3.1 Flow maps . . . 390
9.3.2 Differentiability with respect to initial data . . . 393
9.3.3 The variational equation and transport along flows . . . 395
9.4 Continuation and Global Existence . . . 399
9.4.1 Maximal extensions and the continuation principle . . . 400
9.4.2 A priori estimates and global existence . . . 403

10 Numerical Methods for Ordinary Differential Equations **411**
10.1 From ODEs to Difference Equations . . . 411
10.2 Euler and Runge–Kutta Methods . . . 413
10.2.1 Euler's method . . . 413
10.2.2 Second-order Runge–Kutta methods . . . 414
10.2.3 The classical fourth-order Runge–Kutta method . . . 415
10.3 Accuracy and Convergence . . . 416
10.3.1 Local truncation error . . . 417
10.3.2 Global error and convergence . . . 417
10.3.3 Adaptive step-size control . . . 419
10.4 Stability and Stiffness . . . 421
10.4.1 Error propagation and absolute stability . . . 422

10.4.2 Stiffness and implicit methods . . . 425
10.5 Long-Time Behavior and Structure-Preserving Methods . . . 428
10.5.1 Artificial damping and growth . . . 429
10.5.2 Symplectic methods . . . 431
10.5.3 Order versus structure preservation . . . 434
10.6 Software for ODEs . . . 438
10.6.1 MATLAB . . . 439
10.6.2 Python . . . 439
10.6.3 Julia . . . 440
10.6.4 Choosing a solver . . . 441

Orientation to Part V **447**

11 Dynamical Systems and Long-Time Behavior **449**
11.1 Invariant Sets and ω-Limit Sets . . . 449
11.1.1 Positive orbits and ω-limit sets . . . 451
11.1.2 Basic properties of ω-limit sets . . . 452
11.2 Attractors and Basins of Attraction . . . 455
11.2.1 Attraction to a Set . . . 455
11.2.2 Basins of attraction . . . 456
11.3 Monotone Quantities and Dissipative Structure . . . 459
11.3.1 Conservation and Monotonicity . . . 460
11.3.2 Mechanical Dissipation and Gyroscopic Forces . . . 461
11.3.3 Gradient Systems . . . 463
11.3.4 Three Notions of Dissipation . . . 467
11.4 Lyapunov Functions and Stability . . . 470
11.4.1 Stability of an Equilibrium . . . 470
11.4.2 Lyapunov Functions . . . 471
11.4.3 Lyapunov's Stability Theorem . . . 472
11.4.4 Applications of Lyapunov's Theorem . . . 474
11.5 LaSalle's Invariance Principle . . . 481
11.5.1 The Zero-Dissipation Set . . . 481
11.5.2 LaSalle's Invariance Principle . . . 482
11.5.3 Applications of LaSalle's Principle . . . 483
11.6 Absorbing Sets and Global Attractors . . . 489
11.6.1 Absorbing Sets . . . 490
11.6.2 Uniform Attraction of Sets . . . 491
11.6.3 Construction of the Global Attractor . . . 491
11.6.4 Examples of Global Attractors . . . 493
11.7 Periodic Attractors and Planar Long-Time Dynamics . . . 498
11.7.1 Periodic attractors . . . 498
11.7.2 The Poincaré–Bendixson theorem . . . 499
11.7.3 The Van der Pol oscillator revisited . . . 501
11.7.4 Why the planar theory is special . . . 505
11.8 Hopf Bifurcation and the Birth of Periodic Orbits . . . 506
11.8.1 From an equilibrium to periodic motion . . . 506

11.8.2 The supercritical Hopf normal form . . . 507
11.8.3 A self-excited mechanical oscillator . . . 508
11.8.4 A chemical oscillator: the Brusselator . . . 510
11.8.5 The subcritical Hopf normal form . . . 511
11.8.6 The Hopf bifurcation theorem . . . 512
11.8.7 Recognizing a Hopf bifurcation in a planar system . . . 514
11.9 Higher-Dimensional Attracting Dynamics . . . 516
11.9.1 The Lorenz system and its equilibria . . . 516
11.9.2 Volume contraction and boundedness . . . 518
11.9.3 The Lorenz attractor . . . 520
11.9.4 Beyond the planar theory . . . 523

A Computing Phase Portraits with Python **529**
A.1 The complete Python program . . . 530
A.2 Building the phase portrait . . . 534
A.3 Interpreting the phase portrait . . . 536
A.4 Changing the parameter . . . 537
A.5 A shorter version using `streamplot` . . . 538
A.6 Saving the figure . . . 538
A.7 Further experiments . . . 539

References **541**

Index **549**

Orientation to Part I

Part I introduces ordinary differential equations through four connected viewpoints: modeling, solution methods, oscillation, and phase-plane geometry. The discussion begins with scalar equations and gradually moves toward systems whose behavior can no longer be understood from an explicit formula alone.

Chapter 1 shows how differential equations arise from balances, rates, and feedback mechanisms. Scalar models introduce equilibria, phase lines, stability, and bifurcation. These ideas make it possible to predict long-time behavior directly from the equation, often without finding its solution.

Chapter 2 develops methods for first-order equations, including separation of variables, integrating factors, exact equations, and changes of variables. Vector fields, integral curves, first integrals, and envelopes connect calculation with geometry. The main lesson is that an equation may be studied through formulas, invariant relations, or the geometry of its solution curves.

Chapter 3 turns to second-order linear equations. Mechanical and electrical oscillators introduce natural frequencies, damping, external forcing, resonance, and energy balance. Rewriting a second-order equation as a first-order planar system also provides the first bridge from scalar equations to systems.

Chapter 4 develops this phase-plane viewpoint for nonlinear systems. Nullclines, invariant regions, linearization, stable and unstable directions, separatrices, and periodic orbits reveal behavior that is usually inaccessible through explicit solution formulas. Population, epidemic, chemical, and neuronal models show how common geometric mechanisms arise in different applications.

The chapters are therefore linked by a common set of questions:

$$\begin{array}{cc} \text{How is the equation derived?} & \text{What structures does it preserve?} \\ \text{Which states are stable?} & \text{What is the long-time behavior?} \end{array}$$

Their central progression is

$$\begin{aligned} \text{scalar models} &\longrightarrow \text{first-order methods and geometry} \\ &\longrightarrow \text{linear oscillation} \longrightarrow \text{nonlinear planar dynamics.} \end{aligned}$$

This progression prepares the later parts of the book. The passage from a second-order equation to a first-order system leads naturally to coupled linear systems and modal analysis. Energy and first integrals anticipate the variational and Hamiltonian structures of classical mechanics. Stability, invariant sets, and periodic motion provide the basic language for the later study of existence theory, numerical approximation, and long-time dynamical behavior.

Chapter 1

Modeling and Qualitative Behavior of First-Order ODEs

Ordinary differential equations arise when a quantity changes according to its present state and the external conditions acting on it. In the simplest case, the state is described by a single real-valued function $y(t)$, and the law of motion takes the form

$$\dot{y} = f(t, y).$$

If the right-hand side does not depend explicitly on time, the equation is called autonomous:

$$\dot{y} = f(y).$$

Throughout the book, a dot denotes differentiation with respect to time:

$$\dot{y} = \frac{dy}{dt}.$$

A prime denotes differentiation with respect to the argument of the function. Thus $y'(t) = dy/dt$ when $y = y(t)$, while $y'(x) = dy/dx$ when $y = y(x)$.

In the abstract theory of differential equations, we often use y or $\mathbf{y}$ for the unknown state. In applications, the notation follows the meaning of the model: v may denote velocity, T temperature, c concentration, N population, p price, and K capital.

Even first-order scalar equations already contain many central ideas: equilibrium, stability, relaxation, long-time behavior, and parameter dependence. The purpose of this chapter is not merely to introduce solvable equations. Rather, we begin to view an ordinary differential equation as a model of change.

The main geometric tool is the phase line. For a scalar autonomous equation $y' = f(y)$, the sign of $f(y)$ determines the direction of motion. Equilibria are the zeros of f, and their stability can often be read from the phase line without solving the equation explicitly.

The last part of the chapter studies parameter dependence. As a parameter changes, equilibria may move, appear, disappear, or change stability. Such a qualitative change in the behavior of a differential equation is called a bifurcation.

1.1 What is Mathematical Modeling?

A mathematical model is a simplified description of a phenomenon in mathematical form. It does not reproduce every detail of the real system. Instead, it selects the variables, mechanisms, and assumptions that are essential for the question being studied.

The first step is to choose the state variable. In abstract notation we may write it as $y(t)$. In applications, however, the notation should reflect the meaning of the variable: $T(t)$ for temperature, $N(t)$ for population, $c(t)$ for concentration, $p(t)$ for price, or $K(t)$ for capital.

The second step is to identify the law of change. In an ordinary differential equation, this law is expressed through the derivative:

$$\dot{y} = f(t, y).$$

The function f represents the mechanism of the model. It may describe growth, decay, damping, forcing, reaction, competition, investment, depreciation, or adjustment toward equilibrium.

The third step is to specify the data. For an initial value problem, this means prescribing

$$y(t_0) = y_0.$$

The differential equation gives the rule of motion, while the initial condition selects one particular trajectory.

Parameters are also part of the model. A parameter affects the evolution but is not treated as an unknown state variable. Changing a parameter may change not only the numerical solution, but also the qualitative behavior of the system.

A model should therefore be judged at two levels. At the modeling level, one asks whether the variables and assumptions capture the main mechanism. At the mathematical level, one asks what the equation implies: existence, uniqueness, stability, and long-time behavior.

In this chapter, we begin with first-order scalar equations. They already illustrate a central principle:

$$\text{local law of change} \quad \Longrightarrow \quad \text{global behavior of solutions.}$$

The equation specifies the instantaneous motion, but the main question is often the long-time consequence of that motion.

1.2 Relaxation Processes and Equilibria

A relaxation process is a motion toward a state of balance. The simplest case occurs when the rate of change is proportional to the displacement from the equilibrium. This leads to a first-order linear equation, but the same idea also appears locally in nonlinear models.

1.2.1 Linear relaxation

Let $y(t)$ be a scalar state variable, and let y^* be the equilibrium value toward which the system is driven. If the restoring tendency is proportional to the displacement $y - y^*$, then the model is

$$y' = -\alpha(y - y^*), \qquad \alpha > 0.$$

The constant α is the relaxation rate. The sign of the right-hand side shows the direction of motion. If $y > y^*$, then $y' < 0$, so the solution decreases. If $y < y^*$, then $y' > 0$, so the solution increases. Thus the motion is always directed toward y^*.

This equation can be solved directly. Setting

$$z(t) = y(t) - y^*,$$

we obtain

$$z' = -\alpha z.$$

Equivalently, for $z \neq 0$,

$$\frac{d}{dt} \ln |z(t)| = -\alpha,$$

or

$$d \ln |z| = -\alpha \, dt.$$

Hence

$$z(t) = z(0)e^{-\alpha t} = (y_0 - y^*)e^{-\alpha t},$$

and therefore

$$y(t) = y^* + (y_0 - y^*)e^{-\alpha t}.$$

Thus every solution approaches the equilibrium *exponentially*:

$$\lim_{t \to \infty} y(t) = y^*.$$

The *relaxation time* is defined by

$$\tau = \frac{1}{\alpha}.$$

After one relaxation time, the distance to equilibrium has been multiplied by e^{-1}:

$$|y(\tau) - y^*| = e^{-1}|y_0 - y^*|.$$

Thus τ gives the natural time scale on which the system forgets a large part of its initial displacement.

1.2.2 Examples from mechanics, heat transfer, chemistry, and economics

The same mathematical structure appears in many different contexts. The state variable and the interpretation change from one problem to another, but the mechanism is similar: the state is driven toward a value at which opposing effects balance.

Motion of a raindrop. Let $v(t)$ denote the downward velocity of a raindrop. Gravity increases the velocity, while air resistance opposes the motion. If the drag force is assumed proportional to velocity, Newton's law gives

$$m\dot{v} = mg - \beta v,$$

where m is the mass, g is the gravitational acceleration, and $\beta > 0$ is the drag coefficient. Equivalently,

$$\dot{v} = -\frac{\beta}{m}\left(v - \frac{mg}{\beta}\right).$$

Thus the terminal velocity is

$$v^* = \frac{mg}{\beta},$$

and the solution is

$$v(t) = v^* + (v(0) - v^*)e^{-\beta t/m}.$$

The relaxation rate is β/m, and the relaxation time is m/β.

Newton's law of heating and cooling. Let $T(t)$ be the temperature of a body in an environment with ambient temperature T_a. Newton's law of heating and cooling states that the rate of change of temperature is proportional to the difference between the body temperature and the ambient temperature:

$$\dot{T} = -\alpha(T - T_a), \qquad \alpha > 0.$$

The equilibrium temperature is

$$T^* = T_a.$$

The solution is

$$T(t) = T_a + (T(0) - T_a)e^{-\alpha t}.$$

Thus the body cools or warms exponentially toward the ambient temperature.

A reversible chemical reaction. Consider a reversible reaction

$$A \rightleftharpoons B.$$

Let $c(t)$ be the concentration of B, and suppose the total concentration of A and B is fixed and equal to a. Then the concentration of A is $a - c(t)$. If the forward and backward reaction rates are proportional to the corresponding concentrations, then

$$\dot{c} = k_1(a - c) - k_2 c,$$

where $k_1, k_2 > 0$. This can be written as

$$\dot{c} = -(k_1 + k_2)\left(c - \frac{k_1 a}{k_1 + k_2}\right).$$

Hence the equilibrium concentration is

$$c^* = \frac{k_1 a}{k_1 + k_2},$$

and the relaxation rate is $k_1 + k_2$.

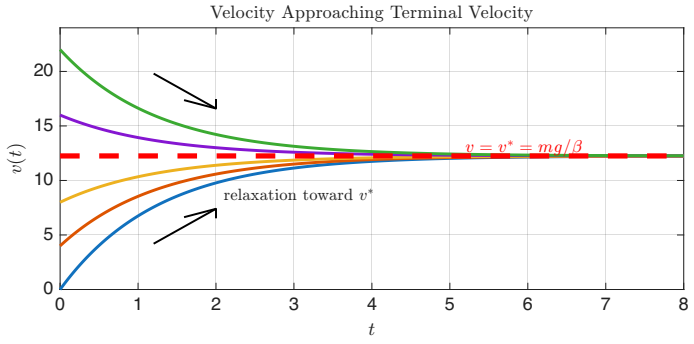


Figure 1.1: Velocity of a falling raindrop approaching the terminal velocity $v^* = mg/\beta$.

A dynamic supply–demand model. Let $p(t)$ be the price of a commodity. Suppose that the price increases when demand exceeds supply and decreases when supply exceeds demand. A simple adjustment model is

$$\dot{p} = \alpha(D(p) - S(p)), \qquad \alpha > 0,$$

where $D(p)$ is the demand and $S(p)$ is the supply. The equilibrium price p^* satisfies

$$D(p^*) = S(p^*).$$

If

$$D(p) = a - bp, \qquad S(p) = c + dp,$$

where $b, d > 0$, then

$$\dot{p} = \alpha(a - c - (b + d)p).$$

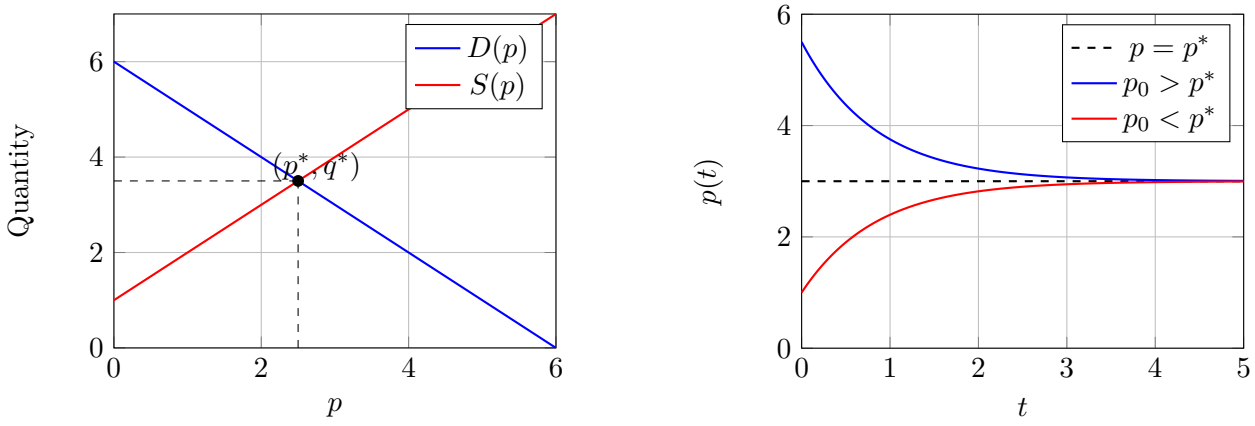
Thus

$$p^* = \frac{a - c}{b + d},$$

and the equation becomes

$$\dot{p} = -\alpha(b + d)(p - p^*).$$

The price relaxes exponentially toward the market-clearing price.



(a) Demand and supply curves. The equilibrium price p^* is determined by the intersection condition $D(p^*) = S(p^*)$.

(b) Dynamic price adjustment. Prices above or below the equilibrium price converge exponentially to p^*.

Figure 1.2: Static and dynamic views of the supply–demand model. The market equilibrium is determined by the intersection of the demand and supply curves, while the dynamic model describes how prices adjust toward the equilibrium.

The Solow growth model. Let $K(t)$ denote the capital stock of an economy. We restrict the model to the physically meaningful region $K \geq 0$. Suppose output is produced according to

$$Y = F(K),$$

and that a fixed fraction s of output is invested, while capital depreciates at rate $\delta > 0$. The capital accumulation equation is

$$\dot{K} = sF(K) - \delta K.$$

For a Cobb–Douglas production function

$$F(K) = AK^{\nu}, \qquad A > 0, \quad 0 < \nu < 1,$$

we obtain

$$\dot{K} = sAK^{\nu} - \delta K.$$

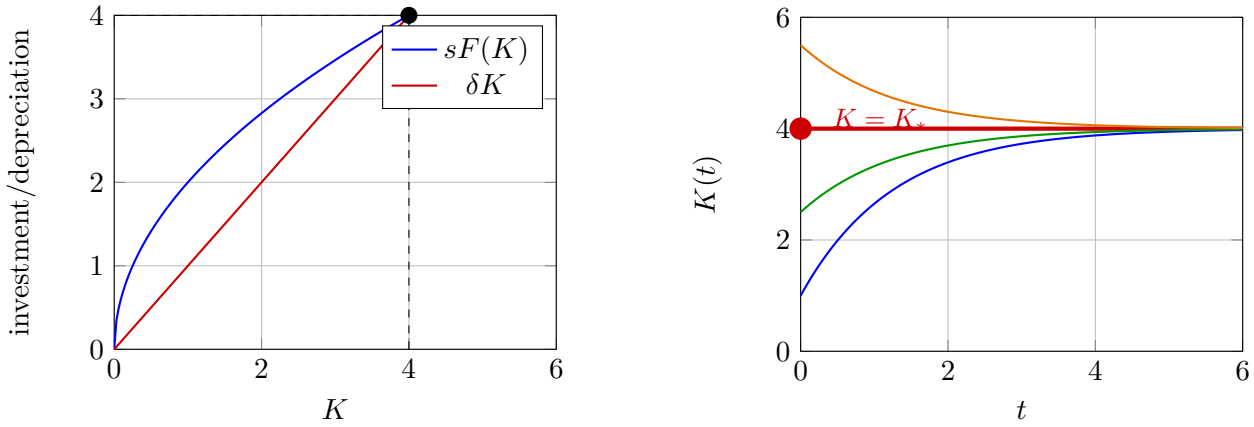
A positive steady state satisfies

$$sA(K^*)^{\nu} = \delta K^*,$$

and hence

$$K^* = \left(\frac{sA}{\delta}\right)^{\frac{1}{1-\nu}}.$$

The equilibrium gives the candidate long-run state of the economy. Its stability and the rate at which nearby solutions approach it will be analyzed in the next section.



(a) The equilibrium capital stock is determined by the intersection $sF(K) = \delta K$. Here $sA = 2$, $\nu = \frac{1}{2}$, and $\delta = 1$, so the positive equilibrium is $K^* = 4$.

(b) Solutions of $\dot{K} = 2\sqrt{K} - K$ are given by $K(t) = \left[2 + \left(\sqrt{K_0} - 2\right)e^{-t/2}\right]^2$. For $K_0 > 0$, they approach the positive equilibrium $K^* = 4$.

Figure 1.3: The Solow growth model. The left panel shows the equilibrium condition, while the right panel illustrates convergence toward the stable equilibrium capital stock.

Exercise 1.1 (Comparing falling objects)**.** Two spherical raindrops fall from rest according to

$$m_i\dot{v}_i = m_i g - \beta_i v_i, \qquad i = 1, 2,$$

where

$$m_2 = 8m_1, \qquad \beta_2 = 4\beta_1.$$

Without first solving the equations, compare their terminal velocities and relaxation times, and

determine which drop reaches one half of its own terminal velocity first. Then normalize each velocity by its terminal value and compare the resulting equations and time scales.

Exercise 1.2 (Estimating a cooling constant)**.** A temperature sensor placed in a chamber records

$$T(0) = 80^\circ\mathrm{C}, \qquad T(5) = 50^\circ\mathrm{C}, \qquad T(10) = 35^\circ\mathrm{C}.$$

Assuming

$$\dot{T} = -\alpha(T - T_a),$$

determine the unknown chamber temperature T_a and cooling rate α, and predict $T(15)$. Explain why two measurements would not determine both unknown parameters.

Hint. For measurements at equally spaced times, the differences $T(t) - T_a$ form a geometric sequence.

Exercise 1.3 (Designing a reversible reaction)**.** For the reversible reaction

$$A \rightleftharpoons B, \qquad [A](t) + [B](t) = a,$$

suppose the equilibrium concentration of B is $3a/4$ and deviations from equilibrium are reduced by one half every 6 minutes. Determine the rate constants k_1 and k_2.

If initially only A is present, find when $[B](t)$ first reaches $a/2$. Can it ever exceed $3a/4$? Justify the latter conclusion qualitatively, without relying only on the explicit solution.

Hint. Use the equilibrium ratio and the relaxation rate $k_1 + k_2$.

Exercise 1.4 (A tax shock in a changing market)**.** Demand and supply are

$$D(p) = A - bp, \qquad S_\tau(p) = C + d(p - \tau),$$

where

$$A > C, \qquad b, d > 0,$$

and $\tau \geq 0$ is a per-unit tax paid by suppliers. The market price satisfies

$$\dot{p} = \alpha\big(D(p) - S_\tau(p)\big), \qquad \alpha > 0.$$

Initially $\tau = 0$ and the market is at equilibrium. At $t = 0$, a tax $\tau > 0$ is imposed. Assume that τ is sufficiently small that the new equilibrium quantity remains positive. Determine the new equilibrium consumer price, producer price $p - \tau$, and traded quantity. What fractions of the tax are borne by consumers and producers?

Find the subsequent price $p(t)$ and its relaxation time. Without using the explicit solution, explain why the price cannot overshoot its new equilibrium.

Hint. Compare the equilibrium before and after replacing

$$S_0(p) = C + dp$$

by

$$S_\tau(p) = C + d(p - \tau).$$

1.3 Phase Lines and Stability

The models in the preceding section describe systems that approach a state of balance. More generally, consider a scalar *autonomous equation*

$$y' = f(y),$$

whose right-hand side depends only on the state y, and not explicitly on time. Such an equation may have several equilibrium states, and different initial conditions may lead to different long-time behavior.

For a scalar autonomous equation, much of this behavior can be determined without solving the equation explicitly. The sign of $f(y)$ tells us whether the state increases or decreases. Recording this information on the real line gives a *phase line*, which provides a geometric picture of the dynamics.

1.3.1 The phase-line method

Equilibria. An equilibrium of

$$y' = f(y)$$

is a constant solution

$$y(t) \equiv y^*.$$

Equivalently, y^* is a zero of the right-hand side:

$$f(y^*) = 0.$$

The equilibria divide the real line into intervals. If the equilibria are isolated and f is continuous, then f has a constant sign on each such interval.

Direction of motion. The sign of $f(y)$ determines the direction in which a solution moves:

$$\begin{cases} f(y) > 0, & \text{the solution increases,} \\ f(y) < 0, & \text{the solution decreases,} \\ f(y) = 0, & \text{the solution remains at an equilibrium.} \end{cases}$$

On the phase line, an arrow points to the right when $f(y) > 0$ and to the left when $f(y) < 0$.

Thus the phase-line method consists of the following steps:

find the zeros of f $\longrightarrow$ determine the sign of f $\longrightarrow$ draw the arrows $\longrightarrow$ describe the solutions.

Stable and unstable equilibria. An equilibrium y^* is called *stable* if solutions that start sufficiently close to y^* remain close to it for all future time.

It is called *asymptotically stable* if it is stable and, in addition, solutions that start sufficiently close to y^* approach y^* as $t \to \infty$.

An equilibrium that is not stable is called *unstable*.

These properties can be read directly from the phase line.

- If the arrows on both sides point toward y^*, then nearby solutions approach the equilibrium, and y^* is asymptotically stable:
$$\longrightarrow \; y^* \; \longleftarrow .$$
- If the arrows on both sides point away from y^*, then nearby solutions move away from the equilibrium, and y^* is unstable:
$$\longleftarrow \; y^* \; \longrightarrow .$$
- If the arrows point in the same direction on both sides,
$$\longrightarrow \; y^* \; \longrightarrow \qquad \text{or} \qquad \longleftarrow \; y^* \; \longleftarrow,$$
then solutions approach the equilibrium from one side and move away from it on the other. The equilibrium is called *semistable*. It is not stable because solutions starting arbitrarily close on one side move away.

Example 1.1. Consider
$$y' = y(y-1)(2-y).$$
The equilibria are
$$y = 0, \qquad y = 1, \qquad y = 2.$$
The sign of the right-hand side is given by

y	$(-\infty, 0)$	0	$(0,1)$	1	$(1,2)$	2	$(2,\infty)$
y'	$+$	0	$-$	0	$+$	0	$-$

Hence the phase line has the form
$$\longrightarrow \; 0 \; \longleftarrow \qquad \longleftarrow \; 1 \; \longrightarrow \qquad \longrightarrow \; 2 \; \longleftarrow .$$
Therefore $y = 0$ and $y = 2$ are asymptotically stable, while $y = 1$ is unstable.

The phase line also gives the long-time behavior. Solutions with initial values $y_0 < 1$ move toward 0, whereas solutions with $y_0 > 1$ move toward 2, except for the equilibrium solution $y(t) \equiv 1$.

The phase line is a one-dimensional phase portrait. It does not provide an explicit formula for $y(t)$, but it shows the direction of motion, the stability of the equilibria, and the possible long-time behavior of all solutions.

When uniqueness holds, a nonconstant solution cannot cross an equilibrium. Consequently, each interval between consecutive equilibria is invariant, and every solution is monotone as long as it remains in that interval.

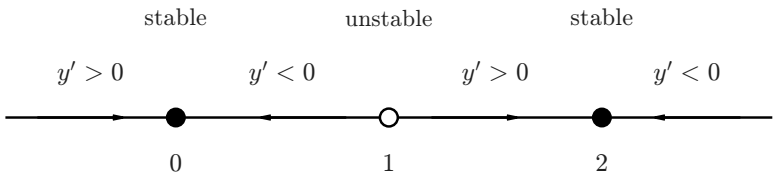


Figure 1.4: Phase line for $y' = y(y-1)(2-y)$. The equilibria $y = 0$ and $y = 2$ are asymptotically stable, while $y = 1$ is unstable.

1.3.2 Local stability and linearization

The phase-line method uses the sign of f on whole intervals. When f is differentiable, the behavior near a single equilibrium can often be determined more quickly by linearization.

Linearization near an equilibrium. Let y^* be an equilibrium, so that

$$f(y^*) = 0.$$

Since f is differentiable at y^* and $f(y^*) = 0$,

$$f(y) = f'(y^*)(y - y^*) + o\big(|y - y^*|\big) \qquad \text{as } y \to y^*.$$

Thus, near y^*, the nonlinear equation

$$y' = f(y)$$

is approximated by the linear equation

$$y' = f'(y^*)(y - y^*).$$

If

$$f'(y^*) < 0,$$

set

$$\alpha = -f'(y^*) > 0.$$

The linearized equation becomes

$$y' = -\alpha(y - y^*),$$

whose solutions approach y^* exponentially. The equilibrium is therefore locally asymptotically stable.

If

$$f'(y^*) > 0,$$

small perturbations grow exponentially in the linearized equation, and the equilibrium is unstable.

We therefore obtain the derivative test

$$\begin{array}{l|l} f'(y^*) < 0 & y^* \text{ is locally asymptotically stable,} \\ f'(y^*) > 0 & y^* \text{ is unstable.} \end{array}$$

This criterion can also be understood from the phase line. If $f'(y^*) < 0$, then f changes from positive to negative as y passes through y^*, so the arrows point toward the equilibrium. If $f'(y^*) > 0$, then f changes from negative to positive, so the arrows point away from it.

Local relaxation rate. When $f'(y^*) < 0$, the linearized equation predicts the local rate at which solutions approach the equilibrium. The quantity

$$\alpha = -f'(y^*)$$

is called the *local relaxation rate*, and

$$\tau = \frac{1}{-f'(y^*)}$$

is the corresponding *local relaxation time*.

A larger value of α corresponds to faster relaxation, while a larger value of τ corresponds to slower relaxation.

The conclusion is local: it describes solutions beginning near y^*, but it does not determine the behavior of solutions far from the equilibrium. For an equation with several equilibria, different initial conditions may approach different stable states. The full phase line is needed to determine their basins of attraction.

Example 1.2. Consider the Solow growth model

$$\dot{K} = sAK^\nu - \delta K, \qquad 0 < \nu < 1.$$

We use linearization to study the positive equilibrium and its local relaxation rate. The positive equilibrium is

$$K^* = \left(\frac{sA}{\delta}\right)^{\frac{1}{1-\nu}}.$$

Let

$$F(K) = sAK^\nu - \delta K.$$

Then

$$F'(K) = sA\nu K^{\nu-1} - \delta.$$

Using the equilibrium relation

$$sA(K^*)^{\nu-1} = \delta,$$

we obtain

$$F'(K^*) = \nu\delta - \delta = -(1-\nu)\delta < 0.$$

Therefore K^* is locally asymptotically stable. Near K^*, the equation has the linearized form

$$\dot{K} = -(1-\nu)\delta(K - K^*) + \text{higher-order terms}.$$

Consequently, the local relaxation rate is

$$\alpha = (1-\nu)\delta,$$

and the local relaxation time is

$$\tau = \frac{1}{(1-\nu)\delta}.$$

The boundary state $K = 0$ is also an equilibrium. For small $K > 0$, the term K^ν dominates K, so $\dot{K} > 0$. The phase-line arrows on the positive side point away from $K = 0$. Thus $K = 0$ is unstable relative to the region $K \geq 0$. Since $0 < \nu < 1$, the right-hand side is not differentiable at $K = 0$, so this conclusion is obtained from the sign of the vector field near $K = 0$, rather than from linearization.

When linearization is inconclusive. If

$$f'(y^*) = 0,$$

the linearized equation gives no conclusion. For example, consider

$$y' = -(y - y^*)^3.$$

The right-hand side is positive when $y < y^*$ and negative when $y > y^*$. The arrows therefore point toward y^* from both sides, so the equilibrium is asymptotically stable. Its convergence is slower than that of the linear relaxation equation.

In contrast, for

$$y' = (y - y^*)^3,$$

the right-hand side is negative when $y < y^*$ and positive when $y > y^*$. The arrows point away from y^* on both sides, so the equilibrium is unstable.

The two equations have the same linearization at y^*, namely

$$y' = 0,$$

but opposite stability behavior. Thus, when $f'(y^*) = 0$, higher-order terms or the full phase line must be examined.

Exercise 1.5. For each equation, locate the equilibria and classify them by both the phase line and the derivative test whenever the derivative test applies:

$$y' = y - y^3,$$
$$y' = (y + 1)^2(y - 2),$$
$$y' = \sin y.$$

Exercise 1.6 (Degenerate equilibria and nonlinear relaxation)**.** Consider

$$y' = -(y - y^*)^2$$

and

$$y' = -(y - y^*)^3.$$

For each equation, draw the phase line, solve the initial value problem $y(0) = y_0$, and classify the equilibrium $y = y^*$. Explain why the quadratic case is semistable, whereas the cubic case is

asymptotically stable. Compare the convergence in the cubic case with exponential relaxation

$$z' = -\alpha z, \qquad \alpha > 0.$$

Exercise 1.7 (Exact relaxation in the Solow model)**.** Consider

$$\dot{K} = sAK^{\nu} - \delta K, \qquad K(0) = K_0 > 0, \qquad 0 < \nu < 1.$$

Find the solution explicitly and show that it remains positive and converges to

$$K^* = \left(\frac{sA}{\delta}\right)^{1/(1-\nu)}.$$

Determine the asymptotic rate at which $K(t)$ approaches K^*, and compare it with the rate predicted by linearization.

Hint. Introduce

$$Z = K^{1-\nu}$$

and show that Z satisfies a linear relaxation equation.

1.4 Population Dynamics

Population dynamics is one of the earliest and most important applications of ordinary differential equations. Let $y(t)$ denote the population size at time t. The purpose is to construct simple laws that describe how a population changes and to understand the long-time behavior predicted by these laws.

In the previous section, many examples had the form of relaxation toward an equilibrium. Population models introduce a related but more nonlinear theme: growth may be rapid when the population is small, but environmental limitations, competition, and harvesting may change the qualitative behavior of the system. These models provide concrete applications of the phase-line analysis developed in the preceding section and introduce parameter-dependent behavior in which the number and stability of equilibria may change through a bifurcation.

1.4.1 Malthusian growth

The simplest population model is based on the balance law

$$y' = \text{births} - \text{deaths} + \text{migration}.$$

If migration is neglected and both births and deaths are proportional to the population size, then

$$y' = ry,$$

where r is the net growth rate. The constant r is the birth rate minus the death rate. If $r > 0$, the population grows; if $r < 0$, the population declines.

For the initial condition

$$y(0) = y_0,$$

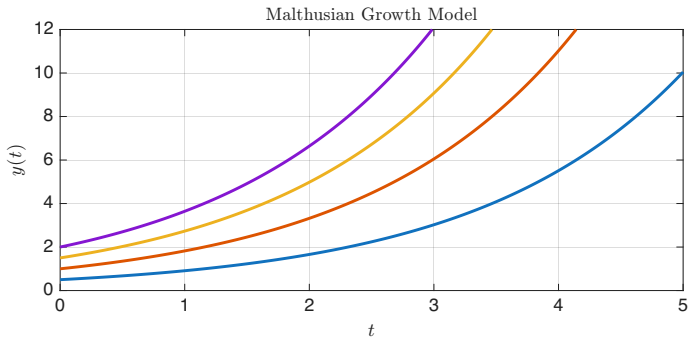


Figure 1.5: Solutions of the Malthusian growth model $y' = ry$ with $r > 0$ for several initial populations. All positive solutions grow exponentially and become unbounded as $t \to \infty$.

the solution is

$$y(t) = y_0 e^{rt}.$$

Thus the model predicts exponential growth when $r > 0$ and exponential decay when $r < 0$. When $r > 0$, all positive solutions become unbounded as $t \to \infty$.

The exponential solution allows us to quantify the growth rate of a population. If the population doubles every t_d units of time, then

$$y(t + t_d) = 2y(t).$$

Substituting $y(t) = y_0 e^{rt}$ gives

$$e^{rt_d} = 2,$$

and therefore

$$t_d = \frac{\log 2}{r}.$$

Equivalently, if the doubling time t_d is known, then

$$r = \frac{\log 2}{t_d}.$$

For example, a population that doubles every ten years has growth rate $r = (\log 2)/10 \approx 0.0693$ per year.

The Malthusian model is useful over time intervals for which resources are not limiting. It is not a global model for most biological populations, because it predicts unlimited exponential growth. This failure is not a minor technical issue; it indicates that the model is missing an important mechanism. Resource limitation and competition must eventually be included.

1.4.2 Logistic growth

The Malthusian model assumes that the net growth rate remains constant regardless of the population size. In reality, resources such as food, water, and living space are limited. As the population grows, competition increases and the effective growth rate decreases.

Let $K > 0$ denote the carrying capacity of the environment. The carrying capacity represents the population level that the environment can sustain in the long run. We want a model with

the following qualitative properties: when $0 < y < K$, the population should increase, and when $y > K$, the population should decrease. A simple equation with these properties is

$$y' = ry\left(1 - \frac{y}{K}\right), \qquad r > 0.$$

This is the logistic growth model.

The factor r is the intrinsic growth rate. The factor $1 - y/K$ represents the reduction of growth due to limited resources. When y is small compared with K, the factor $1 - y/K$ is close to 1, and the equation is close to the Malthusian equation $y' = ry$. When y is near K, the factor $1 - y/K$ is near zero, and growth slows down. When $y > K$, this factor becomes negative, and the population decreases.

The nonlinear term

$$-\frac{r}{K}y^2$$

has an important interpretation. The quantity y^2 measures the scale of pairwise competition among individuals, while r/K measures the strength of this competition. The negative sign reflects the fact that competition reduces the growth rate.

The logistic equation is separable. Separating variables gives

$$\frac{dy}{y(1 - y/K)} = r\,dt.$$

Using the partial fraction identity

$$\frac{1}{y(1 - y/K)} = \frac{1}{y} + \frac{1}{K - y},$$

we obtain

$$\log\left|\frac{y}{K - y}\right| = rt + C.$$

Solving for y gives

$$y(t) = \frac{K}{1 + Ae^{-rt}},$$

where

$$A = \frac{K - y_0}{y_0}.$$

Equivalently,

$$y(t) = \frac{Ky_0}{y_0 + (K - y_0)e^{-rt}}.$$

For every initial value $y_0 > 0$, the solution satisfies

$$\lim_{t\to\infty} y(t) = K.$$

Thus the carrying capacity K is an asymptotically stable equilibrium relative to the region $y > 0$.

The logistic model is one of the basic nonlinear models in mathematical biology. It is still highly simplified, since real populations may involve age structure, spatial motion, seasonal variation, environmental noise, and migration. Nevertheless, it captures a central mechanism: growth is self-limited by competition.

Below are the phase line portrait and the solution curves of the logistic equation.

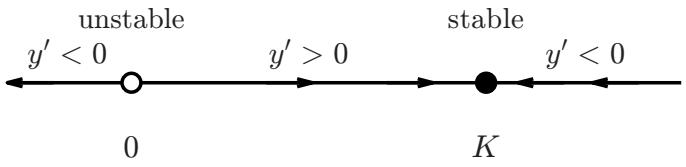


Figure 1.6: Phase line for the logistic equation $y' = ry(1-y/K)$. The equilibrium $y = 0$ is unstable, while $y = K$ is asymptotically stable.

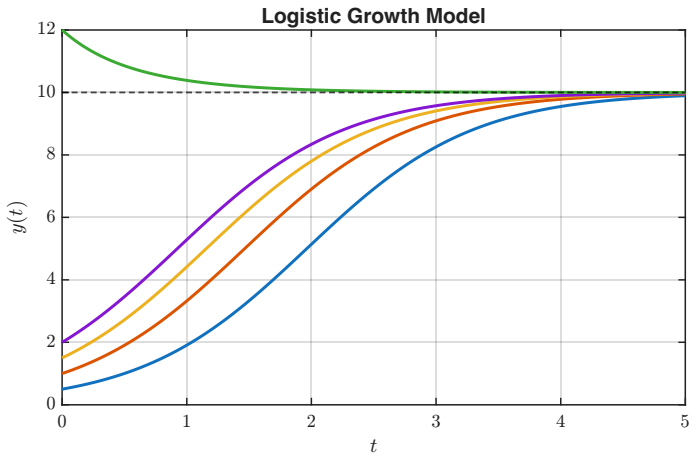


Figure 1.7: Solutions of the logistic growth model $y' = ry\left(1 - \frac{y}{K}\right)$, $r > 0$, for several initial populations, with carrying capacity $K = 10$. In contrast to the Malthusian growth model in Fig. 1.5, positive solutions do not grow without bound. Instead, they approach the carrying capacity K as $t \to \infty$.

1.4.3 The Allee effect and population thresholds

The logistic equation assumes that every sufficiently small positive population grows. Indeed, when $0 < y \ll K$,

$$y' = ry\left(1 - \frac{y}{K}\right) \approx ry > 0.$$

For some populations, however, survival becomes difficult when the population is too small. Individuals may have difficulty finding mates, defending themselves, hunting cooperatively, or maintaining sufficient genetic diversity. The resulting reduction in the per-capita growth rate at low population levels is called the *Allee effect*.

Let A satisfy

$$0 < A < K,$$

where A represents a threshold population. We seek a model in which the population decreases when $0 < y < A$, increases when $A < y < K$, and decreases when $y > K$. The simplest modification of the logistic equation with these properties is

$$y' = ry\left(1 - \frac{y}{K}\right)\left(\frac{y}{A} - 1\right), \qquad r > 0.$$

This is a population model with a strong Allee effect.

The three equilibria are

$$y = 0, \qquad y = A, \qquad y = K.$$

The signs of the right-hand side are

$$\begin{cases} y' < 0, & 0 < y < A, \\ y' > 0, & A < y < K, \\ y' < 0, & y > K. \end{cases}$$

Therefore $y = 0$ and $y = K$ are locally asymptotically stable, while $y = A$ is unstable. The equilibrium A separates two qualitatively different outcomes:

$$\begin{cases} 0 < y(0) < A & \Longrightarrow y(t) \to 0, \\ y(0) > A & \Longrightarrow y(t) \to K. \end{cases}$$

Thus a population below the Allee threshold tends toward extinction, whereas one above the threshold tends toward the carrying capacity.

The equation can be understood as a product of three factors. The factor y ensures that extinction is an equilibrium. The logistic factor $1-y/K$ accounts for competition at large population levels. The additional factor

$$\frac{y}{A} - 1$$

changes sign at $y = A$ and represents the adverse effect of a population that is too small.

The equation is separable. Since

$$y' = \frac{r}{AK} y(K - y)(y - A),$$

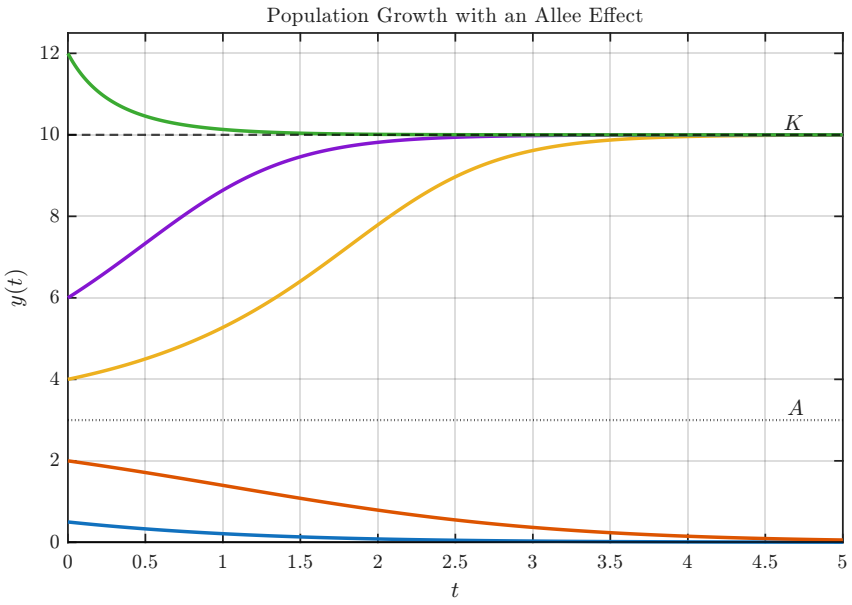


Figure 1.8: Solutions of the population model with a strong Allee effect, $y' = ry\left(1-\frac{y}{K}\right)\left(\frac{y}{A}-1\right), \quad 0 < A < K$, for several initial populations. The equilibrium A is the Allee threshold. Solutions with $0 < y(0) < A$ approach extinction, whereas solutions with $y(0) > A$ approach the carrying capacity K.

we obtain

$$\frac{AK\,dy}{y(K-y)(y-A)} = r\,dt.$$

Using partial fractions and integrating gives

$$-\log|y| + \frac{K}{K-A}\log|y-A| - \frac{A}{K-A}\log|K-y| = rt + C.$$

This relation determines the solution implicitly. Although the equation can be integrated by separation of variables, the phase line describes its long-time behavior more clearly than the resulting implicit relation.

The Allee model illustrates bistability: both extinction and the carrying capacity are stable, and the eventual outcome depends on which side of the threshold A contains the initial population. This threshold behavior also explains why a severely depleted population may fail to recover even after the environmental conditions improve.

Exercise 1.8 (A weak Allee effect)**.** Consider the weak Allee-effect model

$$y' = ry\left(1-\frac{y}{K}\right)\left(1+\frac{y}{A}\right), \qquad 0 < A < K, \qquad r > 0.$$

Without using an explicit solution, do the following:

1. Show that the per-capita growth rate (i.e. y'/y) is positive for $0 < y < K$ and initially increases with y. Explain why this represents a weak, rather than strong, Allee effect.

2. Construct the phase line and show that every solution with $y(0) > 0$ remains positive for all $t > 0$.

3. Prove that
$$\lim_{t\to\infty} y(t) = K$$
for every $y(0) > 0$. Distinguish the cases
$$0 < y(0) < K, \qquad y(0) = K, \qquad y(0) > K.$$

1.5 Elementary Bifurcations in One Dimension

The phase line describes the qualitative behavior of a scalar autonomous equation when its parameters are fixed. Many models, however, contain parameters that describe external conditions or control mechanisms. As such a parameter varies, the equilibria may move, appear, disappear, or change stability.

A parameter-dependent scalar equation has the form
$$y' = f(y;\mu),$$
where μ is a real parameter. For each fixed value of μ, the equation has its own phase line, and the equilibria are the zeros of $f(\cdot;\mu)$. As μ varies, these equilibria trace curves in the (μ, y)-plane. The resulting picture is called a *bifurcation diagram*.

A bifurcation diagram may be viewed as a family of phase lines placed side by side. The parameter is plotted on the horizontal axis, while the equilibrium state is plotted on the vertical axis. For a fixed parameter value, a vertical line in the diagram represents the corresponding phase line. Arrows on that line indicate the direction of the flow: upward when $y' > 0$ and downward when $y' < 0$. Stable equilibrium branches are represented by solid curves and unstable branches by dashed curves.

A *bifurcation* occurs when the qualitative structure of the phase line changes as the parameter passes through a critical value. Typically this happens when an equilibrium becomes degenerate. If
$$f(y_*;\mu_*) = 0, \qquad f_y(y_*;\mu_*) \neq 0,$$
then y_* is a simple equilibrium and persists for nearby values of the parameter. Thus a local bifurcation can occur only when
$$f(y_*;\mu_*) = 0, \qquad f_y(y_*;\mu_*) = 0.$$
At such a point, the usual derivative test is inconclusive, and the number or stability of equilibria may change.

We now examine the basic one-dimensional bifurcations through concrete models. The first two arise naturally from logistic population growth under different harvesting laws.

1.5.1 Saddle-node bifurcation

A saddle-node bifurcation is the mechanism by which a stable and an unstable equilibrium collide and disappear, or are created together. A basic example is provided by logistic growth with constant-yield harvesting.

Logistic growth with constant harvesting. Suppose that individuals are removed from the population at a constant rate $H > 0$. The logistic equation becomes

$$y' = ry\left(1 - \frac{y}{K}\right) - H,$$

where $r > 0$ is the intrinsic growth rate, $K > 0$ is the carrying capacity, and H is the amount harvested per unit time.

An equilibrium population satisfies

$$ry\left(1 - \frac{y}{K}\right) - H = 0,$$

or equivalently,

$$\frac{r}{K}y^2 - ry + H = 0.$$

Hence the number of positive equilibria depends on the harvesting parameter H.

The equilibrium condition may also be written as

$$H = ry\left(1 - \frac{y}{K}\right).$$

Thus, if the harvesting parameter H is plotted horizontally and the equilibrium population y vertically, the equilibrium branches form a folded curve in the (H, y)-plane. For sufficiently small H, there are two positive equilibria. As H increases, the lower equilibrium moves upward while the upper equilibrium moves downward. At a critical harvesting rate the two branches meet, and beyond this value no positive equilibrium remains.

For a fixed value of H, the corresponding vertical line in the bifurcation diagram is a phase line. The arrows on that line show the sign of

$$f_H(y) = ry\left(1 - \frac{y}{K}\right) - H$$

and therefore indicate the stability of the equilibrium branches.

The critical harvesting rate is found by requiring the quadratic equation for the equilibria to have a repeated root. Its discriminant vanishes when

$$r^2 - \frac{4rH}{K} = 0.$$

Therefore

$$H_c = \frac{rK}{4}.$$

The corresponding equilibrium is

$$y_c = \frac{K}{2}.$$

Thus

$$\begin{cases} 0 < H < H_c, & \text{two positive equilibria,} \\ H = H_c, & \text{one positive degenerate equilibrium,} \\ H > H_c, & \text{no positive equilibrium.} \end{cases}$$

When $0 < H < H_c$, the two positive equilibria are

$$y_{\pm} = \frac{K}{2} \pm \sqrt{\frac{K^2}{4} - \frac{HK}{r}}.$$

Since

$$f_H(y) < 0 \quad \text{for } 0 < y < y_-,$$

$$f_H(y) > 0 \quad \text{for } y_- < y < y_+,$$

and

$$f_H(y) < 0 \quad \text{for } y > y_+,$$

the smaller equilibrium y_- is unstable, while the larger equilibrium y_+ is asymptotically stable.

The lower equilibrium y_- therefore acts as a threshold population. If

$$0 < y(0) < y_-,$$

then the population decreases and reaches zero in finite time. In the biological model, this time is interpreted as the extinction time, and the solution is terminated at $y = 0$. If

$$y(0) > y_-,$$

then the population approaches the sustainable equilibrium y_+.

At $H = H_c$, the two equilibria merge at

$$y_c = \frac{K}{2}.$$

This equilibrium is semistable. Solutions starting above y_c approach it, whereas solutions starting below y_c move away from it and reach extinction.

For $H > H_c$,

$$f_H(y) < 0$$

for every $y > 0$. Hence every positive population decreases and eventually reaches extinction.

Normal form. The essential structure of the saddle-node bifurcation can be separated from the particular parameters of the harvesting model. Recall that

$$H_c = \frac{rK}{4}, \qquad y_c = \frac{K}{2}.$$

Introduce the shifted parameter and variable

$$\mu = H_c - H, \qquad z = y - y_c.$$

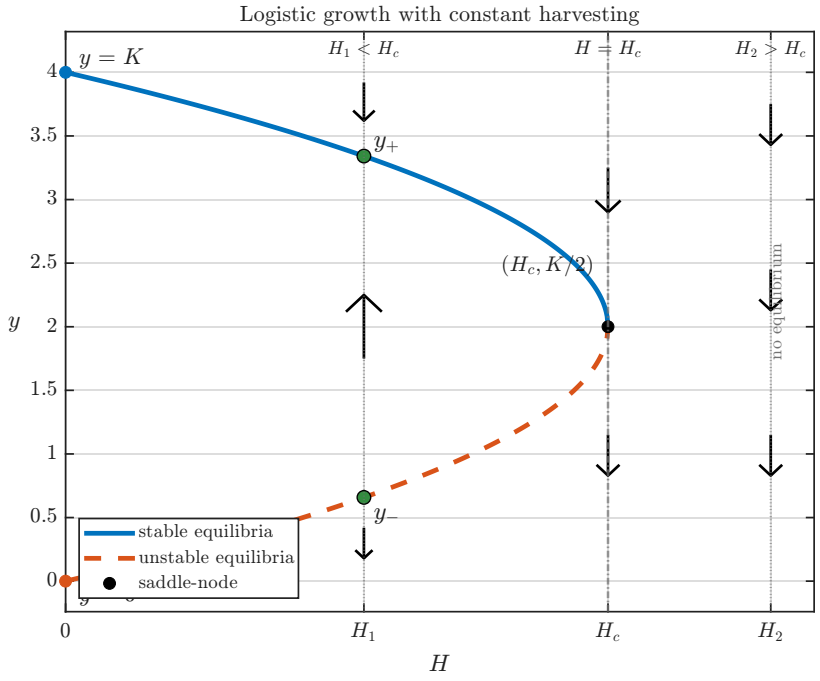


Figure 1.9: Bifurcation diagram for logistic growth with constant harvesting. For $H < H_c$, the lower equilibrium branch is unstable and the upper branch is stable. At $H = H_c$, the two equilibria merge in a saddle-node bifurcation. For $H > H_c$, no positive equilibrium remains. The arrows on representative vertical lines $H =$ constant indicate the direction of the flow for the corresponding phase lines.

Substituting them into the harvesting equation gives

$$z' = \mu - \frac{r}{K} z^2.$$

After rescaling the dependent variable and time, and relabeling the variables, we obtain the standard saddle-node normal form

$$y' = \mu - y^2.$$

The equilibria satisfy

$$\mu - y^2 = 0.$$

Thus

$$y_{\pm} = \pm\sqrt{\mu}$$

when $\mu > 0$, while there is no real equilibrium when $\mu < 0$. At $\mu = 0$, the two equilibria merge into the degenerate equilibrium $y = 0$.

To determine their stability, let

$$f(y; \mu) = \mu - y^2.$$

Then

$$f_y(y; \mu) = -2y.$$

Hence, for $\mu > 0$,

$$f_y(y_-; \mu) = 2\sqrt{\mu} > 0, \qquad f_y(y_+; \mu) = -2\sqrt{\mu} < 0.$$

Therefore the lower branch y_- is unstable and the upper branch y_+ is asymptotically stable.

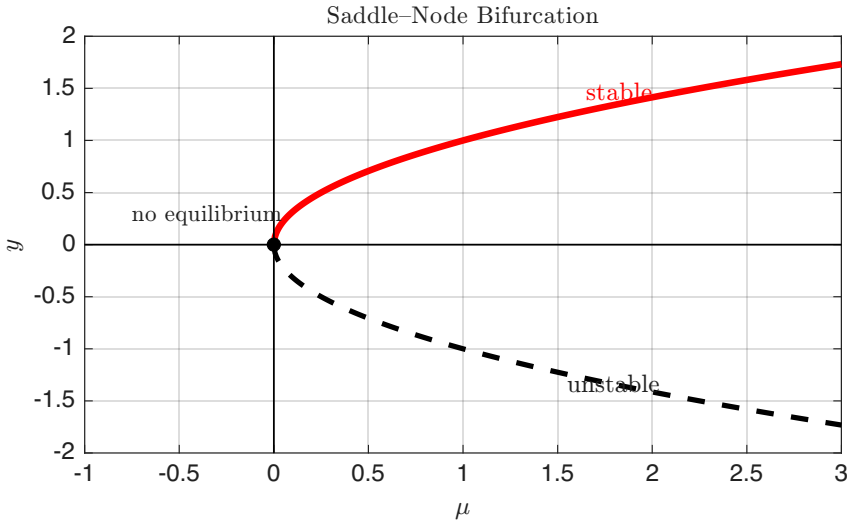


Figure 1.10: Saddle-node bifurcation for $y' = \mu - y^2$. For $\mu > 0$, there are two equilibria: the lower branch $y_- = -\sqrt{\mu}$ is unstable and the upper branch $y_+ = \sqrt{\mu}$ is stable. At $\mu = 0$, the two equilibria merge, while for $\mu < 0$ no equilibrium remains. Stable branches are shown by solid curves and unstable branches by dashed curves.

The name *saddle-node bifurcation* originates from higher-dimensional systems, where a saddle and a node collide and disappear. In one dimension, the corresponding event is the collision of an unstable equilibrium and a stable equilibrium. It is also called a *fold bifurcation*, since the equilibrium curve folds at the bifurcation point.

Slow variation of the harvesting parameter. The bifurcation diagram describes equilibria for fixed values of H, but it also helps interpret a harvesting policy that changes slowly in time. Suppose, for example, that

$$H = H(\varepsilon t), \qquad 0 < \varepsilon \ll 1,$$

so that the harvesting rate changes on a time scale much longer than the population relaxation time. Starting near the stable upper branch, the population then has time to remain close to the instantaneous equilibrium $y_+(H)$ as H increases, at least away from the critical value.

As H approaches H_c, the stable equilibrium and the unstable threshold move together. Beyond H_c, no positive instantaneous equilibrium remains, and the population eventually leaves the neighborhood of the former stable branch and begins to decline. The departure need not occur exactly when $H = H_c$; slowly varying systems may exhibit a delay beyond the static bifurcation value.

This loss of a stable state at a critical parameter value is one of the basic mechanisms behind threshold behavior and sudden transitions in nonlinear dynamical systems.

1.5.2 Transcritical bifurcation

A different type of bifurcation occurs when two equilibrium branches cross and exchange stability. This is called a *transcritical bifurcation*. A simple biological example is again provided by the logistic model, but with a different harvesting law.

Logistic growth with a constant per-capita harvesting rate. In the preceding example, a fixed amount H was removed per unit time, independently of the population size. Since the vector field satisfies

$$y' = f_H(0) = -H < 0,$$

a solution reaching $y = 0$ would continue into negative values, which is biologically meaningless. Naturally, harvesting must cease when the population becomes extinct. One way to avoid this difficulty is to assume that individuals are harvested at a constant per-capita rate $h > 0$, so that the total harvesting rate is hy and therefore vanishes as $y \to 0$. The logistic equation becomes

$$y' = ry\left(1 - \frac{y}{K}\right) - hy.$$

Here h has the same physical dimension as the intrinsic growth rate r.

Factoring the equation gives

$$y' = y\left(r - h - \frac{r}{K}y\right).$$

Hence there are two equilibrium branches,

$$y = 0$$

and

$$y_*(h) = K\left(1 - \frac{h}{r}\right).$$

The second equilibrium is positive when $0 \le h < r$, reaches zero when $h = r$, and becomes negative when $h > r$. The negative part of this branch has no biological interpretation, but its mathematical continuation is useful for displaying the bifurcation.

To determine stability, define

$$f(y;h) = y\left(r - h - \frac{r}{K}y\right).$$

Then

$$f_y(y;h) = r - h - \frac{2r}{K}y.$$

Along the extinction branch,

$$f_y(0;h) = r - h,$$

so $y = 0$ is unstable for $h < r$ and asymptotically stable for $h > r$. Along the nonzero branch,

$$f_y(y_*(h);h) = -(r - h),$$

so $y_*(h)$ is asymptotically stable for $h < r$ and unstable for $h > r$.

Thus the two equilibrium branches meet at

$$(h, y) = (r, 0)$$

and exchange stability there. This is the characteristic structure of a transcritical bifurcation. At the critical value $h = r$, the equation becomes

$$y' = -\frac{r}{K}y^2.$$

Thus $y = 0$ is semistable on the whole real line and asymptotically stable relative to the biologically relevant region $y \geq 0$. Positive solutions approach zero algebraically rather than exponentially.

The biological interpretation is particularly simple. When

$$0 \leq h < r,$$

the per-capita harvesting rate is smaller than the intrinsic growth rate. Extinction is unstable, and every positive population approaches the sustainable equilibrium

$$y_*(h) = K\left(1 - \frac{h}{r}\right).$$

As h increases, this equilibrium decreases continuously and reaches zero at the critical value

$$h_c = r.$$

When $h > r$, harvesting exceeds the intrinsic growth rate at low population density, and extinction becomes asymptotically stable. Every positive population then approaches zero.

Normal form. The transcritical structure of the harvesting model becomes particularly clear after a change of parameters and variables. Set

$$\mu = r - h.$$

Then

$$y' = \mu y - \frac{r}{K}y^2.$$

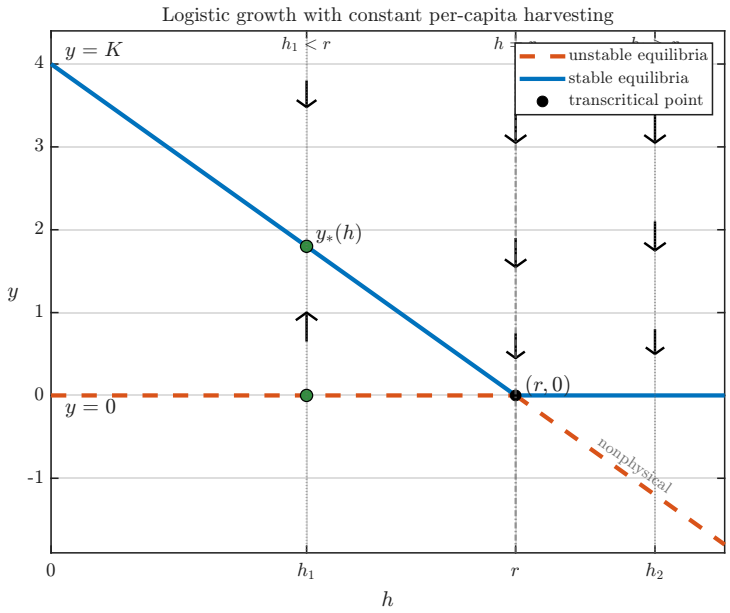


Figure 1.11: Bifurcation diagram for logistic growth with a constant per-capita harvesting rate. The extinction branch $y = 0$ and the nonzero branch $y = K(1 - h/r)$ intersect at $h = r$ and exchange stability. Solid curves represent stable equilibria and dashed curves unstable equilibria. The arrows on representative vertical lines $h =$ constant indicate the direction of the flow. The portion of the nonzero branch with $y < 0$ has no biological interpretation.

Introducing

$$z = \frac{r}{K} y$$

gives

$$z' = \mu z - z^2.$$

Relabeling z as y, we obtain the standard transcritical normal form

$$y' = \mu y - y^2.$$

Its equilibrium branches are

$$y = 0, \qquad y = \mu.$$

For $\mu < 0$, the branch $y = 0$ is stable and $y = \mu$ is unstable. For $\mu > 0$, their stability is reversed. At $\mu = 0$, the two branches intersect and exchange stability.

The essential feature of a transcritical bifurcation is therefore an *exchange of stability* between two equilibrium branches. In the harvesting model, these branches represent extinction and the nonzero population state.

1.5.3 Bistability and hysteresis

The preceding bifurcations describe the creation, disappearance, or exchange of stability of equilibria. Another important phenomenon occurs when two stable equilibria coexist for the same value of a parameter. The state approached by the system then depends on its initial condition and, when

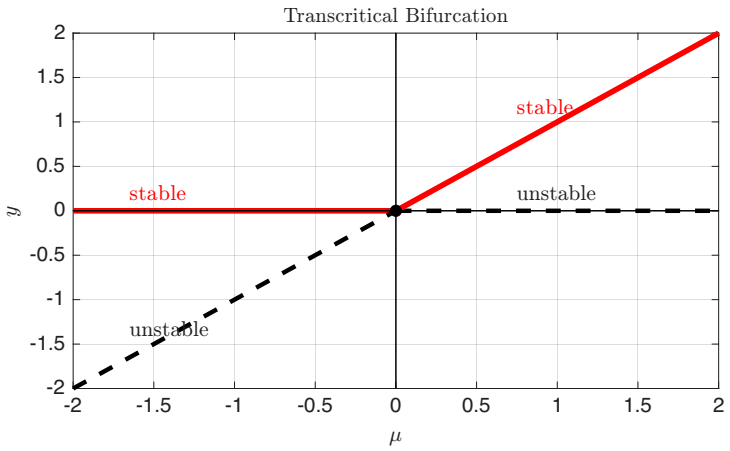


Figure 1.12: Transcritical bifurcation for $y' = \mu y - y^2$. The equilibrium branches $y = 0$ and $y = \mu$ intersect at the origin and exchange stability. Stable branches are shown by solid curves and unstable branches by dashed curves. The arrows on representative vertical lines $\mu =$ constant indicate the direction of the flow.

the parameter is varied, may also depend on the previous history of the system. This phenomenon is called *bistability* and is closely related to *hysteresis*.

An Allee population with proportional harvesting. Consider a population with an Allee effect and suppose that individuals are harvested at a constant per-capita rate $h > 0$. The population equation is

$$y' = ry\left(1 - \frac{y}{K}\right)\left(\frac{y}{A} - 1\right) - hy, \qquad 0 < A < K.$$

Factoring out y gives

$$y' = y\left[r\left(1 - \frac{y}{K}\right)\left(\frac{y}{A} - 1\right) - h\right].$$

Thus extinction,

$$y = 0,$$

is an equilibrium for every value of h.

The positive equilibria satisfy

$$r\left(1 - \frac{y}{K}\right)\left(\frac{y}{A} - 1\right) = h.$$

Equivalently,

$$y^2 - (K + A)y + KA\left(1 + \frac{h}{r}\right) = 0,$$

and hence

$$y_\pm = \frac{K + A \pm \sqrt{(K - A)^2 - \dfrac{4KAh}{r}}}{2}.$$

There are two distinct positive equilibria when

$$0 \le h < h_c, \qquad h_c = \frac{r(K - A)^2}{4KA}.$$

At $h = h_c$, they merge at

$$y_c = \frac{K + A}{2}.$$

For $0 \le h < h_c$, the three equilibria have the stability pattern

$$y = 0 \text{ stable}, \qquad y_- \text{ unstable}, \qquad y_+ \text{ stable}.$$

Thus two stable states coexist: extinction and the positive population y_+. The unstable equilibrium y_- separates their basins of attraction:

$$\begin{cases} 0 < y(0) < y_-, & \Longrightarrow y(t) \to 0, \\ y(0) > y_-, & \Longrightarrow y(t) \to y_+. \end{cases}$$

This coexistence of two stable equilibria for the same parameter value is called *bistability*.

As h increases, the unstable threshold y_- moves upward while the stable positive equilibrium y_+ moves downward. At $h = h_c$, these two branches merge and disappear in a saddle-node bifurcation. For $h > h_c$, only the extinction equilibrium remains.

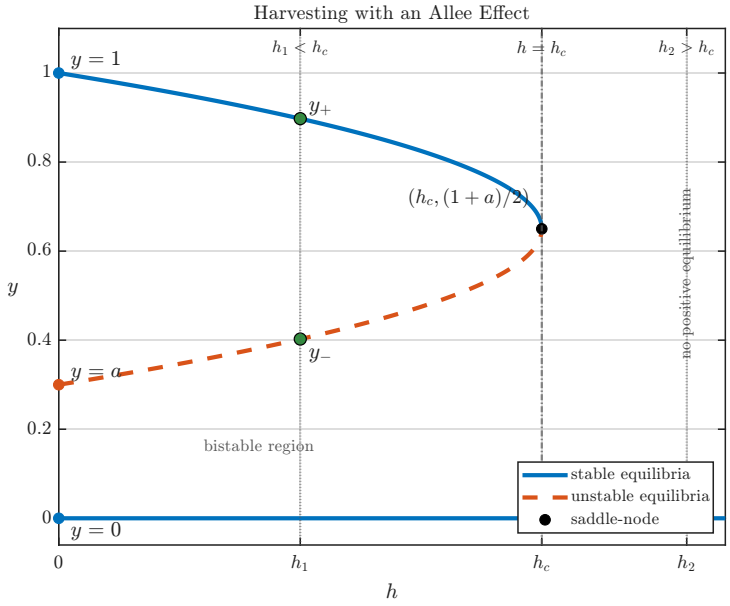


Figure 1.13: Bifurcation diagram for an Allee population with constant per-capita harvesting. For $h < h_c$, the extinction state $y = 0$ and the upper branch y_+ are stable, while the intermediate branch y_- is unstable. At $h = h_c$, the two positive equilibria merge and disappear. For $h > h_c$, only the extinction state remains. The arrows on representative vertical lines $h =$ constant indicate the direction of the flow. Note that, in the diagram, the population and harvesting rate have been nondimensionalized, with y scaled by K, h scaled by r, and $a = A/K$.

The bifurcation diagram also explains why recovery after a population collapse may be difficult. Suppose that h changes slowly compared with the population dynamics. Starting near the stable positive branch, the population follows $y_+(h)$ as harvesting increases. When h reaches h_c, this equilibrium disappears and the population moves toward extinction.

If the harvesting rate is subsequently reduced below h_c, the positive stable equilibrium reappears. However, the population is now near the stable extinction state and therefore does not automatically return to $y_+(h)$. Recovery requires the population to cross the unstable threshold $y_-(h)$, for example through sufficiently large restocking.

Thus, for a fixed value $0 < h < h_c$, the observed population may be near either $y = 0$ or $y_+(h)$, depending on its previous history. This dependence of the state of a system on the history of a varying parameter is called *hysteresis*.

A canonical model for hysteresis. The population model above illustrates bistability and shows why recovery after a collapse may not simply reverse the path toward collapse. A more symmetric situation, with transitions in both directions, leads to the classical S-shaped hysteresis diagram.

A familiar example is the snap-through of a thin flexible object. A curved lid or thin sheet may rest in either of two shapes. As an applied force is slowly increased, the object first deforms while remaining in one stable shape and then suddenly snaps to the other. When the force is reversed, it may not snap back at the same value of the force. Such snap-through phenomena provide a simple physical picture of bistability and hysteresis.

The simplest scalar equation displaying the same qualitative geometry is

$$y' = \mu + y - y^3,$$

where y represents the state of the system and μ is a control parameter. Its equilibria satisfy

$$\mu = y^3 - y,$$

which defines an S-shaped equilibrium curve in the (μ, y)-plane. For a range of values of μ, the upper and lower branches represent two stable states, separated by an unstable middle branch.

The fold points occur where

$$\frac{d\mu}{dy} = 3y^2 - 1 = 0.$$

Hence

$$y = \pm\frac{1}{\sqrt{3}},$$

and the corresponding parameter values are

$$\mu_\pm = \mp\frac{2}{3\sqrt{3}}.$$

Thus three equilibria exist when

$$|\mu| < \frac{2}{3\sqrt{3}},$$

while only one equilibrium exists outside this interval.

To determine stability, let

$$f(y;\mu) = \mu + y - y^3.$$

Then

$$f_y(y;\mu) = 1 - 3y^2.$$

An equilibrium is asymptotically stable when

$$|y| > \frac{1}{\sqrt{3}},$$

and unstable when

$$|y| < \frac{1}{\sqrt{3}}.$$

Consequently, in the three-equilibrium region, the upper and lower branches are stable while the middle branch is unstable.

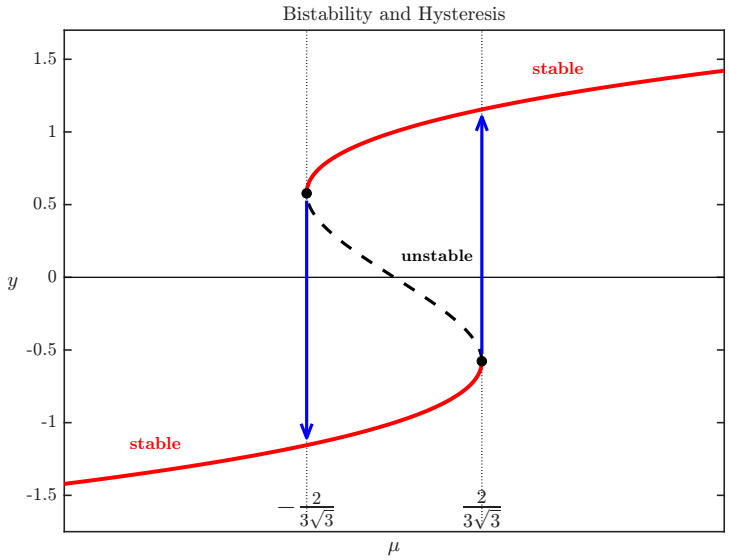


Figure 1.14: Bistability and hysteresis for $y' = \mu + y - y^3$. The upper and lower branches are stable, while the middle branch is unstable. Between the two fold points, two stable equilibria coexist. The arrows on representative vertical lines $\mu =$ constant indicate the direction of the flow.

The S-shaped diagram displays a complete hysteresis cycle. Suppose that μ varies slowly and that the system initially lies on the lower stable branch. As μ increases, the state follows this branch until the right fold is reached. The lower stable equilibrium then disappears, and the state makes a rapid transition to the upper stable branch.

If μ is subsequently decreased, the state remains on the upper stable branch rather than jumping back at the same value of μ. It continues along that branch until the left fold is reached, where the upper equilibrium disappears and the state jumps to the lower branch. Thus the forward and backward transitions occur at different parameter values.

This is the classical hysteresis loop. A slowly varying parameter can produce sudden transitions between stable states, and reversing the parameter does not immediately reverse the transition. The state depends not only on the present value of the parameter but also on the path by which that value was reached. For a parameter varying at a nonzero rate, a small delay beyond each static fold may occur.

1.5.4 Pitchfork bifurcation

A pitchfork bifurcation occurs naturally in systems with symmetry. As a parameter passes through a critical value, a symmetric equilibrium may change stability while two new symmetric equilibrium branches appear. Elastic buckling provides a classical example.

Buckling of an elastic rod. Consider a slender elastic rod compressed by an axial load P. Let y represent the amplitude of its lateral deflection, with $y = 0$ corresponding to the straight configuration. Positive and negative values of y represent buckling in the two opposite transverse directions.

For small loads, the straight configuration is stable. When the compressive load exceeds a critical value P_c, however, the straight configuration becomes unstable and two stable buckled configurations appear. Near the critical load, this behavior can be represented by the reduced equation

$$y' = (P - P_c)y - \beta y^3, \qquad \beta > 0.$$

This equation should be interpreted as a simple evolution law for the buckling amplitude rather than as the full equation of motion of the elastic rod. Its equilibria represent the possible static configurations.

The equilibrium equation is

$$(P - P_c)y - \beta y^3 = 0,$$

or

$$y\left[(P - P_c) - \beta y^2\right] = 0.$$

Thus the straight configuration

$$y = 0$$

exists for every value of P, while the two buckled configurations

$$y_\pm = \pm\sqrt{\frac{P - P_c}{\beta}}$$

exist when $P > P_c$.

To determine stability, define

$$f(y; P) = (P - P_c)y - \beta y^3.$$

Then

$$f_y(y; P) = (P - P_c) - 3\beta y^2.$$

At the straight configuration,

$$f_y(0; P) = P - P_c.$$

Hence $y = 0$ is asymptotically stable for $P < P_c$ and unstable for $P > P_c$.

At the buckled equilibria,

$$f_y(y_\pm; P) = -2(P - P_c) < 0,$$

so both y_- and y_+ are asymptotically stable when $P > P_c$.

Thus

$$\begin{cases} P < P_c, & \text{one stable straight configuration,} \\ P = P_c, & \text{one degenerate equilibrium,} \\ P > P_c, & \text{one unstable straight configuration and two stable buckled configurations.} \end{cases}$$

The two buckled branches are symmetric about $y = 0$. This reflects the symmetry of the idealized rod: there is no preferred direction of lateral buckling. The equilibrium diagram therefore has the characteristic shape of a pitchfork.

Normal form. The essential bifurcation structure is obtained by introducing

$$\mu = P - P_c$$

and rescaling the dependent variable to remove the coefficient β. This gives the standard pitchfork normal form

$$y' = \mu y - y^3.$$

The equilibria satisfy

$$y(\mu - y^2) = 0.$$

Thus

$$y = 0$$

is an equilibrium for every μ, while

$$y_\pm = \pm\sqrt{\mu}$$

exist for $\mu > 0$.

Let

$$f(y; \mu) = \mu y - y^3.$$

Then

$$f_y(y; \mu) = \mu - 3y^2.$$

Along the branch $y = 0$,

$$f_y(0; \mu) = \mu,$$

so the zero equilibrium is stable for $\mu < 0$ and unstable for $\mu > 0$. Along the two nonzero branches,

$$f_y(y_\pm; \mu) = -2\mu < 0,$$

so both are stable for $\mu > 0$.

This is called a *supercritical pitchfork bifurcation*. Its defining feature is the symmetry

$$f(-y; \mu) = -f(y; \mu).$$

If y is an equilibrium, then $-y$ is also an equilibrium. The two nonzero branches therefore emerge symmetrically from the trivial branch at the critical parameter value.

The symmetry is essential. In an ideal elastic rod, the two buckling directions are equivalent. If a small asymmetry is introduced, such as an initial imperfection or a small transverse force, this symmetry is broken and the perfect pitchfork structure is generally lost. Thus the pitchfork bifurcation is particularly characteristic of systems possessing an underlying symmetry.

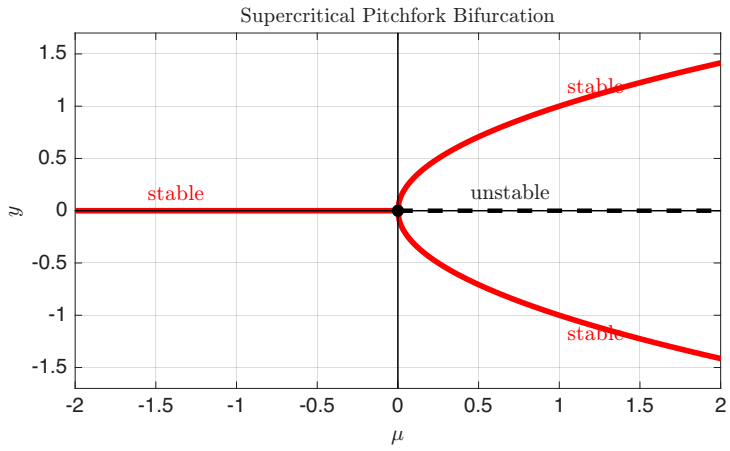


Figure 1.15: Pitchfork bifurcation for $y' = \mu y - y^3$. For $\mu < 0$, the equilibrium $y = 0$ is stable. At $\mu = 0$, it loses stability, and for $\mu > 0$ two stable equilibria $y = \pm\sqrt{\mu}$ appear. Stable branches are shown by solid curves and the unstable branch by a dashed curve. The arrows on representative vertical lines $\mu =$ constant indicate the direction of the flow.

A two-phase material. The same pitchfork structure appears in models of phase transitions. Let y be an order parameter that distinguishes two phases of a material. For example, $y > 0$ and $y < 0$ may represent two symmetry-related phases, while $y = 0$ represents a symmetric state.

A simple energy describing such a system is

$$V(y;\mu) = -\frac{\mu}{2}y^2 + \frac{1}{4}y^4.$$

The energy is symmetric under the transformation $y \mapsto -y$. If the system relaxes toward states of lower energy, a simple evolution law is the gradient equation

$$y' = -V_y(y;\mu).$$

Since

$$V_y(y;\mu) = -\mu y + y^3,$$

we obtain

$$y' = \mu y - y^3.$$

Thus the same equation that describes the reduced buckling amplitude also describes a simple symmetric phase-transition model.

When $\mu < 0$, the potential V has a single minimum at $y = 0$, corresponding to the symmetric phase. At $\mu = 0$, the character of this equilibrium changes. When $\mu > 0$, $y = 0$ becomes a local maximum and two minima appear at

$$y_\pm = \pm\sqrt{\mu}.$$

These two minima correspond to the two stable phases. The pitchfork bifurcation therefore represents a symmetry-breaking transition from one symmetric state to two symmetry-related stable states.

Subcritical pitchfork. The pitchfork considered above is called *supercritical* because the two stable nonzero branches appear on the side where the zero equilibrium has become unstable. A different possibility is the *subcritical pitchfork*, represented locally by

$$y' = \mu y + y^3.$$

The equilibria are

$$y = 0,$$

together with

$$y_\pm = \pm\sqrt{-\mu}, \qquad \mu < 0.$$

For

$$f(y;\mu) = \mu y + y^3,$$

we have

$$f_y(y;\mu) = \mu + 3y^2.$$

Thus $y = 0$ is stable for $\mu < 0$ and unstable for $\mu > 0$, while

$$f_y(y_\pm;\mu) = -2\mu > 0$$

for $\mu < 0$. Hence the two nonzero branches are unstable.

The cubic normal form describes the local structure near the bifurcation point, but by itself it does not provide stable finite-amplitude states after the zero equilibrium loses stability. In physical applications, higher-order terms may stabilize the system. A simple model is

$$y' = \mu y + y^3 - y^5.$$

The corresponding potential is

$$V(y;\mu) = -\frac{\mu}{2}y^2 - \frac{1}{4}y^4 + \frac{1}{6}y^6,$$

so that again

$$y' = -V_y(y;\mu).$$

The nonzero equilibria satisfy

$$\mu + y^2 - y^4 = 0.$$

Setting

$$z = y^2$$

gives

$$z^2 - z - \mu = 0,$$

and hence

$$y^2 = \frac{1 \pm \sqrt{1+4\mu}}{2}.$$

The two nonzero branches meet in folds when

$$\mu = -\frac{1}{4}.$$

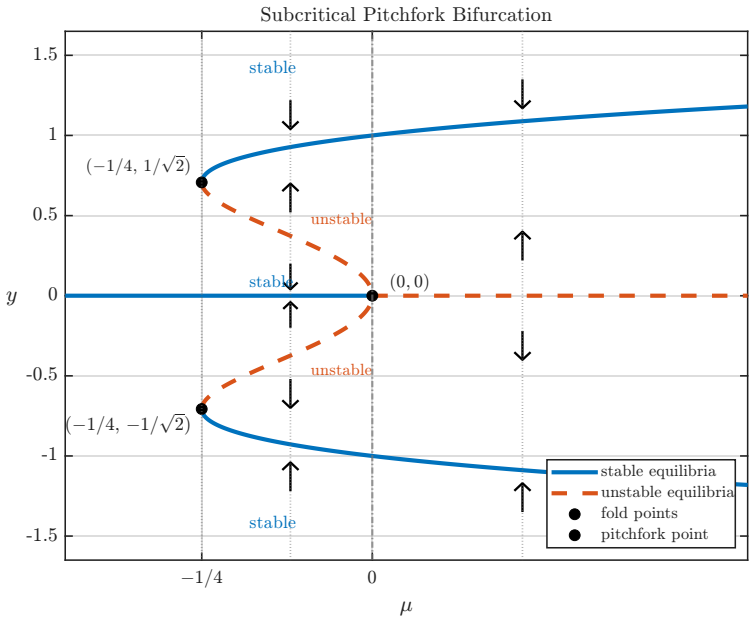


Figure 1.16: Subcritical pitchfork with nonlinear saturation for $y' = \mu y + y^3 - y^5$. The zero branch is stable for $\mu < 0$ and loses stability at $\mu = 0$. The small-amplitude nonzero branches are unstable, while the outer finite-amplitude branches are stable. The folds at $\mu = -1/4$ create a region of bistability for $-1/4 < \mu < 0$. Solid curves represent stable equilibria and dashed curves unstable equilibria.

For

$$-\frac{1}{4} < \mu < 0,$$

the zero equilibrium and a pair of finite-amplitude equilibria are stable, with unstable equilibria separating their basins of attraction. The system is therefore bistable in this parameter range.

This example also connects the subcritical pitchfork with the hysteresis phenomenon discussed in the preceding subsection. Suppose that μ varies slowly and the system initially lies near $y = 0$. As μ increases through zero, the zero equilibrium loses stability and the state moves rapidly to one of the stable finite-amplitude branches. If μ is then decreased, the system remains on that branch past $\mu = 0$ and returns to the zero state only when the fold at

$$\mu = -\frac{1}{4}$$

is reached. Thus the transition occurs at different parameter values in the forward and reverse directions, producing hysteresis.

The supercritical and subcritical pitchforks therefore describe qualitatively different transitions. In a supercritical pitchfork, stable states emerge continuously from the symmetric state at the critical parameter value. In a subcritical pitchfork, stable finite-amplitude states may coexist with the symmetric state before it loses stability, allowing a discontinuous jump and hysteresis.

1.5.5 Interpretation in applications

The examples in this section illustrate a common principle. A parameter in a differential equation may represent an external condition, a control variable, or a physical property of the system. As the parameter varies, the equilibria and their stability may change. A bifurcation occurs when this change alters the qualitative structure of the dynamics.

The examples above exhibit several distinct mechanisms:

phenomenon	example	characteristic change
saddle-node	constant-yield harvesting	two equilibria merge and disappear
transcritical	per-capita harvesting	two branches exchange stability
supercritical pitchfork	elastic buckling, phase transition	two stable symmetric states appear
subcritical pitchfork	nonlinear buckling or phase transition	jump and possible hysteresis

Bistability and hysteresis are somewhat different concepts. Bistability means that two stable states coexist for the same parameter value, while hysteresis describes the resulting dependence of the observed state on the history of a varying parameter.

Parameters and time scales. A bifurcation diagram is constructed by treating the parameter as fixed. For example, for

$$y' = f(y; \mu),$$

each vertical line $\mu =$ constant in the bifurcation diagram represents a different autonomous ODE and its phase line. The equilibrium branches describe how the long-time states of this family of equations depend on μ.

In applications, however, a parameter may itself change with time. If this change is sufficiently slow compared with the relaxation of the state variable, we may write schematically

$$\mu = \mu(\varepsilon t), \qquad 0 < \varepsilon \ll 1.$$

The state can then remain close to a stable equilibrium branch while the parameter changes slowly. When that branch loses stability or disappears, the state may undergo a much faster transition to another attracting state. This separation of time scales explains how gradual changes in external conditions can produce apparently sudden changes in the observed system.

Critical transitions and tipping. The saddle-node harvesting model gives the simplest example. Increasing the harvesting amount gradually moves the sustainable population along its stable branch. At the critical value, that branch disappears, and the population can no longer remain near its previous equilibrium. The system then undergoes a rapid transition toward extinction.

Similar threshold phenomena occur in many applications. A slowly changing load may cause an elastic structure to buckle, a gradual change of external conditions may induce a transition between material phases, and changes in ecological parameters may lead to a transition between different population states. The critical parameter value at which such a qualitative change occurs is often called a *tipping point*.

Not every bifurcation produces such a sudden jump. In the transcritical harvesting model, for example, the stable positive population decreases continuously to zero as the critical harvesting rate is approached. In a supercritical pitchfork, the new stable states also emerge continuously from the original equilibrium. By contrast, saddle-node and subcritical bifurcations can produce

abrupt transitions because a stable state may cease to exist or lose stability without a nearby stable continuation.

Bistability and memory. When two stable equilibria coexist, knowledge of the parameter alone is not sufficient to determine the state of the system. The initial condition and the previous evolution also matter. An unstable equilibrium typically separates the basins of attraction of the two stable states.

If the parameter is varied slowly, this bistability can lead to hysteresis. The system may remain on one stable branch until that branch terminates and then jump to another. Reversing the parameter does not necessarily reverse the transition at the same parameter value. In this sense, the present state retains information about the path by which it was reached.

The role of normal forms. The concrete models in this section contain parameters with particular physical or biological meanings. Nevertheless, near a bifurcation point many different models have the same leading-order mathematical structure. This motivates the use of simple normal forms such as

$$y' = \mu - y^2, \qquad y' = \mu y - y^2, \qquad y' = \mu y - y^3.$$

They represent, respectively, the local structures of saddle-node, transcritical, and supercritical pitchfork bifurcations.

The value of a normal form is that it separates the essential local dynamics from the details of a particular application. Models arising from quite different physical settings may therefore have qualitatively identical bifurcation diagrams near their critical points.

The one-dimensional examples studied here provide only the simplest bifurcation mechanisms. In higher-dimensional systems, equilibria may also lose stability through the creation of periodic motion, and more complicated dynamical transitions become possible. Nevertheless, the central idea remains the same: changes in parameters can produce qualitative changes in the long-time behavior of a dynamical system.

Exercise 1.9 (Square-root scaling near a saddle-node)**.** For the harvested logistic equation

$$y' = ry\left(1 - \frac{y}{K}\right) - H,$$

let $y_-(H)$ and $y_+(H)$ be the two positive equilibria for

$$H < H_c = \frac{rK}{4}.$$

Show that

$$y_+(H) - y_-(H)$$

is proportional to $\sqrt{H_c - H}$ as $H \uparrow H_c$. Compare this scaling with the saddle-node normal form

$$z' = \mu - z^2.$$

Exercise 1.10 (A hysteresis loop)**.** Consider

$$y' = \mu - y(y-1)(y-2).$$

Find the parameter interval in which three equilibria coexist and classify their stability. Describe what happens when μ is slowly increased through this interval and then decreased. Sketch the resulting hysteresis loop.

Hint. Set $z = y - 1$ and find the turning points of

$$\mu = z^3 - z.$$

Exercise 1.11 (Robustness of a pitchfork bifurcation)**.** Consider

$$y' = \mu y - y^3 - a y^5, \qquad a > 0.$$

Determine the equilibrium branches and their stability near $(\mu, y) = (0, 0)$. Show that the system has a supercritical pitchfork and that the nonzero branches satisfy

$$y_{\pm}(\mu) = \pm\sqrt{\mu}\left(1 - \frac{a}{2}\mu + O(\mu^2)\right)$$

as $\mu \downarrow 0$. Explain why the fifth-order term changes the branches but not the local bifurcation type.

Homework Problems

The following problems are intended to synthesize the main ideas of this chapter: modeling assumptions, equilibria, relaxation, phase lines, parameter dependence, and bifurcation. In each problem, the interpretation of the mathematics is as important as the calculation.

Homework 1.1 (Imperfect pitchfork and symmetry breaking)**.** Consider

$$y' = \mu y - y^3 + \varepsilon.$$

Determine the saddle-node bifurcation set in the (μ, ε)-plane and identify the regions containing one and three equilibria. For a fixed small $\varepsilon \neq 0$, sketch the equilibrium branches as μ varies and classify their stability. Explain how breaking the symmetry $y \mapsto -y$ unfolds the pitchfork bifurcation.

Hint. At a saddle-node, solve

$$\mu y - y^3 + \varepsilon = 0, \qquad \mu - 3y^2 = 0.$$

Homework 1.2 (The saddle-node bottleneck)**.** Consider

$$y' = \mu + y^2.$$

Analyze the equilibria for $\mu < 0$, $\mu = 0$, and $\mu > 0$. For $\mu > 0$, find the time required for a solution to travel from $y = -a$ to $y = a$ and show that

$$T(\mu) \sim \frac{\pi}{\sqrt{\mu}} \qquad \text{as } \mu \downarrow 0.$$

Interpret this divergence as a bottleneck near the vanished equilibrium.

Hint. Use

$$T(\mu) = \int_{-a}^{a} \frac{dy}{\mu + y^2}.$$

Homework 1.3 (Nutrient loading and lake eutrophication)**.** A classical phosphorus-recycling model for lake eutrophication combines external nutrient loading, nutrient loss, and a saturating internal recycling mechanism (Carpenter et al., 1999). A nondimensional deterministic version is

$$\dot{x} = \lambda - x + \rho \frac{x^2}{1 + x^2}, \qquad x \geq 0.$$

Here x is the nutrient concentration in a lake, $\lambda \geq 0$ is the external nutrient loading, and $\rho > 0$ measures internal nutrient recycling. The recycling term represents positive feedback: accumulated nutrients promote further release of nutrients from the sediment.

Show that the equilibrium diagram can be written as

$$\lambda = x - \rho \frac{x^2}{1 + x^2}.$$

Determine when this curve has two folds. For

$$\frac{8\sqrt{3}}{9} < \rho < 2,$$

sketch the equilibrium diagram, classify the stability of its branches, and identify the interval of nutrient loading for which two stable equilibria coexist.

Describe what happens when λ is slowly increased and then decreased. Explain why restoring a eutrophic lake may require reducing the nutrient loading below the value at which eutrophication originally occurred.

Such coexistence of clear and eutrophic states, together with abrupt transitions and hysteresis, is an important example of alternative stable states in ecosystems (Scheffer et al., 2001).

Hint. At a fold,

$$1 = \frac{2\rho x}{(1 + x^2)^2}.$$

Study the minimum of

$$\frac{(1+x^2)^2}{2x}, \qquad x > 0.$$

The sign of the slope of the equilibrium curve determines stability.

Historical Notes and Further Reading

Historical notes. Pierre-François Verhulst introduced the logistic equation as a modification of unrestricted exponential population growth (Verhulst, 1838). He later gave a fuller treatment of the logistic curve (Verhulst, 1845). It remains one of the basic models of self-limited population growth.

The Allee effect is named after Warder Clyde Allee, whose studies of animal aggregation in the 1920s and 1930s emphasized that low population density may reduce survival or reproduction (Allee, 1927, 1931). The strong Allee-effect equation used in this chapter is a simple phenomenological representation of this mechanism: extinction and a positive population state are stable, while an intermediate unstable equilibrium acts as a population threshold.

Harvesting models are central in fisheries management. Schaefer (1954) developed a classical surplus-production model in which the unharvested stock follows logistic growth and the catch is proportional to both the stock and the fishing effort. If the fishing effort is held fixed, the harvest term is proportional to the population, as in hy; a prescribed constant catch rate is represented by H. Thus the parameters in these equations have direct biological or management meanings. As a harvesting parameter varies, equilibria may move, merge, disappear, or change stability. Parameter-dependent ODEs, and hence bifurcation theory, arise naturally from the modeling problem.

Henri Poincaré introduced the term *bifurcation* in his study of equilibrium figures of rotating fluid masses (Poincaré, 1885). In this chapter, saddle-node, transcritical, and pitchfork bifurcations describe the simplest one-dimensional mechanisms by which the number or stability of equilibria changes.

Further reading. For application-driven introductions to mathematical modeling, see Tung (2007) and Kaper and Engler (2013). Tung develops a broad range of models from the physical and applied sciences. Kaper and Engler use conceptual climate models to show how differential equations, equilibria, multiple stable states, bifurcation, and hysteresis arise from physical mechanisms.

For further treatments of population models and their qualitative analysis, see Edelstein-Keshet (2005) and Murray (2002). For the ecological mechanisms and conservation consequences of the Allee effect, see Courchamp et al. (2008); for fisheries and renewable-resource management, see Clark (2010). For nonlinear elastic instability, snap-through, and related bistability phenomena, see Thompson and Hunt (1973). An accessible introduction to one-dimensional dynamics and elementary bifurcations is given by Strogatz (2024). For classical analytical methods in bifurcation theory, see Chow and Hale (1982). A systematic modern treatment of applied bifurcation theory is given by Kuznetsov (2023).

Chapter 2

Scalar Differential Equations: Methods and Geometry

In Chapter 1, we studied differential equations through models. We saw equilibria, phase lines, and stability. These ideas already give much information without solving the equation explicitly.

In this chapter, we study scalar first-order equations more systematically. The standard form is

$$y' = f(t, y).$$

We begin with classical solution methods. These are mostly calculus methods. They work only for special equations, but they are important. They also give exact formulas against which qualitative and numerical methods can be compared.

After this analytic part, we turn to geometry. A scalar differential equation determines directions in the plane. Solutions are curves tangent to these directions. Some equations come from families of plane curves. Some equations possess first integrals or integrating factors.

Thus the chapter has two complementary themes: computation and geometry.

2.1 Classical Solution Methods

We first review several classical methods for first-order scalar equations. The main classes are separable equations, linear equations, and equations reducible to these forms.

2.1.1 Separation of variables

A first-order equation is called *separable* if it can be written as

$$y' = f(t)g(y).$$

If $g(y) \neq 0$, then

$$\frac{dy}{g(y)} = f(t)\,dt.$$

Integrating both sides gives

$$\int \frac{dy}{g(y)} = \int f(t)\,dt + C.$$

This usually gives an implicit formula for the solution.

Equilibrium solutions must be checked separately. If

$$g(y^*) = 0,$$

then

$$y(t) \equiv y^*$$

is a solution. This solution may be lost when we divide by $g(y)$.

Example 2.1. Consider

$$y' = y(1 - y^2).$$

The equilibrium solutions are

$$y = -1, \qquad y = 0, \qquad y = 1.$$

For non-equilibrium solutions,

$$\frac{dy}{y(1-y^2)} = dt.$$

Using partial fractions,

$$\frac{1}{y(1-y^2)} = \frac{1}{y} + \frac{1}{2(1-y)} - \frac{1}{2(1+y)}.$$

Therefore,

$$\ln|y| - \frac{1}{2}\ln|1-y^2| = t + C.$$

Equivalently,

$$\frac{|y|}{\sqrt{|1-y^2|}} = Ce^t.$$

This is an implicit formula for the non-equilibrium solutions.

For example, if $0 < y_0 < 1$, then the phase-line picture shows that

$$y(t) \to 1 \qquad (t \to \infty).$$

The formula is consistent with this. Indeed, the left-hand side tends to infinity as $y \to 1^-$.

Example 2.2. Consider the initial value problem

$$y' = t^2(1 - y^2), \qquad y(0) = y_0.$$

The equilibrium solutions are

$$y = 1, \qquad y = -1.$$

For non-equilibrium solutions,

$$\frac{dy}{1-y^2} = t^2\, dt.$$

Using

$$\frac{1}{1-y^2} = \frac{1}{2}\left(\frac{1}{1+y} + \frac{1}{1-y}\right),$$

we obtain

$$\frac{1}{2}\ln\left|\frac{1+y}{1-y}\right| = \frac{t^3}{3} + C.$$

Equivalently,

$$\ln\left|\frac{1+y}{1-y}\right| = \frac{2t^3}{3} + C_1.$$

If $|y_0| < 1$, then

$$C_1 = \ln\left(\frac{1+y_0}{1-y_0}\right).$$

Hence

$$\frac{1+y}{1-y} = \frac{1+y_0}{1-y_0}\exp\left(\frac{2t^3}{3}\right).$$

Solving for y, we get

$$y(t) = \frac{\dfrac{1+y_0}{1-y_0}\exp\left(\dfrac{2t^3}{3}\right) - 1}{\dfrac{1+y_0}{1-y_0}\exp\left(\dfrac{2t^3}{3}\right) + 1}.$$

Equivalently,

$$y(t) = \tanh\left(\frac{t^3}{3} + \tanh^{-1} y_0\right), \qquad |y_0| < 1.$$

Thus, if

$$-1 < y_0 < 1,$$

then

$$-1 < y(t) < 1$$

for all t, and

$$y(t) \to 1 \qquad (t \to \infty).$$

This agrees with the phase-line picture.

Example 2.3. Consider

$$y' = \frac{t^2}{1-y^2}.$$

This is separable:

$$(1-y^2)\,dy = t^2\,dt.$$

Integrating gives

$$y - \frac{y^3}{3} = \frac{t^3}{3} + C.$$

Equivalently,

$$-t^3 + 3y - y^3 = C.$$

This is a natural implicit form of the solution. There is no need to solve the cubic equation for y.

Example 2.4. Consider

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

We rewrite it as

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

Assume $x \neq 0$ and $y^2 \neq 1$. Then

$$\frac{y}{y^2-1}\,dy = -\frac{x^2+1}{x}\,dx.$$

Integrating gives

$$\frac{1}{2}\ln|y^2-1| = -\frac{x^2}{2} - \ln|x| + C.$$

Thus

$$\ln|y^2-1| = -x^2 - 2\ln|x| + C.$$

Hence

$$y^2 - 1 = C\frac{e^{-x^2}}{x^2}.$$

Therefore,

$$y^2 = 1 + C\frac{e^{-x^2}}{x^2}.$$

The constant solutions

$$y = 1, \qquad y = -1$$

should be added separately.

Exercise 2.1. Solve the following separable equations. When appropriate, identify any solutions that may be lost during separation.

1. $y' = y(1-y^3)$. Find the equilibrium solutions and write the non-equilibrium solutions in implicit form.
2. $y' = r\tanh y,\ r > 0$. Show that the non-equilibrium solutions satisfy $\sinh y = Ce^{rt}$.
3. Consider the separable equation

$$y' = \frac{y\left(-\frac{1}{2} + \frac{x}{4}\right)}{x\left(2 - \frac{y}{2}\right)}.$$

This equation appears when one eliminates time from a predator–prey system. Separate the variables and obtain an implicit formula for the solution curves.

2.1.2 Linear first-order equations

Another important class of solvable equations is the linear first-order equation

$$y' = a(t)y + b(t),$$

where $a(t)$ and $b(t)$ are given functions. Throughout this subsection, we assume that a and b are continuous on an interval containing the initial time $t = 0$.

The corresponding equation

$$y' = a(t)y$$

is called the *homogeneous equation*. The function $a(t)$ is often called the *growth rate*, while $b(t)$ is called the *source term* or *forcing term*.

As in linear algebra, we first solve the homogeneous equation and then use it to construct the solution of the inhomogeneous equation.

The homogeneous equation

Consider

$$y' = a(t)y.$$

Assuming $y \neq 0$, we separate variables:

$$\frac{y'}{y} = a(t).$$

Integrating this yields

$$\log |y(t)| = \int_0^t a(s)\, ds + C.$$

Exponentiating both sides gives

$$y(t) = Ce^{A(t)},$$

where

$$A(t) = \int_0^t a(s)\, ds.$$

If $y(0) = y_0$, then $C = y_0$, and therefore

$$y(t) = y_0 \exp\left(\int_0^t a(s)\, ds \right).$$

The quantity

$$\exp\left(\int_0^t a(s)\, ds \right)$$

is called the *propagator* of the homogeneous equation.

Example 2.5 (Seasonal growth rate)**.** Let $y(t)$ be the size of a population. Suppose that the growth rate is not constant, but changes periodically with the season:

$$y' = a(t)y, \qquad a(t) = r_0 + r_1 \cos(\omega t), \qquad \omega > 0.$$

Here r_0 is the average growth rate, r_1 is the amplitude of the seasonal variation, and ω is the seasonal frequency.

Assume

$$y(0) = y_0.$$

Then

$$A(t) = \int_0^t a(s)\, ds = \int_0^t \big(r_0 + r_1 \cos(\omega s)\big)\, ds.$$

Hence

$$A(t) = r_0 t + \frac{r_1}{\omega} \sin(\omega t).$$

Therefore,

$$y(t) = y_0 \exp\left(r_0 t + \frac{r_1}{\omega} \sin(\omega t)\right).$$

The factor $e^{r_0 t}$ gives the average exponential growth. The periodic factor

$$\exp\left(\frac{r_1}{\omega} \sin(\omega t)\right)$$

describes seasonal oscillation around this average growth.

We now return to the linear equation

$$y' = a(t)y + b(t). \tag{2.1}$$

There are two standard methods for solving (2.1): the method of variation of constant and the method of integrating factor. Both are based on the propagator of the homogeneous equation.

Method of variation of constants

The homogeneous equation

$$y' = a(t)y$$

has the general solution

$$y_h(t) = Ce^{A(t)}, \qquad A(t) = \int_0^t a(s)\, ds.$$

Instead of treating C as a constant, we allow it to vary with time and seek a solution of the form

$$y(t) = C(t)e^{A(t)}.$$

Differentiating gives

$$y'(t) = C'(t)e^{A(t)} + a(t)C(t)e^{A(t)}.$$

Substituting into the differential equation yields

$$C'(t)e^{A(t)} + a(t)C(t)e^{A(t)} = a(t)C(t)e^{A(t)} + b(t).$$

The terms involving $a(t)$ cancel, leaving

$$C'(t) = b(t)e^{-A(t)}.$$

Integrating from 0 to t, we obtain

$$C(t) = C(0) + \int_0^t b(s)e^{-A(s)}\,ds.$$

Since $y(0) = y_0$ and $A(0) = 0$, we have $C(0) = y_0$. Therefore,

$$C(t) = y_0 + \int_0^t b(s)e^{-A(s)}\,ds.$$

Substituting back into

$$y(t) = C(t)e^{A(t)},$$

we obtain the solution formula

$$y(t) = y_0e^{A(t)} + \int_0^t b(s)e^{A(t)-A(s)}\,ds.$$

This formula is valid for every linear first-order differential equation.

Physical interpretation. The solution formula admits a simple and useful interpretation. The first term

$$y_0e^{A(t)}$$

describes the evolution of the initial condition under the homogeneous dynamics.

For a fixed time $s < t$, the quantity

$$b(s)e^{A(t)-A(s)}$$

represents the contribution generated by the source term at time s. After being created, this contribution evolves according to the homogeneous equation.

The total solution is obtained by adding all contributions from times between 0 and t:

$$y(t) = y_0e^{A(t)} + \int_0^t b(s)e^{A(t)-A(s)}\,ds.$$

That is,

$$\text{solution} = \text{free evolution of initial data} + \text{sum of all evolved source contributions.}$$

This interpretation is called *Duhamel's principle*.

Example 2.6 (Bacteria growth with injection)**.** Let $y(t)$ be the amount of bacteria in a vessel. Suppose bacteria grow at a constant rate r. Suppose also that bacteria are injected into the vessel at rate $b(t)$. Then

$$y' = ry + b(t), \qquad y(0) = y_0.$$

Here

$$a(t) = r, \qquad A(t) = rt.$$

The homogeneous propagator from time s to time t is

$$e^{r(t-s)}.$$

By the variation-of-constants formula,

$$y(t) = y_0 e^{rt} + \int_0^t e^{r(t-s)} b(s)\, ds.$$

The first term is the growth of the bacteria already present at time 0. The integral term is the total contribution of bacteria injected later.

Bacteria injected at time s have time $t - s$ to grow. Thus their amount is multiplied by

$$e^{r(t-s)}.$$

This is the physical meaning of the Duhamel formula.

If $b(t) = b_0$ is constant, then

$$y(t) = y_0 e^{rt} + b_0 \int_0^t e^{r(t-s)}\, ds.$$

If $r \neq 0$, then

$$\int_0^t e^{r(t-s)}\, ds = \frac{e^{rt} - 1}{r}.$$

Hence

$$y(t) = y_0 e^{rt} + \frac{b_0}{r}(e^{rt} - 1).$$

Equivalently,

$$y(t) = \left(y_0 + \frac{b_0}{r}\right) e^{rt} - \frac{b_0}{r}.$$

If $r = 0$, then there is no growth in the vessel and

$$y' = b(t).$$

Hence

$$y(t) = y_0 + \int_0^t b(s)\, ds.$$

Method of integrating factor

For computations, the integrating-factor method is often convenient. Rewrite

$$y' = a(t)y + b(t)$$

as

$$y' - a(t)y = b(t).$$

Let

$$A(t) = \int_0^t a(s)\, ds.$$

Then

$$e^{-A(t)}$$

is an integrating factor, because

$$e^{-A(t)}(y' - a(t)y) = \frac{d}{dt}\left(e^{-A(t)}y\right).$$

Thus

$$\frac{d}{dt}\left(e^{-A(t)}y\right) = e^{-A(t)}b(t).$$

Integrating gives the same formula as variation of constants.

The method of variation of constants explains the structure of the solution. The integrating factor is often the shortest way to compute it.

Example 2.7. Solve

$$y' = \frac{2}{t}y + t - 1, \qquad t > 0.$$

We rewrite the equation as

$$y' - \frac{2}{t}y = t - 1.$$

An integrating factor is

$$\mu(t) = \exp\left(-\int \frac{2}{t}\,dt\right) = t^{-2}.$$

Multiplying the equation by t^{-2} gives

$$t^{-2}y' - 2t^{-3}y = t^{-1} - t^{-2}.$$

Hence

$$\frac{d}{dt}(t^{-2}y) = t^{-1} - t^{-2}.$$

Integrating,

$$t^{-2}y = \ln t + \frac{1}{t} + C.$$

Therefore,

$$y(t) = t^2 \ln t + t + Ct^2.$$

Exercise 2.2. Solve the following linear first-order equations and interpret the propagator when a source term is present.

1. $y' = 3y + e^t, \quad y(0) = 1.$ Use variation of constants.
2. Solve $y' = -2y + \sin t, \quad y(0) = 0.$
3. Solve $y' = \frac{2}{t}y + t^2, \qquad t > 0.$

2.1.3 Reducible nonlinear equations

Some nonlinear first-order equations can be reduced to linear or separable ones by a change of variables. These examples are useful, but they are not the main structural point of the chapter. The important lesson is that the apparent form of an equation may depend strongly on the variables used.

Bernoulli equations. A Bernoulli equation has the form

$$y' + p(t)y = q(t)y^{\alpha}, \qquad \alpha \neq 0, 1.$$

For $y \neq 0$, divide by y^{α} and put

$$z = y^{1-\alpha}.$$

Since $z' = (1-\alpha)y^{-\alpha}y'$, the equation becomes

$$z' + (1-\alpha)p(t)z = (1-\alpha)q(t),$$

which is linear in z. After solving this linear equation, one recovers $y = z^{1/(1-\alpha)}$. Any solution lost by division, especially $y = 0$, must be checked separately.

Example 2.8. Solve

$$y' + y = y^2.$$

Solution. This is a Bernoulli equation with $\alpha = 2$. Put $z = 1/y$. Dividing by y^2 gives

$$\frac{y'}{y^2} + \frac{1}{y} = 1,$$

and hence

$$-z' + z = 1.$$

Thus

$$z' - z = -1.$$

Solving the linear equation gives

$$z = 1 + Ce^t.$$

Therefore

$$y = \frac{1}{1 + Ce^t}.$$

The solution $y = 0$ should be added separately.

□

Riccati equations. A Riccati equation has the form

$$y' = P(t)y^2 + Q(t)y + R(t).$$

In general, there is no universal elementary method for solving it. However, if one particular solution $y_1(t)$ is known, then the substitution

$$y = y_1 + \frac{1}{u}$$

reduces the Riccati equation to a linear equation for u. Indeed, using

$$y_1' = P(t)y_1^2 + Q(t)y_1 + R(t),$$

one obtains

$$u' + (2P(t)y_1 + Q(t))u = -P(t).$$

Thus the Riccati equation becomes solvable once one particular solution is available.

Example 2.9. Solve

$$y' = y^2 - 2y,$$

given the particular solution $y_1 = 2$.

Solution. Here $P = 1$, $Q = -2$, and $R = 0$. Put

$$y = 2 + \frac{1}{u}.$$

The equation for u is

$$u' + (2Py_1 + Q)u = -P.$$

Since $2Py_1 + Q = 2$, we get

$$u' + 2u = -1.$$

Thus

$$u = Ce^{-2t} - \frac{1}{2},$$

and therefore

$$y(t) = 2 + \frac{1}{Ce^{-2t} - 1/2}.$$

The particular solution

$$y(t) \equiv 2$$

is not represented by the substitution $y = 2 + 1/u$ and must be added separately.

□

Homogeneous equations. A first-order equation of the form

$$\frac{dy}{dx} = F\Big(\frac{y}{x}\Big)$$

is reduced to a separable equation by the substitution

$$v = \frac{y}{x}, \qquad y = xv.$$

Then

$$\frac{dy}{dx} = v + x\frac{dv}{dx},$$

and hence

$$x\frac{dv}{dx} = F(v) - v.$$

When $F(v) - v \neq 0$, this gives

$$\frac{dv}{F(v) - v} = \frac{dx}{x}.$$

The exceptional values satisfying $F(v) = v$ correspond to straight-line solutions $y = vx$ and should be checked separately.

This reduction is especially useful when first-order equations arise from plane geometry, because many geometric vector fields have slopes depending only on y/x.

Clairaut equations. A Clairaut equation has the form

$$y = xy' + f(y').$$

Put $p = y'$. Then

$$y = xp + f(p).$$

Differentiating with respect to x gives

$$(x + f'(p))\frac{dp}{dx} = 0.$$

Thus either p is constant, giving the family of straight lines

$$y = Cx + f(C),$$

or the singular branch is determined by

$$x + f'(p) = 0, \qquad y = xp + f(p).$$

Geometrically, this singular branch is often the envelope of the family of straight lines.

Example 2.10 (A Clairaut equation and its envelope)**.** Consider

$$y = xy' - (y')^2.$$

The ordinary solutions are

$$y = Cx - C^2.$$

The singular branch is found from $f(p) = -p^2$, hence $x + f'(p) = x - 2p = 0$. Thus $p = x/2$, and substitution gives

$$y = x\frac{x}{2} - \left(\frac{x}{2}\right)^2 = \frac{x^2}{4}.$$

This parabola is the envelope of the line family $y = Cx - C^2$.

Exercise 2.3. The following problems concern equations that are reducible by a change of variables.

1. Solve the Bernoulli equation
$$y' - 2y = e^t y^2.$$
Add any solution that may be lost when dividing by y^2.

2. Use the substitution $v = y/x$ to solve

$$\frac{dy}{dx} = \frac{2xy}{y^2 - x^2}.$$

Write the answer in implicit form.

2.2 Geometric Viewpoint of First-Order Equations

A first-order equation can be viewed not only as a rule for finding a function $y(x)$, but also as a geometric object in the plane. It assigns tangent directions. Its solutions are curves tangent to these prescribed directions.

In this section we describe three closely related languages:

$$y' = f(x, y), \qquad \mathbf{V} = (u(x, y), v(x, y)), \qquad P(x, y)\,dx + Q(x, y)\,dy = 0.$$

The first is the usual scalar differential equation, the second is a vector field, and the third is a Pfaffian equation. Locally, where the relevant coefficients do not vanish, they describe the same tangent directions and hence the same integral curves. When these curves can be written as

$$\psi(x, y) = C,$$

the function ψ is called a first integral.

Thus solving a first-order equation may be understood as finding the family of plane curves whose tangent directions are prescribed by the equation. This viewpoint also works in the opposite direction: starting from a family of curves, one may derive a differential equation, a vector field, or a Pfaffian equation for which those curves are integral curves. This reverse viewpoint is useful in the study of orthogonal trajectories, envelopes, and other geometric constructions.

2.2.1 Direction fields, vector fields, and Pfaffian equations

A first-order equation

$$\frac{dy}{dx} = f(x, y)$$

assigns a tangent direction at each point where f is defined. The graph of a solution has tangent vector

$$(1, y') = (1, f(x, y)).$$

Thus the equation determines a *direction field* in the plane.

Example 2.11. Consider

$$\frac{dy}{dx} = -\frac{x}{y}, \qquad y \neq 0.$$

Separating variables gives

$$x^2 + y^2 = C.$$

Thus the solution curves are arcs of circles centered at the origin. The apparent singularity on

$y = 0$ comes from representing the curves as graphs over the x-axis. A vector-field formulation will recover the complete circles.

The same geometry can be described more globally by vector fields. Let

$$\mathbf{V} = (u(x, y), v(x, y))$$

be a vector field on a domain in the plane. An integral curve $\boldsymbol{\gamma}(\tau) = (x(\tau), y(\tau))$ satisfies

$$\dot{x} = u(x(\tau), y(\tau)), \qquad \dot{y} = v(x(\tau), y(\tau)).$$

If $u \neq 0$ along a portion of the curve, then x may be used as a local parameter, and

$$\frac{dy}{dx} = \frac{\dot{y}}{\dot{x}} = \frac{v(x, y)}{u(x, y)}.$$

Thus the scalar equation gives a local description of the integral curves of the vector field.

Example 2.12 (A point vortex)**.** Consider the vector field

$$\mathbf{V}(x, y) = \frac{1}{x^2 + y^2}(-y, x), \qquad (x, y) \neq \mathbf{0}.$$

This is, up to a constant factor, the velocity field of a point vortex in an incompressible planar flow. The origin is excluded because the field is singular there. At each point, the velocity is tangent to the circle centered at the origin passing through that point. Its speed is

$$|\mathbf{V}(x, y)| = \frac{1}{\sqrt{x^2 + y^2}}.$$

Thus the flow is faster near the origin and slower far away from the origin.

The integral curves satisfy

$$\dot{x} = -\frac{y}{x^2 + y^2}, \qquad \dot{y} = \frac{x}{x^2 + y^2}.$$

Along such a curve,

$$\frac{d}{d\tau}(x^2 + y^2) = 2x\dot{x} + 2y\dot{y}$$

and hence

$$\frac{d}{d\tau}(x^2 + y^2) = 2x\left(-\frac{y}{x^2 + y^2}\right) + 2y\left(\frac{x}{x^2 + y^2}\right) = 0.$$

Therefore

$$x^2 + y^2 = C.$$

The streamlines are the concentric circles centered at the origin.

If we use x as the independent variable, then where $y \neq 0$,

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

Thus the same geometric curves are described locally by

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

This scalar equation breaks down on the line $y = 0$, although the streamlines pass through the x-axis smoothly. The singularity is caused by the choice of x as the independent variable; it is not a singularity of the geometric curves.

The simpler rotational field

$$\mathbf{V}_0 = (-y, x)$$

has the same streamlines, since

$$\mathbf{V} = \frac{1}{x^2 + y^2}\mathbf{V}_0.$$

The two fields have the same tangent directions, but different speeds. Thus multiplying a vector field by a positive scalar function changes the parametrization of the motion, but not the unparametrized integral curves.

This illustrates an important distinction. A vector field contains both the tangent direction and the speed along the curve. The family of integral curves remembers only the tangent directions. Multiplying a vector field by a positive scalar function changes the speed, but not the geometric curves.

We now introduce the corresponding Pfaffian language. A *differential* 1*-form* on the plane is an expression of the form

$$\eta = P(x, y)\, dx + Q(x, y)\, dy.$$

At each point (x, y), the 1-form η is a linear functional on vectors based at that point. Hence, if

$$\mathbf{V}(x, y) = (u(x, y), v(x, y))$$

is a vector field, then η acts on $\mathbf{V}$ pointwise by

$$\eta(\mathbf{V})(x, y) = P(x, y)u(x, y) + Q(x, y)v(x, y).$$

A *Pfaffian equation* is an equation of the form

$$P(x, y)\, dx + Q(x, y)\, dy = 0.$$

It describes curves whose tangent vector field is annihilated by the 1-form. If $\boldsymbol{\gamma}(\tau) = (x(\tau), y(\tau))$ is such a curve, then

$$P(x, y)\dot{x} + Q(x, y)\dot{y} = 0.$$

For a vector field $\mathbf{V} = (u, v)$, the Pfaffian equation describing the same tangent directions is

$$u\, dy - v\, dx = 0.$$

Indeed, along an integral curve of $\mathbf{V}$, we have $dx = u\, d\tau$ and $dy = v\, d\tau$, and hence

$$u\, dy - v\, dx = uv\, d\tau - vu\, d\tau = 0.$$

For example, for

$$\mathbf{V} = (-y, x),$$

the associated Pfaffian equation is

$$-y\,dy - x\,dx = 0,$$

or equivalently,

$$x\,dx + y\,dy = 0.$$

Integrating, we obtain

$$x^2 + y^2 = C.$$

Thus

$$\psi(x, y) = x^2 + y^2$$

is a first integral, and its level curves $\psi(x, y) = C$ are the integral curves of the vector field.

The Pfaffian form has an advantage over the scalar graph equation. The equation

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

breaks down on $y = 0$, while the Pfaffian equation

$$x\,dx + y\,dy = 0$$

is regular there. It records the tangent directions without choosing x or y as the independent variable.

In summary, the same tangent directions may be written in three forms:

$$\mathbf{V} = (u, v), \qquad \frac{dy}{dx} = \frac{v}{u} \quad (u \neq 0), \qquad u\,dy - v\,dx = 0.$$

The scalar equation is a local graph representation. The vector field shows the direction and speed along the curves. The Pfaffian equation records the same directions without choosing x or y as the independent variable. This form is especially useful when the integral curves are naturally written implicitly.

Exercise 2.4. Consider the following questions on vector fields, flows, and Pfaffian equations.

1. For the uniform flow $\mathbf{V} = (U, 0)$, $U > 0$, find the integral curves, the scalar equation, and the corresponding Pfaffian equation. What changes if U is replaced by a positive function $U(y)$?

2. For the shear flow $\mathbf{V} = (y, 0)$, describe the motion in the regions $y > 0$, $y < 0$, and on $y = 0$. Explain what information is lost if one keeps only the unparametrized curves.

3. Consider the source field

$$\mathbf{V}(x, y) = \frac{1}{x^2 + y^2}(x, y), \qquad (x, y) \neq \mathbf{0}.$$

Find the streamlines. Explain why they are radial lines.

4. Let $\mathbf{V}_1 = (u, v)$ and $\mathbf{V}_2 = \lambda(x, y)(u, v)$, where $\lambda > 0$. Show that the two vector fields have the same unparametrized integral curves. What may change if λ vanishes or changes sign?

2.2.2 Families of integral curves and first integrals

A first-order equation usually has infinitely many solution curves. The general solution is therefore not a single curve, but a one-parameter family of curves in the plane. We write such a family in the form

$$\Psi(x, y, C) = 0,$$

where C is a parameter.

In favorable cases, the parameter can be solved as a function of (x, y):

$$C = \psi(x, y).$$

Then the same family is written in the level-set form

$$\psi(x, y) = C.$$

The function ψ is called a *first integral*. It is constant along every integral curve and labels the curves in the family.

Let

$$\mathbf{V} = (u(x, y), v(x, y))$$

be a vector field. A function $\psi(x, y)$ is a first integral of $\mathbf{V}$ if

$$\frac{d}{d\tau}\psi(x(\tau), y(\tau)) = 0$$

along every integral curve $\boldsymbol{\gamma}(\tau) = (x(\tau), y(\tau))$ of $\mathbf{V}$. Since

$$\dot{x} = u(x, y), \qquad \dot{y} = v(x, y),$$

this condition becomes

$$\psi_x u + \psi_y v = 0.$$

Equivalently,

$$d\psi(\mathbf{V}) = 0, \qquad \nabla\psi \cdot \mathbf{V} = 0.$$

Thus the vector field is tangent to the level curves of ψ.

In fluid mechanics, the same idea appears as the stream function. For a planar incompressible velocity field, a stream function ψ can be chosen locally so that, up to a sign convention,

$$\mathbf{V} = (u, v) = (\psi_y, -\psi_x).$$

Then

$$\nabla\psi \cdot \mathbf{V} = \psi_x\psi_y + \psi_y(-\psi_x) = 0.$$

Hence ψ is constant along the streamlines. In this sense, a stream function is a first integral of the velocity field.

Example 2.13 (A doublet flow)**.** Consider the velocity field

$$\mathbf{V}(x,y) = \left(\frac{x^2 - y^2}{(x^2+y^2)^2}, \frac{2xy}{(x^2+y^2)^2} \right), \qquad (x,y) \neq \mathbf{0}.$$

This is a two-dimensional doublet flow, up to a constant factor. Let

$$\psi(x,y) = \frac{y}{x^2+y^2}.$$

Then

$$\psi_x = -\frac{2xy}{(x^2+y^2)^2}, \qquad \psi_y = \frac{x^2-y^2}{(x^2+y^2)^2}.$$

Therefore

$$\mathbf{V} = (\psi_y, -\psi_x).$$

It follows immediately that

$$\nabla\psi \cdot \mathbf{V} = 0.$$

Thus ψ is a first integral of the doublet flow.

The streamlines are the level curves

$$\psi(x,y) = C,$$

that is,

$$\frac{y}{x^2+y^2} = C.$$

For $C \neq 0$, this can be rewritten as

$$x^2 + y^2 - \frac{1}{C}y = 0,$$

or

$$x^2 + \left(y - \frac{1}{2C} \right)^2 = \frac{1}{4C^2}.$$

Thus the streamlines are circles tangent to the origin, with centers on the y-axis. The case $C = 0$ gives the streamline $y = 0$.

First integrals are naturally connected with Pfaffian equations. If the integral curves are written as

$$\psi(x,y) = C,$$

then differentiating gives

$$d\psi = \psi_x\, dx + \psi_y\, dy = 0$$

along each curve. Thus the same family is represented by the Pfaffian equation

$$\psi_x\, dx + \psi_y\, dy = 0.$$

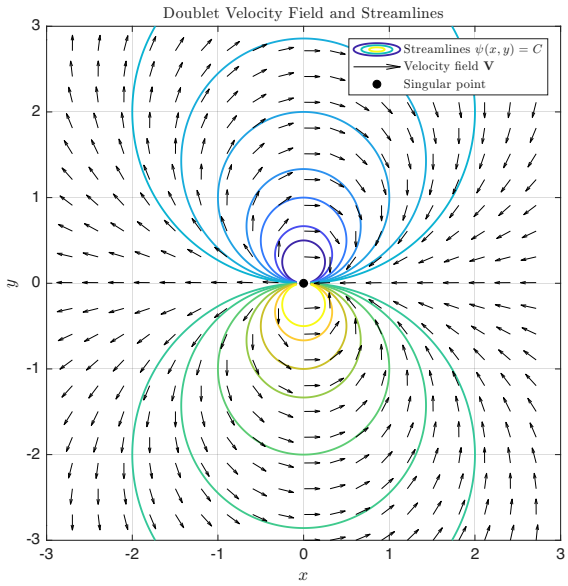


Figure 2.1: Phase portrait of the doublet velocity field $\mathbf{V} = (u, v)$ together with the streamlines $\psi(x, y) = C$, where $\psi(x, y) = \frac{y}{x^2+y^2}$. The streamlines are circles passing through the origin. The origin is a singular point of the velocity field and is omitted from the plot.

For the doublet flow above,

$$d\psi = -\frac{2xy}{(x^2+y^2)^2}\,dx + \frac{x^2-y^2}{(x^2+y^2)^2}\,dy.$$

On the other hand, if

$$\mathbf{V} = (u, v) = \left(\frac{x^2-y^2}{(x^2+y^2)^2}, \frac{2xy}{(x^2+y^2)^2}\right),$$

then the Pfaffian equation associated with the vector field is

$$u\,dy - v\,dx = 0.$$

This is precisely

$$-\frac{2xy}{(x^2+y^2)^2}\,dx + \frac{x^2-y^2}{(x^2+y^2)^2}\,dy = 0,$$

which is the equation $d\psi = 0$.

Conversely, suppose a Pfaffian equation

$$P(x, y)\,dx + Q(x, y)\,dy = 0$$

can be written as

$$P(x, y)\,dx + Q(x, y)\,dy = d\psi.$$

Then its integral curves are

$$\psi(x, y) = C.$$

Thus finding a first integral means finding a function whose differential represents the Pfaffian equation, possibly after multiplication by an integrating factor. This is the topic of the next subsection.

Remark 2.14 (Nonuniqueness of first integrals). A first integral is not unique. If ψ is a first integral and h is a differentiable function, then

$$h(\psi)$$

is also constant along the same integral curves. If $h' \neq 0$, then $h(\psi)$ gives the same family of curves with a different labeling.

Conversely, suppose ψ and ϕ are two first integrals which describe the same regular family of integral curves on a region where $\nabla\psi \neq 0$. Then, locally, the level sets of ψ and ϕ correspond to each other. Hence there exists a one-variable function h such that

$$\phi = h(\psi).$$

In this sense, two first integrals for the same regular family are functionally dependent.

2.2.3 Exact differential forms and integrating factors

We have seen that a first integral ψ gives a Pfaffian equation

$$d\psi = 0.$$

In coordinates,

$$d\psi = \psi_x\,dx + \psi_y\,dy.$$

Thus, if the integral curves of a first-order equation are written as

$$\psi(x, y) = C,$$

then their tangent directions are described by the Pfaffian equation

$$\psi_x\,dx + \psi_y\,dy = 0.$$

This observation suggests the following question. Given a Pfaffian equation

$$P(x, y)\,dx + Q(x, y)\,dy = 0,$$

when does there exist a function $\psi(x, y)$ such that

$$d\psi = P\,dx + Q\,dy?$$

If such a function exists, then the Pfaffian equation becomes

$$d\psi = 0,$$

and the integral curves are simply

$$\psi(x, y) = C.$$

Definition 2.15. A differential 1-form

$$\omega = P(x, y)\,dx + Q(x, y)\,dy$$

is called *exact* on a domain Ω if there exists a function ψ on Ω such that

$$\omega = d\psi.$$

In this case, ψ is called a potential function for ω.

If

$$\omega = d\psi,$$

then

$$P = \psi_x, \qquad Q = \psi_y.$$

Therefore, assuming the functions are continuously differentiable, we must have

$$P_y = \psi_{xy}, \qquad Q_x = \psi_{yx}.$$

Hence

$$P_y = Q_x.$$

This is the basic test for exactness.

Theorem 2.16 (Exactness test)**.** *Let*

$$\omega = P(x, y)\, dx + Q(x, y)\, dy$$

be a continuously differentiable 1*-form on a simply connected domain* $\Omega \subset \mathbb{R}^2$*. Then* ω *is exact on* Ω *if and only if*

$$P_y = Q_x$$

on Ω*.*

Proof. If $\omega = d\psi$, then $P = \psi_x$ and $Q = \psi_y$. Hence

$$P_y = \psi_{xy} = \psi_{yx} = Q_x.$$

This proves necessity.

Conversely, assume that $P_y = Q_x$ on a simply connected domain Ω. Fix a point $(x_0, y_0) \in \Omega$. For any piecewise smooth curve Γ in Ω from (x_0, y_0) to (x, y), define

$$\psi(x, y) = \int_\Gamma P\, dx + Q\, dy.$$

We claim that this line integral is independent of the path. Indeed, if Γ_1 and Γ_2 are two such curves, then the closed curve $\Gamma_1 - \Gamma_2$ bounds a region $D \subset \Omega$, after the usual decomposition into simple closed curves if necessary. By Green's theorem,

$$\int_{\Gamma_1 - \Gamma_2} P\, dx + Q\, dy = \iint_D (Q_x - P_y)\, dA.$$

Since $P_y = Q_x$, the right-hand side is zero. Hence

$$\int_{\Gamma_1} P\, dx + Q\, dy = \int_{\Gamma_2} P\, dx + Q\, dy.$$

Thus ψ is well-defined.

It remains to check that $d\psi = \omega$. If the endpoint is moved from (x, y) to $(x + \Delta x, y)$ along a horizontal segment, then

$$\psi(x + \Delta x, y) - \psi(x, y) = \int_x^{x+\Delta x} P(s, y)\, ds.$$

Dividing by Δx and letting $\Delta x \to 0$, we obtain

$$\psi_x(x, y) = P(x, y).$$

Similarly, moving the endpoint vertically gives

$$\psi_y(x, y) = Q(x, y).$$

Therefore

$$d\psi = P\,dx + Q\,dy,$$

and ω is exact. □

Remark 2.17 (Local and global exactness). The assumption that Ω is simply connected is essential. The condition

$$P_y = Q_x$$

implies that the form is locally exact, but it need not imply the existence of a single potential on the whole domain when the domain contains a hole. For example, on the punctured plane,

$$-\frac{y}{x^2 + y^2}\,dx + \frac{x}{x^2 + y^2}\,dy$$

is locally $d\theta$, but the polar angle θ cannot be chosen as a single-valued continuous function on the entire punctured plane. Indeed, the integral of this form around the unit circle is 2π, whereas the integral of an exact form around every closed curve must be zero.

Example 2.18. Consider the Pfaffian equation

$$(2xy + 1)\,dx + x^2\,dy = 0.$$

Here

$$P(x, y) = 2xy + 1, \qquad Q(x, y) = x^2.$$

Since

$$P_y = 2x, \qquad Q_x = 2x,$$

the form

$$\omega = (2xy + 1)\,dx + x^2\,dy$$

is exact.

We find ψ from

$$\psi_x = 2xy + 1.$$

Integrating with respect to x, we obtain

$$\psi = x^2y + x + h(y).$$

Differentiating with respect to y gives

$$\psi_y = x^2 + h'(y).$$

Since $\psi_y = Q = x^2$, we have

$$h'(y) = 0.$$

Thus we may take

$$\psi(x, y) = x^2 y + x.$$

The integral curves are therefore

$$x^2 y + x = C.$$

Not every Pfaffian equation is exact. However, sometimes it becomes exact after multiplication by a nonzero function.

Definition 2.19. Let

$$\omega = P(x, y)\, dx + Q(x, y)\, dy.$$

A function $\mu(x, y)$ that is nowhere zero on the domain is called an *integrating factor* for ω if

$$\mu\omega = \mu P\, dx + \mu Q\, dy$$

is exact.

If μ is an integrating factor, then there exists a function ψ such that

$$d\psi = \mu(P\, dx + Q\, dy).$$

The Pfaffian equation

$$P\, dx + Q\, dy = 0$$

is then equivalent to

$$d\psi = 0,$$

because multiplying a Pfaffian equation by a nonzero function does not change its tangent directions. Hence the integral curves are again

$$\psi(x, y) = C.$$

Example 2.20. Consider

$$y\, dx - x\, dy = 0.$$

This form is not exact. Indeed,

$$P = y, \qquad Q = -x,$$

and

$$P_y = 1, \qquad Q_x = -1.$$

However, on a region where $x \neq 0$, multiplication by

$$\mu(x, y) = \frac{1}{x^2}$$

gives

$$\mu(y\, dx - x\, dy) = \frac{y}{x^2}\, dx - \frac{1}{x}\, dy.$$

This form is exact, since

$$\frac{\partial}{\partial y}\left(\frac{y}{x^2}\right) = \frac{1}{x^2}, \qquad \frac{\partial}{\partial x}\left(-\frac{1}{x}\right) = \frac{1}{x^2}.$$

We find ψ from

$$\psi_x = \frac{y}{x^2}.$$

Thus

$$\psi = -\frac{y}{x} + h(y).$$

Since

$$\psi_y = -\frac{1}{x} + h'(y)$$

must equal $-1/x$, we get $h'(y) = 0$. Hence

$$\psi = -\frac{y}{x}.$$

The integral curves are

$$\frac{y}{x} = C.$$

The integrating factor is not unique. If $\mu\omega = d\psi$, and if $h'(\psi) \neq 0$, then

$$h'(\psi)\mu\omega = d(h(\psi)).$$

Thus another integrating factor is

$$\widetilde{\mu} = h'(\psi)\mu.$$

This nonuniqueness reflects the fact that a first integral is only a label for the same family of curves.

In practice, finding an integrating factor can be difficult. The most useful cases occur when an integrating factor depends only on x or only on y. Suppose

$$\omega = P\,dx + Q\,dy.$$

If an integrating factor depends only on x, say $\mu = \mu(x)$, then exactness of $\mu\omega$ requires

$$\frac{\partial}{\partial y}(\mu P) = \frac{\partial}{\partial x}(\mu Q).$$

Since $\mu_y = 0$, this becomes

$$\mu P_y = \mu' Q + \mu Q_x.$$

Therefore

$$\frac{\mu'}{\mu} = \frac{P_y - Q_x}{Q}.$$

Thus an integrating factor depending only on x exists if

$$\frac{P_y - Q_x}{Q}$$

is a function of x alone.

Similarly, an integrating factor depending only on y exists if

$$\frac{Q_x - P_y}{P}$$

is a function of y alone.

Exercise 2.5. Consider the following questions on exact forms, integrating factors, and first integrals.

1. For the differential form
$$(x + y)\,dx + 2x\,dy = 0,$$
show that it is not exact. On the region $x > 0$, find an integrating factor depending only on x and determine a first integral.

2. For
$$(x^2 + y^2)\,dx - xy\,dy = 0,$$
verify that $\mu(x, y) = 1/x^3$ is an integrating factor on a region where $x \neq 0$, and find a first integral.

3. Let
$$F(x, y) = x^2y + y^3, \qquad \mathbf{V} = (F_y, -F_x).$$
Show that the integral curves of $\mathbf{V}$ lie on the level sets $F(x, y) = C$. Write the corresponding scalar differential equation where possible and determine the equilibrium points of $\mathbf{V}$.

4. Let $y' = f(x, y)$ be continuously differentiable. Suppose that $F(x, y)$ is a first integral and that $\nabla F(x_0, y_0) \neq 0$. Show that the integral curve through (x_0, y_0) is locally a level set of F, and interpret the result geometrically.

2.3 Families of Plane Curves

In the previous section, we started from a differential equation or a vector field and obtained its integral curves. We now reverse the viewpoint. Suppose a family of plane curves is given first. We ask how to construct a differential equation whose solutions are precisely the curves in the family.

This point of view is useful for several reasons. It shows how differential equations arise from geometry. It also gives a systematic method for studying orthogonal trajectories and envelopes. In these problems, the curves are known before the differential equation is written down.

2.3.1 From curve families to differential equations

A one-parameter family of plane curves is often written as

$$\Psi(x, y, C) = 0,$$

where C is a parameter. The goal is to eliminate C and obtain a differential equation involving x, y, and y'. Equivalently, one may derive a Pfaffian equation

$$P(x, y)\,dx + Q(x, y)\,dy = 0$$

whose integral curves are the members of the family.

The guiding idea is simple. Along a curve in the family, the parameter C is constant. Therefore differentiating

$$\Psi(x, y, C) = 0$$

with respect to x gives

$$\Psi_x(x, y, C) + \Psi_y(x, y, C)y' = 0.$$

Together with the original equation

$$\Psi(x, y, C) = 0,$$

this relation may be used to eliminate C. The result is a differential equation for the family of curves.

Example 2.21 (Radial lines)**.** Consider the family of straight lines through the origin:

$$y = Cx.$$

Differentiating gives

$$y' = C.$$

Eliminating C from

$$y = Cx, \qquad y' = C,$$

we obtain

$$y' = \frac{y}{x}, \qquad x \neq 0.$$

Equivalently, the corresponding Pfaffian equation is

$$x\,dy - y\,dx = 0.$$

This family appears as the integral curves of a radial source field. It will also be useful when we discuss orthogonal trajectories, since radial lines are orthogonal to circles centered at the origin.

Example 2.22 (Streamlines of a doublet)**.** In the previous section we encountered the doublet stream function

$$\psi(x, y) = \frac{y}{x^2 + y^2}.$$

Its streamlines are the level curves

$$\frac{y}{x^2 + y^2} = C.$$

Equivalently,

$$Cx^2 + Cy^2 - y = 0.$$

Differentiating with respect to x, while keeping C fixed, gives

$$C(2x + 2yy') - y' = 0.$$

Using

$$C = \frac{y}{x^2 + y^2},$$

we obtain

$$\frac{y}{x^2 + y^2}(2x + 2yy') - y' = 0.$$

Multiplying by $x^2 + y^2$ gives

$$2xy + 2y^2y' - (x^2 + y^2)y' = 0.$$

Hence

$$(y^2 - x^2)y' + 2xy = 0,$$

or

$$y' = \frac{2xy}{x^2 - y^2}.$$

Thus the family of doublet streamlines satisfies the homogeneous first-order equation

$$\frac{dy}{dx} = \frac{2xy}{x^2 - y^2}.$$

This is the same equation obtained from the velocity field

$$\mathbf{V} = \left(\frac{x^2 - y^2}{(x^2 + y^2)^2}, \frac{2xy}{(x^2 + y^2)^2} \right),$$

because the slope of a streamline is $dy/dx = v/u$.

Sometimes eliminating the parameter produces a differential equation that is not solved for y'.

Example 2.23 (Translated parabolas)**.** Consider the one-parameter family of translated parabolas

$$y = (x - C)^2.$$

Differentiating gives

$$y' = 2(x - C).$$

Thus

$$x - C = \frac{y'}{2}.$$

Substituting this into the original equation gives

$$y = \left(\frac{y'}{2} \right)^2.$$

Therefore

$$(y')^2 = 4y.$$

This equation is not single-valued in y'. It represents two possible slopes:

$$y' = 2\sqrt{y}, \qquad y' = -2\sqrt{y}.$$

This example will be used later when we discuss nonuniqueness. The equation

$$y' = 2\sqrt{y}$$

has solutions that can stay on the envelope $y = 0$ for a while and then leave it along a parabola.

Example 2.24 (A Clairaut family)**.** The Clairaut equation discussed in Section 2.1.3 provides another example. Its one-parameter family of straight-line solutions,

$$y = Cx - C^2,$$

is obtained by setting $C = y'$. The associated singular solution will be revisited below from the geometric viewpoint of envelopes.

The previous examples show that eliminating the parameter may produce different kinds of first-order equations. Sometimes the result is a single-valued slope equation $y' = f(x, y)$; sometimes it is an implicit equation $F(x, y, y') = 0$. The latter occurs naturally for families with several possible tangent directions, such as translated parabolas or Clairaut families.

In the next subsection we use the same idea in a different way. Starting from one family of curves, we construct another family that intersects it orthogonally.

Exercise 2.6. Derive differential equations from the following families of curves. In each problem, eliminate the parameter and interpret the result geometrically.

1. **Cooling curves.** The curves

   $$T(t) = T_a + Ce^{-kt}, \qquad k > 0,$$

   represent Newton cooling with different initial temperatures. Eliminate C and derive the differential equation satisfied by the family.

2. **Logistic curves.** For fixed $r > 0$ and $K > 0$, the curves

   $$N(t) = \frac{K}{1 + Ce^{-rt}}$$

   are logistic solution curves. Eliminate C and recover the logistic equation.

3. **Projectile trajectories.** For

   $$y = Cx - \alpha(1 + C^2)x^2, \qquad \alpha > 0,$$

eliminate C. Explain why the resulting first-order equation is implicit and quadratic in y'.

2.3.2 Orthogonal trajectories

Orthogonal trajectories arise naturally in many applications. In electrostatics, electric field lines are orthogonal to equipotential curves. In heat conduction through an isotropic medium, heat flux lines are orthogonal to isothermal curves. In planar potential flow, streamlines are orthogonal to equipotential curves. Thus the problem of finding orthogonal trajectories is not merely a geometric construction; it is a common way to pass from one physically meaningful family of curves to another.

Mathematically, suppose a family of curves is given by

$$\Psi(x, y, C) = 0.$$

Its tangent direction may be described by a differential equation

$$\frac{dy}{dx} = f(x, y),$$

where this representation is valid. A curve orthogonal to this family must have slope equal to the negative reciprocal:

$$\frac{dy}{dx} = -\frac{1}{f(x, y)}.$$

Solving this new equation gives the family of orthogonal trajectories.

The same idea can be written in Pfaffian form. If the original family has tangent direction represented by the vector field

$$\mathbf{V} = (u, v),$$

then an orthogonal direction is represented by

$$\mathbf{V}^{\perp} = (-v, u).$$

Thus orthogonal trajectories are obtained by rotating the tangent direction by a right angle at each point.

The physical interpretation is especially clear when a scalar potential is present. If $\phi(x, y)$ is an electric potential, then the electric field is

$$\mathbf{E} = -\nabla\phi.$$

The equipotential curves are

$$\phi(x, y) = C.$$

Since $\nabla\phi$ is normal to the level curves of ϕ, the electric field lines are orthogonal to the equipotential curves. Similarly, in isotropic heat conduction, Fourier's law gives

$$\mathbf{q} = -k\nabla T,$$

where T is the temperature and $k > 0$ is the thermal conductivity. Hence heat flux lines are orthogonal to the isothermal curves $T(x, y) = C$.

Example 2.25 (Concentric circles and radial lines)**.** Consider

$$x^2 + y^2 = C.$$

Differentiating,

$$2x + 2yy' = 0.$$

Hence

$$y' = -\frac{x}{y}.$$

The orthogonal trajectories satisfy

$$y' = \frac{y}{x}.$$

Separating variables,

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

Thus

$$y = Cx.$$

So the orthogonal trajectories are the straight lines passing through the origin.

Example 2.26 (Confocal ellipses and hyperbolas)**.** The family of confocal ellipses can be written as

$$\frac{x^2}{a^2} + \frac{y^2}{a^2 - 1} = 1, \qquad a > 1.$$

The common foci are $(\pm 1, 0)$.

Differentiating with respect to x gives

$$\frac{x}{a^2} + \frac{yy'}{a^2 - 1} = 0.$$

Multiplying by $a^2(a^2 - 1)$ gives

$$x(a^2 - 1) + a^2 yy' = 0.$$

Combining this relation with the original equation and eliminating a^2 yields

$$xy\left(y' - \frac{1}{y'}\right) + x^2 - y^2 = 1.$$

This is the differential equation satisfied by the confocal ellipses.

The orthogonal trajectories are obtained by replacing y' by $-1/y'$. Since

$$xy\left(-\frac{1}{y'} + y'\right) + x^2 - y^2 = 1$$

is unchanged under this replacement, the orthogonal trajectories satisfy the same differential equation. They are the confocal hyperbolas

$$\frac{x^2}{a^2} + \frac{y^2}{a^2 - 1} = 1, \qquad 0 < a < 1.$$

Thus confocal ellipses and confocal hyperbolas form an orthogonal coordinate system.

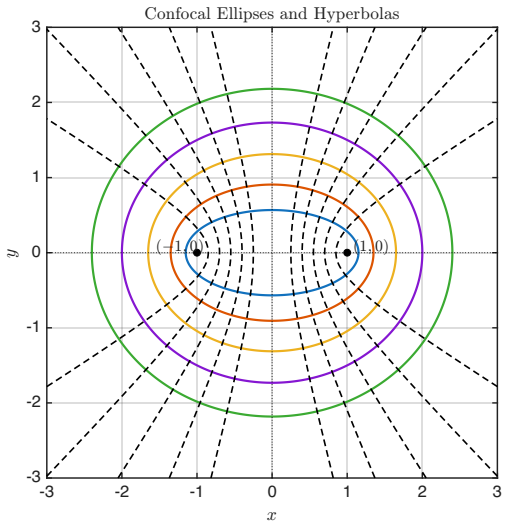


Figure 2.2: Confocal ellipses and hyperbolas with common foci $(\pm 1, 0)$. The two families are orthogonal at every intersection point.

Exercise 2.7. Find the orthogonal trajectories of the following families.

1. $y = Cx^2$.
2. $y^2 = Cx$.
3. $\frac{y}{x^2+y^2} = C$.

2.3.3 Envelopes

A one-parameter family of curves may have another curve tangent to its members. Such a curve is called an envelope. Envelopes appear in many geometric and physical problems. In optics, for example, a family of light rays may have an envelope called a caustic, where neighboring rays concentrate. In acoustics, a family of sound rays may form an envelope where wave energy is focused. Thus the envelope is not merely an auxiliary curve; it often marks a physically observable concentration of rays.

Let

$$\Psi(x, y, C) = 0$$

be a one-parameter family of curves. Candidate points of the envelope are obtained by solving

$$\Psi = 0, \qquad \Psi_C = 0,$$

and the resulting curve must be checked for tangency.

Proposition 2.27. *Suppose a family of curves*

$$\Psi(x, y, C) = 0$$

possesses an envelope. Then points on the envelope satisfy

$$\Psi(x, y, C) = 0, \qquad \Psi_C(x, y, C) = 0.$$

Proof. Let the envelope be parametrized by

$$(x(C), y(C)).$$

Since this point lies on the member curve with parameter C,

$$\Psi(x(C), y(C), C) = 0.$$

Differentiating with respect to C gives

$$\Psi_x x' + \Psi_y y' + \Psi_C = 0.$$

At a point of tangency, the tangent vector $(x'(C), y'(C))$ is tangent to the curve

$$\Psi(x, y, C) = 0.$$

Therefore,

$$\Psi_x x' + \Psi_y y' = 0.$$

Subtracting, we obtain

$$\Psi_C(x(C), y(C), C) = 0.$$

□

Example 2.28 (Envelope of a family of lines)**.** Consider the family of straight lines

$$\frac{x}{C} + \frac{y}{1 - C} = 1, \qquad 0 < C < 1.$$

Equivalently,

$$\Psi(x, y, C) = C^2 + (-x + y - 1)C + x = 0.$$

The envelope is determined by

$$\Psi = 0, \qquad \Psi_C = 0.$$

Thus

$$\begin{cases} C^2 + (-x + y - 1)C + x = 0, \\ 2C + (-x + y - 1) = 0. \end{cases}$$

Solving for x and y in terms of C, we obtain

$$x = C^2, \qquad y = (C - 1)^2, \qquad 0 < C < 1.$$

Hence the envelope is the arc of the parabola

$$(y - x - 1)^2 = 4x$$

corresponding to

$$0 < C < 1.$$

Its endpoints $(0, 1)$ and $(1, 0)$ are obtained as limiting points when $C \to 0^+$ and $C \to 1^-$, respectively.

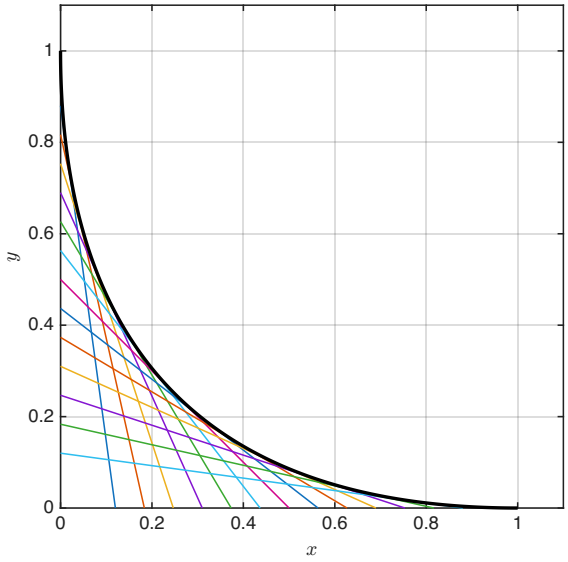


Figure 2.3: A family of straight lines and its envelope. Each line is tangent to the envelope at one point. The envelope is parametrized by $x = C^2$, $y = (C-1)^2$, $0 < C < 1$, and is the corresponding arc of the parabola $(y - x - 1)^2 = 4x$.

Example 2.29 (Envelope of a Clairaut family)**.** We return to the Clairaut family

$$y = Cx - C^2.$$

Previously, we eliminated the parameter C to obtain the Clairaut equation

$$y = xy' - (y')^2.$$

Here we use the same parameter in a different way: we look for the curve tangent to the members of the family.

Write the family as

$$F(x, y, C) = y - Cx + C^2 = 0.$$

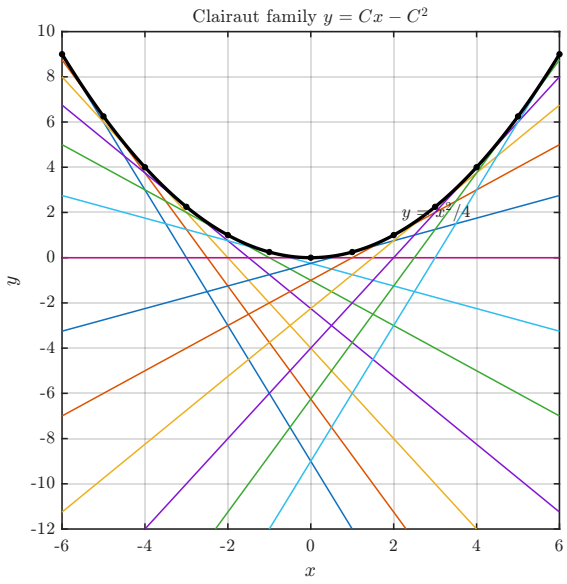


Figure 2.4: The Clairaut family $y = Cx - C^2$. Each value of C gives a straight line. The envelope of the family is the parabola $y = x^2/4$, which touches each line at the point $(2C, C^2)$.

An envelope is obtained, when it exists, by solving

$$F(x, y, C) = 0, \qquad F_C(x, y, C) = 0,$$

and then eliminating C. Since

$$F_C = -x + 2C,$$

we get

$$C = \frac{x}{2}.$$

Substituting this into the family gives

$$y = \frac{x}{2}x - \left(\frac{x}{2}\right)^2 = \frac{x^2}{4}.$$

Hence the envelope is the parabola

$$y = \frac{x^2}{4}.$$

We can also see the tangency directly. The line corresponding to C is

$$y = Cx - C^2.$$

The envelope point corresponding to the same parameter is

$$x = 2C, \qquad y = C^2.$$

At this point, the slope of the envelope

$$y = \frac{x^2}{4}$$

is

$$y' = \frac{x}{2} = C,$$

which is exactly the slope of the line. Therefore each line in the family is tangent to the envelope.

The associated Clairaut equation is

$$y = xy' - (y')^2.$$

Its ordinary solutions are the straight lines

$$y = Cx - C^2,$$

while the envelope

$$y = \frac{x^2}{4}$$

is a singular solution. Thus the Clairaut equation provides a simple example in which a differential equation contains both a one-parameter family of ordinary solutions and an additional solution obtained as their envelope.

Example 2.30 (Envelope of translated parabolas)**.** Consider the family

$$y = (x - C)^2.$$

Let

$$\Psi(x, y, C) = y - (x - C)^2.$$

The envelope is determined by

$$\Psi = 0, \qquad \Psi_C = 0.$$

Since

$$\Psi_C = 2(x - C),$$

we get

$$x = C.$$

Substituting into $\Psi = 0$ gives

$$y = 0.$$

Therefore the envelope is the x-axis:

$$y = 0.$$

Each parabola is tangent to the x-axis at $(C, 0)$.

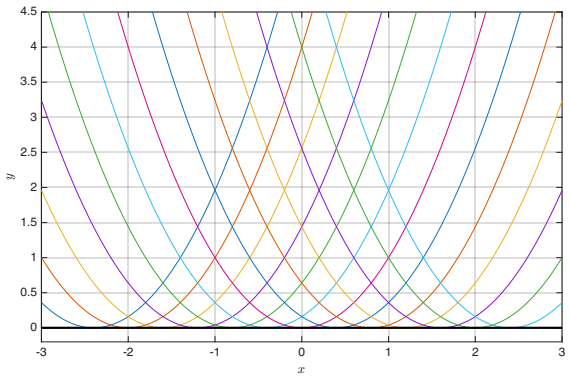


Figure 2.5: The family of parabolas $y = (x - C)^2$ and its envelope $y = 0$. Each parabola is tangent to the x-axis at $(C, 0)$.

Example 2.31 (The cycloid and its evolute)**.** Consider the cycloid

$$x = a(\theta - \sin\theta), \qquad y = a(1 - \cos\theta), \qquad 0 < \theta < 2\pi,$$

where $a > 0$. We want to find the envelope of the family of normal lines to this curve. This envelope is called the evolute of the cycloid.

The tangent vector is

$$(x'(\theta), y'(\theta)) = \big(a(1 - \cos\theta),\, a\sin\theta\big).$$

Hence a normal direction is

$$\big(-\sin\theta,\, 1 - \cos\theta\big).$$

Therefore the normal line at the point corresponding to θ is

$$F(X, Y, \theta) = 0,$$

where

$$F(X, Y, \theta) = \big(X - a(\theta - \sin\theta)\big)(1 - \cos\theta) + \big(Y - a(1 - \cos\theta)\big)\sin\theta.$$

The evolute is the envelope of this one-parameter family of normal lines. Thus it is obtained by solving

$$F(X, Y, \theta) = 0, \qquad F_\theta(X, Y, \theta) = 0.$$

A direct calculation gives

$$X = a(\theta + \sin\theta), \qquad Y = -a(1 - \cos\theta).$$

Thus the evolute of the cycloid is

$$X = a(\theta + \sin\theta), \qquad Y = a(\cos\theta - 1).$$

This curve is again a cycloid, translated and reflected in the plane.

This example shows that an evolute is an envelope: it is the envelope of the family of normal lines. The cycloid is special because its evolute is again a cycloid. This geometric fact is the basis of the classical construction of the cycloidal pendulum.

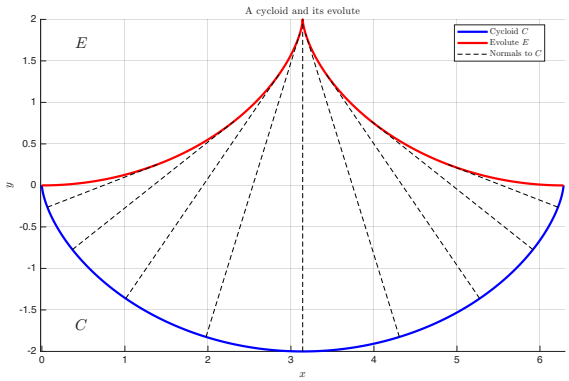


Figure 2.6: A cycloid and its evolute. The evolute (red curve) is the envelope of the normal lines to the original cycloid (blue curve). The evolute is again a cycloid, translated and reflected in the plane.

Remark 2.32. A family of curves need not possess an envelope. For example, confocal ellipses and confocal hyperbolas form orthogonal families, but they do not have envelopes in the above sense.

Exercise 2.8 (Envelopes of curve families)**.** For each of the following one-parameter families, determine whether a regular envelope exists. If it does, find it.

$$y = Cx + \frac{1}{C}, \qquad C \neq 0,$$

$$y = (x - C)^2 + C,$$

and

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

Interpret any degenerate solution of the envelope equations.

2.4 Existence and Uniqueness: A Preview

In the previous sections, we studied first-order equations from several points of view. We solved some equations explicitly, represented others by first integrals, and interpreted solution curves geometrically as families of plane curves. We now turn to a more basic question.

Given an initial value problem

$$y' = f(t, y), \qquad y(t_0) = y_0,$$

does a solution exist? If it exists, is it uniquely determined by the initial condition?

These questions are essential. A differential equation is meant to describe an evolution law. If no solution exists, the law is not realized. If two different solutions start from the same initial state, then the law does not determine a unique future.

The complete theory of existence and uniqueness will be developed later for systems of differential equations. Here we give a preview in the scalar case. The main ideas are local existence and uniqueness, maximal intervals of existence, non-crossing of solution curves, and the geometric meaning of nonuniqueness.

2.4.1 The local existence and uniqueness theorem

We begin with a standard local theorem. The version stated here is not the most general one, but it is simple and sufficient for most equations encountered in this chapter.

Theorem 2.33 (Local existence and uniqueness)**.** *Suppose that $f(t, y)$ and $f_y(t, y)$ are continuous in a neighborhood of (t_0, y_0). Then the initial value problem*

$$y' = f(t, y), \qquad y(t_0) = y_0,$$

has a unique solution on some interval

$$(t_0 - \varepsilon, t_0 + \varepsilon)$$

for some $\varepsilon > 0$.

We shall not prove this theorem here. Its meaning, however, is important. If f and f_y are continuous near the initial point, then the equation has one and only one local solution passing through that point. Thus, under this regularity assumption, the initial condition determines the solution at least for a short time.

The conclusion is local. The theorem does not say that the solution exists for all time. It only asserts existence and uniqueness on some possibly small interval around t_0.

The theorem is often used in this way. One checks that f and f_y are continuous near the initial point, and then one knows that the initial value problem has a unique local solution.

2.4.2 Maximal intervals of existence

A local solution can often be extended. Suppose a solution has been found on an interval. If the solution remains finite and the differential equation is still defined near the endpoint, then the local theorem can be applied again to continue the solution.

Repeating this process gives the largest interval on which the solution is defined. This interval is called the *maximal interval of existence*.

The need for maximal intervals can already be seen in a simple separable equation.

Example 2.34 (Finite-time blow-up)**.** Consider

$$y' = y^2, \qquad y(0) = y_0.$$

If $y_0 = 0$, then

$$y(t) \equiv 0$$

is the solution. Suppose $y_0 \neq 0$. Separating variables gives

$$\frac{dy}{y^2} = dt.$$

Thus

$$-\frac{1}{y} = t + C.$$

Using $y(0) = y_0$, we obtain

$$C = -\frac{1}{y_0}.$$

Therefore

$$-\frac{1}{y} = t - \frac{1}{y_0},$$

and hence

$$y(t) = \frac{y_0}{1 - y_0 t}.$$

If $y_0 > 0$, the denominator vanishes at

$$t = \frac{1}{y_0}.$$

Thus

$$\lim_{t \to (1/y_0)^-} y(t) = +\infty.$$

The solution cannot be continued beyond this time as a finite-valued solution. Its maximal interval of existence is

$$\left(-\infty, \frac{1}{y_0}\right).$$

If $y_0 < 0$, then $1/y_0 < 0$, and the maximal interval of existence is

$$\left(\frac{1}{y_0}, \infty\right).$$

This example shows that local existence does not imply global existence. Even when the right-hand side is smooth everywhere, a solution may become unbounded in finite time.

The general principle is the following. A solution can fail to extend beyond a finite endpoint only if one of two things happens:

1. the solution becomes unbounded;

2. the graph $(t, y(t))$ reaches the boundary of the region where the differential equation is defined.

For equations defined on all of $\mathbb{R}^2$, the second obstruction is absent. In that case, boundedness on finite time intervals is the main mechanism for continuing a solution.

Principle 2.1 (Continuation by boundedness)**.** *Suppose $f(t, y)$ is defined and sufficiently regular on all of $\mathbb{R}^2$, and let $y : (\alpha, \beta) \to \mathbb{R}$ be a maximal solution. If $\beta < \infty$ and $y(t)$ remains bounded as*

$t \to \beta^-$, then the solution can be continued beyond β. Therefore a finite right endpoint can occur only if

$$|y(t)| \to \infty \qquad \text{as } t \to \beta^-.$$

The analogous statement holds at a finite left endpoint.

We state this as a principle rather than a theorem because the precise form will be proved later for systems. For now, the key idea is simple: a bounded solution remains in a compact part of the (t, y)-plane on each finite time interval. Since the equation is regular there, the local existence theorem can be applied again at the endpoint.

2.4.3 Non-crossing and barriers

Uniqueness has a direct geometric consequence. If uniqueness holds, two different solution curves cannot cross in the (t, y)-plane.

Indeed, suppose two solution curves pass through the same point

$$(t_0, y_0).$$

Then both solve the same initial value problem

$$y' = f(t, y), \qquad y(t_0) = y_0.$$

By uniqueness, they must coincide near $t = t_0$. Therefore two distinct solution curves cannot cross.

This non-crossing property is especially useful for autonomous equations

$$y' = f(y).$$

If $y = \bar{y}$ satisfies

$$f(\bar{y}) = 0,$$

then

$$y(t) \equiv \bar{y}$$

is an equilibrium solution. Under uniqueness, no other solution can cross this horizontal solution curve. Equilibrium solutions therefore act as barriers.

Example 2.35 (The logistic equation)**.** Consider

$$y' = ry\left(1 - \frac{y}{K}\right), \qquad r > 0, \quad K > 0.$$

The right-hand side is smooth, so uniqueness holds. The equilibrium solutions are

$$y = 0, \qquad y = K.$$

These two horizontal solution curves divide the (t, y)-plane into three regions:

$$y < 0, \qquad 0 < y < K, \qquad y > K.$$

Since solutions cannot cross equilibrium curves, if

$$0 < y(0) < K,$$

then

$$0 < y(t) < K$$

for all t in the interval of existence. Thus the interval $(0, K)$ is invariant.

Moreover, in this interval,

$$ry\left(1 - \frac{y}{K}\right) > 0.$$

Hence the solution is increasing. Since it is also bounded above by K, it cannot blow up in finite positive time. Therefore the solution exists for all positive time.

This conclusion is qualitative. It follows from uniqueness, barriers, and boundedness. We do not need the explicit solution formula.

The same reasoning applies to many scalar autonomous equations. Equilibria divide the line into intervals, and uniqueness prevents a solution from moving from one interval to another by crossing an equilibrium.

2.4.4 Nonuniqueness and envelopes

The uniqueness theorem requires regularity. If the regularity condition fails, then an initial value problem may have more than one solution. The following example shows this in a way that connects directly with the envelope construction of the previous section.

Example 2.36 (A nonunique initial value problem)**.** Consider

$$y' = 2\sqrt{y}, \qquad y(0) = 0,$$

where the equation is considered for $y \geq 0$.

Solution. The constant function

$$y(t) \equiv 0$$

is a solution.

There are also nontrivial solutions. For $y > 0$, separation of variables gives

$$\frac{dy}{2\sqrt{y}} = dt.$$

Hence

$$\sqrt{y} = t + C.$$

Thus

$$y = (t + C)^2.$$

For positive time, the solution may leave the origin along

$$y = t^2.$$

More generally, for any $a \geq 0$, define

$$y_a(t) = \begin{cases} 0, & t \leq a, \\ (t-a)^2, & t \geq a. \end{cases}$$

Then y_a is continuously differentiable. For $t < a$,

$$y_a'(t) = 0 = 2\sqrt{y_a(t)}.$$

For $t > a$,

$$y_a'(t) = 2(t-a), \qquad 2\sqrt{y_a(t)} = 2(t-a).$$

At $t = a$, both the left and right derivatives are zero, so the equation is also satisfied there. Hence y_a is a solution.

If $a \geq 0$, then

$$y_a(0) = 0.$$

Thus the same initial value problem has infinitely many solutions.

□

The source of the nonuniqueness is the failure of the Lipschitz condition. Here

$$f(y) = 2\sqrt{y}.$$

Although f is continuous at $y = 0$, its derivative

$$f'(y) = \frac{1}{\sqrt{y}}$$

becomes unbounded as $y \to 0^+$. Therefore the local existence and uniqueness theorem does not apply.

There is also a geometric explanation. As shown in Section 2.3.3, the family of translated parabolas

$$y = (t-a)^2$$

has the t-axis

$$y = 0$$

as its envelope. Here this envelope is also the equilibrium solution of the differential equation.

The nonunique solutions can now be interpreted geometrically. A solution may remain on the envelope $y = 0$ for an arbitrary amount of time and then leave along one of the parabolas

$$y = (t-a)^2.$$

The point $(a, 0)$ is the tangency point between the parabola and the envelope. Since the tangency is smooth, the delayed solution remains continuously differentiable.

The same family also satisfies the implicit differential equation

$$(y')^2 = 4y.$$

This equation has two branches:

$$y' = 2\sqrt{y}, \qquad y' = -2\sqrt{y}.$$

The branch $y' = 2\sqrt{y}$ corresponds to the right halves of the parabolas, while the branch $y' = -2\sqrt{y}$ corresponds to the left halves. The failure of uniqueness comes from the possibility of staying on the envelope before entering one of the branches.

This example illustrates the central point of the section. Existence and uniqueness are not automatic consequences of writing down a differential equation. Under sufficient regularity, an initial condition determines a unique local solution. Without such regularity, solution curves may merge, separate, or follow an envelope before leaving it.

Exercise 2.9. Suppose $y(t) > 0$ satisfies

$$y' \geq y^2, \qquad y(0) = y_0 > 0.$$

Show that $y(t)$ cannot be extended as a finite-valued solution beyond

$$t = \frac{1}{y_0}.$$

Homework Problems

The following problems combine the analytic and geometric viewpoints developed in this chapter. In each problem, justify both the calculation and the geometric interpretation.

Homework 2.1 (A nonautonomous Riccati equation)**.** Consider

$$y' = (y - x)^2 + 1, \qquad y(0) = y_0.$$

Find the solution for every $y_0 \in \mathbb{R}$ and determine its maximal interval of existence. Describe the finite-time blow-up when it occurs, and identify any solution lost during the reduction.

Hint. The function $y_1(x) = x$ is a particular solution. Use

$$y = x + \frac{1}{u}.$$

Homework 2.2 (A vector field and its first integral)**.** Consider

$$\mathbf{V}(x, y) = \left(x(1 - y^2), y(1 - x^2)\right).$$

Find all equilibria and invariant coordinate axes. On each quadrant, derive a first integral and use its level curves to describe the integral curves of the vector field.

Hint. Write the equation for the integral curves in Pfaffian form and use the integrating factor

$$\mu(x, y) = \frac{1}{xy}.$$

Homework 2.3 (The envelope of waves from a moving source)**.** A source moves along the x-axis with constant speed $v > c$, where c is the propagation speed of the waves. At time T, the wave emitted at time τ has wavefront

$$(x - v\tau)^2 + y^2 = c^2(T - \tau)^2, \qquad 0 \le \tau \le T.$$

Find the envelope of these circular wavefronts and show that it consists of two straight lines meeting at the current source position $(vT, 0)$. Find the angle between the envelope and the direction of motion.

Hint. Apply the envelope conditions

$$F(x, y, \tau) = 0, \qquad F_\tau(x, y, \tau) = 0.$$

Homework 2.4 (Shapes of relaxation toward equilibrium)**.** Let

$$y' = f(y), \qquad y(0) = y_0 < b,$$

where

$$f(y) > 0 \quad \text{for } y_0 \le y < b, \qquad f(b) = 0,$$

and

$$f(y) \sim c(b - y)^p \qquad \text{as } y \to b^-, \qquad c > 0, \quad p > 0.$$

For $0 < p < 1$, $p = 1$, and $p > 1$, determine the leading time profile of $b - y(t)$. Decide whether the solution reaches b in finite time or approaches it as $t \to \infty$, and sketch the three qualitatively different solution shapes.

Hint. Use separation of variables and compare

$$t = \int_{y_0}^{y(t)} \frac{ds}{f(s)}$$

with the corresponding integral for $c(b - s)^p$.

Historical Notes and Further Reading

Historical notes. Many of the solvable first-order equations studied in this chapter arose during the early development of differential calculus. An equation now called the Bernoulli equation appeared in the work of Bernoulli (1695); its reduction to a linear equation is an early example of transforming a nonlinear equation by a change of variables. Riccati (1724) studied a quadratic class of first-order equations that now bears his name. Clairaut (1734) investigated equations not explicitly solved for the derivative, including the form

$$y = xy' + f(y').$$

Its family of straight-line solutions and its singular envelope display particularly clearly the connection between analytic solution methods and plane geometry.

The language of Pfaffian equations takes its name from the work of Pfaff (1815) on first-order differential relations in several variables. A Pfaffian equation records tangent directions without requiring one coordinate to be chosen as the independent variable and therefore provides a natural connection among scalar differential equations, vector fields, first integrals, exact forms, and integrating factors. The cycloid example also has an important physical origin: in his study of pendulum clocks, Huygens (1673) used the geometry of the cycloid and its evolute in the construction of an isochronous pendulum.

The final section of this chapter passes from formal integration to the questions of existence, uniqueness, maximal intervals, and non-crossing of solution curves. These analytic foundations are developed systematically in Chapter 9.

Further reading. For a classical treatment of first-order equations and their methods of integration, including Bernoulli, Riccati, and Clairaut equations, see Ince (1956). For a geometric approach based on vector fields, integral curves, first integrals, and applications, see Arnold (1992). A rigorous modern account of explicitly solvable equations, existence and uniqueness, continuation, and qualitative methods is given by Teschl (2012).

Chapter 3

Second-Order Linear Equations and Linear Oscillators

In the previous chapters, we studied first-order equations from the viewpoints of modeling, solution methods, and geometry. We now turn to second-order differential equations. The new feature is that the evolution is no longer described by a law depending only on the present value of the unknown. For a second-order equation, the acceleration may depend on both y and $\dot{y}$. Thus the natural state consists of two components:

$$(y, \dot{y}).$$

For a mechanical oscillator, these are position and velocity. For an electrical oscillator, they correspond to charge and current.

The most important examples come from mechanical and electrical oscillators. A mass attached to a spring and dashpot, and a series RLC circuit driven by an external voltage, both lead to equations of the form

$$a\ddot{y} + b\dot{y} + cy = f(t), \qquad a \neq 0,$$

where $y(t)$ is the unknown state and $f(t)$ is an external forcing term. The three terms on the left represent three basic mechanisms:

$$a\ddot{y} \quad \text{inertia or storage}, \qquad b\dot{y} \quad \text{dissipation}, \qquad cy \quad \text{restoring effect}.$$

In a mechanical oscillator these correspond to mass, damping, and spring stiffness. In an electrical oscillator they correspond to inductance, resistance, and inverse capacitance.

This chapter has two purposes. The first is to develop the basic solution theory of second-order linear equations with constant coefficients. The second is to interpret these solutions dynamically. The roots of the characteristic equation determine whether the motion grows, decays, oscillates, or returns to equilibrium without oscillation. External forcing gives rise to a transient response, a synchronized response, and, under appropriate conditions, resonance.

Thus the guiding theme of the chapter is

$$\text{linear equation} \longrightarrow \text{characteristic roots} \longrightarrow \text{oscillation, damping, and resonance.}$$

3.1 Physical Origins of Second-Order Linear Equations

Second-order linear equations arise naturally when a system possesses inertia, dissipation, and a restoring mechanism. The most basic examples are mechanical oscillators and electrical circuits. Although their physical meanings are different, their mathematical structures are the same.

3.1.1 The spring–mass–dashpot system

Consider a mass attached to a spring and a dashpot, as shown in Figure 3.1. Let $y(t)$ denote the displacement of the mass from its equilibrium position. The spring pulls the mass back toward equilibrium, the dashpot resists the motion, and an external force $f(t)$ may also act on the mass.

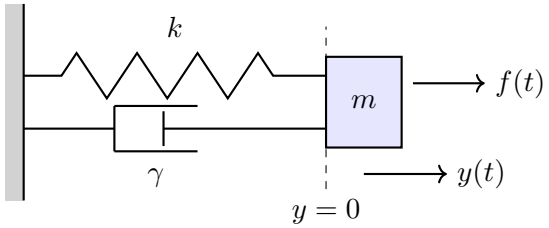


Figure 3.1: A spring–mass–dashpot system. The spring produces the restoring force $-ky$, the dashpot produces the damping force $-\gamma\dot{y}$, and the external force is denoted by $f(t)$.

Newton's law gives

$$m\ddot{y} = F,$$

where $m > 0$ is the mass and F is the total force acting on the mass. The forces are modeled as follows.

- **Inertia.** The term

  $$m\ddot{y}$$

  measures the resistance of the mass to changes in velocity. A larger mass requires a larger force to produce the same acceleration.

- **Restoring force.** By Hooke's law, the spring force is proportional to the displacement and acts in the opposite direction:

  $$F_{\text{spring}} = -ky, \qquad k > 0.$$

  The constant k is the spring stiffness. When $y > 0$, the force is negative; when $y < 0$, the force is positive. In both cases, the spring pulls the mass back toward the equilibrium position.

- **Friction or damping.** The dashpot produces a force proportional to the velocity and opposite to the direction of motion:

  $$F_{\text{damp}} = -\gamma\dot{y}, \qquad \gamma \geq 0.$$

  The constant γ is the damping coefficient. This term dissipates mechanical energy.

- **External forcing.** The function
$$f(t)$$
represents an applied force. It may be periodic, impulsive, or more general.

The force balance is therefore

$$m\ddot{y} = -\gamma\dot{y} - ky + f(t).$$

Equivalently,

$$m\ddot{y} + \gamma\dot{y} + ky = f(t).$$

This is the basic equation of a forced damped oscillator.

3.1.2 The RLC circuit

A series RLC circuit consists of a resistor, an inductor, and a capacitor connected in one loop. Let $Q(t)$ denote the charge stored on one plate of the capacitor. In a lumped circuit model, this single scalar variable describes the electric energy stored in the capacitor. The current in the loop is the rate of change of this charge:

$$I(t) = \dot{Q}(t).$$

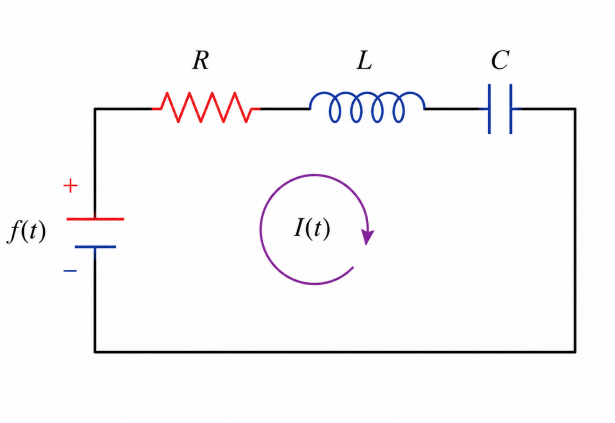


Figure 3.2: A series RLC circuit. The resistor dissipates energy, the capacitor stores electric energy, and the inductor stores magnetic energy.

We first recall two basic circuit laws.

- **Kirchhoff's current law.** At each junction, the total current flowing into the junction equals the total current flowing out. In a series circuit, this means that the same current flows through the resistor, the inductor, and the capacitor.
- **Kirchhoff's voltage law.** The algebraic sum of the oriented voltage drops around a closed loop equals the applied voltage.

Choose the direction of the current I as the positive orientation of the loop. The voltage drops across the three circuit elements are as follows.

- **Resistor.** By Ohm's law, the voltage drop across the resistor is

$$V_R = RI, \qquad R > 0,$$

  where R is the resistance. This term represents dissipation.

- **Capacitor.** The voltage drop across the capacitor is

$$V_C = \frac{Q}{C}, \qquad C > 0,$$

  where C is the capacitance. The capacitor stores electric energy.

- **Inductor.** The voltage drop across the inductor is

$$V_L = L\dot{I}, \qquad L > 0,$$

  where L is the inductance. The inductor resists changes in current and stores magnetic energy.

If the circuit is driven by an applied voltage $\mathcal{E}(t)$, then Kirchhoff's voltage law gives

$$V_L + V_R + V_C = \mathcal{E}(t).$$

Therefore,

$$L\dot{I} + RI + \frac{1}{C}Q = \mathcal{E}(t).$$

Since $I = \dot{Q}$, we obtain

$$L\ddot{Q} + R\dot{Q} + \frac{1}{C}Q = \mathcal{E}(t).$$

This is a second-order linear equation for the charge $Q(t)$.

The correspondence with the mechanical oscillator is

Mechanical oscillator	Electrical circuit
m	L
γ	R
k	$1/C$
y	Q
$\dot{y}$	$I = \dot{Q}$
$f(t)$	$\mathcal{E}(t)$

Thus the series RLC circuit has the same mathematical structure as the spring–mass–dashpot system:

$$m\ddot{y} + \gamma\dot{y} + ky = f(t), \qquad L\ddot{Q} + R\dot{Q} + \frac{1}{C}Q = \mathcal{E}(t).$$

3.1.3 The standard oscillator equation

Motivated by these examples, we study the general second-order linear equation with constant coefficients:

$$a\ddot{y} + b\dot{y} + cy = f(t), \qquad a \neq 0.$$

For the physical oscillators considered above, the coefficients satisfy

$$a > 0, \qquad b \geq 0, \qquad c > 0.$$

Under these sign conditions, the three terms on the left represent the three basic mechanisms:

term	role
$a\ddot{y}$	inertia or storage
$b\dot{y}$	dissipation
cy	restoring effect

The function $f(t)$ is the external forcing term.

The homogeneous equation is

$$a\ddot{y} + b\dot{y} + cy = 0.$$

It describes the free motion of the system. The inhomogeneous equation describes the response to the input $f(t)$.

For an initial value problem, two initial conditions are prescribed:

$$y(0) = y_0, \qquad \dot{y}(0) = v_0.$$

In this chapter we usually take the initial time to be 0. This causes no loss of generality for constant-coefficient equations, since one may replace t by $t - t_0$ if the initial time is t_0.

It is useful to introduce the differential operator

$$D = \frac{d}{dt}.$$

Then the equation can be written as

$$P(D)y = f(t), \qquad P(D) = aD^2 + bD + c.$$

The corresponding polynomial

$$P(\lambda) = a\lambda^2 + b\lambda + c$$

is called the characteristic polynomial. Its roots determine the fundamental behavior of the homogeneous equation. In the following sections, we shall see how these roots lead to exponential decay, growth, oscillation, damping, and resonance.

Exercise 3.1 (Two capacitors in series)**.** Consider a series circuit consisting of an inductor L, a resistor R, and two capacitors with capacitances C_1 and C_2. The circuit is driven by an external voltage source $\mathcal{E}(t)$.

Assume that both capacitors are initially uncharged. Let $Q(t)$ denote the charge that has passed through the series branch since $t = 0$, and let

$$I(t) = \dot{Q}(t)$$

be the current. The two capacitors then carry charges of equal magnitude $Q(t)$.

1. Explain why the same current passes through all components of the circuit.
2. The voltage drops across the two capacitors are

$$V_{C_1} = \frac{Q}{C_1}, \qquad V_{C_2} = \frac{Q}{C_2}.$$

   Use Kirchhoff's voltage law to derive the equation

$$L\dot{I} + RI + \frac{Q}{C_1} + \frac{Q}{C_2} = \mathcal{E}(t).$$

3. Rewrite the equation as a second-order equation for $Q(t)$.
4. Show that the two capacitors in series are equivalent to one capacitor C_{eq} satisfying

$$\frac{1}{C_{\text{eq}}} = \frac{1}{C_1} + \frac{1}{C_2}.$$

3.2 Homogeneous Motion and Characteristic Roots

We first study the homogeneous equation

$$a\ddot{y} + b\dot{y} + cy = 0, \qquad a \neq 0.$$

This equation describes the free motion of the system, without external forcing. For the mechanical oscillator, this means that no external force is applied. For the RLC circuit, this means that no external voltage source is present.

The central idea is that the homogeneous equation is governed by an algebraic equation. Its roots determine the elementary modes of the motion.

3.2.1 Exponential modes and the characteristic equation

We look for solutions of the form

$$y(t) = e^{\lambda t}.$$

Then

$$\dot{y} = \lambda e^{\lambda t}, \qquad \ddot{y} = \lambda^2 e^{\lambda t}.$$

Substituting into the homogeneous equation gives

$$a\lambda^2 e^{\lambda t} + b\lambda e^{\lambda t} + ce^{\lambda t} = 0.$$

Since $e^{\lambda t} \neq 0$, we obtain

$$a\lambda^2 + b\lambda + c = 0.$$

This algebraic equation is called the *characteristic equation*, and

$$P(\lambda) = a\lambda^2 + b\lambda + c$$

is called the *characteristic polynomial*.

Thus the differential equation

$$a\ddot{y} + b\dot{y} + cy = 0$$

is reduced to the algebraic equation

$$P(\lambda) = 0.$$

Each root of P gives an exponential mode of the homogeneous equation.

3.2.2 Linear independence and the Wronskian

For a homogeneous linear equation, the set of solutions is a vector space. Indeed, if y_1 and y_2 are solutions, then every linear combination

$$C_1 y_1 + C_2 y_2$$

is also a solution.

For every pair of initial values

$$y(0) = y_0, \qquad \dot{y}(0) = v_0,$$

the existence and uniqueness theorem gives exactly one solution. Hence the linear map

$$y \longmapsto \big(y(0), \dot{y}(0)\big)$$

from the solution space to $\mathbb{R}^2$ is one-to-one and onto. Therefore the solution space is two-dimensional. To write the general solution, we need two linearly independent solutions.

For two solutions y_1 and y_2, their Wronskian is defined by

$$W[y_1, y_2](t) = \begin{vmatrix} y_1(t) & y_2(t) \\ \dot{y}_1(t) & \dot{y}_2(t) \end{vmatrix} = y_1(t)\dot{y}_2(t) - \dot{y}_1(t)y_2(t).$$

It is the determinant of the two state vectors

$$(y_1(t), \dot{y}_1(t)), \qquad (y_2(t), \dot{y}_2(t)).$$

Hence a nonzero Wronskian means that the two solutions carry independent initial data.

Proposition 3.1 (Evolution of the Wronskian)**.** *Let y_1 and y_2 be two solutions of*

$$a\ddot{y} + b\dot{y} + cy = 0, \qquad a \neq 0.$$

Let

$$W(t) = W[y_1, y_2](t).$$

Then

$$\dot{W}(t) = -\frac{b}{a} W(t).$$

Consequently,

$$W(t) = W(0) \exp\left(-(b/a)t\right).$$

In particular, the Wronskian is either identically zero or never zero.

Proof. By definition,

$$W = y_1\dot{y}_2 - \dot{y}_1 y_2.$$

Differentiating gives

$$\begin{aligned}\dot{W} &= \dot{y}_1\dot{y}_2 + y_1\ddot{y}_2 - \ddot{y}_1 y_2 - \dot{y}_1\dot{y}_2 \\ &= y_1\ddot{y}_2 - \ddot{y}_1 y_2.\end{aligned}$$

Since y_1 and y_2 both solve the equation,

$$\ddot{y}_j = -\frac{b}{a}\dot{y}_j - \frac{c}{a}y_j, \qquad j = 1, 2.$$

Therefore

$$\begin{aligned}\dot{W} &= y_1\left(-\frac{b}{a}\dot{y}_2 - \frac{c}{a}y_2\right) - \left(-\frac{b}{a}\dot{y}_1 - \frac{c}{a}y_1\right)y_2 \\ &= -\frac{b}{a}\left(y_1\dot{y}_2 - \dot{y}_1 y_2\right) \\ &= -\frac{b}{a}W.\end{aligned}$$

Solving this first-order equation for W gives

$$W(t) = W(0)\exp\left(-\frac{b}{a}t\right).$$

□

The proposition shows that it is enough to check the Wronskian at one time. If

$$W[y_1, y_2](0) \neq 0,$$

then y_1 and y_2 are linearly independent and form a fundamental set of solutions. Hence the general solution is

$$y(t) = C_1 y_1(t) + C_2 y_2(t).$$

For example, suppose the characteristic equation has two distinct real roots λ_1 and λ_2. Then

$$y_1(t) = e^{\lambda_1 t}, \qquad y_2(t) = e^{\lambda_2 t}$$

are two solutions. Their Wronskian is

$$\begin{aligned}W[y_1, y_2](t) &= \begin{vmatrix} e^{\lambda_1 t} & e^{\lambda_2 t} \\ \lambda_1 e^{\lambda_1 t} & \lambda_2 e^{\lambda_2 t}\end{vmatrix} \\ &= (\lambda_2 - \lambda_1)e^{(\lambda_1+\lambda_2)t}.\end{aligned}$$

Since $\lambda_1 \neq \lambda_2$, this Wronskian is nonzero for all t. Thus

$$e^{\lambda_1 t}, \qquad e^{\lambda_2 t}$$

form a fundamental set of solutions.

3.2.3 Classification of characteristic roots

Assume that $a, b, c \in \mathbb{R}$ with $a \neq 0$. The form of the solution depends on the roots of

$$P(\lambda) = a\lambda^2 + b\lambda + c.$$

Let

$$\Delta = b^2 - 4ac$$

be the discriminant.

Distinct real roots. If $\Delta > 0$, then the characteristic equation has two distinct real roots λ_1 and λ_2. The two exponential modes

$$e^{\lambda_1 t}, \qquad e^{\lambda_2 t}$$

are linearly independent, and the general solution is

$$y(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}.$$

Repeated real root. If $\Delta = 0$, then the characteristic equation has a repeated real root λ. Thus

$$P(\lambda) = 0, \qquad P'(\lambda) = 0.$$

The function $e^{\lambda t}$ is one solution, because

$$P(D)e^{\lambda t} = P(\lambda)e^{\lambda t} = 0.$$

To find a second solution, we use the identity

$$P(D)e^{\lambda t} = P(\lambda)e^{\lambda t}.$$

Differentiate both sides with respect to λ. Since D acts on t and not on λ, we obtain

$$P(D)\left(te^{\lambda t}\right) = P'(\lambda)e^{\lambda t} + P(\lambda)te^{\lambda t}.$$

For a repeated root, $P(\lambda) = P'(\lambda) = 0$. Hence

$$P(D)\left(te^{\lambda t}\right) = 0.$$

Therefore $te^{\lambda t}$ is also a solution.

The two functions

$$e^{\lambda t}, \qquad te^{\lambda t}$$

are linearly independent, since their Wronskian is

$$W(t) = \begin{vmatrix} e^{\lambda t} & te^{\lambda t} \\ \lambda e^{\lambda t} & (1+\lambda t)e^{\lambda t} \end{vmatrix} = e^{2\lambda t} \neq 0.$$

Thus the general solution is

$$y(t) = (C_1 + C_2 t)e^{\lambda t}.$$

This calculation also has a useful interpretation. The solution $te^{\lambda t}$ appears when two exponential modes coalesce. Indeed,

$$\frac{\partial}{\partial \lambda} e^{\lambda t} = te^{\lambda t}.$$

Thus $te^{\lambda t}$ is the derivative of the exponential mode with respect to the characteristic root.

Complex conjugate roots. If $\Delta < 0$, then the characteristic equation has complex conjugate roots

$$\lambda = \alpha + i\omega, \qquad \bar{\lambda} = \alpha - i\omega, \qquad \omega > 0.$$

The corresponding complex exponential solutions are

$$e^{(\alpha+i\omega)t}, \qquad e^{(\alpha-i\omega)t}.$$

Using Euler's formula,

$$e^{(\alpha+i\omega)t} = e^{\alpha t}\big(\cos\omega t + i\sin\omega t\big).$$

Taking real and imaginary parts gives two real-valued solutions:

$$e^{\alpha t}\cos\omega t, \qquad e^{\alpha t}\sin\omega t.$$

Therefore the general real-valued solution is

$$y(t) = e^{\alpha t}\big(C_1\cos\omega t + C_2\sin\omega t\big).$$

The number ω is the angular frequency of the oscillation. The factor $e^{\alpha t}$ controls the amplitude. If $\alpha < 0$, the oscillation is damped. If $\alpha = 0$, the oscillation has constant amplitude. If $\alpha > 0$, the oscillation grows exponentially.

Summary. The three cases are summarized as follows:

case	roots	general solution
$\Delta > 0$	$\lambda_1, \lambda_2 \in \mathbb{R},\ \lambda_1 \neq \lambda_2$	$C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}$
$\Delta = 0$	$\lambda \in \mathbb{R}$ repeated	$(C_1 + C_2 t)e^{\lambda t}$
$\Delta < 0$	$\alpha \pm i\omega,\ \omega > 0$	$e^{\alpha t}\big(C_1\cos\omega t + C_2\sin\omega t\big)$

Thus the characteristic roots give the natural modes of the system. In the next section, we interpret these modes as undamped, damped, critically damped, and overdamped motion for linear oscillators.

Exercise 3.2. For each of the following equations, determine the roots of the characteristic equation and describe the qualitative behavior of a typical solution without solving for the constants:

$$\ddot{y} - 2\dot{y} - 3y = 0, \qquad \ddot{y} + 4\dot{y} + 4y = 0, \qquad \ddot{y} + 2\dot{y} + 5y = 0.$$

Exercise 3.3 (Recovering an equation from its solutions)**.** Find a second-order homogeneous linear equation with constant coefficients whose general solution is, respectively,

$$y(t) = (C_1 + C_2 t)e^{-2t}$$

and

$$y(t) = e^{3t}\big(C_1\cos t + C_2\sin t\big).$$

Exercise 3.4. Let $\lambda = \alpha + i\omega$, where $\omega > 0$.

1. Compute
$$W(e^{\lambda t}, e^{\bar{\lambda} t}) \qquad \text{and} \qquad W(e^{\alpha t}\cos\omega t, e^{\alpha t}\sin\omega t).$$
2. Explain why the nonvanishing of these Wronskians gives two linearly independent solutions over $\mathbb{C}$ and $\mathbb{R}$, respectively.

3.3 Linear Oscillators

We now interpret the characteristic roots for the standard oscillator

$$m\ddot{y} + \gamma\dot{y} + ky = 0, \qquad m > 0, \quad \gamma \geq 0, \quad k > 0.$$

This equation describes free motion: there is no external force. The parameters have the following meanings:

- m is the mass, or inertia;
- γ is the damping coefficient;
- k is the spring stiffness, or restoring coefficient.

The characteristic equation is
$$m\lambda^2 + \gamma\lambda + k = 0.$$
Its roots are
$$\lambda = \frac{-\gamma \pm \sqrt{\gamma^2 - 4mk}}{2m}.$$
Thus the behavior of the oscillator is determined by the discriminant
$$\Delta = \gamma^2 - 4mk.$$

It is useful to introduce the natural frequency
$$\omega_0 = \sqrt{\frac{k}{m}}.$$
In the absence of damping, ω_0 is the angular frequency of oscillation. The equation can then be written as
$$\ddot{y} + 2\beta\dot{y} + \omega_0^2 y = 0, \qquad \beta = \frac{\gamma}{2m}.$$
Here β measures the strength of damping. The characteristic equation becomes
$$\lambda^2 + 2\beta\lambda + \omega_0^2 = 0.$$
The roots are
$$\lambda = -\beta \pm \sqrt{\beta^2 - \omega_0^2}.$$

3.3.1 Undamped oscillation

First consider the case without damping:

$$\gamma = 0.$$

Then the equation is

$$m\ddot{y} + ky = 0,$$

or

$$\ddot{y} + \omega_0^2 y = 0.$$

The characteristic equation is

$$\lambda^2 + \omega_0^2 = 0,$$

with roots

$$\lambda = \pm i\omega_0.$$

Therefore the general solution is

$$y(t) = C_1 \cos \omega_0 t + C_2 \sin \omega_0 t.$$

Equivalently, it may be written in amplitude–phase form

$$y(t) = A \cos(\omega_0 t - \phi),$$

where $A \geq 0$ is the amplitude and ϕ is the phase.

The motion is periodic. The angular frequency is ω_0, and the period is

$$T = \frac{2\pi}{\omega_0}.$$

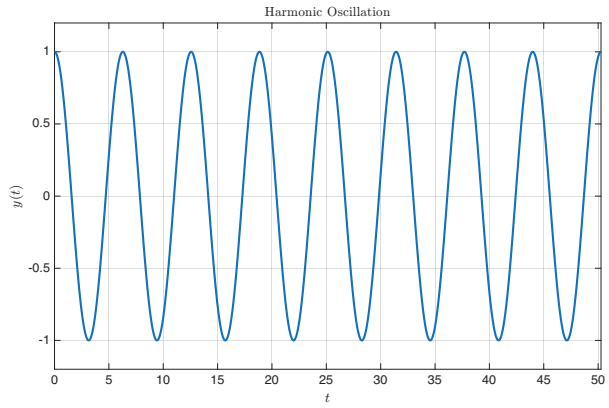


Figure 3.3: Undamped harmonic oscillation. The solution oscillates with constant amplitude and angular frequency ω_0.

3.3.2 Damped oscillation

Now assume that

$$\gamma > 0.$$

The equation is

$$m\ddot{y} + \gamma\dot{y} + ky = 0.$$

The damping term $\gamma\dot{y}$ removes energy from the system. The nature of the return to equilibrium depends on the sign of

$$\Delta = \gamma^2 - 4mk.$$

Underdamped motion. If

$$\gamma^2 < 4mk,$$

then the characteristic roots are complex:

$$\lambda = -\frac{\gamma}{2m} \pm i\sqrt{\frac{k}{m} - \frac{\gamma^2}{4m^2}}.$$

Set

$$\alpha = \frac{\gamma}{2m}, \qquad \omega_d = \sqrt{\frac{k}{m} - \frac{\gamma^2}{4m^2}}.$$

Then

$$\lambda = -\alpha \pm i\omega_d,$$

and the general solution is

$$y(t) = e^{-\alpha t}\big(C_1 \cos\omega_d t + C_2 \sin\omega_d t\big).$$

This is an oscillation whose amplitude decays exponentially. The number ω_d is the damped angular frequency. Since

$$\omega_d < \omega_0,$$

damping lowers the frequency of oscillation.

Critical damping. If

$$\gamma^2 = 4mk,$$

then the characteristic equation has a repeated real root

$$\lambda = -\frac{\gamma}{2m}.$$

The general solution is

$$y(t) = (C_1 + C_2 t)e^{-\frac{\gamma}{2m}t}.$$

This is called critically damped motion. It is the boundary between oscillatory and non-oscillatory decay.

Overdamped motion. If

$$\gamma^2 > 4mk,$$

then the characteristic equation has two distinct real roots

$$\lambda_1 = \frac{-\gamma - \sqrt{\gamma^2 - 4mk}}{2m}, \qquad \lambda_2 = \frac{-\gamma + \sqrt{\gamma^2 - 4mk}}{2m}.$$

Both roots are negative, and the general solution is

$$y(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}.$$

The solution decays to equilibrium without oscillation.

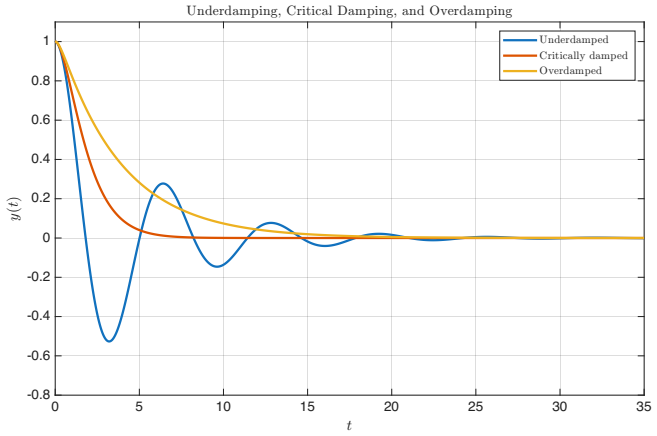


Figure 3.4: Comparison of underdamped, critically damped, and overdamped motion for a damped oscillator.

The three damping regimes are summarized as follows:

condition	roots	motion
$\gamma^2 < 4mk$	$-\alpha \pm i\omega_d$	decaying oscillation
$\gamma^2 = 4mk$	$-\dfrac{\gamma}{2m}$ repeated	critical damping
$\gamma^2 > 4mk$	$\lambda_1, \lambda_2 < 0$	non-oscillatory decay

Thus damping changes the character of the motion. Weak damping allows oscillation with decreasing amplitude. Critical damping gives the borderline case. Strong damping removes oscillation and produces a sum of decaying exponentials.

Exercise 3.5. 1. Solve the initial value problem

$$\ddot{y} - \dot{y} - 2y = 0, \qquad y(0) = \alpha, \qquad \dot{y}(0) = 2.$$

2. Determine the value of α for which

$$y(t) \to 0 \qquad (t \to \infty).$$

Exercise 3.6. Consider the differential equation

$$\ddot{y} - (2\alpha - 1)\dot{y} + \alpha(\alpha - 1)y = 0.$$

1. Determine all values of α for which every solution satisfies

$$y(t) \to 0 \qquad (t \to \infty).$$

2. Determine all values of α for which every nontrivial solution is unbounded as

$$t \to \infty.$$

3.4 Forced Motion and Duhamel's Principle

We now return to the inhomogeneous equation

$$a\ddot{y} + b\dot{y} + cy = f(t), \qquad a \neq 0.$$

The function $f(t)$ represents an external input. In a mechanical oscillator, it is an applied force. In an RLC circuit, it is an applied voltage.

The solution has two parts. One part comes from the initial data and the natural modes of the homogeneous equation. The other part comes from the forcing term $f(t)$. Thus we write

$$y(t) = y_h(t) + y_p(t),$$

where

$$a\ddot{y}_h + b\dot{y}_h + cy_h = 0$$

and

$$a\ddot{y}_p + b\dot{y}_p + cy_p = f(t).$$

For an initial value problem, we choose y_h to satisfy the prescribed initial data and choose y_p to have zero initial data. With this convention, y_h is called the *free response*, and y_p is called the *forced response*.

3.4.1 General solution of the inhomogeneous equation

Let y_1 and y_2 be a fundamental pair of solutions of the homogeneous equation

$$a\ddot{y} + b\dot{y} + cy = 0.$$

Then the homogeneous solution is

$$y_h(t) = C_1 y_1(t) + C_2 y_2(t).$$

If $y_p(t)$ is any one particular solution of the inhomogeneous equation, then the general solution is

$$y(t) = C_1 y_1(t) + C_2 y_2(t) + y_p(t).$$

Indeed, if y and y_p both solve the inhomogeneous equation, then their difference satisfies

$$a(\ddot{y} - \ddot{y}_p) + b(\dot{y} - \dot{y}_p) + c(y - y_p) = 0.$$

Thus $y - y_p$ is a homogeneous solution. This is why one particular solution, together with the general homogeneous solution, gives all solutions.

For an initial value problem,

$$y(0) = y_0, \qquad \dot{y}(0) = v_0,$$

we often choose the homogeneous part y_h to satisfy the initial conditions:

$$y_h(0) = y_0, \qquad \dot{y}_h(0) = v_0.$$

Then the forced part y_p is chosen to have zero initial data:

$$y_p(0) = 0, \qquad \dot{y}_p(0) = 0.$$

With this convention, y_h describes the free response due to the initial state, while y_p describes the response caused only by the forcing term $f(t)$.

3.4.2 Variation of constants

We now derive a particular solution. The idea is to replace the constants in the homogeneous solution by functions of time:

$$y_p(t) = u_1(t)y_1(t) + u_2(t)y_2(t).$$

Then

$$\dot{y}_p = \dot{u}_1 y_1 + u_1 \dot{y}_1 + \dot{u}_2 y_2 + u_2 \dot{y}_2.$$

To avoid second derivatives of u_1, u_2 in the next step, we impose the auxiliary condition

$$\dot{u}_1 y_1 + \dot{u}_2 y_2 = 0.$$

Then

$$\dot{y}_p = u_1 \dot{y}_1 + u_2 \dot{y}_2.$$

Differentiating once more gives

$$\ddot{y}_p = \dot{u}_1 \dot{y}_1 + u_1 \ddot{y}_1 + \dot{u}_2 \dot{y}_2 + u_2 \ddot{y}_2.$$

Substituting into

$$a\ddot{y}_p + b\dot{y}_p + c y_p = f(t),$$

and using that y_1, y_2 solve the homogeneous equation, we obtain

$$a\big(\dot{u}_1 \dot{y}_1 + \dot{u}_2 \dot{y}_2\big) = f(t).$$

Therefore $\dot{u}_1$ and $\dot{u}_2$ satisfy the system

$$\begin{cases} \dot{u}_1 y_1 + \dot{u}_2 y_2 = 0, \\ \dot{u}_1 \dot{y}_1 + \dot{u}_2 \dot{y}_2 = \dfrac{f(t)}{a}. \end{cases}$$

Using the Wronskian

$$W(t) = y_1(t)\dot{y}_2(t) - \dot{y}_1(t)y_2(t),$$

we solve for

$$\dot{u}_1(t) = -\frac{y_2(t)f(t)}{aW(t)}, \qquad \dot{u}_2(t) = \frac{y_1(t)f(t)}{aW(t)}.$$

Thus one particular solution is

$$y_p(t) = -y_1(t)\int_0^t \frac{y_2(s)f(s)}{aW(s)}\,ds + y_2(t)\int_0^t \frac{y_1(s)f(s)}{aW(s)}\,ds.$$

It satisfies

$$y_p(0) = 0, \qquad \dot{y}_p(0) = 0.$$

3.4.3 Duhamel's formula

The preceding formula can be written in a more symmetric form. Define

$$G(t,s) = \frac{y_1(s)y_2(t) - y_2(s)y_1(t)}{aW(s)}, \qquad t \geq s.$$

Then

$$y_p(t) = \int_0^t G(t,s)f(s)\,ds.$$

Therefore the solution of the forced equation is

$$y(t) = y_h(t) + \int_0^t G(t,s)f(s)\,ds.$$

This is Duhamel's principle for second-order linear equations.

The meaning is simple: the response at time t is obtained by summing the effects of the forcing at all earlier times s. The kernel $G(t,s)$ measures how an impulse applied at time s influences the solution at time t.

3.4.4 Impulse Response and Green's Function

For the constant-coefficient equation

$$a\ddot{y} + b\dot{y} + cy = f(t), \qquad a \neq 0,$$

the kernel in Duhamel's formula depends only on the time difference. Thus, for $t \geq s$, we may write

$$G(t,s) = g(t-s),$$

where g is called the *impulse response*. The zero-initial-data response then takes the convolution form

$$y_p(t) = \int_0^t g(t-s)f(s)\,ds.$$

We now determine the equation and initial conditions satisfied by g. Differentiating once gives

$$\dot{y}_p(t) = g(0)f(t) + \int_0^t \dot{g}(t-s)f(s)\,ds.$$

To obtain

$$\dot{y}_p(0) = 0$$

for every source f, we require

$$g(0) = 0.$$

With this condition,

$$\dot{y}_p(t) = \int_0^t \dot{g}(t-s)f(s)\,ds.$$

Differentiating once more gives

$$\ddot{y}_p(t) = \dot{g}(0)f(t) + \int_0^t \ddot{g}(t-s)f(s)\,ds.$$

Therefore,

$$\begin{aligned} a\ddot{y}_p + b\dot{y}_p + cy_p = a\dot{g}(0)f(t) \\ + \int_0^t \big(a\ddot{g}(t-s) + b\dot{g}(t-s) + cg(t-s)\big)f(s)\,ds. \end{aligned}$$

Hence the convolution formula solves the forced equation for every f if

$$a\ddot{g} + b\dot{g} + cg = 0, \qquad t > 0,$$

and

$$g(0) = 0, \qquad \dot{g}(0) = \frac{1}{a}.$$

Thus the impulse response is obtained by solving a homogeneous equation with special initial data:

$$\boxed{\begin{aligned} a\ddot{g} + b\dot{g} + cg = 0, \qquad t > 0, \\ g(0) = 0, \\ \dot{g}(0) = \frac{1}{a}. \end{aligned}}$$

Once g is known, the solution of

$$a\ddot{y} + b\dot{y} + cy = f(t), \qquad y(0) = 0, \qquad \dot{y}(0) = 0,$$

is

$$y(t) = \int_0^t g(t-s)f(s)\,ds.$$

Example 3.2 (Green function and step response)**.** Consider

$$\ddot{y} + 3\dot{y} + 2y = f(t), \qquad y(0) = 0, \qquad \dot{y}(0) = 0.$$

The impulse response satisfies

$$\ddot{g} + 3\dot{g} + 2g = 0, \qquad g(0) = 0, \qquad \dot{g}(0) = 1.$$

The characteristic equation is

$$\lambda^2 + 3\lambda + 2 = 0,$$

with roots

$$\lambda = -1, \qquad \lambda = -2.$$

Hence

$$g(t) = C_1 e^{-t} + C_2 e^{-2t}.$$

The initial conditions give

$$C_1 + C_2 = 0,$$

and

$$-C_1 - 2C_2 = 1.$$

Therefore

$$C_1 = 1, \qquad C_2 = -1,$$

and

$$g(t) = e^{-t} - e^{-2t}.$$

The zero-initial-data response to a general source is therefore

$$y(t) = \int_0^t \left(e^{-(t-s)} - e^{-2(t-s)} \right) f(s)\, ds.$$

Now suppose that the source is switched on at time $t = t_0$:

$$f(t) = H(t - t_0),$$

where H is the Heaviside step function. For $0 \le t < t_0$,

$$y(t) = 0.$$

For $t \ge t_0$,

$$\begin{aligned} y(t) &= \int_{t_0}^t \left(e^{-(t-s)} - e^{-2(t-s)} \right) ds \\ &= \int_0^{t-t_0} \left(e^{-\tau} - e^{-2\tau} \right) d\tau \\ &= \frac{1}{2} - e^{-(t-t_0)} + \frac{1}{2} e^{-2(t-t_0)}. \end{aligned}$$

Thus

$$y(t) = H(t-t_0)\left[\frac{1}{2} - e^{-(t-t_0)} + \frac{1}{2}e^{-2(t-t_0)}\right].$$

The limiting response is

$$\lim_{t\to\infty} y(t) = \frac{1}{2}.$$

This agrees with the equilibrium relation

$$2y = 1$$

for a constant unit source. The step response therefore consists of a transient part, which decays exponentially, and the steady response $1/2$.

The name *impulse response* comes from an idealized source concentrated at a single instant. Let

$$H(t) = \begin{cases} 0, & t < 0, \\ 1, & t > 0. \end{cases}$$

In the sense of distributions,

$$\frac{d}{dt}H(t) = \delta(t),$$

where δ is the Dirac delta.

Define the causal Green's function by

$$G(t,s) = H(t-s)g(t-s).$$

Since

$$g(0) = 0,$$

distributional differentiation with respect to t gives

$$\frac{\partial G}{\partial t} = H(t-s)\dot{g}(t-s),$$

and

$$\frac{\partial^2 G}{\partial t^2} = H(t-s)\ddot{g}(t-s) + \dot{g}(0)\delta(t-s).$$

Therefore,

$$\begin{aligned} a\frac{\partial^2 G}{\partial t^2} + b\frac{\partial G}{\partial t} + cG &= H(t-s)\big(a\ddot{g} + b\dot{g} + cg\big)(t-s) \\ &\quad + a\dot{g}(0)\delta(t-s) \\ &= \delta(t-s). \end{aligned}$$

Thus $G(t,s)$ is the causal solution of

$$aG_{tt} + bG_t + cG = \delta(t-s).$$

This gives another interpretation of Duhamel's formula:

$$y_p(t) = \int_0^t G(t,s)f(s)\,ds.$$

A general source may be viewed as a continuous superposition of impulses applied at earlier times. The Green function $G(t,s)$ describes the response at time t to a unit impulse applied at time s.

Exercise 3.7 (Eliminating residual vibration)**.** An undamped oscillator is initially at rest and is acted on by a constant force for a finite time:

$$\ddot{y} + \omega_0^2 y = f(t), \qquad y(0) = \dot{y}(0) = 0,$$

where

$$f(t) = \begin{cases} F_0, & 0 \le t \le T, \\ 0, & t > T. \end{cases}$$

Use Duhamel's formula to determine the motion. Find all durations $T > 0$ for which the oscillator returns to rest after the force is removed. Interpret the result in terms of the natural period.

3.5 Sinusoidal Forcing, Beating, Resonance, and Synchronization

We now study a special but important forcing:

$$m\ddot{y} + \gamma\dot{y} + ky = F_0 \cos \omega t.$$

Here F_0 is the amplitude of the applied force, and ω is the forcing frequency. The main question is how the system responds when the external frequency ω is close to the natural frequency of the oscillator.

To see the central mechanism clearly, we first consider the undamped case.

3.5.1 Undamped forced oscillation

Consider

$$m\ddot{y} + ky = F_0 \cos \omega t, \qquad m > 0, \quad k > 0, \quad \omega \ge 0.$$

Let

$$\omega_0 = \sqrt{\frac{k}{m}}$$

be the natural frequency. Dividing the equation by m, we obtain

$$\ddot{y} + \omega_0^2 y = \frac{F_0}{m} \cos \omega t.$$

Assume first that

$$\omega \neq \omega_0.$$

We look for a particular solution of the form

$$y_p(t) = A \cos \omega t.$$

Substitution gives

$$(-\omega^2 + \omega_0^2) A \cos \omega t = \frac{F_0}{m} \cos \omega t.$$

Therefore

$$A = \frac{F_0/m}{\omega_0^2 - \omega^2}.$$

Hence

$$y_p(t) = \frac{F_0/m}{\omega_0^2 - \omega^2} \cos \omega t.$$

The general solution is

$$y(t) = C_1 \cos \omega_0 t + C_2 \sin \omega_0 t + \frac{F_0/m}{\omega_0^2 - \omega^2} \cos \omega t.$$

Thus the motion contains two frequencies: the natural frequency ω_0 and the forcing frequency ω.

Beating and the resonant limit

Consider the undamped initial value problem

$$\ddot{y} + \omega_0^2 y = \frac{F_0}{m} \cos \omega t, \qquad y(0) = 0, \qquad \dot{y}(0) = 0.$$

When $\omega \neq \omega_0$, the solution is

$$y(t) = \frac{F_0/m}{\omega_0^2 - \omega^2} \left(\cos \omega t - \cos \omega_0 t \right).$$

Using the identity

$$\cos A - \cos B = -2 \sin \frac{A+B}{2} \sin \frac{A-B}{2},$$

we obtain

$$y(t) = \frac{2F_0/m}{\omega^2 - \omega_0^2} \sin \left(\frac{\omega + \omega_0}{2} t \right) \sin \left(\frac{\omega - \omega_0}{2} t \right).$$

When ω is close to ω_0, the factor

$$\sin \left(\frac{\omega + \omega_0}{2} t \right)$$

oscillates rapidly, while

$$\sin \left(\frac{\omega - \omega_0}{2} t \right)$$

varies slowly. Thus the solution is a rapid oscillation whose amplitude is modulated by a slowly varying envelope. This phenomenon is called *beating*; see Figure 3.5 (a).

Since the absolute value of the envelope repeats when the slow factor changes sign, the beat period is

$$T_{\text{beat}} = \frac{2\pi}{|\omega - \omega_0|}.$$

The case $\omega = \omega_0$ is qualitatively different. The forcing then has the same frequency as a homogeneous solution, so the ansatz

$$y_p(t) = A \cos \omega_0 t$$

cannot produce a particular solution. We instead try

$$y_p(t) = At \sin \omega_0 t.$$

Substitution gives

$$\ddot{y}_p + \omega_0^2 y_p = 2A\omega_0 \cos \omega_0 t.$$

Hence

$$A = \frac{F_0}{2m\omega_0},$$

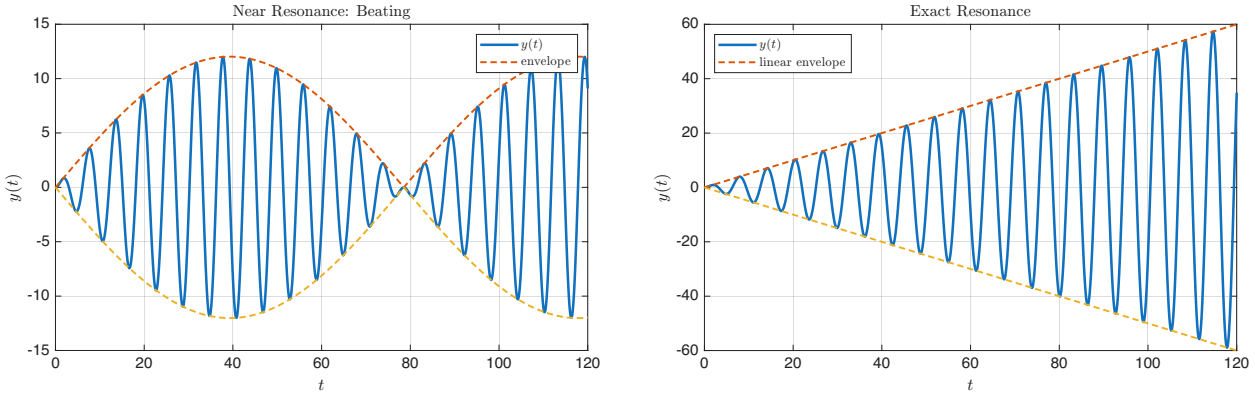
and therefore

$$y_p(t) = \frac{F_0}{2m\omega_0} t \sin \omega_0 t.$$

For zero initial data, this is the resonant solution:

$$y(t) = \frac{F_0}{2m\omega_0} t \sin \omega_0 t.$$

Its amplitude grows linearly in time. This is *resonance*; see Figure 3.5 (b).



(a) Near resonance: $\omega \approx \omega_0$.

(b) Exact resonance: $\omega = \omega_0$.

Figure 3.5: Beating and resonance for the undamped forced oscillator. Left: when the forcing frequency is close to the natural frequency, the solution exhibits beating. Right: when the forcing frequency equals the natural frequency, the amplitude grows linearly in time.

Resonance may also be viewed as the limiting case of beating. Let

$$y_\omega(t) = \frac{F_0/m}{\omega_0^2 - \omega^2} \left(\cos \omega t - \cos \omega_0 t\right).$$

As $\omega \to \omega_0$, both the numerator and denominator tend to zero. Differentiating with respect to ω, we obtain

$$\lim_{\omega \to \omega_0} \frac{\cos \omega t - \cos \omega_0 t}{\omega_0^2 - \omega^2} = \lim_{\omega \to \omega_0} \frac{-t \sin \omega t}{-2\omega} = \frac{t \sin \omega_0 t}{2\omega_0}.$$

Therefore

$$\lim_{\omega \to \omega_0} y_\omega(t) = \frac{F_0}{2m\omega_0} t \sin \omega_0 t.$$

Thus the resonant solution is the limit of the beating response as the forcing frequency approaches the natural frequency. The slowly varying beat envelope stretches over longer and longer time scales, and in the limit it becomes the linearly growing resonant amplitude.

This is the same structural phenomenon that appeared for repeated characteristic roots. When two natural modes coalesce, the generalized mode contains the factor t. Here the forcing frequency coalesces with the natural frequency, and the forced response again acquires the factor t.

3.5.2 Damped forced oscillation

Now consider the damped forced oscillator

$$m\ddot{y} + \gamma\dot{y} + ky = F_0 \cos \omega t, \qquad m > 0, \quad \gamma > 0, \quad k > 0, \quad F_0 > 0, \quad \omega \geq 0.$$

Unlike the undamped case, every solution remains bounded. Unlike the undamped case, the response does not grow without bound. Damping removes energy from the system and produces a steady periodic response.

It is convenient to use complex notation. We replace

$$F_0 \cos \omega t$$

by the complex force

$$F_0 e^{i\omega t}$$

and look for a complex particular solution

$$z_p(t) = Ae^{i\omega t}.$$

Substitution into

$$m\ddot{z} + \gamma\dot{z} + kz = F_0 e^{i\omega t}$$

gives

$$\big(-m\omega^2 + i\gamma\omega + k\big)A = F_0.$$

Hence

$$A = \frac{F_0}{k - m\omega^2 + i\gamma\omega}.$$

Taking the real part, we obtain the periodic particular solution

$$y_{\text{sync}}(t) = \Re\left(\frac{F_0 e^{i\omega t}}{k - m\omega^2 + i\gamma\omega}\right).$$

It oscillates with the same frequency as the forcing, although its amplitude and phase are modified. We therefore call y_{sync} the *synchronized solution*. When $\gamma > 0$, every homogeneous solution decays, so every solution approaches $y_{\text{sync}}(t)$.

The amplitude of the synchronized response is

$$R(\omega) = \frac{F_0}{\sqrt{(k - m\omega^2)^2 + \gamma^2\omega^2}}.$$

The function $R(\omega)$ is the frequency-response amplitude of the damped oscillator. To see its structure more clearly, introduce the natural frequency

$$\omega_0 = \sqrt{\frac{k}{m}},$$

the nondimensional forcing frequency

$$\Omega = \frac{\omega}{\omega_0},$$

and the nondimensional damping parameter

$$\widetilde{\gamma} = \frac{\gamma}{m\omega_0}.$$

Then

$$k - m\omega^2 = m\omega_0^2(1-\Omega^2), \qquad \gamma\omega = m\omega_0^2\widetilde{\gamma}\,\Omega.$$

Hence

$$R(\omega) = \frac{F_0}{m\omega_0^2}\frac{1}{\sqrt{(1-\Omega^2)^2 + \widetilde{\gamma}^2\Omega^2}}.$$

Since $m\omega_0^2 = k$, this may be written as

$$R(\omega) = \frac{F_0}{k}\,R_{\widetilde{\gamma}}(\Omega),$$

where

$$R_{\widetilde{\gamma}}(\Omega) = \frac{1}{\sqrt{(1-\Omega^2)^2 + \widetilde{\gamma}^2\Omega^2}}.$$

Thus $R_{\widetilde{\gamma}}(\Omega)$ is the normalized frequency response. See Figure 3.6.

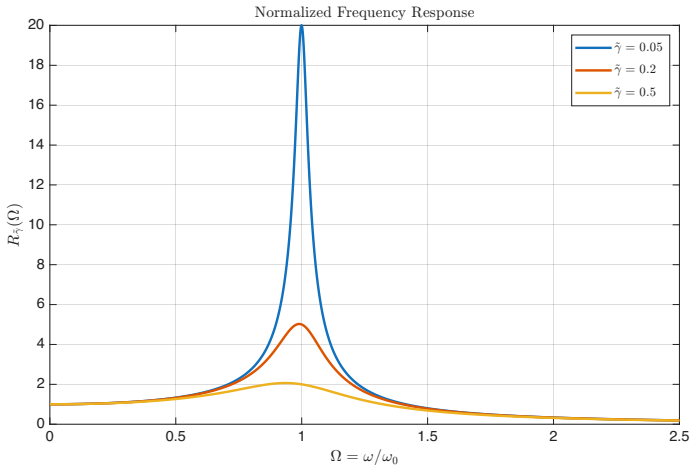


Figure 3.6: Normalized frequency response curves $R_{\widetilde{\gamma}}(\Omega)$ for several nondimensional damping parameters $\widetilde{\gamma}$, where $R_{\widetilde{\gamma}}(\Omega) = 1/\sqrt{(1-\Omega^2)^2 + \widetilde{\gamma}^2\Omega^2}$ and $\Omega = \omega/\omega_0$. Smaller damping produces a sharper resonant peak near $\Omega = 1$.

The phase lag $\phi(\omega) \in [0, \pi]$ is determined by *

$$\phi(\omega) = \operatorname{atan2}\big(\gamma\omega, k - m\omega^2\big) = \operatorname{atan2}\big(\widetilde{\gamma}\,\Omega, 1 - \Omega^2\big),$$

*The angle $\theta = \operatorname{atan2}(y, x)$ is the oriented angle measured counterclockwise from the positive x-axis to the ray from the origin through the point (x, y), with $-\pi < \theta \le \pi$.

so that

$$\boxed{y_{\text{sync}}(t) = R(\omega)\cos(\omega t - \phi(\omega)).}$$

The displacement amplitude is largest when

$$(1-\Omega^2)^2 + \widetilde{\gamma}^2\Omega^2$$

is smallest. Differentiating with respect to Ω gives the normalized displacement-resonance frequency

$$\Omega_{\text{res}}^2 = 1 - \frac{\widetilde{\gamma}^2}{2},$$

provided

$$\widetilde{\gamma}^2 < 2.$$

Therefore the displacement-resonance frequency is

$$\omega_{\text{res}} = \omega_0\sqrt{1 - \frac{\widetilde{\gamma}^2}{2}}.$$

Thus damping shifts the resonance frequency below the undamped natural frequency ω_0. If the damping is sufficiently strong, namely if

$$\widetilde{\gamma}^2 \geq 2,$$

the displacement amplitude has no maximum at a positive forcing frequency.

In summary, an undamped oscillator exhibits true resonance: forcing at the natural frequency produces linearly growing amplitude. A damped oscillator exhibits a finite resonant peak: the steady-state amplitude is large near the natural frequency, but remains bounded because energy is dissipated.

Exercise 3.8. Solve the initial value problem

$$\ddot{y} + 4y = 3\cos 2t + \cos t, \qquad y(0) = 2, \qquad \dot{y}(0) = 1.$$

3.6 Phase Portraits of Planar Linear Systems

A second-order equation naturally contains two state variables: the unknown function and its velocity. Thus the state of the motion is described by the pair

$$(y, \dot{y}).$$

This leads naturally from a scalar second-order equation to a first-order system in the plane.

In this section, we study planar linear systems

$$\dot{\mathbf{y}} = A\mathbf{y}, \qquad A \in \mathbb{R}^{2\times 2}.$$

The purpose is not yet to develop the general theory of linear systems. That will be done later using matrix exponentials and the full structure of the matrix A. Here the emphasis is geometric: eigenvalues and eigenvectors determine invariant directions and organize the basic phase portraits, including nodes, saddles, spirals, and centers.

This viewpoint also prepares for the nonlinear systems of the next chapter. Near an equilibrium $\mathbf{y}_*$, a nonlinear system

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$$

is approximated, after writing

$$\mathbf{z} = \mathbf{y} - \mathbf{y}_*,$$

by the linear system

$$\dot{\mathbf{z}} = D\mathbf{f}(\mathbf{y}_*)\mathbf{z}.$$

Thus the planar linear phase portraits developed here provide the basic models for studying nonlinear dynamics near equilibria. Nodes, saddles, and spirals persist as local models when the linearized eigenvalues have nonzero real parts. A linear center, however, is a borderline case, and the nonlinear terms may change its behavior.

3.6.1 From a second-order equation to a first-order system

Consider the homogeneous second-order equation

$$a\ddot{y} + b\dot{y} + cy = 0, \qquad a \neq 0.$$

Introduce the state variables

$$y^1 = y, \qquad y^2 = \dot{y}.$$

Then

$$\dot{y}^1 = \dot{y} = y^2.$$

Also, from the equation,

$$\ddot{y} = -\frac{b}{a}\dot{y} - \frac{c}{a}y.$$

Therefore

$$\dot{y}^2 = -\frac{c}{a}y^1 - \frac{b}{a}y^2.$$

Thus the second-order equation is equivalent to the first-order system

$$\begin{cases} \dot{y}^1 = y^2, \\ \dot{y}^2 = -\dfrac{c}{a}y^1 - \dfrac{b}{a}y^2. \end{cases}$$

In matrix form,

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}, \qquad \mathbf{y} = \begin{bmatrix} y^1 \\ y^2 \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} 0 & 1 \\ -c/a & -b/a \end{bmatrix}.$$

The phase plane is the (y^1, y^2)-plane. Since $y^1 = y$ and $y^2 = \dot{y}$, each point in the phase plane records both the position and the velocity of the original second-order equation.

For the mechanical oscillator

$$m\ddot{y} + \gamma\dot{y} + ky = 0,$$

the corresponding first-order system is

$$\frac{d}{dt}\begin{bmatrix} y \\ \dot{y} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -k/m & -\gamma/m \end{bmatrix}\begin{bmatrix} y \\ \dot{y} \end{bmatrix}.$$

Thus the damping regimes of the oscillator are also phase portraits of a planar linear system.

The characteristic polynomial of the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -c/a & -b/a \end{bmatrix}$$

is

$$\begin{aligned} \det(\lambda \mathbf{I} - \mathbf{A}) &= \det \begin{bmatrix} \lambda & -1 \\ c/a & \lambda + b/a \end{bmatrix} \\ &= \lambda^2 + \frac{b}{a}\lambda + \frac{c}{a}. \end{aligned}$$

Multiplying by a, we recover the characteristic polynomial of the second-order equation:

$$a\lambda^2 + b\lambda + c.$$

Thus the characteristic roots of the scalar equation are exactly the eigenvalues of its phase-plane matrix.

3.6.2 From a planar linear system to a second-order equation

Conversely, a planar linear system often reduces to a scalar second-order equation. Consider

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}, \qquad \mathbf{A} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y^1 \\ y^2 \end{bmatrix}.$$

In components,

$$\begin{cases} \dot{y}^1 = \alpha y^1 + \beta y^2, \\ \dot{y}^2 = \gamma y^1 + \delta y^2. \end{cases}$$

Assume first that $\beta \neq 0$. Then the first equation gives

$$y^2 = \frac{1}{\beta}(\dot{y}^1 - \alpha y^1).$$

Differentiating the first equation,

$$\ddot{y}^1 = \alpha \dot{y}^1 + \beta \dot{y}^2.$$

Using the second equation,

$$\ddot{y}^1 = \alpha \dot{y}^1 + \beta(\gamma y^1 + \delta y^2).$$

Substitute

$$y^2 = \frac{1}{\beta}(\dot{y}^1 - \alpha y^1)$$

to obtain

$$\begin{aligned} \ddot{y}^1 &= \alpha \dot{y}^1 + \beta\gamma y^1 + \delta(\dot{y}^1 - \alpha y^1) \\ &= (\alpha + \delta)\dot{y}^1 + (\beta\gamma - \alpha\delta) y^1. \end{aligned}$$

Hence

$$\ddot{y}^1 - (\alpha + \delta)\dot{y}^1 + (\alpha\delta - \beta\gamma) y^1 = 0.$$

That is,

$$\ddot{y}^1 - \operatorname{tr}(\mathbf{A})\dot{y}^1 + \det(\mathbf{A})y^1 = 0.$$

The same conclusion is obtained from the characteristic polynomial. The matrix $\mathbf{A}$ has characteristic equation

$$\det(\lambda\mathbf{I} - \mathbf{A}) = 0,$$

namely

$$\lambda^2 - \operatorname{tr}(\mathbf{A})\lambda + \det(\mathbf{A}) = 0.$$

This is precisely the characteristic equation of

$$\ddot{y}^1 - \operatorname{tr}(\mathbf{A})\dot{y}^1 + \det(\mathbf{A})y^1 = 0.$$

If $\beta = 0$ but $\gamma \neq 0$, the same argument may be carried out using the second component. If $\beta = \gamma = 0$, the two scalar equations are already decoupled.

Thus, whenever one component is coupled to the other so that it can be eliminated, a planar linear system may be represented by a scalar second-order equation. Conversely, every scalar second-order linear equation has a canonical planar first-order representation.

3.6.3 Eigenvalues and invariant directions

Consider the planar linear system

$$\dot{\mathbf{y}} = A\mathbf{y}, \qquad A \in \mathbb{R}^{2\times 2}.$$

Suppose that $\mathbf{v} \neq 0$ is an eigenvector of A with eigenvalue λ:

$$A\mathbf{v} = \lambda\mathbf{v}.$$

Then the line

$$\ell = \{s\mathbf{v} : s \in \mathbb{R}\}$$

is invariant under the flow. Indeed, if

$$\mathbf{y}(0) = C\mathbf{v},$$

then

$$\mathbf{y}(t) = Ce^{\lambda t}\mathbf{v}$$

is a solution, since

$$\frac{d}{dt}\left(Ce^{\lambda t}\mathbf{v}\right) = C\lambda e^{\lambda t}\mathbf{v} = Ce^{\lambda t}A\mathbf{v} = A\left(Ce^{\lambda t}\mathbf{v}\right).$$

Thus a real eigenvector determines an invariant straight line in the phase plane. The sign of the corresponding eigenvalue determines the direction of motion along this line:

$$\lambda < 0 \quad \Longrightarrow \quad \mathbf{y}(t) \to 0 \quad \text{as } t \to \infty,$$

while

$$\lambda > 0 \quad \Longrightarrow \quad |\mathbf{y}(t)| \to \infty \quad \text{along the eigenline.}$$

If A has two distinct real eigenvalues λ_1, λ_2 with corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2$, then

$$\mathbf{y}(t) = C_1 e^{\lambda_1 t}\mathbf{v}_1 + C_2 e^{\lambda_2 t}\mathbf{v}_2.$$

The two eigenlines are invariant, and the signs of λ_1 and λ_2 determine whether the motion is attracting, repelling, or of saddle type.

When the eigenvalues are complex,

$$\lambda = \alpha \pm i\omega, \qquad \omega > 0,$$

there are no real eigendirections. The corresponding real solutions contain the factor $e^{\alpha t}$ together with oscillatory terms. Hence trajectories rotate around the origin, while the sign of α determines whether they move inward, move outward, or remain bounded.

Repeated real eigenvalues require a separate distinction. If there are two linearly independent eigenvectors, every direction is generated by ordinary exponential modes. If there is only one eigenvector, the generalized mode $te^{\lambda t}$ produces trajectories that become tangent to the single eigendirection.

These observations lead to the basic phase portraits described in the next subsection.

3.6.4 Basic phase portraits

The phase portrait of

$$\dot{\mathbf{y}} = A\mathbf{y}$$

is determined by the eigenvalues of A and, in the repeated-root case, by the number of linearly independent eigenvectors.

Distinct real eigenvalues. Suppose that

$$\lambda_1 \neq \lambda_2$$

are real. Then there are two invariant eigenlines.

If $\lambda_1 < 0$, $\lambda_2 < 0$, every trajectory approaches the origin as $t \to \infty$. The origin is a *stable node*.

If $\lambda_1 > 0$, $\lambda_2 > 0$, every nonzero trajectory moves away from the origin. The origin is an *unstable node*.

If the eigenvalues have opposite signs, one eigenline is attracting and the other is repelling. The origin is a *saddle*.

The three typical phase portraits associated with distinct real eigenvalues are shown in Figure 3.7.

Complex eigenvalues. Suppose that

$$\lambda = \alpha \pm i\omega, \qquad \omega > 0.$$

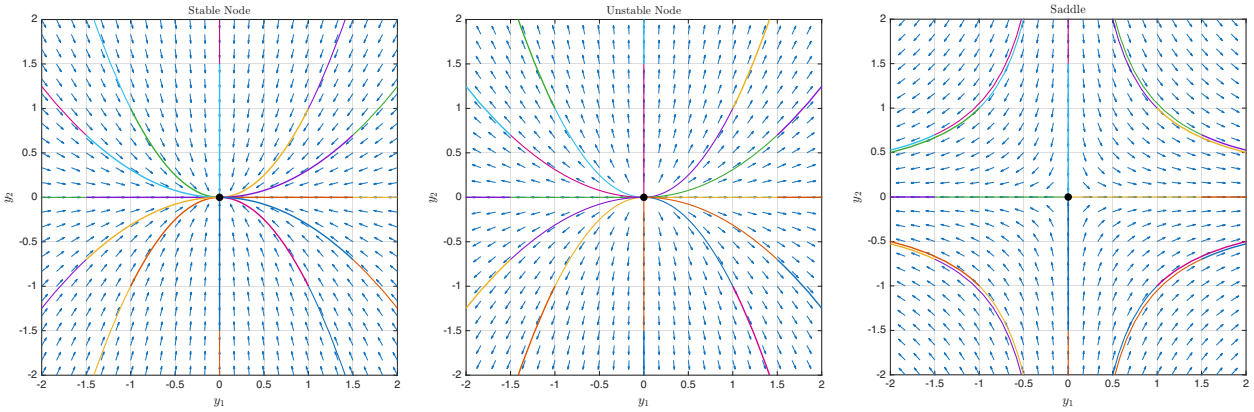
Trajectories rotate around the origin.

If $\alpha < 0$, the trajectories spiral inward, and the origin is a stable spiral.

If $\alpha > 0$, the trajectories spiral outward, and the origin is an unstable spiral.

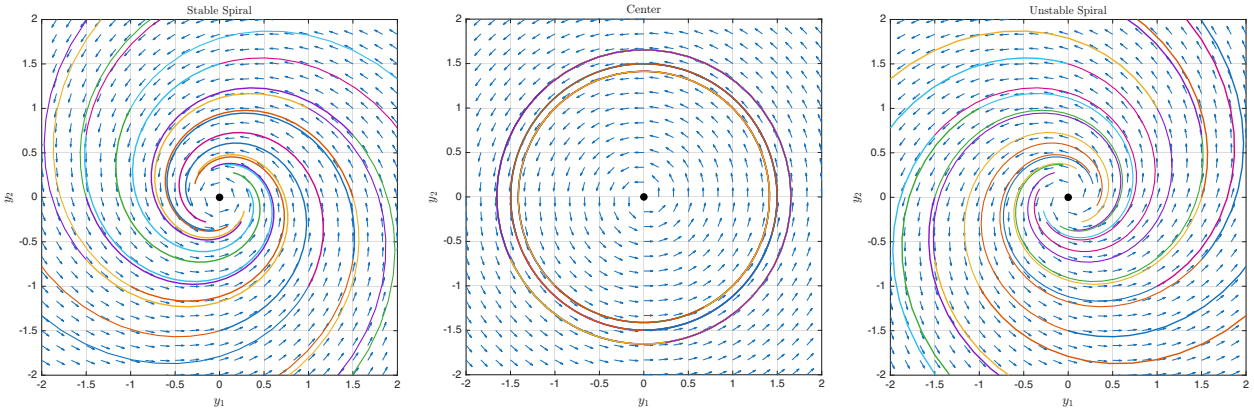
If $\alpha = 0$, the eigenvalues are purely imaginary. The trajectories are closed curves surrounding the origin, and the origin is a center.

The corresponding phase portraits are shown in Figure 3.8.



(a) Stable node (b) Unstable node (c) Saddle

Figure 3.7: Phase portraits associated with distinct real eigenvalues. (a) Stable node: both eigenvalues are negative. (b) Unstable node: both eigenvalues are positive. (c) Saddle: the eigenvalues have opposite signs.



(a) Stable spiral (b) Center (c) Unstable spiral

Figure 3.8: Phase portraits associated with complex eigenvalues. (a) Stable spiral: the eigenvalues have negative real part and nonzero imaginary part, so trajectories spiral into the origin. (b) Center: the eigenvalues are purely imaginary, and the trajectories are closed curves around the origin. (c) Unstable spiral: the eigenvalues have positive real part and nonzero imaginary part, so trajectories spiral away from the origin.

Repeated eigenvalues. Finally, suppose that $\mathbf{A}$ has a repeated nonzero real eigenvalue

$$\lambda_1 = \lambda_2 = \lambda, \qquad \lambda \neq 0.$$

There are two possibilities.

If $\mathbf{A}$ has two linearly independent eigenvectors, then $\mathbf{A} = \lambda \mathbf{I}$. The phase portrait is a proper node, also called a star node. Hence every line through the origin is invariant, and every trajectory is a straight ray. If $\lambda < 0$, these rays are attracting; if $\lambda > 0$, they are repelling.

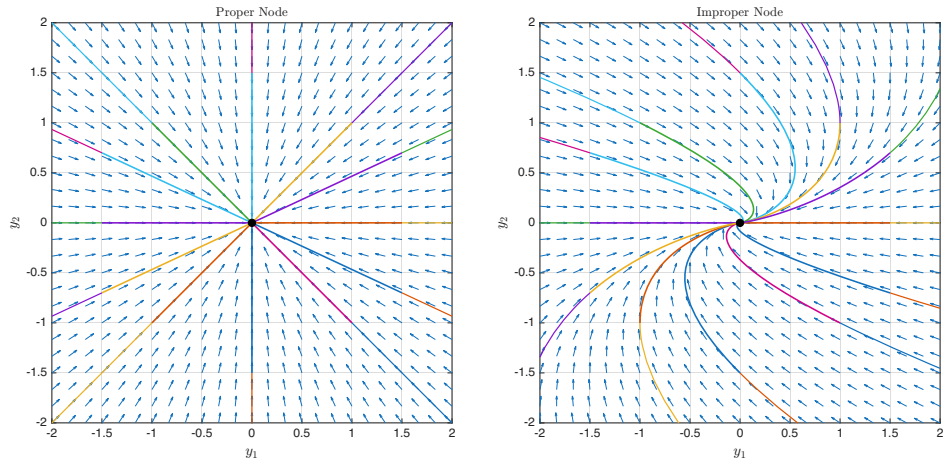
If $\mathbf{A}$ has only one linearly independent eigenvector, then the phase portrait is an *improper node*. In this case there is only one eigendirection. The remaining trajectories are curved, and they become tangent to the eigendirection as they approach or leave the origin. This is the planar linear-system analogue of the repeated-root solution

$$(C_1 + C_2 t)e^{\lambda t}$$

for a scalar second-order equation.

If the repeated eigenvalue is $\lambda = 0$, the system is degenerate. When $\mathbf{A} = \mathbf{0}$, every point is an equilibrium. When $\mathbf{A} \neq \mathbf{0}$, the matrix is nilpotent and the trajectories form a shear flow rather than a node.

The two typical phase portraits associated with repeated eigenvalues are shown in Figure 3.9.



(a) Proper repeated eigenvalue (b) Improper repeated eigenvalue

Figure 3.9: Phase portraits with repeated eigenvalues. In the proper case there are two independent eigenvectors and all trajectories are straight rays. In the improper case there is only one eigenvector; trajectories become tangent to the eigendirection.

3.6.5 Trace-determinant classification

The preceding phase portraits can be summarized by two scalar quantities: the trace and the determinant of the matrix. Let

$$T = \operatorname{tr}(\mathbf{A}), \qquad D = \det(\mathbf{A}).$$

The characteristic polynomial of $\mathbf{A}$ is

$$p(\lambda) = \det(\lambda \mathbf{I} - \mathbf{A}) = \lambda^2 - T\lambda + D.$$

Thus the eigenvalues are determined by

$$\lambda^2 - T\lambda + D = 0.$$

The discriminant is

$$\Delta = T^2 - 4D.$$

The signs of D, T, and Δ determine the phase portrait except on the repeated-root boundary $\Delta = 0$, where one must also determine the number of linearly independent eigenvectors.

First, if

$$D < 0,$$

then the product of the eigenvalues is negative. Hence the eigenvalues are real and have opposite signs. The origin is a saddle.

Now suppose

$$D > 0.$$

Then the eigenvalues have the same sign if they are real, and have the same real part if they are complex. The trace determines stability: negative trace corresponds to attraction, while positive trace corresponds to repulsion.

The curve

$$\Delta = 0$$

is the parabola

$$D = \frac{T^2}{4}.$$

It separates the region of real eigenvalues from the region of complex eigenvalues. Below the parabola,

$$\Delta > 0,$$

and the eigenvalues are real. Above the parabola,

$$\Delta < 0,$$

and the eigenvalues are complex.

The classification is summarized as follows:

condition	eigenvalues	phase portrait
$D < 0$	$\lambda_1 < 0 < \lambda_2$	saddle
$D > 0,\ \Delta > 0,\ T < 0$	$\lambda_1, \lambda_2 < 0$	stable node
$D > 0,\ \Delta > 0,\ T > 0$	$\lambda_1, \lambda_2 > 0$	unstable node
$D > 0,\ \Delta < 0,\ T < 0$	$\Re\lambda < 0$	stable spiral
$D > 0,\ \Delta < 0,\ T > 0$	$\Re\lambda > 0$	unstable spiral
$D > 0,\ \Delta < 0,\ T = 0$	$\lambda = \pm i\omega$	center

On the boundary $\Delta = 0$ away from the origin $(T, D) = (0, 0)$, the matrix has a repeated real eigenvalue. The node is proper if there are two independent eigenvectors, and improper if there is only one eigenvector.

On the line $D = 0$, at least one eigenvalue is zero. These systems are degenerate and are not classified as nodes, saddles, spirals, or centers. In the diagonalizable case, the zero eigenspace consists of equilibria; in the defective case, nilpotent behavior may occur.

Away from the degenerate line $D = 0$, the trace-determinant plane gives a compact summary of the basic planar linear phase portraits.

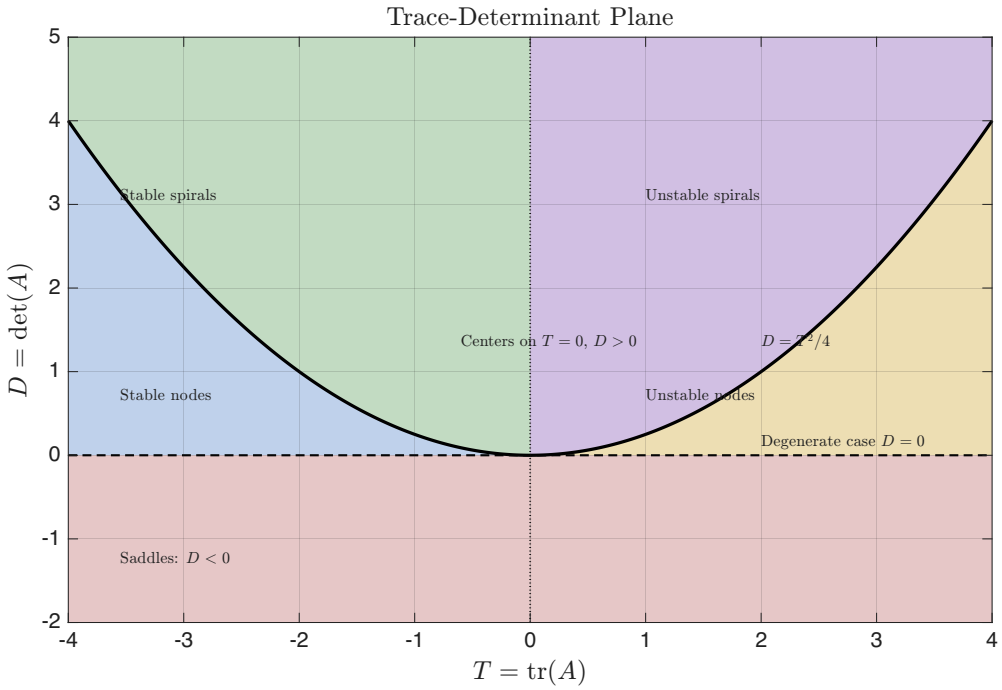


Figure 3.10: Trace-determinant classification of planar linear systems. Here $T = \mathrm{tr}(A)$, $D = \det(A)$, and the parabola $D = T^2/4$ separates real eigenvalues from complex eigenvalues.

For the phase-plane matrix obtained from

$$a\ddot{y} + b\dot{y} + cy = 0,$$

we have

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -c/a & -b/a \end{bmatrix}.$$

Therefore

$$T = -\frac{b}{a}, \qquad D = \frac{c}{a}.$$

The characteristic equation

$$\lambda^2 - T\lambda + D = 0$$

becomes

$$\lambda^2 + \frac{b}{a}\lambda + \frac{c}{a} = 0,$$

or equivalently

$$a\lambda^2 + b\lambda + c = 0.$$

Thus the trace-determinant classification is exactly the phase-plane version of the characteristic-root classification for second-order linear equations.

Closing remark

A scalar second-order equation becomes a planar first-order system when its state is represented by position and velocity. The characteristic roots then become the eigenvalues of the phase-plane matrix, and their classification determines the geometry of the trajectories.

This viewpoint prepares for the nonlinear systems of the next chapter. Near an equilibrium, a nonlinear planar system is approximated by the linear system determined by its Jacobian matrix. The nodes, saddles, and spirals described here will therefore serve as the basic local models for nonlinear phase portraits.

Exercise 3.9. For each matrix below, compute

$$T = \operatorname{tr}(\mathbf{A}), \qquad D = \det(\mathbf{A}), \qquad \Delta = T^2 - 4D,$$

and classify the equilibrium at the origin.

$$\mathbf{A}_1 = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}, \qquad \mathbf{A}_2 = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}, \qquad \mathbf{A}_3 = \begin{bmatrix} -1 & -2 \\ 2 & -1 \end{bmatrix}.$$

Exercise 3.10. Consider the parameter-dependent system

$$\dot{\mathbf{y}} = \begin{bmatrix} \mu & -1 \\ 1 & \mu \end{bmatrix} \mathbf{y}.$$

1. Find the eigenvalues.
2. Compute $T(\mu)$ and $D(\mu)$.
3. Classify the origin for $\mu < 0$, $\mu = 0$, and $\mu > 0$.
4. Locate the curve $(T(\mu), D(\mu))$ in the trace-determinant plane.

Homework Problems

The following problems synthesize the analytic, physical, and geometric viewpoints developed in this chapter.

Homework 3.1. Consider

$$\ddot{y} + \alpha\dot{y} + \omega_0^2 y = 0, \qquad \alpha, \omega_0 > 0.$$

In the critical case $\alpha = 2\omega_0$, one solution is

$$y^*(t) = te^{-\omega_0 t}.$$

For $\alpha < 2\omega_0$, construct a solution $y_\alpha(t)$ such that

$$y_\alpha(t) \longrightarrow y^*(t)$$

as $\alpha \to (2\omega_0)^-$.

Homework 3.2 (Energy input and average dissipated power)**.** Consider the forced damped oscillator

$$m\ddot{y} + \gamma\dot{y} + ky = F_0 \cos \omega t, \qquad m, \gamma, k, F_0 > 0,$$

with mechanical energy

$$E(t) = \frac{m}{2}\dot{y}(t)^2 + \frac{k}{2}y(t)^2.$$

For the synchronized periodic solution:

1. Use the energy identity
$$\dot{E} = -\gamma\dot{y}^2 + F_0\dot{y}\cos\omega t$$
to show that the average energy input equals the average energy dissipation over one forcing period.

2. Determine the average dissipated power as a function of ω, find the frequency at which it is largest, and compare this frequency with the displacement-resonance frequency.

Hint. Use

$$R(\omega) = \frac{F_0}{\sqrt{(k - m\omega^2)^2 + \gamma^2\omega^2}}$$

and

$$\langle \dot{y}^2 \rangle = \frac{1}{2}\omega^2 R(\omega)^2.$$

Homework 3.3 (Vibration caused by a moving support)**.** A mass is attached to a support by a spring and a damper. Let $x(t)$ be the displacement of the mass and let $z(t)$ be the prescribed displacement of the support. The equation of motion is

$$m\ddot{x} + \gamma(\dot{x} - \dot{z}) + k(x - z) = 0, \qquad m, \gamma, k > 0.$$

1. Introduce the relative displacement $y = x - z$ and derive its differential equation.

2. Suppose
$$z(t) = A\cos\omega t.$$

Determine the amplitudes of the synchronized relative motion $y(t)$ and the absolute motion $x(t)$.

3. Examine the limits as $\omega \to 0$ and $\omega \to \infty$. Explain when the mass follows the support and when the system acts as a vibration isolator.

Hint. For the absolute displacement, show that the amplitude ratio is

$$\frac{X(\omega)}{A} = \frac{\sqrt{k^2 + \gamma^2\omega^2}}{\sqrt{(k - m\omega^2)^2 + \gamma^2\omega^2}}.$$

Homework 3.4 (Oscillator parameters in the trace–determinant plane)**.** Consider

$$m\ddot{y} + \gamma\dot{y} + ky = 0, \qquad m > 0, \qquad \gamma, k \in \mathbb{R}.$$

1. Fix $\gamma > 0$ and decrease k from a value satisfying $4mk > \gamma^2$ to a negative value. Trace the resulting path in the trace–determinant plane and describe, in order, the phase portraits encountered. Treat the case $k = 0$ separately.

2. Fix $k > 0$ and let γ decrease from a sufficiently large positive value to a sufficiently large negative value. Describe the corresponding changes in the phase portrait.

3. Relate the systems with damping coefficients γ and $-\gamma$ by the transformation $t \mapsto -t$, and explain the resulting reversal of stability.

Hint. In phase-plane form,

$$T = -\frac{\gamma}{m}, \qquad D = \frac{k}{m},$$

and critical damping corresponds to $T^2 = 4D$.

Historical Notes and Further Reading

Historical notes. The linear oscillator has its physical origin in the study of elasticity and vibration. Robert Hooke stated the proportional relation between the force applied to a spring and its extension in the seventeenth century, publishing it in 1678 in the form commonly called Hooke's law (Hooke, 1678). Combined with Newton's law of motion, this relation leads to the basic spring–mass equation. The inclusion of friction or viscous resistance gives the damped oscillator, while an applied force produces the forced equation studied in this chapter. Hooke's work on springs therefore provides one of the earliest physical foundations for the second-order linear ODE. Electrical circuits provide a second realization of the same mathematics. Kirchhoff formulated his circuit laws in 1845, extending Ohm's ideas to networks containing several loops (Kirchhoff, 1845). For a series RLC circuit, Kirchhoff's voltage law gives a second-order equation for the charge. Resistance, inductance, and inverse capacitance play roles analogous to damping, mass, and spring stiffness in the mechanical oscillator. Thus mechanical and electrical systems with different physical meanings share the same characteristic roots, transient modes, resonance, and steady response.

The superposition principle made it possible to represent the response to a general source as a sum of elementary responses. In his work on heat conduction during the 1830s, Jean-Marie Duhamel developed the method now called Duhamel's principle, in which a time-dependent source is decomposed into contributions introduced at successive times (Duhamel, 1833). In the present chapter, the same idea appears as the variation-of-constants formula for a forced oscillator.

George Green's 1828 essay on electricity and magnetism introduced methods based on a response kernel associated with a differential operator (Green, 1828). The term *Green function* was adopted later, but the underlying idea is the one used here: determine the response to an impulse, and obtain the response to a general forcing by integration. This connects the impulse response, convolution, and Duhamel's principle, and it will be developed further through Laplace transforms in Chapter 6.

Further reading. For a classical treatment of second-order linear equations, characteristic roots, forcing, and variation of parameters, see Ince (1956). The mechanics of free, damped, and forced vibration is developed in greater physical detail by Den Hartog (1985). For electrical circuit models and their relation to differential equations, see Dorf and Svoboda (2014). A broader treatment of Green functions and linear response methods is given by Stakgold and Holst (2011).

Chapter 4

Nonlinear Systems in the Plane

In Chapter 3, we studied planar linear systems. There the phase portrait is largely determined by eigenvalues and eigenvectors, and the local and global pictures coincide because the vector field is linear. For nonlinear systems, the situation is different. Linearization still gives the first approximation near an equilibrium, but it does not determine the whole phase portrait. Away from equilibria, nonlinear effects may bend trajectories, create invariant regions, produce separatrices, or generate closed orbits.

The purpose of this chapter is to develop the elementary phase-plane method for nonlinear systems in two dimensions. We shall use equilibria, linearization, nullclines, invariant regions, and first integrals to understand the qualitative behavior of solutions. These tools allow us to study systems that cannot be solved explicitly.

The examples in this chapter fall into two broad classes. The first consists of interacting-component models arising from population dynamics, epidemiology, and chemical reactions. The second consists of nonlinear oscillators, including conservative, dissipative, and self-sustained systems. Together they illustrate how coupled differential equations arise from modeling and how the geometry of the associated vector field reveals the qualitative behavior of the system.

The guiding question is therefore no longer only whether explicit formulas for $x(t)$ and $y(t)$ can be found, but how the motion is organized in the phase plane: where the equilibria lie, which regions are invariant, how trajectories approach or leave equilibria, and whether closed orbits, separatrices, or other distinguished structures organize the dynamics.

The analysis in this chapter is intentionally elementary. We use equilibria, linearization, nullclines, invariant regions, invariant curves, first integrals, and numerical phase portraits to describe the qualitative behavior of planar systems. Local conclusions obtained from linearization are rigorous. Global phase portraits often suggest the long-time behavior of solutions, but a complete proof may require deeper results, such as the Poincaré–Bendixson theorem and the theory of limit cycles. These results will be developed in Chapter 11.

4.1 Planar Autonomous Systems

An autonomous system is often written in the abstract form

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}),$$

where $\mathbf{y}$ is the state of the system and $\mathbf{f}$ is a vector field on the phase space. In this chapter the phase space is two-dimensional, so we write

$$\mathbf{y} = (x, y), \qquad \mathbf{f}(\mathbf{y}) = \mathbf{f}(x, y) = (f(x, y), g(x, y)),$$

and the system takes the coordinate form

$$\dot{x} = f(x, y), \qquad \dot{y} = g(x, y).$$

The abstract notation emphasizes the dynamical system on phase space, while the coordinate notation emphasizes the geometry of the phase plane. We shall use both viewpoints, but concrete planar models will usually be written in terms of the state variables x and y.

A solution is a differentiable curve

$$t \mapsto \mathbf{y}(t) = (x(t), y(t))$$

satisfying

$$\dot{\mathbf{y}}(t) = \mathbf{f}(\mathbf{y}(t)).$$

Thus the state point moves in the phase plane with velocity prescribed by the vector field. The main object is no longer only the separate time graphs of $x(t)$ and $y(t)$, but the motion of the state point $(x(t), y(t))$ in the phase plane.

4.1.1 Vector fields and phase portraits

At each point (x, y), the system assigns the vector

$$\mathbf{f}(x, y) = (f(x, y), g(x, y)).$$

Drawing representative vectors gives the vector field of the system. A solution curve must be tangent to this field at every point:

$$\frac{d}{dt}(x(t), y(t)) = \mathbf{f}(x(t), y(t)).$$

A collection of representative solution curves, together with their direction of motion, is called a *phase portrait*. In practice, a phase portrait can be obtained in two complementary ways. Qualitatively, one locates equilibria and nullclines, determines the signs of $\dot{x}$ and $\dot{y}$ in the resulting regions, and uses linearization and invariant sets to infer the geometry of the trajectories. Numerically, one chooses several initial conditions, solves the system, and plots the parametric curves

$$(x(t), y(t)).$$

The phase portrait displays qualitative information that may not be apparent from the separate graphs of $x(t)$ and $y(t)$. It shows the geometry and direction of the orbits, how solutions approach or leave equilibria, whether they remain in a prescribed region, and how special orbits organize the motion.

If $\mathbf{f}$ is continuously differentiable, then the solution through each initial point is locally unique. Consequently, two distinct orbits cannot cross in the phase plane. Indeed, if two orbits passed

through the same point, then the corresponding solutions would solve the same initial value problem and hence would coincide.

A solution and its orbit contain different information. The solution

$$t \mapsto \mathbf{y}(t)$$

records both the path and its time parametrization. Its orbit is the set of points traced in phase space and records only the geometric path. Thus two vector fields may have the same orbits but traverse them with different speeds.

4.1.2 Equilibria and local behavior

Equilibria. Consider the autonomous system

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}), \qquad \mathbf{y} \in \mathbb{R}^2,$$

where $\mathbf{f}$ is continuously differentiable. An *equilibrium* is a point $\mathbf{y}_*$ at which the vector field vanishes:

$$\mathbf{f}(\mathbf{y}_*) = \mathbf{0}.$$

If $\mathbf{y}_* = (x_*, y_*)$, this means

$$f(x_*, y_*) = 0, \qquad g(x_*, y_*) = 0.$$

The corresponding solution is constant:

$$\mathbf{y}(t) \equiv \mathbf{y}_*.$$

Thus equilibria represent the fixed states of the system.

Linearization. To determine how nearby solutions behave, we compare the nonlinear system with a linear system. Write

$$\mathbf{y} = \mathbf{y}_* + \boldsymbol{\eta}.$$

Since $\mathbf{y}_*$ is constant,

$$\dot{\boldsymbol{\eta}} = \mathbf{f}(\mathbf{y}_* + \boldsymbol{\eta}).$$

By the first-order Taylor expansion,

$$\mathbf{f}(\mathbf{y}_* + \boldsymbol{\eta}) = \mathbf{f}(\mathbf{y}_*) + D\mathbf{f}(\mathbf{y}_*)\boldsymbol{\eta} + \mathbf{r}(\boldsymbol{\eta}),$$

where

$$\frac{|\mathbf{r}(\boldsymbol{\eta})|}{|\boldsymbol{\eta}|} \longrightarrow 0 \qquad \text{as } \boldsymbol{\eta} \longrightarrow \mathbf{0}.$$

Since $\mathbf{f}(\mathbf{y}_*) = \mathbf{0}$, we obtain

$$\dot{\boldsymbol{\eta}} = \mathbf{A}\boldsymbol{\eta} + \mathbf{r}(\boldsymbol{\eta}), \qquad \mathbf{A} = D\mathbf{f}(\mathbf{y}_*).$$

Neglecting the higher-order term gives the linear system

$$\dot{\boldsymbol{\eta}} = \mathbf{A}\boldsymbol{\eta},$$

called the *linearization* of the nonlinear system at $\mathbf{y}_*$.

For a planar system

$$\dot{x} = f(x, y),$$
$$\dot{y} = g(x, y),$$

the Jacobian matrix is

$$D\mathbf{f}(x, y) = \begin{bmatrix} f_x(x, y) & f_y(x, y) \\ g_x(x, y) & g_y(x, y) \end{bmatrix}.$$

Hence, at an equilibrium (x_*, y_*), the linearization is

$$\frac{d}{dt}\begin{bmatrix} \xi \\ \eta \end{bmatrix} = D\mathbf{f}(x_*, y_*)\begin{bmatrix} \xi \\ \eta \end{bmatrix},$$

where

$$x = x_* + \xi, \qquad y = y_* + \eta.$$

The planar linear systems in Chapter 3 were classified by their eigenvalues, or equivalently by the trace and determinant of the system matrix. We may therefore apply that classification to the Jacobian

$$\mathbf{A} = D\mathbf{f}(\mathbf{y}_*).$$

Writing

$$T = \operatorname{tr}(\mathbf{A}), \qquad D = \det(\mathbf{A}),$$

the eigenvalues satisfy

$$\lambda^2 - T\lambda + D = 0.$$

Thus

$$\lambda_\pm = \frac{T \pm \sqrt{T^2 - 4D}}{2}.$$

Hyperbolic and nonhyperbolic equilibria. An equilibrium $\mathbf{y}_*$ is called *hyperbolic* if no eigenvalue of $D\mathbf{f}(\mathbf{y}_*)$ has zero real part. Thus saddles, nodes, and spirals whose eigenvalues have nonzero real parts are hyperbolic. Equilibria with a zero eigenvalue or a pair of purely imaginary eigenvalues are *nonhyperbolic*.

Stability of a hyperbolic equilibrium. For a hyperbolic equilibrium, the basic local behavior of the nonlinear system is determined by the corresponding linearized system. In particular:

- if $D < 0$, the equilibrium is a saddle and is unstable;
- if $D > 0$ and $T < 0$, the equilibrium is locally asymptotically stable;
- if $D > 0$ and $T > 0$, the equilibrium is unstable.

The sign of $T^2 - 4D$ further distinguishes nodes from spirals, exactly as in Chapter 3.

Principle 4.1 (Linearization principle: Hartman–Grobman theorem)**.** *Let $\mathbf{y}_*$ be a hyperbolic equilibrium of*

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}).$$

Then, in a sufficiently small neighborhood of $\mathbf{y}_*$*, the nonlinear system has the same qualitative orbit structure as its linearization*

$$\dot{\boldsymbol{\eta}} = D\mathbf{f}(\mathbf{y}_*)\boldsymbol{\eta}.$$

More precisely, there is a homeomorphism between neighborhoods of the two equilibria (that is, a continuous, one-to-one, and onto map with continuous inverse) that conjugates the nonlinear flow to the linearized flow. Thus it maps trajectories of the nonlinear system to trajectories of the linearized system while preserving the direction of time. Corresponding trajectories need not have the same geometric shape.

In particular, the stability and basic local behavior of the equilibrium are determined by the eigenvalues of the Jacobian matrix.

The hyperbolicity assumption is essential. If the Jacobian has a zero eigenvalue or a pair of purely imaginary eigenvalues, the equilibrium is *nonhyperbolic*. In this case the linearized system alone may not determine the stability or local phase portrait; higher-order nonlinear terms or additional structure, such as a conserved quantity, may be decisive. We return to this issue at the end of this subsection.

Stable and unstable curves near a saddle. For a linear saddle

$$\dot{\boldsymbol{\eta}} = \mathbf{A}\boldsymbol{\eta},$$

there are two distinguished invariant lines: the stable eigenspace and the unstable eigenspace. Initial data on the stable eigenspace approach the equilibrium as $t \to \infty$, while initial data on the unstable eigenspace approach the equilibrium as $t \to -\infty$.

For a nonlinear saddle, these straight invariant lines are generally deformed into curves. Near a hyperbolic saddle equilibrium $\mathbf{y}_*$, there is a stable curve tangent to the stable eigendirection and an unstable curve tangent to the unstable eigendirection. Before describing these curves more precisely, we introduce the general notion of an invariant curve.

Invariant curves. A curve Γ in the phase plane is called *invariant* if every trajectory starting on Γ remains on Γ for as long as the solution exists. For a regular curve, invariance is equivalent to the vector field being tangent to the curve at every point.

If an invariant curve is given implicitly by

$$F(x, y) = 0,$$

then along the curve,

$$\frac{d}{dt}F(x(t), y(t)) = F_x(x, y)f(x, y) + F_y(x, y)g(x, y) = 0.$$

Thus the directional derivative of F along the vector field vanishes on $F = 0$.

If instead the curve is written locally as

$$y = h(x),$$

then invariance requires

$$\dot{y} = h'(x)\dot{x}.$$

For the system

$$\dot{x} = f(x, y), \qquad \dot{y} = g(x, y),$$

this gives the invariance equation

$$g(x, h(x)) = h'(x) f(x, h(x)).$$

This equation provides a useful way to determine invariant curves directly.

The stable and unstable curves of a hyperbolic saddle are important examples. More precisely, the local stable curve consists of initial points whose forward trajectories remain near $\mathbf{y}_*$ and approach it:

$$W^s_{\text{loc}}(\mathbf{y}_*) = \{\mathbf{y}_0 : \mathbf{y}(t; \mathbf{y}_0) \to \mathbf{y}_* \text{ as } t \to \infty, \quad \mathbf{y}(t; \mathbf{y}_0) \text{ remains near } \mathbf{y}_*\}.$$

Similarly, the local unstable curve is

$$W^u_{\text{loc}}(\mathbf{y}_*) = \{\mathbf{y}_0 : \mathbf{y}(t; \mathbf{y}_0) \to \mathbf{y}_* \text{ as } t \to -\infty, \quad \mathbf{y}(t; \mathbf{y}_0) \text{ remains near } \mathbf{y}_*\}.$$

These curves are locally invariant and are tangent at $\mathbf{y}_*$ to the stable and unstable eigenvectors of the Jacobian matrix $D\mathbf{f}(\mathbf{y}_*)$.

The corresponding global stable and unstable sets are defined by

$$W^s(\mathbf{y}_*) = \{\mathbf{y}_0 : \mathbf{y}(t; \mathbf{y}_0) \to \mathbf{y}_* \text{ as } t \to \infty\},$$

and

$$W^u(\mathbf{y}_*) = \{\mathbf{y}_0 : \mathbf{y}(t; \mathbf{y}_0) \to \mathbf{y}_* \text{ as } t \to -\infty\}.$$

The local stable and unstable curves are the portions of the global stable and unstable sets near the equilibrium. In particular,

$$W^s_{\text{loc}}(\mathbf{y}_*) \subset W^s(\mathbf{y}_*), \qquad W^u_{\text{loc}}(\mathbf{y}_*) \subset W^u(\mathbf{y}_*).$$

The local curves describe the geometry near the saddle, while their global extensions may form separatrices, basin boundaries, homoclinic orbits, or heteroclinic connections.

Example 4.1 (A curved unstable curve)**.** Consider

$$\dot{x} = x, \qquad \dot{y} = -y + x^2.$$

The origin is an equilibrium, and

$$D\mathbf{f}(0, 0) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Hence the origin is a saddle. The unstable eigendirection is the x-axis, while the stable eigendirection is the y-axis.

The line

$$x = 0$$

is invariant, since $\dot{x} = 0$ there. Along this line,

$$\dot{y} = -y,$$

so every point on it approaches the origin as $t \to \infty$. Hence

$$W^s_{\mathrm{loc}}(0,0) = \{(0,y) : |y| < \delta\},$$

for sufficiently small $\delta > 0$. The unstable curve is not a straight line. Seek it in the form

$$y = h(x), \qquad h(0) = 0, \qquad h'(0) = 0.$$

The condition $h'(0) = 0$ expresses tangency to the unstable eigendirection.

For the graph $y = h(x)$ to be invariant, the vector field must be tangent to the graph. Along the graph,

$$\dot{y} = h'(x)\dot{x}.$$

Using the differential equations,

$$h'(x)x = -h(x) + x^2.$$

Thus

$$xh'(x) + h(x) = x^2,$$

or

$$\frac{d}{dx}\big(xh(x)\big) = x^2.$$

Integrating gives

$$xh(x) = \frac{x^3}{3} + C.$$

For the curve to pass smoothly through the origin, we must have $C = 0$. Therefore

$$h(x) = \frac{x^2}{3}.$$

Hence the unstable curve is

$$W^u_{\mathrm{loc}}(0,0) = \left\{ \left(x, \frac{x^2}{3} \right) : |x| < \delta \right\},$$

for sufficiently small $\delta > 0$. It is tangent to the x-axis at the origin but bends upward away from the equilibrium. The linearized system determines its tangent direction, while the nonlinear term determines its curvature.

Stable and unstable curves are therefore special invariant curves associated with hyperbolic saddles. Their local behavior is controlled by linearization, but their global extensions can be much more complicated. They may form boundaries between regions with different qualitative behavior, connect equilibria, or return to the same equilibrium.

Later we will see another important source of invariant curves: if a system has a first integral, its level sets are invariant. Both mechanisms—stable and unstable curves on the one hand, and invariant level sets on the other—lead naturally to the separatrices that organize global phase portraits.

Nonhyperbolic equilibria and the limits of linearization. The linearization principle requires hyperbolicity. If the Jacobian has an eigenvalue with zero real part, higher-order nonlinear terms may determine the local dynamics.

Two basic situations occur in the plane. A zero eigenvalue leaves the motion in one direction undetermined by the linear part. A pair of purely imaginary eigenvalues leaves open whether the nonlinear equilibrium is a center, a spiral sink, or a spiral source.

Example 4.2 (A zero eigenvalue)**.** Consider the two systems

$$\dot{x} = -x^3, \qquad \dot{y} = -y,$$

and

$$\dot{x} = x^3, \qquad \dot{y} = -y.$$

For both systems, the origin is an equilibrium and

$$D\mathbf{f}(0,0) = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}.$$

Thus both have the same linearization,

$$\dot{x} = 0, \qquad \dot{y} = -y,$$

with eigenvalues

$$0, \qquad -1.$$

The linearized system gives no information about the motion in the x-direction.

The nonlinear terms resolve the ambiguity. For

$$\dot{x} = -x^3,$$

solutions on both sides of the origin move toward $x = 0$, so the origin is asymptotically stable. For

$$\dot{x} = x^3,$$

solutions with $x \neq 0$ move away from $x = 0$, so the origin is unstable.

Thus the same linearization can correspond to opposite stability behavior. When a zero eigenvalue is present, the first nonvanishing nonlinear term in the corresponding direction may determine the dynamics.

Example 4.3 (Purely imaginary eigenvalues)**.** Consider the family

$$\dot{x} = -y + \alpha x(x^2 + y^2), \qquad \dot{y} = x + \alpha y(x^2 + y^2),$$

where α is a constant. For every α, the origin is an equilibrium and

$$D\mathbf{f}(0,0) = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

The eigenvalues are

$$\lambda_{\pm} = \pm i.$$

Thus every member of the family has the same linearization, namely a linear center.

To determine the nonlinear behavior, introduce polar coordinates

$$x = r\cos\theta, \qquad y = r\sin\theta.$$

Using

$$r\dot{r} = x\dot{x} + y\dot{y},$$

we obtain

$$\begin{aligned} r\dot{r} &= x\big(-y + \alpha x(x^2+y^2)\big) + y\big(x + \alpha y(x^2+y^2)\big) \\ &= \alpha(x^2+y^2)^2 \\ &= \alpha r^4. \end{aligned}$$

Hence

$$\dot{r} = \alpha r^3.$$

Similarly,

$$\dot{\theta} = 1.$$

Therefore

$\alpha < 0$	spiral sink
$\alpha = 0$	center
$\alpha > 0$	spiral source.

Thus purely imaginary eigenvalues do not imply that the nonlinear system has a center. Higher-order terms may cause nearby trajectories to spiral inward or outward even though the linearized system consists of closed circles.

These examples show that nonhyperbolicity is not itself a dynamical type. Rather, it signals that additional information is required. Depending on the problem, this information may come from higher-order terms, a first integral, an energy or Lyapunov function, polar coordinates, or parameter dependence.

The models studied later illustrate several of these possibilities. For the simple pendulum and the conservative Duffing oscillator, purely imaginary eigenvalues occur at potential minima, and the conserved energy establishes the center structure. In parameter-dependent systems, loss of hyperbolicity may instead signal a change in the phase portrait; this leads to bifurcation theory developed later.

Exercise 4.1 (A parameter-dependent equilibrium)**.** For

$$\dot{x} = y, \qquad \dot{y} = -\alpha x - \gamma y - x^3, \qquad \gamma > 0,$$

find and classify all hyperbolic equilibria as α varies. At what value of α is hyperbolicity lost, and how does the equilibrium structure change there?

4.1.3 Global phase-plane geometry

Nullclines For the planar system

$$\dot{x} = f(x, y), \qquad \dot{y} = g(x, y),$$

the *x-nullcline* is the set

$$f(x, y) = 0,$$

and the *y-nullcline* is the set

$$g(x, y) = 0.$$

On the x-nullcline, $\dot{x} = 0$, so the vector field is vertical. On the y-nullcline, $\dot{y} = 0$, so the vector field is horizontal. Equilibria are the intersections of the two nullclines.

The nullclines divide the phase plane into regions in which the signs of $\dot{x}$ and $\dot{y}$ are fixed:

sign of f	sign of g	direction of motion
$f > 0$	$g > 0$	right and upward
$f > 0$	$g < 0$	right and downward
$f < 0$	$g > 0$	left and upward
$f < 0$	$g < 0$	left and downward

This information does not determine the exact slope or speed of a trajectory, but it often reveals the large-scale organization of the phase portrait.

Nullclines also have a direct modeling interpretation. In a population model, an x-nullcline separates regions where the population x increases from regions where it decreases. In an epidemic model, a nullcline may mark the threshold between growth and decay of infection. Thus nullclines express both the geometry and the mechanisms of the model.

A nullcline need not be an orbit. On the x-nullcline the vector field is vertical, so a solution will generally cross it unless the nullcline itself is invariant. Similarly, a solution usually crosses a y-nullcline horizontally. This distinction is important when interpreting a nullcline diagram.

Exercise 4.2 (Nullclines are not trajectories)**.** For

$$\dot{x} = y - x^2, \qquad \dot{y} = x - y,$$

find the nullclines and equilibria. Determine the vector-field direction on each nullcline, decide whether either is invariant, classify the equilibria, and sketch the qualitative direction field.

Invariant regions. A set D in the phase plane is called *positively invariant* if every solution starting in D remains in D for all future time for which it exists:

$$\mathbf{y}(0) \in D \quad \Longrightarrow \quad \mathbf{y}(t) \in D \qquad \text{for all } t \geq 0 \text{ for which the solution exists.}$$

Similarly, D is negatively invariant if solutions remain in D for backward time, and invariant if it is both positively and negatively invariant.

Invariant regions are especially important in applications, since the model may be meaningful only in part of the phase plane. Population and concentration variables should remain nonnegative, while proportions in an epidemic model should remain between 0 and the total population.

A common way to establish positive invariance is to examine the vector field along the boundary. If the vector field points inward or is tangent to the boundary at every boundary point, then a solution cannot cross the boundary outward.

For example, consider the first quadrant

$$Q = \{(x, y) : x \geq 0,\ y \geq 0\}.$$

Along the boundary $x = 0$, leaving Q would require $\dot{x} < 0$. Along the boundary $y = 0$, leaving Q would require $\dot{y} < 0$. Therefore a sufficient condition for positive invariance is

$$f(0, y) \geq 0 \quad (y \geq 0), \qquad g(x, 0) \geq 0 \quad (x \geq 0).$$

In many biological models one has the stronger identities

$$f(0, y) = 0, \qquad g(x, 0) = 0,$$

so that the coordinate axes are themselves invariant.

Invariant regions also provide a priori control of solutions.

If a solution remains in a compact subset of the domain, then it can be continued for all forward time. In particular, this holds when the solution remains in a bounded positively invariant region whose closure is contained in the domain. Boundedness alone does not determine the eventual behavior, but it sharply restricts the possibilities.

Nullclines and invariant regions play different roles. Nullclines indicate where one component of the velocity vanishes and divide the phase plane according to the direction of motion. Invariant regions constrain the orbit itself. Together with equilibria and linearization, they form the basic toolkit for constructing nonlinear phase portraits.

Exercise 4.3 (An invariant triangle)**.** Let

$$D = \{(x, y) : x \geq 0,\ y \geq 0,\ x + y \leq 1\}.$$

Show that D is positively invariant for $\dot{x} = f(x, y)$, $\dot{y} = g(x, y)$ if

$$f(0, y) \geq 0, \qquad g(x, 0) \geq 0, \qquad f + g \leq 0 \quad \text{on } x + y = 1.$$

Interpret each condition geometrically.

Invariant curves. Invariant curves were introduced earlier in connection with the stable and unstable curves of a saddle. Here we emphasize their global role: they constrain the motion and may divide the phase plane into dynamically distinct regions or form boundaries of invariant regions.

Several important invariant curves have already appeared. The stable and unstable curves of a hyperbolic saddle are invariant. Coordinate axes may also be invariant, as in many population models where $x = 0$ or $y = 0$ represents extinction of one species. In conservative systems, level curves of the energy provide another important class of invariant curves.

Invariant curves may arise from the local geometry near equilibria, from symmetry, from special algebraic structure, or from conservation laws. A particularly systematic source is provided by first integrals.

First integrals. Consider the planar system

$$\dot{x} = f(x, y), \qquad \dot{y} = g(x, y).$$

A differentiable function $H(x, y)$ is called a *first integral* if it is constant along every solution:

$$H(x(t), y(t)) = \text{constant}.$$

Equivalently,

$$\frac{d}{dt} H(x(t), y(t)) = H_x(x, y) f(x, y) + H_y(x, y) g(x, y) = 0.$$

Thus the vector field is tangent to the level sets

$$H(x, y) = C.$$

Every regular level curve of a first integral is therefore an invariant curve.

A first integral can reveal much of the phase portrait without requiring an explicit formula for the solution. Once the level sets of H are known, the possible trajectories are confined to these curves. Closed level curves may correspond to periodic orbits, while critical level sets may pass through equilibria and form separatrices.

Conservative mechanical systems are a particularly rich source of first integrals. For

$$\ddot{x} + V'(x) = 0,$$

or

$$\dot{x} = y, \qquad \dot{y} = -V'(x),$$

the mechanical energy

$$E(x, y) = \frac{1}{2} y^2 + V(x)$$

satisfies

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

Thus every trajectory lies on an energy level

$$E(x, y) = C.$$

The shape of the potential V then determines much of the geometry of the phase portrait. Strict isolated local minima of V are surrounded locally by closed energy curves, while critical energy levels through saddle equilibria often form separatrices. The simple pendulum and the conservative Duffing oscillator, studied later, are basic examples of this structure.

First integrals also occur in nonmechanical systems. The classical Lotka–Volterra predator–prey model possesses a first integral whose level curves form a family of closed orbits in the positive quadrant. Thus first integrals provide a common geometric language for systems arising from quite different applications.

A simple closed regular level curve of a first integral is a periodic orbit provided that it contains no equilibrium. Such periodic orbits typically occur in continuous families of level sets. They should be distinguished from a *limit cycle*, which is an isolated periodic orbit.

This distinction is important. Conservative mechanical systems often possess families of closed energy curves, but these are not limit cycles. By contrast, the Van der Pol oscillator has an isolated attracting periodic orbit even though it has no conserved mechanical energy. Limit cycles will be studied more systematically later.

Critical level sets of a first integral are especially important. When such a level set contains a saddle equilibrium, it may include branches of the stable and unstable curves of that saddle and separate regions with qualitatively different motions. This leads naturally to the notion of a separatrix, discussed next.

Separatrices and connecting orbits. When stable or unstable curves separate regions with different qualitative behavior, they are called *separatrices*.

Separatrices are closely related to connecting orbits. A trajectory $\mathbf{y}(t)$ is called a *heteroclinic orbit* if it connects two distinct equilibria:

$$\lim_{t\to-\infty}\mathbf{y}(t)=\mathbf{y}_-,\qquad \lim_{t\to\infty}\mathbf{y}(t)=\mathbf{y}_+,\qquad \mathbf{y}_-\neq\mathbf{y}_+.$$

Thus a heteroclinic orbit lies in

$$W^u(\mathbf{y}_-)\cap W^s(\mathbf{y}_+).$$

A trajectory is called a *homoclinic orbit* if it leaves and returns to the same equilibrium:

$$\lim_{t\to-\infty}\mathbf{y}(t)=\mathbf{y}_*,\qquad \lim_{t\to\infty}\mathbf{y}(t)=\mathbf{y}_*.$$

Thus a homoclinic orbit lies in

$$W^u(\mathbf{y}_*)\cap W^s(\mathbf{y}_*).$$

In planar phase portraits, connecting orbits often involve saddle points, since saddles have one-dimensional stable and unstable curves.

Connecting orbits are exceptional trajectories, but they strongly organize the surrounding dynamics. They may form basin boundaries, delimit trapped regions, or mark transitions between different modes of motion. Because they are constructed from stable and unstable curves, their existence is a global question and cannot generally be decided from linearization alone.

Exercise 4.4 (Constructing a phase portrait)**.** Consider in the first quadrant

$$\dot{x}=x(2-x-y),\qquad \dot{y}=y(x-y).$$

Use nullclines, invariance, linearization, and boundedness to construct the phase portrait and describe the long-time behavior of interior solutions. Is the line $x=y$ invariant?

4.2 Interacting Population and Reaction Models

We now apply the phase-plane language to several elementary models. In each case, the equations are derived from simple assumptions, and the dynamics are studied through equilibria, nullclines, invariant regions, and special structures such as first integrals.

The examples illustrate several different forms of interaction: predation, competition, infection, and chemical reaction. Although the interpretations are different, the same geometric methods apply to all of them.

4.2.1 Predator–prey dynamics

Let $x(t)$ denote the prey population and $y(t)$ the predator population. The classical Lotka–Volterra predator–prey model is

$$\dot{x} = ax - bxy, \qquad \dot{y} = -cy + dxy,$$

where

$$a, b, c, d > 0.$$

In the absence of predators, the prey population grows exponentially:

$$\dot{x} = ax.$$

In the absence of prey, the predator population decays exponentially:

$$\dot{y} = -cy.$$

The product xy represents the frequency of encounters between prey and predators. Such encounters decrease the prey population and increase the predator population, giving the terms $-bxy$ and dxy.

The positive coexistence equilibrium is

$$(x_*, y_*) = \left(\frac{c}{d}, \frac{a}{b}\right).$$

We scale the populations by this equilibrium:

$$X = \frac{x}{x_*} = \frac{d}{c}x, \qquad Y = \frac{y}{y_*} = \frac{b}{a}y,$$

and scale time by

$$\tau = at.$$

The system becomes

$$\frac{dX}{d\tau} = X(1 - Y), \qquad \frac{dY}{d\tau} = \gamma Y(X - 1),$$

where

$$\gamma = \frac{c}{a}.$$

Thus the four dimensional parameters reduce to one dimensionless parameter.

Renaming (X, Y, τ) as (x, y, t), we obtain

$$\dot{x} = x(1 - y), \qquad \dot{y} = \gamma y(x - 1), \qquad \gamma > 0.$$

The nullclines are

$$x = 0, \qquad y = 1,$$

for the first equation, and

$$y = 0, \qquad x = 1,$$

for the second. Hence the equilibria in the closed first quadrant are

$$(0, 0), \qquad (1, 1).$$

The origin represents extinction of both species, while $(1,1)$ represents coexistence in the scaled variables.

The Jacobian matrix is

$$D\mathbf{f}(x,y) = \begin{bmatrix} 1-y & -x \\ \gamma y & \gamma(x-1) \end{bmatrix}.$$

At the origin,

$$D\mathbf{f}(0,0) = \begin{bmatrix} 1 & 0 \\ 0 & -\gamma \end{bmatrix},$$

whose eigenvalues are 1 and $-\gamma$. Thus $(0,0)$ is a saddle.

At the coexistence equilibrium,

$$D\mathbf{f}(1,1) = \begin{bmatrix} 0 & -1 \\ \gamma & 0 \end{bmatrix},$$

whose eigenvalues are

$$\lambda_{\pm} = \pm i\sqrt{\gamma}.$$

The equilibrium is therefore nonhyperbolic, and linearization alone does not determine the behavior of the nonlinear system near $(1,1)$. We will instead use the geometry of the vector field and a conserved quantity.

The coordinate axes are invariant, since

$$\dot{x} = 0 \quad \text{on } x = 0, \qquad \dot{y} = 0 \quad \text{on } y = 0.$$

Consequently, the closed first quadrant is positively invariant.

In the positive quadrant,

$$\dot{x} > 0 \iff y < 1, \qquad \dot{x} < 0 \iff y > 1,$$

while

$$\dot{y} > 0 \iff x > 1, \qquad \dot{y} < 0 \iff x < 1.$$

The motion is therefore counterclockwise around $(1,1)$. When predators are few, the prey population grows. The abundance of prey then causes the predator population to grow. As predators become numerous, the prey population declines, after which the predator population also declines.

The system possesses a first integral. Define

$$H(x,y) = \gamma(x - \log x) + y - \log y, \qquad x > 0, \quad y > 0.$$

Along a positive solution,

$$\begin{aligned} \frac{d}{dt}H(x(t),y(t)) &= \gamma\left(1 - \frac{1}{x}\right)\dot{x} + \left(1 - \frac{1}{y}\right)\dot{y} \\ &= \gamma(x-1)(1-y) + \gamma(y-1)(x-1) \\ &= 0. \end{aligned}$$

Hence

$$H(x(t),y(t)) = \text{constant},$$

and every positive orbit lies on a level curve of H.

The function H has a unique critical point at $(1,1)$. Moreover,

$$H_{xx} = \frac{\gamma}{x^2} > 0, \qquad H_{yy} = \frac{1}{y^2} > 0,$$

so $(1,1)$ is its strict minimum. Also,

$$H(x,y) \longrightarrow \infty$$

as $x \to 0^+$, $y \to 0^+$, $x \to \infty$, or $y \to \infty$. Therefore every nonminimum level set of H in the positive quadrant is a closed curve surrounding $(1,1)$.

Since the vector field does not vanish on such a level curve, every positive non-equilibrium solution is periodic. Thus the classical Lotka–Volterra model predicts persistent population oscillations rather than convergence to the coexistence equilibrium. In particular, the conserved quantity confirms that the nonhyperbolic equilibrium $(1,1)$ is a center, a conclusion that could not be obtained from linearization alone.

For the representative case $\gamma = 1$,

$$\dot{x} = x(1-y), \qquad \dot{y} = y(x-1),$$

and

$$H(x,y) = x - \log x + y - \log y.$$

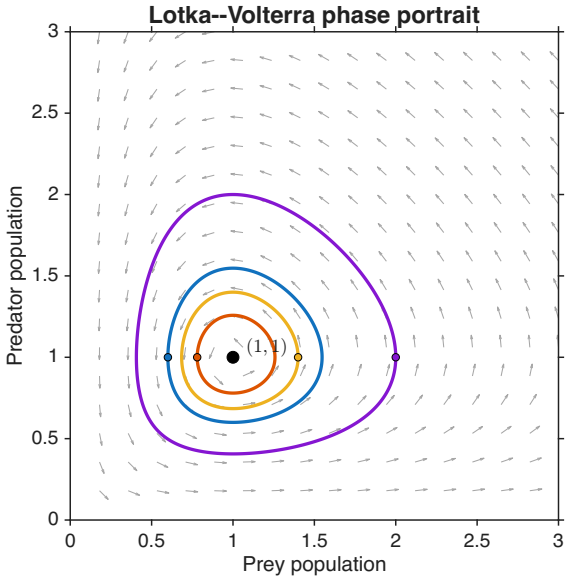


Figure 4.1: Phase portrait of the normalized Lotka–Volterra predator–prey system with $\gamma = 1$. The orbits in the positive quadrant are closed level curves of $H(x,y) = x - \log x + y - \log y$ surrounding the coexistence equilibrium $(1,1)$.

Prey self-limitation. The classical Lotka–Volterra model assumes unlimited prey growth in the absence of predators. Replacing exponential growth by logistic growth gives

$$\dot{x} = x(1-x) - axy, \qquad \dot{y} = -by + cxy,$$

where $a, b, c > 0$.

The nontrivial nullclines are

$$1 - x - ay = 0, \qquad x = \frac{b}{c}.$$

Hence a positive coexistence equilibrium exists precisely when

$$\frac{b}{c} < 1,$$

and is given by

$$x_* = \frac{b}{c}, \qquad y_* = \frac{1 - \frac{b}{c}}{a}.$$

The condition $b/c < 1$ means that the prey population required to sustain the predators lies below the prey carrying capacity.

To determine the local behavior near the coexistence equilibrium, the Jacobian matrix is

$$D\mathbf{f}(x,y) = \begin{bmatrix} 1 - 2x - ay & -ax \\ cy & -b + cx \end{bmatrix}.$$

At (x_*, y_*), the equilibrium relations

$$-b + cx_* = 0, \qquad 1 - x_* - ay_* = 0$$

give

$$D\mathbf{f}(x_*, y_*) = \begin{bmatrix} -x_* & -ax_* \\ cy_* & 0 \end{bmatrix}.$$

Its trace and determinant are

$$T = -x_* < 0, \qquad D = acx_*y_* > 0.$$

Thus the coexistence equilibrium is hyperbolic and locally asymptotically stable. More precisely, if

$$\Delta = T^2 - 4D,$$

it is a stable node when $\Delta > 0$ and a stable spiral when $\Delta < 0$. When $\Delta = 0$, the linearization has a repeated negative eigenvalue and the equilibrium remains locally asymptotically stable.

The logistic term therefore changes the local behavior near the positive coexistence equilibrium. In the classical Lotka–Volterra model, this equilibrium is a nonlinear center surrounded by a family of closed orbits. With prey self-limitation, it becomes locally asymptotically stable. Depending on the parameters, nearby solutions approach it either without oscillation or through damped oscillations.

This comparison illustrates an important modeling principle: a seemingly small change in the growth law can change the phase portrait from persistent conservative oscillations to dissipative convergence.

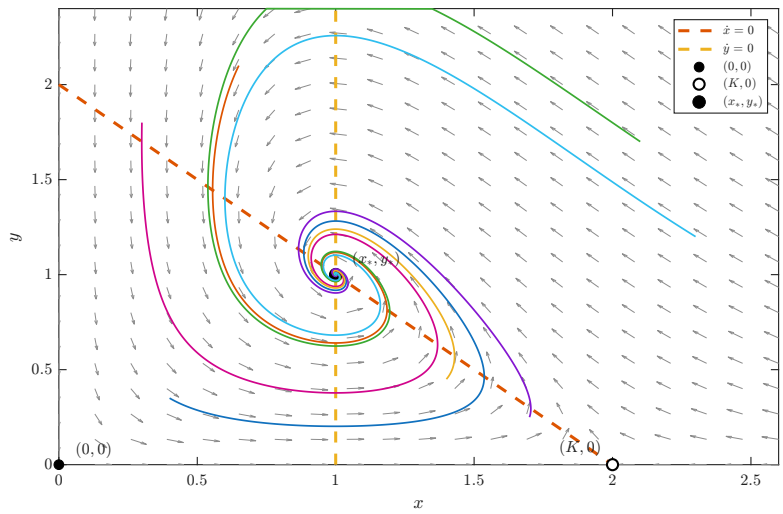


Figure 4.2: Phase portrait of the predator–prey model with prey self-limitation $\dot{x} = x\left(1 - \frac{x}{K}\right) - axy, \quad \dot{y} = -by + cxy$, for the parameter values $K = 2, a = \frac{1}{2}, b = 1, c = 1$. The nontrivial nullclines are $\dot{x} = 0 : y = 2 - x$, and $\dot{y} = 0 : x = 1$. Their intersection gives the positive coexistence equilibrium $(x_*, y_*) = (1, 1)$. The other equilibria are $(0, 0)$ and $(2, 0)$. The displayed trajectories approach the coexistence equilibrium, illustrating how prey self-limitation destroys the conservative family of closed orbits of the classical Lotka–Volterra system.

4.2.2 Competition models

We next consider two species competing for the same limited resources. Let $X(t)$ and $Y(t)$ denote their populations. In the absence of the other species, each population is assumed to grow logistically. A standard model is

$$\dot{X} = r_1 X \left(1 - \frac{X + \alpha_{12} Y}{K_1}\right), \qquad \dot{Y} = r_2 Y \left(1 - \frac{Y + \alpha_{21} X}{K_2}\right),$$

where $r_1, r_2 > 0$ are intrinsic growth rates, $K_1, K_2 > 0$ are carrying capacities, and $\alpha_{12}, \alpha_{21} > 0$ measure interspecies competition.

We scale the populations by their carrying capacities:

$$x = \frac{X}{K_1}, \qquad y = \frac{Y}{K_2}.$$

Then

$$\dot{x} = r_1 x(1 - x - s_1 y), \qquad \dot{y} = r_2 y(1 - y - s_2 x),$$

where

$$s_1 = \frac{\alpha_{12} K_2}{K_1}, \qquad s_2 = \frac{\alpha_{21} K_1}{K_2}.$$

The parameter s_1 measures the competitive effect of species y on species x, relative to the self-limitation of species x. Similarly, s_2 measures the competitive effect of species x on species y. The terms $-x^2$ and $-y^2$ represent competition within each species, while $-s_1 xy$ and $-s_2 xy$ represent competition between the two species.

Scaling time by $\tau = r_1 t$ gives

$$\frac{dx}{d\tau} = x(1 - x - s_1 y), \qquad \frac{dy}{d\tau} = \rho y(1 - y - s_2 x), \qquad \rho = \frac{r_2}{r_1} > 0.$$

Renaming τ as t, we study

$$\dot{x} = x(1 - x - s_1 y), \qquad \dot{y} = \rho y(1 - y - s_2 x).$$

The closed first quadrant

$$Q = \{(x, y) : x \geq 0,\ y \geq 0\}$$

is positively invariant. Indeed,

$$\dot{x} = 0 \quad \text{on } x = 0, \qquad \dot{y} = 0 \quad \text{on } y = 0.$$

Thus nonnegative initial populations remain nonnegative.

The nullclines are

$$\dot{x} = 0 \quad \Longleftrightarrow \quad x = 0 \quad \text{or} \quad x + s_1 y = 1,$$

and

$$\dot{y} = 0 \quad \Longleftrightarrow \quad y = 0 \quad \text{or} \quad s_2 x + y = 1.$$

The two nontrivial nullclines are therefore the straight lines

$$x + s_1 y = 1, \qquad s_2 x + y = 1.$$

Their intercepts are

$$x + s_1 y = 1 : \qquad (1, 0), \quad \left(0, \frac{1}{s_1}\right),$$

and

$$s_2 x + y = 1 : \qquad \left(\frac{1}{s_2}, 0\right), \quad (0, 1).$$

The boundary equilibria are

$$(0, 0), \qquad (1, 0), \qquad (0, 1).$$

The equilibrium $(1, 0)$ represents survival of species x alone, while $(0, 1)$ represents survival of species y alone.

An interior coexistence equilibrium is obtained by solving

$$x + s_1 y = 1, \qquad s_2 x + y = 1.$$

If $s_1 s_2 \neq 1$, the solution is

$$x_* = \frac{1 - s_1}{1 - s_1 s_2}, \qquad y_* = \frac{1 - s_2}{1 - s_1 s_2}.$$

It is biologically meaningful only when

$$x_* > 0, \qquad y_* > 0.$$

We can now combine the nullcline geometry with linearization. The Jacobian matrix is

$$D\mathbf{f}(x,y) = \begin{bmatrix} 1-2x-s_1y & -s_1x \\ -\rho s_2 y & \rho(1-2y-s_2x) \end{bmatrix}.$$

At $(1,0)$,

$$D\mathbf{f}(1,0) = \begin{bmatrix} -1 & -s_1 \\ 0 & \rho(1-s_2) \end{bmatrix},$$

so the eigenvalues are

$$-1, \qquad \rho(1-s_2).$$

Thus $(1,0)$ is locally asymptotically stable when $s_2 > 1$ and is a saddle when $s_2 < 1$. Similarly,

$$D\mathbf{f}(0,1) = \begin{bmatrix} 1-s_1 & 0 \\ -\rho s_2 & -\rho \end{bmatrix},$$

whose eigenvalues are

$$1-s_1, \qquad -\rho.$$

Hence $(0,1)$ is locally asymptotically stable when $s_1 > 1$ and is a saddle when $s_1 < 1$.

These calculations have a direct biological interpretation. The sign of $1-s_2$ determines whether species y can invade the state $(1,0)$ in which only species x survives, while the sign of $1-s_1$ determines whether species x can invade the state $(0,1)$ in which only species y survives.

At a positive coexistence equilibrium, the equilibrium relations

$$1-x_*-s_1y_* = 0, \qquad 1-y_*-s_2x_* = 0$$

simplify the Jacobian to

$$D\mathbf{f}(x_*,y_*) = \begin{bmatrix} -x_* & -s_1x_* \\ -\rho s_2 y_* & -\rho y_* \end{bmatrix}.$$

Its trace and determinant are

$$T = -x_* - \rho y_* < 0$$

and

$$D = \rho x_* y_* (1-s_1s_2).$$

These formulas, together with the relative positions of the nullclines, lead to four qualitatively different cases.

Case 1: $s_1 < 1$, $s_2 < 1$, weak competition. Suppose

$$s_1 < 1, \qquad s_2 < 1.$$

Then

$$1 - s_1s_2 > 0,$$

and the coexistence equilibrium lies in the positive quadrant:

$$x_* = \frac{1-s_1}{1-s_1s_2} > 0, \qquad y_* = \frac{1-s_2}{1-s_1s_2} > 0.$$

Moreover,

$$T < 0, \qquad D > 0,$$

so the coexistence equilibrium is locally asymptotically stable. Each species limits itself more strongly than it limits the other species, and both species can invade when rare. The phase portrait indicates convergence toward coexistence.

Case 2: $s_1 < 1$, $s_2 > 1$, species x excludes species y. Suppose

$$s_1 < 1, \qquad s_2 > 1.$$

There is no positive coexistence equilibrium. Since $s_1 < 1$, the equilibrium $(0, 1)$ is a saddle, while $s_2 > 1$ implies that $(1, 0)$ is locally asymptotically stable.

The same conclusion can be read directly from the invasion conditions. Near $(0, 1)$, with $x > 0$ small,

$$\dot{x} = x(1 - x - s_1 y) \approx x(1 - s_1) > 0,$$

so species x can invade a population consisting almost entirely of species y. Near $(1, 0)$, with $y > 0$ small,

$$\dot{y} = \rho y(1 - y - s_2 x) \approx \rho y(1 - s_2) < 0,$$

so species y cannot invade a population consisting almost entirely of species x. The attracting boundary equilibrium is therefore

$$(1, 0).$$

Species x survives, while species y becomes extinct.

Case 3: $s_1 > 1$, $s_2 < 1$, species y excludes species x. Suppose

$$s_1 > 1, \qquad s_2 < 1.$$

Again there is no positive coexistence equilibrium. Now $(1, 0)$ is a saddle, while $(0, 1)$ is locally asymptotically stable.

Near $(1, 0)$, with $y > 0$ small,

$$\dot{y} \approx \rho y(1 - s_2) > 0,$$

so species y can invade. Near $(0, 1)$, with $x > 0$ small,

$$\dot{x} \approx x(1 - s_1) < 0,$$

so species x cannot invade. The attracting boundary equilibrium is

$$(0, 1).$$

Species y survives, while species x becomes extinct.

Case 4: $s_1 > 1$, $s_2 > 1$, strong competition. Suppose

$$s_1 > 1, \qquad s_2 > 1.$$

Then

$$1 - s_1 s_2 < 0,$$

and the coexistence equilibrium again lies in the positive quadrant. At this equilibrium,

$$D = \rho x_* y_* (1 - s_1 s_2) < 0,$$

so it is a hyperbolic saddle. On the other hand, the boundary equilibria

$$(1,0), \qquad (0,1)$$

are both locally asymptotically stable.

The stable curve of the coexistence saddle is a separatrix. It separates the first quadrant into two basins of attraction: initial data on one side approach $(1,0)$, while initial data on the other side approach $(0,1)$. Initial data lying on the stable curve itself approach the coexistence saddle. Thus the stable curve is the boundary between two different long-time outcomes.

The surviving species therefore depends on the initial populations. This behavior is called *bistability*. In Figure 4.4(d), the stable curve of the interior saddle appears as the basin boundary separating the two attracting boundary equilibria.

Case 5: boundary cases. The parameter values $s_1 = 1$ or $s_2 = 1$ separate the four generic competition regimes. If $s_1 = 1$ and $s_2 \neq 1$, the coexistence equilibrium collides with $(0,1)$, which becomes nonhyperbolic. Similarly, if $s_2 = 1$ and $s_1 \neq 1$, the coexistence equilibrium collides with $(1,0)$.

If

$$s_1 = s_2 = 1,$$

the two nontrivial nullclines coincide:

$$x + y = 1.$$

Every point of this line segment in the first quadrant is then an equilibrium. Thus the equality cases are degenerate transition cases between the four generic phase portraits.

Summary. The competition model has four generic regimes, determined by the invasion conditions, together with degenerate boundary cases when $s_1 = 1$ or $s_2 = 1$. Species x can invade $(0,1)$ precisely when $s_1 < 1$, while species y can invade $(1,0)$ precisely when $s_2 < 1$:

	species x invades	species y invades
$s_1 < 1,\ s_2 < 1$	yes	yes
$s_1 < 1,\ s_2 > 1$	yes	no
$s_1 > 1,\ s_2 < 1$	no	yes
$s_1 > 1,\ s_2 > 1$	no	no

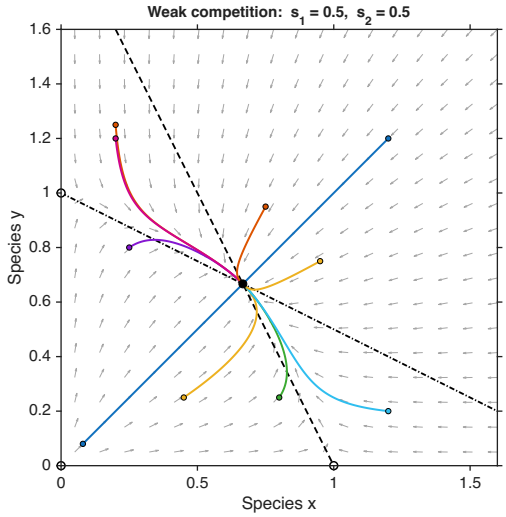
Mutual invasion leads to coexistence, one-sided invasion to competitive exclusion, and failure of both species to invade to bistability. The equality cases $s_1 = 1$ or $s_2 = 1$ form the nongeneric boundaries between these four phase portraits.

Exercise 4.5 (The strong-competition separatrix)**.** For

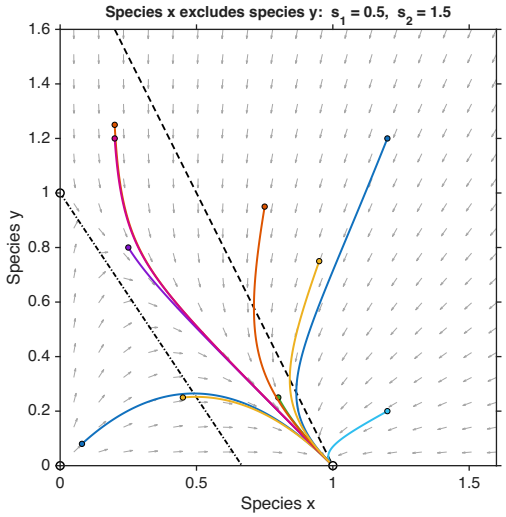
$$\dot{x} = x(1 - x - s_1 y), \qquad \dot{y} = \rho y(1 - y - s_2 x), \qquad s_1, s_2 > 1,$$

let (x_*, y_*) be the coexistence saddle and $\lambda_s < 0$ its stable eigenvalue. Show that a stable eigenvector $(1, m_s)^T$ has

$$m_s = -\frac{x_* + \lambda_s}{s_1 x_*}.$$

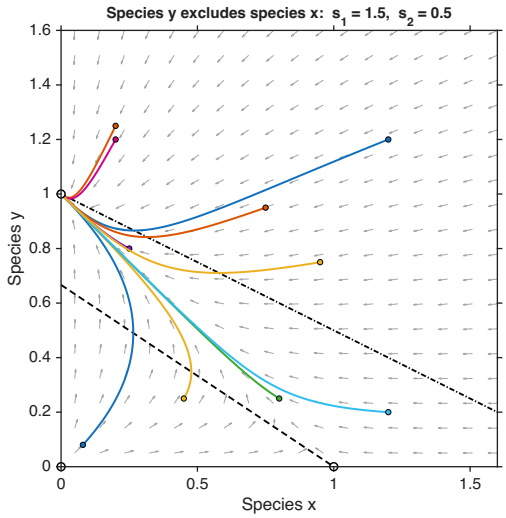


(a) $s_1 < 1$, $s_2 < 1$: coexistence

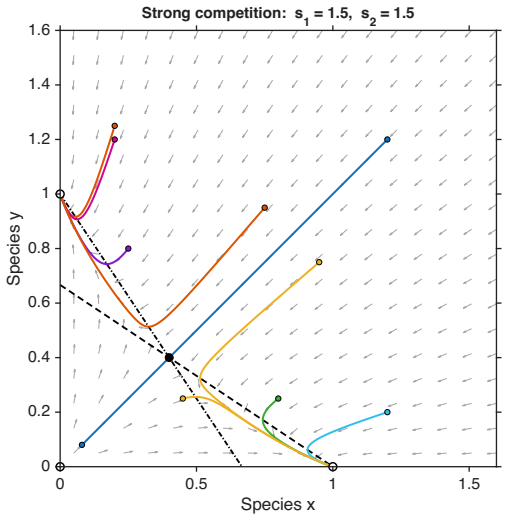


(b) $s_1 < 1$, $s_2 > 1$: species x excludes species y

Figure 4.3: Phase portraits of the competition model in the cases $s_1 < 1$. The dark dashed lines are the nullclines. When $s_2 < 1$, both species can invade and the coexistence equilibrium is attracting. When $s_2 > 1$, only species x can invade, and the equilibrium $(1, 0)$ is attracting.



(c) $s_1 > 1$, $s_2 < 1$: species y excludes species x



(d) $s_1 > 1$, $s_2 > 1$: bistability

Figure 4.4: Phase portraits of the competition model in the cases $s_1 > 1$. The dark dashed lines are the nullclines. When $s_2 < 1$, only species y can invade, and the equilibrium $(0, 1)$ is attracting. When $s_2 > 1$, neither species can invade the other, and the system is bistable. In panel (d), for the symmetric parameter choice $s_1 = s_2 = 1.5$, the blue line $x = y$ is the separatrix between the two basins of attraction.

Hence find the tangent to the separatrix at (x_*, y_*). When $s_1 = s_2$, determine when the separatrix is the diagonal $y = x$. Otherwise, derive its differential equation and explain why it separates the basins of $(1, 0)$ and $(0, 1)$.

Exercise 4.6 (A continuum of equilibria)**.** This is the competition model with $s_1 = s_2 = 1$ and $\rho = 1$:

$$\dot{x} = x(1 - x - y), \qquad \dot{y} = y(1 - x - y), \qquad x(0), y(0) > 0.$$

Show that x/y is constant and that $z = x + y$ satisfies a logistic equation. Find the limiting point on $x + y = 1$ in terms of the initial data, and sketch the invariant rays and the line of equilibria. Interpret biologically why neither species gains a competitive advantage and why the final population shares depend on the initial proportions.

4.2.3 Epidemic models

Epidemic models provide a natural example of interacting populations, where individuals move between compartments according to their disease status.

The SIR model. Let $S(t)$ denote the susceptible population, $I(t)$ the infected population, and $R(t)$ the removed population. The removed class includes individuals who have recovered, been isolated, or otherwise ceased to participate in transmission.

The classical SIR model is

$$\dot{S} = -\beta SI, \qquad \dot{I} = \beta SI - \gamma I, \qquad \dot{R} = \gamma I,$$

where $\beta > 0$ is the transmission parameter and $\gamma > 0$ is the removal rate. The mean time spent in the infected class is therefore $1/\gamma$.

The transmission term βSI is based on a mass-action assumption: the number of new infections is proportional to the number of encounters between susceptible and infected individuals. Each such infection removes one individual from the susceptible class and adds one individual to the infected class. Infected individuals leave the infected class at rate γI.

Conservation law and invariant region. The total population is conserved:

$$\frac{d}{dt}(S + I + R) = 0.$$

Thus, if

$$S(0) + I(0) + R(0) = N,$$

then

$$S(t) + I(t) + R(t) = N$$

for all time. Since

$$R = N - S - I,$$

the essential dynamics take place in the (S, I)-plane:

$$\dot{S} = -\beta SI, \qquad \dot{I} = I(\beta S - \gamma).$$

The physically meaningful region is the triangle

$$D = \{(S, I) : S \geq 0,\ I \geq 0,\ S + I \leq N\}.$$

This region is positively invariant. On the boundary $S = 0$,

$$\dot{S} = 0,$$

and on the boundary $I = 0$,

$$\dot{I} = 0.$$

Along the third boundary $S + I = N$,

$$\frac{d}{dt}(S + I) = -\gamma I \leq 0.$$

Hence the vector field is tangent to the boundary or points into the triangle. Solutions starting with nonnegative susceptible, infected, and removed populations therefore remain physically meaningful.

The nullclines and the epidemic threshold. The nullclines are

$$\dot{S} = 0 \quad \Longleftrightarrow \quad S = 0 \quad \text{or} \quad I = 0,$$

and

$$\dot{I} = 0 \quad \Longleftrightarrow \quad I = 0 \quad \text{or} \quad S = \frac{\gamma}{\beta}.$$

The vertical line

$$S = \frac{\gamma}{\beta}$$

is the *epidemic threshold*. Indeed,

$$\dot{I} > 0 \quad \Longleftrightarrow \quad S > \frac{\gamma}{\beta},$$

while

$$\dot{I} < 0 \quad \Longleftrightarrow \quad S < \frac{\gamma}{\beta}.$$

Thus the infected population grows while the susceptible population is above the threshold and decreases after the susceptible population falls below it.

The reproduction number $\mathcal{R}_0$. It is customary to introduce the basic reproduction number

$$\mathcal{R}_0 = \frac{\beta N}{\gamma}.$$

This is the reproduction number for a fully susceptible population of size N. At a general time t, when only $S(t)$ individuals remain susceptible, the corresponding effective reproduction number is

$$\mathcal{R}_{\text{eff}}(t) = \frac{\beta S(t)}{\gamma}.$$

Since

$$\dot{I} = I\big(\beta S - \gamma\big) = \gamma I\big(\mathcal{R}_{\text{eff}} - 1\big),$$

we have

$$\dot{I} > 0 \quad \Longleftrightarrow \quad \mathcal{R}_{\text{eff}}(t) > 1.$$

In particular, the infected population initially grows precisely when

$$\frac{\beta S(0)}{\gamma} > 1.$$

Thus $\mathcal{R}_0 > 1$ indicates the potential for an epidemic outbreak in a fully susceptible population; for a partially susceptible initial population, the relevant condition is

$$\mathcal{R}_{\text{eff}}(0) > 1.$$

The equilibria. The equilibria of the reduced system satisfy

$$-\beta SI = 0, \qquad I(\beta S - \gamma) = 0.$$

Hence every point on the line

$$I = 0$$

is an equilibrium. This continuum of equilibria represents disease-free states. The value of S records the susceptible population remaining after the epidemic has ended.

The direction of motion is particularly simple. Since

$$\dot{S} = -\beta SI \leq 0,$$

the susceptible population is nonincreasing. Whenever $I > 0$, the trajectory moves strictly to the left in the phase plane. If initially

$$S(0) > \frac{\gamma}{\beta},$$

the infected population first increases. It reaches its maximum when the trajectory crosses

$$S = \frac{\gamma}{\beta},$$

and then decreases. Thus the phase portrait explains the characteristic rise and fall of an epidemic without requiring explicit formulas for $S(t)$ and $I(t)$.

First integral. The reduced system also possesses a first integral. For $S > 0$ and $I > 0$,

$$\frac{dI}{dS} = \frac{I(\beta S - \gamma)}{-\beta SI} = -1 + \frac{\gamma}{\beta S}.$$

Hence

$$\frac{d}{dS}(S + I) = \frac{\gamma}{\beta S},$$

and integration gives

$$S + I - \frac{\gamma}{\beta} \log S = C.$$

Thus the phase curves are level curves of

$$H(S, I) = S + I - \frac{\gamma}{\beta} \log S.$$

These level curves are not closed. In contrast with the Lotka–Volterra model, the susceptible population is monotone, so the motion cannot return to its initial state. Instead, a solution with $I(0) > 0$ approaches a point

$$(S_\infty, 0)$$

on the disease-free line.

The limiting susceptible population is determined implicitly by the first integral:

$$S_\infty - \frac{\gamma}{\beta} \log S_\infty = S_0 + I_0 - \frac{\gamma}{\beta} \log S_0.$$

Equivalently,

$$S_\infty - S_0 - I_0 = \frac{\gamma}{\beta} \log \left(\frac{S_\infty}{S_0} \right).$$

This relation is called the final-size relation. For an outbreak that begins above the transmission threshold, this relation shows that the epidemic ends after the susceptible population has fallen below the threshold, rather than because every susceptible individual has become infected. If S_∞ is nevertheless close to zero, then nearly the entire population has been infected during the course of the epidemic.

The SIR model therefore exhibits a form of threshold dynamics different from the predator–prey and competition models. There is no positive endemic equilibrium in the model without births and deaths. Instead, every epidemic trajectory approaches a disease-free equilibrium whose location depends on the initial data.

Exercise 4.7 (SIR model with demography)**.** For

$$\dot{S} = \mu N - \beta SI - \mu S, \qquad \dot{I} = \beta SI - (\gamma + \mu) I,$$

find the disease-free and endemic equilibria and determine the threshold for persistence of infection in $S, I \geq 0$, $S + I \leq N$.

4.2.4 Chemical reaction systems–the Brusselator

Chemical reaction systems provide another natural source of nonlinear differential equations. The state variables are concentrations, and the nonlinear terms arise from the law of mass action.

Mass-action kinetics. For an elementary reaction, the reaction rate is proportional to the product of the concentrations of the reactants, with each concentration raised to its stoichiometric coefficient. Thus

$$X + Y \longrightarrow \text{products}$$

has a rate proportional to xy, while

$$2X + Y \longrightarrow \text{products}$$

has a rate proportional to $x^2 y$.

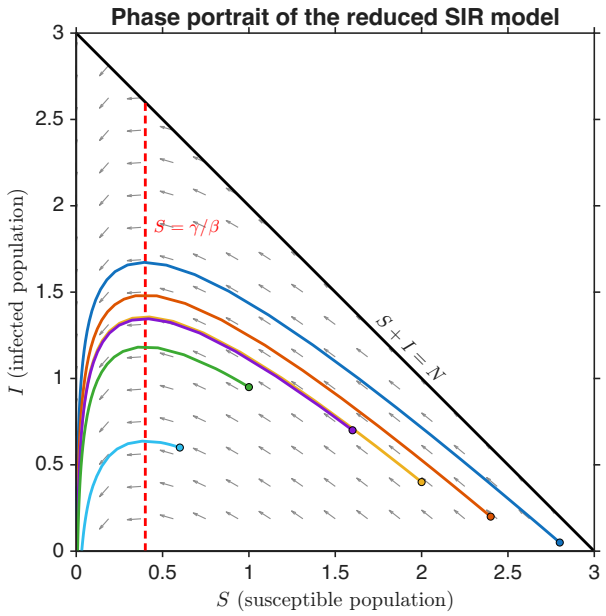


Figure 4.5: Phase portrait of the reduced SIR system in the (S, I)-plane, with $N = 3$ and $\gamma/\beta = 0.4$. The physically relevant region is $S \geq 0$, $I \geq 0$, and $S + I \leq N$, since $R = N - S - I \geq 0$. The dashed line $S = \gamma/\beta$ separates the region where the infected population grows from the region where it decays.

The Brusselator. We consider the Brusselator, a simple model of an open chemical reaction system. The term *open* means that some species are supplied from external reservoirs and some products are removed. A standard reaction scheme is

$$\begin{aligned} A &\longrightarrow X, \\ B + X &\longrightarrow Y + D, \\ 2X + Y &\longrightarrow 3X, \\ X &\longrightarrow E. \end{aligned}$$

The concentrations of A and B are maintained at fixed values, while the products D and E are removed. The dynamic variables are therefore the concentrations

$$x(t) = [X], \qquad y(t) = [Y].$$

We derive the model by adding the contributions of the four reactions.

The reaction

$$A \longrightarrow X$$

supplies X at a constant rate. After nondimensionalization, its contribution to $\dot{x}$ is

$$+A.$$

The reaction

$$B + X \longrightarrow Y + D$$

consumes one molecule of X and produces one molecule of Y. Since the concentration of B is fixed, its mass-action rate is proportional to x. In normalized variables, the rate is Bx, so the contribution is

$$-Bx \quad \text{to } \dot{x}, \qquad +Bx \quad \text{to } \dot{y}.$$

The autocatalytic reaction

$$2X + Y \longrightarrow 3X$$

has mass-action rate x^2y. Its net effect is to produce one molecule of X and consume one molecule of Y. Hence it contributes

$$+x^2y \quad \text{to } \dot{x}, \qquad -x^2y \quad \text{to } \dot{y}.$$

Finally,

$$X \longrightarrow E$$

removes X at a rate proportional to x, contributing $-x$ to $\dot{x}$.

Adding these contributions gives the Brusselator system

$$\begin{cases} \dot{x} = A - (B+1)x + x^2y, \\ \dot{y} = Bx - x^2y, \end{cases}$$

where

$$A > 0, \qquad B > 0.$$

The parameter A represents the external feed of X, while B controls the conversion of X into Y. The nonlinear term x^2y represents autocatalytic production of X.

Invariant region. The closed first quadrant

$$Q = \{(x, y) : x \geq 0,\ y \geq 0\}$$

is positively invariant. Along $x = 0$,

$$\dot{x} = A > 0,$$

so the vector field points into the first quadrant. Along $y = 0$,

$$\dot{y} = Bx \geq 0.$$

Thus solutions beginning with nonnegative concentrations remain physically meaningful.

Nullclines. The nullclines are obtained from

$$A - (B+1)x + x^2y = 0$$

and

$$Bx - x^2y = 0.$$

For $x > 0$, they are

$$y = \frac{(B+1)x - A}{x^2}$$

and

$$y = \frac{B}{x}.$$

The signs of the derivatives can be read directly from the nullclines. Since

$$\dot{y} = x(B - xy),$$

we have

$$\dot{y} > 0 \quad \Longleftrightarrow \quad y < \frac{B}{x},$$

and

$$\dot{y} < 0 \quad \Longleftrightarrow \quad y > \frac{B}{x}.$$

Similarly,

$$\dot{x} > 0 \quad \Longleftrightarrow \quad y > \frac{(B+1)x - A}{x^2},$$

while

$$\dot{x} < 0 \quad \Longleftrightarrow \quad y < \frac{(B+1)x - A}{x^2}.$$

Thus the two nullclines divide the positive quadrant into regions in which the concentrations increase or decrease. Their geometry already suggests a rotational motion around the positive equilibrium.

The equilibrium. The equilibrium is the intersection of the two nullclines. Since $A > 0$, an equilibrium cannot have $x = 0$. From

$$Bx - x^2 y = 0$$

we obtain

$$y = \frac{B}{x}.$$

Substitution into the first equation gives

$$A - (B+1)x + Bx = 0,$$

and therefore

$$x = A.$$

Hence the unique positive equilibrium is

$$(x_*, y_*) = \left(A, \frac{B}{A}\right).$$

Linearization. To determine its local behavior, we linearize the system. For

$$f(x,y) = A - (B+1)x + x^2 y, \qquad g(x,y) = Bx - x^2 y,$$

the Jacobian matrix is

$$D\mathbf{f}(x,y) = \begin{pmatrix} -(B+1) + 2xy & x^2 \\ B - 2xy & -x^2 \end{pmatrix}.$$

At the equilibrium $(A, B/A)$,

$$D\mathbf{f}\left(A, \frac{B}{A}\right) = \begin{pmatrix} B-1 & A^2 \\ -B & -A^2 \end{pmatrix}.$$

Its trace and determinant are

$$T = B - 1 - A^2, \qquad D = A^2 > 0.$$

Therefore the eigenvalues satisfy

$$\lambda^2 - (B - 1 - A^2)\lambda + A^2 = 0,$$

and hence

$$\lambda_\pm = \frac{B - 1 - A^2 \pm \sqrt{(B - 1 - A^2)^2 - 4A^2}}{2}.$$

The sign of the trace determines the stability of the equilibrium.

- If $B < 1 + A^2$, both eigenvalues have negative real part, and the equilibrium is locally stable.
- If $B > 1 + A^2$, both eigenvalues have positive real part, and the equilibrium is locally unstable.
- At $B = 1 + A^2$, the eigenvalues are $\lambda_\pm = \pm iA$, so the equilibrium is nonhyperbolic and linearization alone no longer determines its nonlinear behavior.

In the Brusselator, this critical value marks the onset of oscillatory behavior and is associated with a Hopf bifurcation. The mechanism and the resulting periodic orbit will be studied systematically in Chapter 11.

Limit cycle. As B increases through $1 + A^2$, the equilibrium loses stability, and numerical phase portraits indicate the emergence of a stable periodic orbit. The concentrations then do not settle to constant values. Instead, they undergo sustained oscillations caused by the interaction of feed, removal, conversion, and autocatalytic feedback.

The periodic orbit in the Brusselator differs fundamentally from the closed orbits of the Lotka–Volterra model. In the Lotka–Volterra system, the closed orbits form a continuous family of level curves of a first integral. In the Brusselator, the periodic orbit is isolated and attracts nearby trajectories. An isolated periodic orbit is called a *limit cycle*. Thus the Brusselator possesses an attracting limit cycle, rather than a conservative family of periodic orbits.

This example illustrates how sustained oscillations can arise in an open dissipative system. Material is continually supplied and removed, while the nonlinear reaction mechanism organizes the motion into a persistent cycle. The existence, stability, and bifurcation of limit cycles will be studied systematically in Chapter 11.

Exercise 4.8 (Generalized autocatalysis)**.** Replace $2X + Y \to 3X$ in the Brusselator by $mX + Y \to (m + 1)X$, $m \geq 2$. Derive the mass-action system and its positive equilibrium. Compute the Jacobian trace and determinant there, determine the stability threshold in B, and compare it with the case $m = 2$.

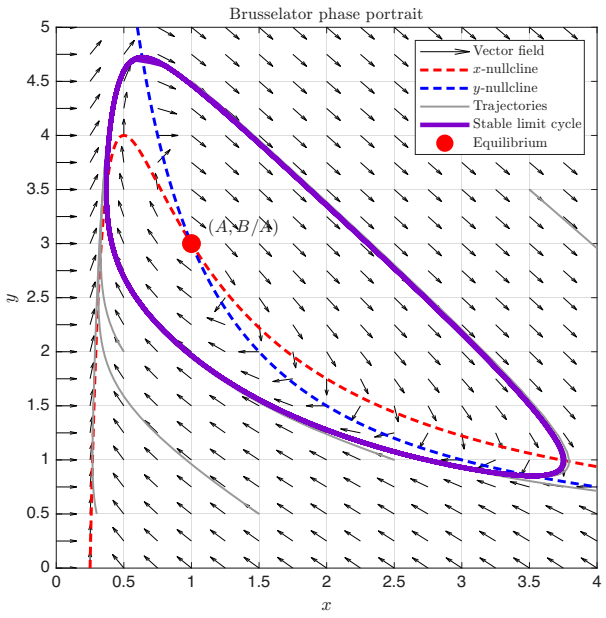


Figure 4.6: Phase portrait of the Brusselator with $A = 1$ and $B = 3$. The positive equilibrium $(1, 3)$ is unstable, and nearby trajectories approach an attracting periodic orbit.

Exercise 4.9 (Material balance on a periodic orbit)**.** For the Brusselator,

$$\dot{x} = A - (B+1)x + x^2 y, \qquad \dot{y} = Bx - x^2 y,$$

let $(x(t), y(t))$ be T-periodic. Integrate the equations over one period to show that

$$\frac{1}{T}\int_0^T x(t)\,dt = A, \qquad \frac{1}{T}\int_0^T x(t)^2 y(t)\,dt = AB.$$

Interpret these identities as mean material balances.

Summary of the modeling examples. The models in this section arise from different applications, but their nonlinearities have a common origin: interaction. In the predator–prey and competition systems, products such as xy represent encounters between populations. In the SIR model, βSI represents transmission through contact. In the Brusselator, $x^2 y$ arises from mass-action kinetics for an autocatalytic reaction.

The same phase-plane tools apply throughout. Equilibria describe steady states, nullclines determine directional regions, and invariant sets enforce physical constraints such as positivity. Nevertheless, the resulting dynamics are quite different: conservative periodic motion, convergence to a coexistence state, competitive exclusion, bistability, epidemic thresholds, and attracting chemical oscillations.

The next section organizes these examples according to their underlying dynamical structure: conservative, dissipative, and oscillatory systems.

4.3 Conservative, Dissipative, and Oscillatory Systems

The phase-plane method provides a natural framework for nonlinear oscillators. If $x(t)$ denotes position and

$$y = \dot{x},$$

then a second-order equation

$$\ddot{x} = F(x, \dot{x})$$

becomes the planar system

$$\dot{x} = y, \qquad \dot{y} = F(x, y).$$

The state of the oscillator is therefore represented by a point (x, y) in the phase plane, and its motion by an orbit.

Periodic motion and closed orbits have already appeared in the preceding examples. For oscillators, however, such behavior can arise through rather different mechanisms. In a conservative mechanical system there is often an energy $E(x, y)$ satisfying

$$\frac{dE}{dt} = 0,$$

so that trajectories are confined to the level sets $E(x, y) = C$. Closed regular energy levels then give periodic motions, and conservative systems often possess continuous families of such closed orbits.

In a dissipative system, by contrast, an energy-like quantity may satisfy

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

Trajectories then move toward lower energy levels and may approach equilibria; the closed energy curves of the conservative system are typically destroyed.

A third mechanism occurs in self-sustained oscillators. The system may gain energy in one part of the phase plane and lose energy in another, producing an isolated periodic orbit. Such an isolated closed orbit is called a *limit cycle*; unlike the continuous families of closed orbits in conservative systems, a limit cycle may attract or repel nearby trajectories.

Not every nonlinear oscillator possesses a useful mechanical energy. Nullclines, invariant regions, and different time scales may instead organize the phase portrait. In the examples below, we combine linearization near equilibria with whatever additional structure is available—conservation, dissipation, nonlinear energy balance, or nullcline geometry—to understand the global motion.

We begin with the simple pendulum, the basic nonlinear mechanical oscillator. It illustrates conserved energy, centers and saddles, periodic motion, and separatrices. We then add damping, consider the Duffing oscillator, and turn to self-sustained and fast–slow oscillations.

The purpose here is geometric. A systematic treatment of Hamiltonian structure, attractors, limit cycles, and periodic-orbit theory will be developed in later chapters.

4.3.1 The simple pendulum as a conservative system

Consider a point mass m attached to a massless rigid rod of length L. Let $\theta(t)$ denote the angular displacement from the downward vertical position. The tangential component of gravity is

$$-mg \sin \theta.$$

Since the tangential acceleration is $L\ddot{\theta}$, Newton's law gives

$$mL\ddot{\theta} = -mg\sin\theta,$$

or

$$\ddot{\theta} + \frac{g}{L}\sin\theta = 0.$$

The mass cancels, so the motion is independent of m.

Introducing the dimensionless time

$$\tau = \sqrt{\frac{g}{L}}\,t,$$

we have

$$\frac{d^2\theta}{dt^2} = \frac{g}{L}\frac{d^2\theta}{d\tau^2}.$$

Hence the dimensionless pendulum equation is

$$\frac{d^2\theta}{d\tau^2} + \sin\theta = 0.$$

Renaming the dimensionless time τ as t and the angular displacement θ as x, we obtain

$$\ddot{x} + \sin x = 0.$$

The restoring force can be written in the form

$$-\sin x = -V'(x),$$

where

$$V(x) = 1 - \cos x.$$

The function V is the *potential energy* of the normalized pendulum, up to an additive constant. Thus the pendulum equation has the mechanical form

$$\ddot{x} = -V'(x).$$

Introducing

$$y = \dot{x},$$

we obtain the planar system

$$\dot{x} = y, \qquad \dot{y} = -\sin x.$$

The total mechanical energy is

$$E(x,y) = \frac{1}{2}y^2 + V(x) = \frac{1}{2}y^2 + 1 - \cos x.$$

Along a solution,

$$\begin{aligned}\frac{d}{dt}E(x(t),y(t)) &= y\dot{y} + V'(x)\dot{x}\\ &= y(-\sin x) + \sin x\, y\\ &= 0.\end{aligned}$$

Hence

$$E(x(t), y(t)) = \text{constant},$$

and every orbit lies on a level curve

$$E(x, y) = C.$$

The pendulum is therefore a conservative system.

The nullclines are

$$y = 0, \qquad \sin x = 0.$$

Thus the equilibria are

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

To determine their local behavior, the Jacobian matrix is

$$D\mathbf{f}(x, y) = \begin{bmatrix} 0 & 1 \\ -\cos x & 0 \end{bmatrix}.$$

At a downward equilibrium

$$(2k\pi, 0),$$

we have

$$D\mathbf{f}(2k\pi, 0) = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix},$$

whose eigenvalues are

$$\lambda_{\pm} = \pm i.$$

Thus the equilibrium is nonhyperbolic. The linearized system is a center, but linearization alone does not imply that the nonlinear equilibrium is a center.

Here the conserved energy provides the missing information. Since

$$V(x) = 1 - \cos x$$

has a strict local minimum at $x = 2k\pi$, the energy

$$E(x, y) = \frac{1}{2}y^2 + V(x)$$

has a strict local minimum at $(2k\pi, 0)$. Nearby energy levels are closed curves surrounding the equilibrium. Since the vector field does not vanish on these curves, they are periodic orbits. Hence each downward equilibrium is a center.

At an upward equilibrium

$$((2k+1)\pi, 0),$$

we have

$$D\mathbf{f}((2k+1)\pi, 0) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

whose eigenvalues are

$$\lambda_{\pm} = \pm 1.$$

Thus every upward equilibrium is a hyperbolic saddle. This agrees with the potential-energy picture, since V has a local maximum at $x = (2k+1)\pi$.

The energy of an upward equilibrium is

$$E((2k+1)\pi, 0) = 2.$$

The level set

$$E(x, y) = 2$$

contains the stable and unstable curves of the saddle equilibria and forms the separatrices in the phase portrait.

These separatrices divide the phase plane into different kinds of motion. For

$$0 < E < 2,$$

the energy level is a closed curve surrounding a downward equilibrium, and the pendulum oscillates back and forth. This motion is called *libration*.

For

$$E > 2,$$

the pendulum has enough energy to pass repeatedly through the upward position and rotates continuously. In the phase plane these rotational trajectories are not closed curves in the (x, y) plane if x is regarded as a real-valued coordinate, since x increases or decreases by multiples of 2π. If the angular variable is regarded modulo 2π, however, they correspond to periodic motion on the cylindrical phase space.

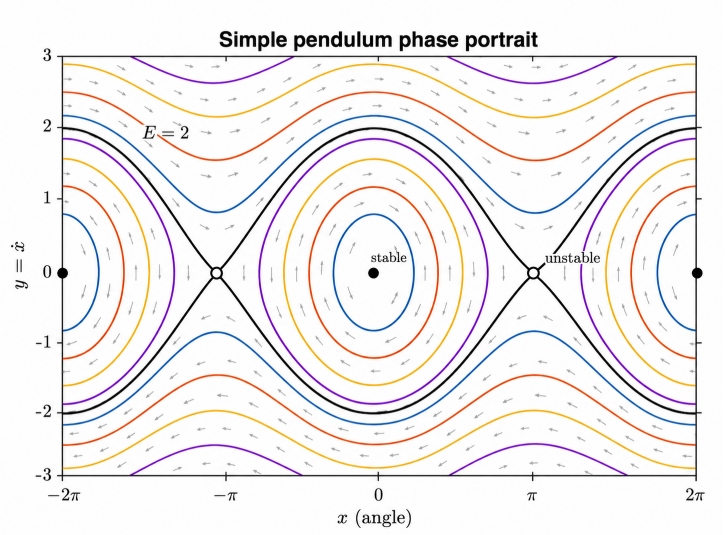


Figure 4.7: Phase portrait of the simple pendulum $\dot{x} = y$, $\dot{y} = -\sin x$. The trajectories are level curves of $E(x, y) = \frac{1}{2}y^2 + 1 - \cos x$. The separatrix $E = 2$ divides oscillatory libration from rotational motion.

The simple pendulum illustrates a general class of conservative mechanical systems. Consider

$$\ddot{x} + V'(x) = 0,$$

or equivalently

$$\dot{x} = y, \qquad \dot{y} = -V'(x).$$

The mechanical energy

$$E(x, y) = \frac{1}{2}y^2 + V(x)$$

satisfies

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

Thus the phase portrait is organized by the level curves of the energy.

An equilibrium has the form

$$(x_*, 0), \qquad V'(x_*) = 0.$$

The Jacobian matrix there is

$$\begin{bmatrix} 0 & 1 \\ -V''(x_*) & 0 \end{bmatrix},$$

so the eigenvalues satisfy

$$\lambda^2 + V''(x_*) = 0.$$

If

$$V''(x_*) < 0,$$

the equilibrium is a hyperbolic saddle. If

$$V''(x_*) > 0,$$

the eigenvalues are purely imaginary, so linearization is inconclusive; however, the strict local minimum of the energy implies that nearby level curves are closed, and the equilibrium is a center.

Thus the shape of the mechanical potential V provides a direct geometric picture of the local and global dynamics.

4.3.2 The damped simple pendulum as a dissipative system

The simple pendulum is conservative: its trajectories remain on level curves of the mechanical energy. We now introduce friction by adding a linear damping force proportional to the velocity. The normalized equation becomes

$$\ddot{x} + \gamma\dot{x} + \sin x = 0, \qquad \gamma > 0.$$

Equivalently,

$$\ddot{x} + \gamma\dot{x} + V'(x) = 0, \qquad V(x) = 1 - \cos x.$$

Thus the restoring force is unchanged, but the damping term $\gamma\dot{x}$ removes mechanical energy from the system.

Introducing

$$y = \dot{x},$$

we obtain the planar system

$$\dot{x} = y, \qquad \dot{y} = -\gamma y - \sin x.$$

The nullclines are

$$y = 0,$$

and

$$y = -\frac{1}{\gamma}\sin x.$$

Their intersections give the equilibria

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

Thus damping does not change the locations of the equilibria, but it changes their local behavior.

The Jacobian matrix is

$$D\mathbf{f}(x, y) = \begin{bmatrix} 0 & 1 \\ -\cos x & -\gamma \end{bmatrix}.$$

At a downward equilibrium

$$(2k\pi, 0),$$

we have

$$D\mathbf{f}(2k\pi, 0) = \begin{bmatrix} 0 & 1 \\ -1 & -\gamma \end{bmatrix}.$$

The characteristic equation is

$$\lambda^2 + \gamma\lambda + 1 = 0,$$

so

$$\lambda_{\pm} = \frac{-\gamma \pm \sqrt{\gamma^2 - 4}}{2}.$$

Both eigenvalues have negative real part. Hence every downward equilibrium is locally asymptotically stable. More precisely,

$$\begin{cases} 0 < \gamma < 2, & \text{stable spiral,} \\ \gamma = 2, & \text{repeated stable node,} \\ \gamma > 2, & \text{stable node.} \end{cases}$$

Thus weak damping produces decaying oscillations, whereas sufficiently strong damping produces a nonoscillatory return toward equilibrium.

At an upward equilibrium

$$((2k+1)\pi, 0),$$

we have

$$D\mathbf{f}((2k+1)\pi, 0) = \begin{bmatrix} 0 & 1 \\ 1 & -\gamma \end{bmatrix}.$$

Its characteristic equation is

$$\lambda^2 + \gamma\lambda - 1 = 0,$$

and therefore

$$\lambda_{\pm} = \frac{-\gamma \pm \sqrt{\gamma^2 + 4}}{2}.$$

One eigenvalue is positive and the other negative, so every upward equilibrium remains a hyperbolic saddle.

The most important change from the conservative pendulum is revealed by the mechanical energy

$$E(x,y) = \frac{1}{2}y^2 + V(x) = \frac{1}{2}y^2 + 1 - \cos x.$$

Along a solution,

$$\begin{aligned}\frac{d}{dt}E(x(t),y(t)) &= y\dot{y} + V'(x)\dot{x} \\ &= y(-\gamma y - \sin x) + \sin x\, y \\ &= -\gamma y^2.\end{aligned}$$

Hence

$$\dot{E} = -\gamma y^2 \leq 0.$$

The mechanical energy is therefore nonincreasing and is strictly decreasing whenever $y \neq 0$.

This energy law gives a geometric interpretation of dissipation. For the conservative pendulum,

$$\dot{E} = 0,$$

so each trajectory remains on one energy level. For the damped pendulum,

$$\dot{E} < 0$$

whenever the pendulum is moving, so trajectories cross the conservative energy levels in the direction of decreasing energy.

Consequently, the closed energy curves surrounding the downward equilibria are no longer periodic orbits. A trajectory may spiral across them as its energy decreases, eventually approaching a downward equilibrium. Thus damping destroys the continuous family of periodic motions present in the conservative pendulum.

There is also a useful general principle behind this example. For a mechanical system

$$\ddot{x} + \gamma\dot{x} + V'(x) = 0, \qquad \gamma > 0,$$

the energy

$$E(x,y) = \frac{1}{2}y^2 + V(x)$$

satisfies

$$\dot{E} = -\gamma y^2 \leq 0.$$

Thus linear damping converts a conservative mechanical system into a dissipative one. The shape of the potential still determines the locations and basic types of the equilibria, while dissipation determines the direction in which trajectories move across the energy landscape.

The conservative and damped pendula therefore have the same potential and the same equilibrium locations, but very different dynamics. In the conservative system, nearby solutions move forever on closed energy curves. In the damped system, energy loss selects a preferred direction of time and drives nearby solutions toward stable equilibria.

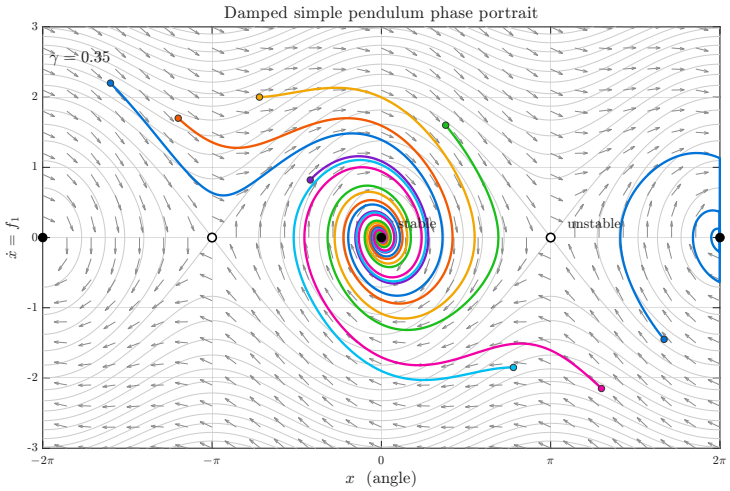


Figure 4.8: Phase portrait of the damped pendulum $\dot{x} = y$, $\dot{y} = -\gamma y - \sin x$. Since $\dot{E} = -\gamma y^2 \leq 0$, trajectories cross the conservative energy levels in the direction of decreasing energy. The downward equilibria are asymptotically stable, while the upward equilibria remain saddles.

4.3.3 The Duffing oscillator

The equation is named after the German engineer Georg Duffing, who studied nonlinear forced vibrations in his 1918 monograph *Erzwungene Schwingungen bei veränderlicher Eigenfrequenz und ihre technische Bedeutung*. His work showed how nonlinear elasticity changes the frequency and resonance behavior of an oscillator.

The Duffing oscillator introduces a nonlinear restoring force. It arose as a model for mechanical systems whose restoring force is not accurately described by Hooke's law when the displacement is no longer small. If the restoring force is smooth and symmetric under $x \mapsto -x$, its expansion near the equilibrium contains only odd powers:

$$F(x) = -\alpha x - \beta x^3 + \text{higher-order terms}.$$

Retaining the first nonlinear correction and normalizing the mass to one gives the unforced and undamped Duffing equation

$$\ddot{x} + \alpha x + \beta x^3 = 0.$$

The linear term describes the Hookean restoring force, while the cubic term models the leading nonlinear correction. When $\beta > 0$, the spring is *hardening*: the restoring force grows more rapidly than a linear force as the amplitude increases. When $\beta < 0$, it is *softening*. The sign of α also affects the equilibrium structure.

Using the mechanical potential introduced for the simple pendulum, we write

$$\ddot{x} + V'(x) = 0.$$

For the Duffing equation,

$$V'(x) = \alpha x + \beta x^3,$$

so we may choose

$$V(x) = \frac{\alpha}{2}x^2 + \frac{\beta}{4}x^4.$$

Thus the conserved mechanical energy is

$$E(x, \dot{x}) = \frac{1}{2}\dot{x}^2 + \frac{\alpha}{2}x^2 + \frac{\beta}{4}x^4.$$

Different choices of α and β produce different potential landscapes and therefore different phase portraits. Of particular interest is

$$\alpha < 0, \qquad \beta > 0,$$

for which the potential has two local minima separated by a local maximum. This is called a *double-well potential*.

As a normalized representative of this case, consider

$$\ddot{x} + x^3 - x = 0,$$

or equivalently,

$$\ddot{x} = x - x^3.$$

Its potential is

$$V(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2.$$

Introducing

$$y = \dot{x},$$

we obtain the planar system

$$\dot{x} = y, \qquad \dot{y} = x - x^3.$$

The conserved energy is

$$E(x, y) = \frac{1}{2}y^2 + \frac{1}{4}x^4 - \frac{1}{2}x^2.$$

Indeed,

$$\begin{aligned} \frac{d}{dt}E(x(t), y(t)) &= y\dot{y} + (x^3 - x)\dot{x} \\ &= y(x - x^3) + (x^3 - x)y \\ &= 0. \end{aligned}$$

The equilibria satisfy

$$y = 0, \qquad x - x^3 = 0,$$

and hence are

$$(-1, 0), \qquad (0, 0), \qquad (1, 0).$$

Their local behavior can be examined by linearization. The Jacobian matrix is

$$D\mathbf{f}(x, y) = \begin{bmatrix} 0 & 1 \\ 1 - 3x^2 & 0 \end{bmatrix}.$$

At the origin,

$$D\mathbf{f}(0, 0) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

whose eigenvalues are

$$\lambda_{\pm} = \pm 1.$$

Thus $(0,0)$ is a hyperbolic saddle.

At either equilibrium $(\pm 1, 0)$,

$$D\mathbf{f}(\pm 1, 0) = \begin{bmatrix} 0 & 1 \\ -2 & 0 \end{bmatrix},$$

whose eigenvalues are

$$\lambda_{\pm} = \pm i\sqrt{2}.$$

These equilibria are nonhyperbolic, so linearization alone does not determine their nonlinear behavior.

The potential supplies the missing information. Since

$$V(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2$$

has strict local minima at $x = \pm 1$, the energy

$$E(x,y) = \frac{1}{2}y^2 + V(x)$$

has strict local minima at

$$(-1,0) \qquad \text{and} \qquad (1,0).$$

Nearby energy levels are closed curves surrounding these equilibria. Therefore both are centers, and nearby solutions are periodic.

The origin is a hyperbolic saddle. It corresponds to the local maximum of the potential. This saddle energy is

$$E(0,0) = 0.$$

Hence the critical energy level

$$E(x,y) = 0$$

contains the separatrices through the saddle. On this level,

$$y^2 = x^2 - \frac{1}{2}x^4 = x^2\left(1 - \frac{x^2}{2}\right).$$

Thus the separatrix consists of two loops through the origin, reaching the turning points $x = \pm\sqrt{2}$. Each loop leaves the saddle along one branch of its unstable curve and returns to the same saddle along a branch of its stable curve. The two loops are therefore *homoclinic orbits*.

Inside the left loop, solutions oscillate about $(-1,0)$, while inside the right loop they oscillate about $(1,0)$. For $E > 0$, the energy level crosses both wells and the motion passes repeatedly from one well to the other. Thus the homoclinic separatrix is the boundary between motion confined to a single potential well and motion crossing the central energy barrier.

The separatrix divides the motion into several regimes. For

$$-\frac{1}{4} < E < 0,$$

the energy level consists of two closed curves, one surrounding $(-1, 0)$ and the other surrounding $(1, 0)$. The corresponding solutions oscillate periodically within one of the two potential wells.

At

$$E = 0,$$

the two separatrix loops meet at the saddle $(0, 0)$. These trajectories form the boundary between motion confined to a single well and motion that crosses between the two wells.

For

$$E > 0,$$

the energy level is a single closed curve surrounding all three equilibria. The particle has enough energy to cross the potential barrier at $x = 0$ and moves repeatedly between the two wells.

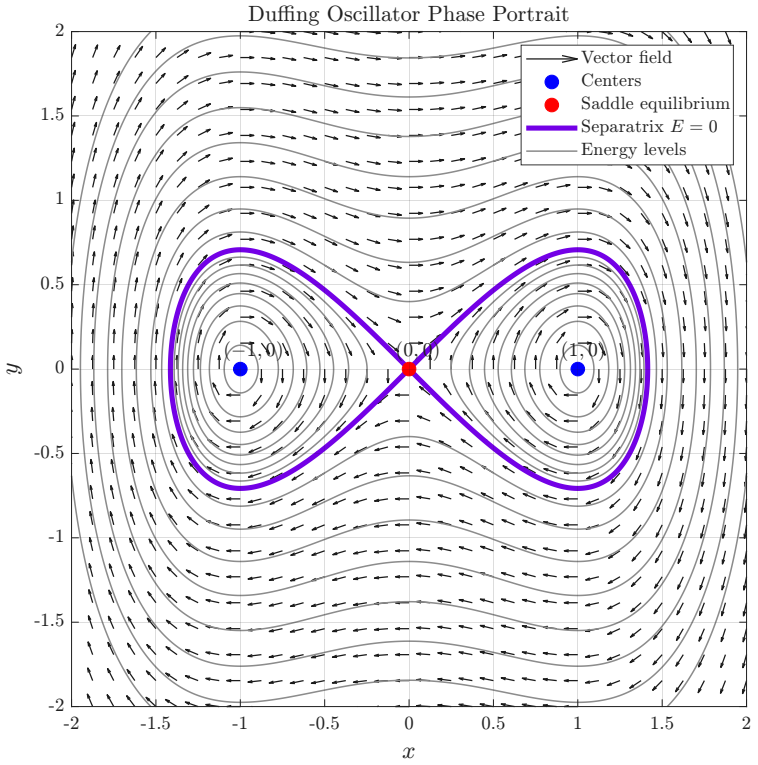


Figure 4.9: Phase portrait of the conservative Duffing oscillator $\dot{x} = y$, $\dot{y} = x - x^3$. The double-well potential produces two centers at $(\pm 1, 0)$ and a saddle at the origin. The separatrix $E = 0$ divides oscillations confined to a single well from periodic motion crossing between the two wells.

The Duffing oscillator shows how changing the mechanical potential can alter the global topology of the phase portrait. The simple pendulum has a periodic sequence of potential wells and barriers, whereas the normalized double-well Duffing oscillator has two wells separated by a single barrier. In both cases, linearization determines the saddle equilibria, while the conserved energy is needed to establish the center structure at equilibria with purely imaginary eigenvalues.

Exercise 4.10 (Duffing potential geometry)**.** For the Duffing system

$$\dot{x} = y, \qquad \dot{y} = -\alpha x - \beta x^3, \qquad V(x) = \frac{\alpha}{2}x^2 + \frac{\beta}{4}x^4,$$

analyze the cases $\alpha > 0, \beta > 0$; $\alpha < 0, \beta > 0$; and $\alpha > 0, \beta < 0$. Classify the equilibria from V, then use the energy levels $E = \frac{1}{2}y^2 + V(x)$ to sketch each phase portrait and identify bounded oscillations, separatrices, and escape orbits.

4.3.4 The Van der Pol oscillator

The Van der Pol oscillator originated in the study of self-sustained oscillations in electrical circuits containing vacuum-tube amplifiers. The nonlinear element in such a circuit amplifies small signals but damps large ones, allowing the circuit to settle into an oscillation of preferred amplitude.

After nondimensionalization, the model takes the form

$$\ddot{x} - \mu(1 - x^2)\dot{x} + x = 0, \qquad \mu > 0.$$

Here $x(t)$ represents a scaled voltage or current in the oscillator, t denotes scaled time, and μ measures the strength of the nonlinear gain and damping.

Introducing

$$y = \dot{x},$$

we obtain the planar system

$$\dot{x} = y, \qquad \dot{y} = \mu(1 - x^2)y - x.$$

The origin is the unique equilibrium.

The Jacobian matrix is

$$D\mathbf{f}(x, y) = \begin{bmatrix} 0 & 1 \\ -1 - 2\mu xy & \mu(1 - x^2) \end{bmatrix}.$$

At the origin,

$$D\mathbf{f}(0, 0) = \begin{bmatrix} 0 & 1 \\ -1 & \mu \end{bmatrix}.$$

Thus

$$T = \mu > 0, \qquad D = 1, \qquad \Delta = \mu^2 - 4.$$

It follows that the origin is an unstable spiral when

$$0 < \mu < 2,$$

a repeated unstable node when

$$\mu = 2,$$

and an unstable node when

$$\mu > 2.$$

Thus nearby nonconstant solutions move away from the origin.

The nullclines are

$$y = 0$$

and

$$\mu(1 - x^2)y - x = 0.$$

Away from $x = \pm 1$, the latter can be written as

$$y = \frac{x}{\mu(1 - x^2)}.$$

Linearization explains the local instability of the origin, but it does not explain what prevents trajectories from moving indefinitely outward. To understand the competing effects of gain and dissipation, it is useful to place the Van der Pol equation in a broader class.

A *Liénard equation* has the form

$$\ddot{x} + f(x)\dot{x} + g(x) = 0.$$

Here $g(x)$ represents a restoring force, while $f(x)$ represents a state-dependent damping coefficient. If

$$G(x) = \int_0^x g(s)\, ds,$$

then

$$E(x, \dot{x}) = \frac{1}{2}\dot{x}^{\,2} + G(x)$$

satisfies

$$\begin{aligned} \frac{dE}{dt} &= \dot{x}\,\ddot{x} + g(x)\dot{x} \\ &= -f(x)\dot{x}^{\,2}. \end{aligned}$$

Thus the sign of $f(x)$ determines whether energy is dissipated or supplied.

For the Van der Pol oscillator,

$$f(x) = \mu(x^2 - 1), \qquad g(x) = x,$$

so

$$G(x) = \frac{1}{2}x^2$$

and

$$E(x, y) = \frac{1}{2}x^2 + \frac{1}{2}y^2.$$

Moreover,

$$\dot{E} = \mu(1 - x^2)y^2.$$

Hence the system gains energy when $|x| < 1$ and loses energy when $|x| > 1$, except when $y = 0$. Small oscillations are therefore amplified, while large oscillations are damped. This balance between energy gain and dissipation is the basic mechanism of self-sustained oscillation.

The damped pendulum considered previously also belongs to the Liénard class, with

$$f(x) = \gamma > 0, \qquad g(x) = \sin x.$$

In that case the damping has one sign and the mechanical energy decreases monotonically. The Van der Pol oscillator differs precisely in that its damping coefficient changes sign.

The local instability of the origin, together with the change in sign of the energy balance, suggests the selection of an intermediate oscillation amplitude. In fact, the Van der Pol oscillator possesses an isolated attracting periodic orbit, as illustrated in Figure 4.10. Unlike the closed orbits of a conservative system, which occur in continuous families of energy levels, this periodic orbit is isolated. It is therefore a *limit cycle*.

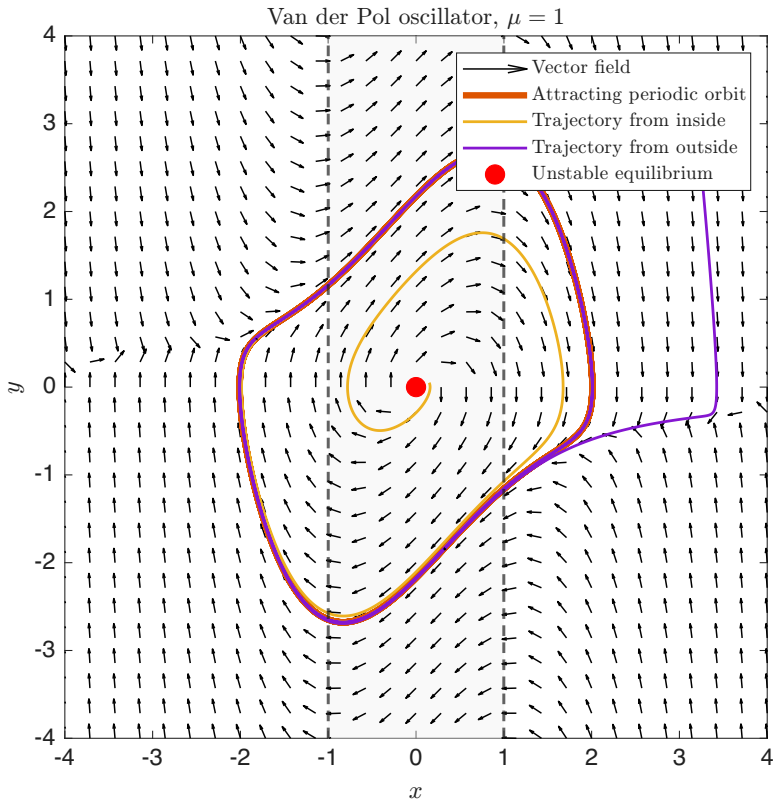


Figure 4.10: Phase portrait of the Van der Pol oscillator $\dot{x} = y, \dot{y} = (1 - x^2)y - x$, with $\mu = 1$. The dashed vertical lines $x = \pm 1$ separate the region $|x| < 1$, where $\dot{E} = (1 - x^2)y^2 > 0$ for $y \neq 0$, from the regions $|x| > 1$, where $\dot{E} < 0$. Thus trajectories near the unstable equilibrium $(0, 0)$ are driven outward, while trajectories of sufficiently large amplitude lose energy. The displayed trajectories from inside and outside approach the same isolated attracting periodic orbit, the Van der Pol limit cycle.

4.3.5 The FitzHugh–Nagumo system

The FitzHugh–Nagumo system is a simplified planar model of neuronal excitability. It was introduced as a reduction of the much more detailed Hodgkin–Huxley equations, with the aim of retaining the basic geometric mechanism of excitation and recovery while reducing the dynamics to two variables. The model separates the evolution into a fast activation process and a slow recovery

process. Thus it provides a convenient setting for studying thresholds, action-potential-like pulses, refractory behavior, and relaxation oscillations.

A common form of the model is

$$\dot{x} = x - \frac{x^3}{3} - y + I, \qquad \dot{y} = \varepsilon(x + a - by),$$

where

$$0 < \varepsilon \ll 1.$$

Here x is a fast voltage-like variable and y is a slow recovery variable. The parameter I represents an applied input, while a and b control the recovery dynamics. The small parameter ε expresses the separation of time scales between the rapid excitation and the slower recovery.

The nullclines are

$$y = x - \frac{x^3}{3} + I$$

and

$$y = \frac{x + a}{b}.$$

Thus the x-nullcline is cubic and the y-nullcline is a straight line. Their intersections are the equilibria. More explicitly, an equilibrium (x_*, y_*) satisfies

$$y_* = \frac{x_* + a}{b},$$

where x_* is a root of

$$x_* - \frac{x_*^3}{3} - \frac{x_* + a}{b} + I = 0.$$

Equivalently,

$$\frac{x_*^3}{3} + \left(\frac{1}{b} - 1\right) x_* + \frac{a}{b} - I = 0.$$

When $0 < b < 1$, the left-hand side is strictly increasing, because

$$\frac{d}{dx}\left[\frac{x^3}{3} + \left(\frac{1}{b} - 1\right) x + \frac{a}{b} - I\right] = x^2 + \frac{1}{b} - 1 > 0.$$

Hence there is exactly one equilibrium for every value of I.

The stability of this equilibrium depends on its location. The linearization at (x_*, y_*) is

$$J_* = \begin{pmatrix} 1 - x_*^2 & -1 \\ \varepsilon & -\varepsilon b \end{pmatrix}.$$

Therefore

$$\operatorname{tr} J_* = 1 - x_*^2 - \varepsilon b, \qquad \det J_* = \varepsilon(1 - b + b x_*^2).$$

For $0 < b < 1$, the determinant is positive. Hence the sign of the trace determines whether the equilibrium is locally attracting or repelling:

$$x_*^2 > 1 - \varepsilon b \quad \Longrightarrow \quad \text{locally asymptotically stable},$$

while

$$x_*^2 < 1 - \varepsilon b \quad \Longrightarrow \quad \text{unstable.}$$

Thus changing the applied input I moves the intersection of the nullclines and can change the qualitative behavior of the system.

The small parameter ε also gives the phase portrait a characteristic fast–slow geometry. The variable x changes rapidly, whereas y changes slowly. Away from the cubic nullcline, the fast variable moves quickly. Near the cubic nullcline, the fast motion becomes small and the slower recovery variable controls the evolution.

Two important regimes are shown in Figure 4.11.

For

$$a = 0.7, \qquad b = 0.8, \qquad \varepsilon = 0.08, \qquad I = 0,$$

the unique equilibrium is approximately

$$(x_*, y_*) = (-1.199, -0.624).$$

It lies on the left outer branch of the cubic nullcline and is locally asymptotically stable. A sufficiently small perturbation therefore returns directly to the equilibrium. A larger perturbation, however, may cross an effective threshold in the phase plane and produce a large excursion before returning. This behavior is called *excitability*. The large excursion models an excitation pulse, even though the eventual long-time state is still the stable equilibrium.

For the same values of a, b, and ε, but with

$$I = 0.5,$$

the equilibrium moves to approximately

$$(x_*, y_*) = (-0.805, -0.131).$$

At this point,

$$\operatorname{tr} J_* > 0,$$

so the equilibrium is unstable. Numerically, trajectories are attracted to a periodic orbit. Because ε is small, this periodic motion has the form of a *relaxation oscillation*: the trajectory moves slowly near one outer branch of the cubic nullcline, jumps rapidly toward the other branch, drifts slowly again, and then jumps back.

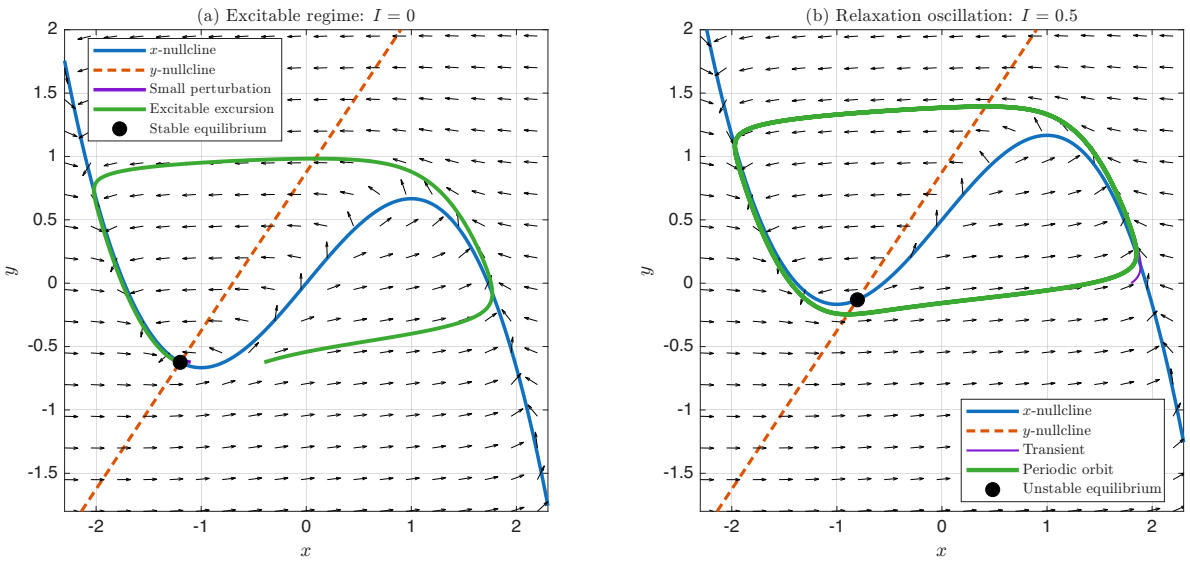


Figure 4.11: Two dynamical regimes of the FitzHugh–Nagumo system with $a = 0.7$, $b = 0.8$, and $\varepsilon = 0.08$. (a) For $I = 0$, the unique equilibrium is stable. A small perturbation returns directly to the equilibrium, whereas a sufficiently large perturbation produces a large excitable excursion before returning. (b) For $I = 0.5$, the equilibrium is unstable, and the displayed trajectory approaches an attracting periodic orbit with the characteristic slow–fast structure of a relaxation oscillation. In both panels, the cubic curve is the x-nullcline and the straight line is the y-nullcline.

The two panels show how the input I changes the dynamics. Moving the nullcline intersection from a stable branch of the cubic into an unstable regime changes the long-time behavior from return to an excitable equilibrium to sustained periodic motion.

Unlike the pendulum and Duffing oscillator, this system has no conserved mechanical energy. Its dynamics are instead organized by its nullclines, the stability of their intersection, and the separation of fast and slow time scales.

Exercise 4.11 (Fast and slow dynamics in the FitzHugh–Nagumo system)**.** Consider

$$\dot{x} = x - \frac{x^3}{3} - y + I, \qquad \dot{y} = \varepsilon(x + a - by), \qquad 0 < \varepsilon \ll 1.$$

Set $\varepsilon = 0$ and show that the outer branches of the cubic x-nullcline are attracting for the fast dynamics, while the middle branch is repelling. With the slow time $\tau = \varepsilon t$, assume that the motion follows

$$y = x - \frac{x^3}{3} + I$$

and derive

$$(1 - x^2)\frac{dx}{d\tau} = x + a - b\left(x - \frac{x^3}{3} + I\right).$$

Explain the singularities at $x = \pm 1$ and use the fast–slow geometry to describe a relaxation oscillation.

Summary

The models in this section exhibit four distinct phase-plane mechanisms. In conservative systems, energy levels are invariant and closed orbits occur in families. In dissipative systems, energy decreases and trajectories may approach stable equilibria. In self-sustained oscillators, the effective damping changes sign: the oscillator energy increases in one region of the phase plane and decreases in another. The balance between these effects selects a preferred amplitude and produces an attracting limit cycle. In fast–slow systems, the dynamics is organized by nonlinear nullclines and widely separated time scales.

4.4 From Phase Portraits to Time-Dependent Solutions

The phase portrait is the central object of this chapter. It describes the geometry of solutions in the phase plane: equilibria, nullclines, invariant regions, separatrices, and closed orbits. However, the phase portrait does not always give the time-dependent functions $x(t)$ and $y(t)$ explicitly.

The distinction is simple but important. An orbit is a geometric curve in the phase plane. A solution is a time-parametrized curve. Once an orbit is known, one may return to the original differential equation to recover the motion along that orbit.

For a planar system

$$\dot{x} = f(x, y), \qquad \dot{y} = g(x, y),$$

the orbit may be studied by eliminating time:

$$\frac{dy}{dx} = \frac{\dot{y}}{\dot{x}} = \frac{g(x, y)}{f(x, y)}.$$

This equation gives the geometric relation between x and y. If the orbit is known locally as

$$y = Y(x),$$

then the original equation gives

$$\frac{dx}{dt} = f(x, Y(x)).$$

Thus the time parametrization is recovered from the scalar equation

$$t - t_0 = \int_{x_0}^{x} \frac{d\xi}{f(\xi, Y(\xi))}.$$

This is the point where the phase-plane method returns to the scalar methods of Chapter 2.

4.4.1 Lotka–Volterra: closed orbits and periodic solutions

Consider the Lotka–Volterra system

$$\dot{x} = x(1 - y), \qquad \dot{y} = y(x - 1).$$

The phase portrait is organized by the conserved quantity

$$H(x, y) = x - \log x + y - \log y.$$

Thus each orbit in the first quadrant lies on a level curve

$$H(x, y) = C.$$

The equilibrium $(1, 1)$ is surrounded by closed level curves. Hence solutions starting on these level curves are periodic.

The equation

$$H(x, y) = C$$

describes the orbit geometrically. It does not usually give elementary formulas for $x(t)$ and $y(t)$. Nevertheless, the time evolution is determined by the original system. Along a local branch $y = Y(x)$ of the level curve, one has

$$\frac{dx}{dt} = x(1 - Y(x)).$$

Thus

$$t - t_0 = \int_{x_0}^{x} \frac{d\xi}{\xi(1 - Y(\xi))}.$$

In practice, one often uses the phase portrait to understand the periodic motion and numerical integration to display the time-parametrized solution.

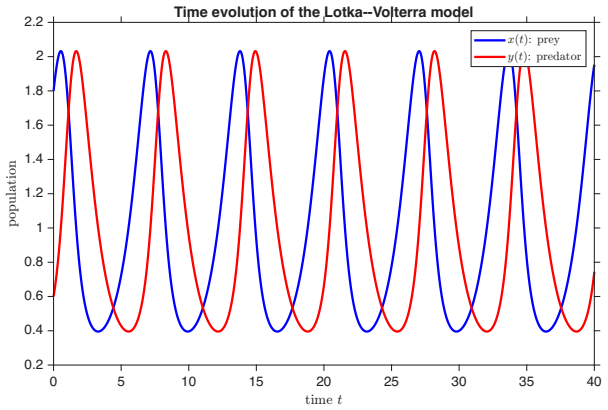


Figure 4.12: Time evolution of the Lotka–Volterra system $\dot{x} = x(1 - y)$, $\dot{y} = y(x - 1)$. The solution oscillates periodically because the corresponding phase-plane orbit is a closed level curve of $H(x, y) = x - \log x + y - \log y$.

This example shows how a phase portrait can give a strong qualitative solution even when explicit formulas for $x(t)$ and $y(t)$ are not available. The closed orbit in the phase plane corresponds to periodic time evolution.

4.4.2 SIR: phase-plane orbits and epidemic time evolution

Consider the reduced SIR system

$$\dot{S} = -\beta SI, \qquad \dot{I} = I(\beta S - \gamma),$$

with

$$\dot{R} = \gamma I.$$

The phase portrait in the (S, I)-plane is obtained by eliminating time:

$$\frac{dI}{dS} = \frac{I(\beta S - \gamma)}{-\beta S I} = -1 + \frac{\gamma}{\beta S}.$$

Therefore

$$I + S - \frac{\gamma}{\beta} \log S = C.$$

This equation gives the orbit in the (S, I)-plane.

The time evolution is obtained by returning to the original equations. If the orbit is written as $I = I(S)$, then

$$\frac{dS}{dt} = -\beta S I(S),$$

and hence

$$t - t_0 = -\int_{S_0}^{S} \frac{d\sigma}{\beta \sigma I(\sigma)}.$$

Once $S(t)$ and $I(t)$ are known, the removed population is determined by

$$R(t) = N - S(t) - I(t),$$

or equivalently by integrating $\dot{R} = \gamma I$.

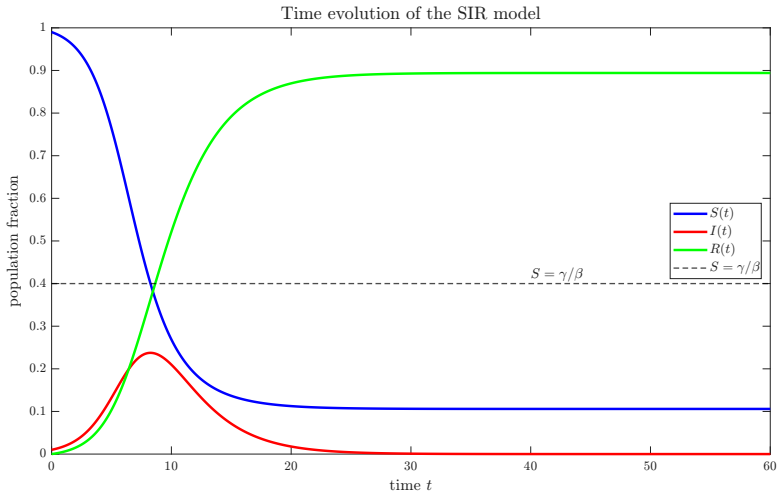


Figure 4.13: Time evolution of the SIR model. The phase portrait describes the orbit in the (S, I)-plane, while this plot shows the time-parametrized solution $S(t)$, $I(t)$, and $R(t)$. The infected population increases while $S > \gamma/\beta$ and decreases after S crosses the threshold.

This example is important in epidemic modeling. The phase portrait identifies the threshold

$$S = \frac{\gamma}{\beta}$$

and shows why infection first grows and then decays. The time-dependent plot shows when this happens and how the total population is distributed among susceptible, infected, and removed classes.

4.4.3 FitzHugh–Nagumo: excitation pulses and relaxation oscillations

The FitzHugh–Nagumo system considered earlier,

$$\dot{x} = x - \frac{x^3}{3} - y + I, \qquad \dot{y} = \varepsilon(x + a - by),$$

provides another illustration of the distinction between an orbit in the phase plane and its time parametrization. Here x is the fast voltage-like variable and y is the slow recovery variable. The phase portrait describes the trajectory as a curve in the (x, y)-plane, whereas the functions $x(t)$ and $y(t)$ describe how that curve is traversed in time.

Recall the two parameter regimes illustrated previously in Figure 4.11. We keep

$$a = 0.7, \qquad b = 0.8, \qquad \varepsilon = 0.08,$$

and vary only the applied input I.

For $I = 0$, the system has a stable equilibrium

$$(x_*, y_*) \approx (-1.199, -0.624).$$

A small perturbation produces only a small response and the solution returns directly to equilibrium. A sufficiently large perturbation, however, sends the trajectory on a large excursion through the phase plane before it returns to the same equilibrium. In the time-dependent variable $x(t)$, this excursion appears as a large isolated pulse. This is the characteristic signature of an *excitable system*: the long-time state is an equilibrium, but a sufficiently large disturbance can produce a substantial transient response.

For $I = 0.5$, the equilibrium

$$(x_*, y_*) \approx (-0.805, -0.131)$$

is unstable. The phase portrait in Figure 4.11 shows a trajectory approaching an attracting periodic orbit. Consequently, after an initial transient, $x(t)$ undergoes sustained oscillations rather than returning to an equilibrium.

Because ε is small, these oscillations are strongly non-sinusoidal. Along the slow portions of the phase-plane orbit, the solution changes relatively slowly. When the trajectory reaches a region where it leaves one branch of the cubic nullcline and moves rapidly toward another, $x(t)$ changes abruptly. Thus the slow and fast portions of the phase-plane orbit appear in the time-dependent solution as alternating intervals of slow evolution and rapid transitions. This is the time-dependent signature of a *relaxation oscillation*.

This comparison shows information that is not immediately visible from the phase-plane curves alone. In the excitable regime, both the small and large responses eventually approach the same equilibrium, but their transient time histories are very different. In the oscillatory regime, the closed phase-plane orbit corresponds to a periodic function of time. Moreover, the nonuniform speed along the orbit explains the characteristic shape of the relaxation oscillation: long slow stages are connected by short rapid transitions.

Thus the phase portrait and the time-dependent solution provide complementary descriptions of the same dynamics. The phase portrait reveals the geometric organization of trajectories, while the time parametrization reveals the duration, rate, and temporal pattern of the observed response.

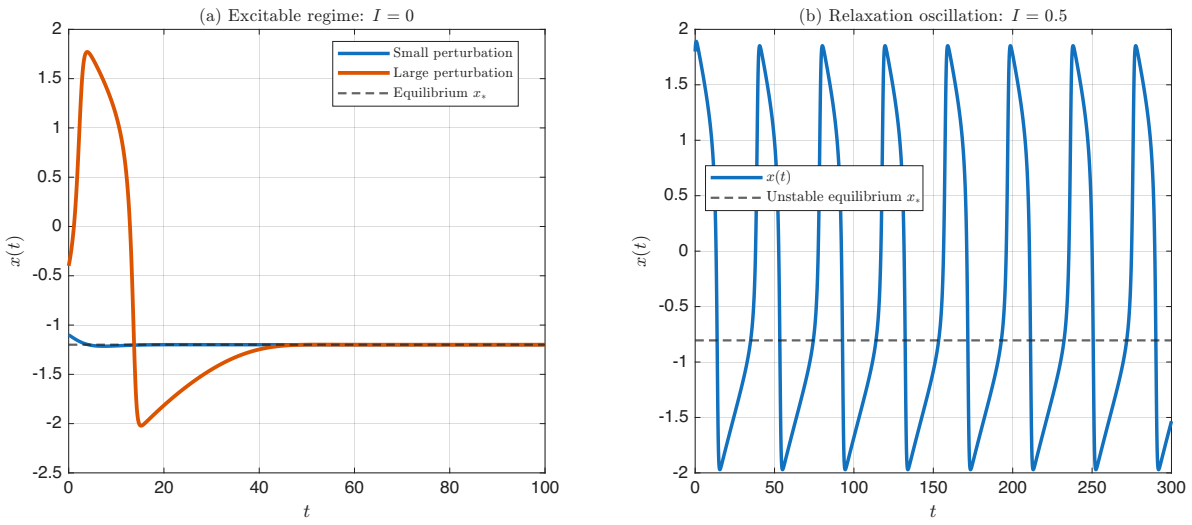


Figure 4.14: Time-dependent solutions of the FitzHugh–Nagumo system corresponding to the phase portraits in Figure 4.11, with $a = 0.7$, $b = 0.8$, and $\varepsilon = 0.08$. (a) For $I = 0$, a small perturbation decays directly to the stable equilibrium, whereas a larger perturbation produces a large excitation pulse before returning to equilibrium. (b) For $I = 0.5$, the equilibrium is unstable and, after an initial transient, $x(t)$ approaches a sustained relaxation oscillation. The intervals of slow evolution separated by rapid transitions are the time-dependent manifestation of the fast–slow geometry seen in the phase plane.

4.4.4 What has been solved?

The phase-plane method does not always produce explicit formulas for the solution. Its main achievement is geometric. It reveals how the plane is organized by equilibria, nullclines, invariant regions, separatrices, closed orbits, and attracting cycles.

This geometric information is often enough to answer the main qualitative questions. It tells us which equilibria are stable, which regions are physically meaningful, which initial data lead to different long-time outcomes, and whether oscillatory behavior is possible.

When the actual time dependence is needed, one must return to the original differential equation. If the orbit is known as a relation between the phase variables, then the motion along the orbit can be reduced to a scalar equation, as in Chapter 2. Thus the phase portrait and the time-parametrized solution are complementary:

the phase portrait gives the geometry of motion, the differential equation gives the timing.

This is the main lesson of planar nonlinear systems. Nonlinear equations are rarely solved by formulas alone. They are understood by combining local linearization, global phase-plane geometry, invariant quantities, numerical solutions, and scalar methods for the time evolution along an orbit.

Homework Problems

Homework 4.1 (Local invariant curves)**.** Consider

$$\dot{x} = x + y^2, \qquad \dot{y} = -y + x^2.$$

1. Classify the origin and find the stable and unstable eigendirections.
2. Use the invariance condition to determine the leading terms of the local stable and unstable curves in the forms

$$x = ay^2 + O(y^3), \qquad y = bx^2 + O(x^3).$$

3. Explain why linearization determines the tangents to these curves but not the coefficients a and b.

Homework 4.2 (An explicit heteroclinic connection)**.** Consider

$$\dot{x} = y, \qquad \dot{y} = x^3 - x.$$

1. Find and classify all equilibria. Verify that

$$H(x, y) = \frac{1}{2}y^2 + \frac{1}{2}x^2 - \frac{1}{4}x^4$$

is conserved.

2. Find the energy level containing both saddle points. Show that its two connecting branches satisfy

$$y = \pm\frac{1 - x^2}{\sqrt{2}}, \qquad -1 \le x \le 1,$$

and determine the direction of motion on each branch.

3. Parametrize the branch from $(-1, 0)$ to $(1, 0)$ and verify its limits as $t \to \pm\infty$.

Homework 4.3 (A predator–prey model with saturation)**.** Consider

$$\dot{x} = x(1 - x) - \frac{axy}{1 + x}, \qquad \dot{y} = y\left(\frac{bx}{1 + x} - d\right), \qquad a > 0, \quad b > d > 0.$$

1. Show that the first quadrant is positively invariant. Use

$$W = x + \frac{a}{b}y$$

to prove that every trajectory starting there is bounded.

2. Find the nullclines and equilibria. Determine precisely when a positive coexistence equilibrium exists.

3. Classify the equilibria when they are hyperbolic and sketch the phase portrait in the coexistence case. Explain how saturation changes the prey nullcline from that of the logistic predator–prey model.

Homework 4.4 (A symmetric genetic toggle switch)**.** Consider

$$\dot{x} = \frac{\alpha}{1+y^2} - x, \qquad \dot{y} = \frac{\alpha}{1+x^2} - y, \qquad \alpha > 0.$$

1. Show that the square $[0, \alpha]^2$ is positively invariant.

2. At an equilibrium, prove that

$$(x - y)(1 - xy) = 0.$$

   Use this identity to find all equilibria and show that two asymmetric ones appear exactly when $\alpha > 2$.

3. Classify the equilibria for $\alpha \neq 2$. For $\alpha > 2$, sketch the phase portrait in the invariant square and interpret the stable curve of the symmetric saddle as a switching threshold.

Homework 4.5 (Chemostat threshold)**.** A chemostat is a continuously stirred vessel into which fresh nutrient solution flows while an equal volume of the well-mixed culture is removed. Let $S(t)$ be the substrate concentration and $X(t)$ the biomass concentration. The parameter S_{in} is the substrate concentration in the inflow, D is the dilution rate, m is the maximum specific growth rate, and K is the half-saturation constant. A unit biomass yield is assumed.

The model is

$$\dot{S} = D(S_{\text{in}} - S) - \frac{mS}{K+S}X, \qquad \dot{X} = X\left(\frac{mS}{K+S} - D\right),$$

where all parameters are positive.

1. Show that the first quadrant is positively invariant and that $Z = S + X$ satisfies

$$\dot{Z} = D(S_{\text{in}} - Z).$$

   What does this imply about boundedness and the long-time material balance?

2. Find the washout equilibrium and determine the exact condition for a positive equilibrium to exist. Compute that equilibrium.

3. Classify both equilibria whenever they are hyperbolic. Draw the nullclines and describe how the threshold separates washout from persistence.

Projects

The following projects are intended as a selection menu. Students may choose one project for an extended investigation; they are not expected to complete all five. The projects emphasize different combinations of modeling, phase-plane analysis, and numerical computation. Project 4.5 is more specialized and may be assigned as an advanced option.

Project 4.1 (Predator–prey model with harvesting)**.** Study the system

$$\dot{x} = x(1-x) - axy - h\frac{x}{x+\varepsilon}, \qquad \dot{y} = -by + cxy - k\frac{y}{y+\varepsilon},$$

where $x, y \geq 0$ are the prey and predator populations, $h, k \geq 0$ are constant harvesting rates, and $\varepsilon > 0$ regularizes the harvest terms to preserve positivity.

1. Show that the first quadrant is positively invariant. Find the boundary equilibria and the critical prey-harvest rate $h_c(\varepsilon)$ at which the two nonzero prey equilibria merge.
2. Reduce the coexistence-equilibrium equations to one equation in x. For one stated parameter set, locate all positive roots numerically and classify the corresponding equilibria.
3. With h just below $h_c(\varepsilon)$, compare phase portraits for $k = 0$ and for one positive value of k. Explain how harvesting the predator changes the prey's long-time behavior.

Project 4.2 (Mutualism and saturation)**.** Compare the two mutualism models

$$\dot{x} = x(1 - x + \alpha y), \qquad \dot{y} = y(1 - y + \beta x),$$

and

$$\dot{x} = x\left(1 - x + \frac{\alpha y}{1+y}\right), \qquad \dot{y} = y\left(1 - y + \frac{\beta x}{1+x}\right).$$

Here $x, y \geq 0$ are the populations of two species that benefit from one another, while $\alpha, \beta > 0$ measure the strengths of the mutualistic effects. In the second model, the benefit supplied by either species saturates when that species is abundant.

1. Find and classify the equilibria of the unsaturated model. Explain the role of the threshold $\alpha\beta = 1$.
2. Prove that every solution of the saturated model in the first quadrant is bounded, and analyze its positive equilibrium using nullclines and linearization.
3. For parameter values on both sides of $\alpha\beta = 1$, compare numerical phase portraits of the two models and explain the effect of saturation.

Project 4.3 (The Selkov model of glycolytic oscillations)**.** Consider the Selkov model

$$\dot{x} = -x + ay + x^2 y, \qquad \dot{y} = b - ay - x^2 y, \qquad a, b > 0.$$

Here $x, y \geq 0$ are scaled concentrations of two chemical intermediates in glycolysis. The linear and autocatalytic terms describe their conversion, while b represents a constant supply and a controls a linear reaction rate.

1. Show that the first quadrant is positively invariant, find the unique positive equilibrium, and verify that $\frac{d}{dt}(x+y) = b - x$.

2. Draw the nullclines and determine from the Jacobian when the equilibrium is locally asymptotically stable.

3. Choose parameters on both sides of the stability boundary. Compare the phase portraits and time series, and relate any oscillation observed to the Brusselator.

Project 4.4 (Damped motion in a double-well potential)**.** Study the damped Duffing system

$$\dot{x} = y, \qquad \dot{y} = x - x^3 - \gamma y, \qquad \gamma > 0,$$

which models a particle in the double-well potential $V(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2$ under linear damping. Here x is the displacement of the particle, y is its velocity, and γ is the damping coefficient. The minima of the potential represent two stable resting configurations. Let

$$E(x, y) = \frac{1}{2}y^2 + V(x).$$

1. Find and classify all equilibria. Show that $\dot{E} = -\gamma y^2 \leq 0$ and deduce that every forward solution is bounded.

2. Numerically approximate the stable curve of the saddle at the origin and verify that it separates the basins of $(-1, 0)$ and $(1, 0)$.

3. Compare the phase portraits for $\gamma = 0$ and for one positive value of γ. Explain what becomes of the conservative homoclinic loops.

Project 4.5 (A dynamic Shockley–Read–Hall recombination model)**.** Let n and p be free-electron and free-hole densities, and let n_T be the density of occupied traps. The physical region is $n, p \geq 0$, $0 \leq n_T \leq N_T$. Consider

$$\dot{n} = G - k_n n(N_T - n_T), \qquad \dot{p} = G - k_p p n_T,$$

$$\dot{n}_T = k_n n(N_T - n_T) - k_p p n_T,$$

where all parameters are positive. Here G is the electron–hole pair generation rate, N_T is the total trap density, and k_n, k_p are electron and hole capture-rate constants. The factors $n(N_T - n_T)$ and pn_T represent capture by empty and occupied traps, respectively.

1. Show that the physical region is positively invariant and that $p - n - n_T = C$ is conserved. Use this relation to reduce the model to a planar system in (n, n_T).

2. Derive a scalar equation for the positive equilibrium, prove that it is unique, and classify it from the Jacobian whenever it is hyperbolic.

3. For $N_T = 1$, $k_n = k_p = 1$, $C = 0$, and $G = 2/9$, plot the nullclines and phase portrait. Then vary G and describe how the equilibrium changes.

Historical Notes and Further Reading

Historical notes. Chapter 1 introduced nonlinear behavior through scalar models; the present chapter turns to systems in which two state variables interact. The modern qualitative theory of differential equations has one of its foundations in Poincaré's memoirs on curves defined by differential equations, where the geometry of trajectories becomes central when explicit formulas are unavailable (Poincaré, 1881). The Hartman–Grobman theorem, due independently to Grobman and Hartman, justifies the use of linearization near hyperbolic equilibria; see Hartman (2002). Closely related predator–prey equations were developed independently by Lotka (1925) and Volterra (1926), while the classical susceptible–infected–removed epidemic model goes back to Kermack and McKendrick (1927).

Nonlinear oscillation provided another major source of planar systems. van der Pol (1926) introduced the model that now bears his name in the study of self-sustained electrical oscillations and emphasized relaxation oscillations. The Duffing oscillator, the Brusselator, and the FitzHugh–Nagumo system extend the same phase-plane viewpoint to nonlinear mechanics, chemical kinetics, and neurophysiology. Together, these examples illustrate the central theme of the chapter: when explicit formulas are unavailable, phase-plane geometry reveals how the motion is organized.

Further reading. For an accessible treatment of planar systems, phase portraits, and nonlinear oscillations, see Strogatz (2024). Biological models, including predator–prey, competition, epidemic, and excitable systems, are developed further in Edelstein-Keshet (2005) and Murray (2002). The systematic theory of invariant sets, periodic attractors, and planar long-time behavior is continued in Chapter 11.

Orientation to Part II

Part II studies how linear systems are assembled and how their coupled motions can be separated into simpler components. A complex model is often built from elementary units—masses, springs, circuit elements, or vertices of a network—whose local interactions are encoded in a matrix. This leads naturally to systems of the form

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y} + \mathbf{f}(t).$$

The matrix $\mathbf{A}$ records how the components are coupled. Its eigenvectors identify modes that evolve independently, while its eigenvalues determine their growth, decay, and oscillation rates. When ordinary eigenvectors do not form a basis, generalized eigenvectors provide the additional modes needed to describe every solution. Thus the analysis reverses the modeling process: after assembling a system from interacting components, we decompose its motion into elementary modal contributions.

Matrix exponentials and Duhamel's principle describe the resulting evolution in time. The Laplace transform provides a complementary operator viewpoint, converting differential equations into algebraic equations involving the resolvent. Convolution, Green's functions, and impulses then describe how a linear system responds to external inputs.

The central progression is

$$\begin{aligned}\text{elementary components} &\longrightarrow \text{coupled linear system} \longrightarrow \text{independent modes} \\ &\longrightarrow \text{evolution operators} \longrightarrow \text{resolvents and transforms.}\end{aligned}$$

Chapter 5

Linear Systems and Modal Analysis

Many physical, biological, and engineering models consist of interacting components. Spring–mass networks, electrical circuits, chemical reaction chains, and diffusion on graphs often lead to linear systems

$$\mathbf{y}' = \mathbf{A}\mathbf{y},$$

where the matrix $\mathbf{A}$ records both the interaction law and the connection structure.

The dynamics of a constant-coefficient linear system are organized by its modes. Eigenvectors produce exponential modes, complex eigenvalues produce oscillatory pairs, and Jordan chains produce generalized modes with polynomial factors. These modes determine growth, decay, oscillation, and stability.

We begin with models of coupled systems and the general theory of linear equations. We then develop modal decomposition, Jordan structure, and linear stability. Finally, the matrix exponential $e^{t\mathbf{A}}$ represents the complete evolution in a single operator.

5.1 Modeling Linear Dynamical Systems

A system with many components is modeled by first identifying the state variables of the individual components and then describing how the components interact. These interactions are often local: each component communicates only with a small number of neighboring components. The resulting connectivity may form a chain or, more generally, a graph.

Once the interaction law is specified, the coupling coefficients are assembled into a matrix. Thus the system matrix is not introduced as an abstract algebraic object; it records both the connection pattern and the physical laws governing the interactions.

The examples in this section illustrate several recurring mechanisms. Spring–mass systems and coupled RLC circuits lead naturally to second-order systems. Chemical reaction chains and diffusion on graphs are naturally first-order. Nearest-neighbor interactions produce tridiagonal matrices and discrete Laplacians, while arbitrary network connections lead to graph Laplacians.

After suitable state variables are introduced, all these models can be written as first-order systems of the form

$$\mathbf{y}' = \mathbf{A}\mathbf{y}.$$

This common form is the starting point for the abstract linear theory developed in the next section.

5.1.1 Coupled spring–mass systems

The coupled spring–mass system is one of the most fundamental models in mechanics. It illustrates how a complex dynamical system is assembled from simple components. Each mass interacts only with its neighboring masses through Hooke's law, so the governing equations are determined by the connection pattern of the springs. This provides a typical example of a *serial connection*, in which local interactions lead naturally to a sparse coefficient matrix.

Two coupled masses Consider two masses m_1 and m_2 connected to fixed walls by springs of stiffness k_1 and k_2, respectively, and coupled to each other by a spring of stiffness k_{12}. Let $x^1(t)$ and $x^2(t)$ denote their displacements from equilibrium, with positive displacement directed toward the right.

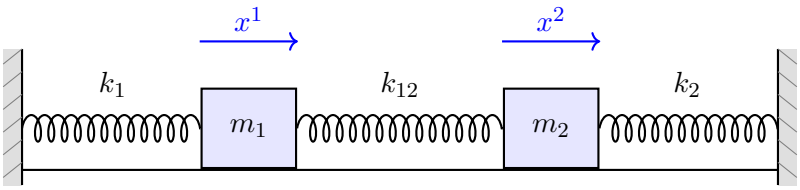


Figure 5.1: Two masses connected to fixed walls by springs of stiffness k_1 and k_2 and coupled by a spring of stiffness k_{12}. The displacements x^1 and x^2 are measured from equilibrium and are positive toward the right.

When the first mass is displaced by x^1 and the second by x^2, the extension of the coupling spring is $x^2 - x^1$. Hence the coupling spring exerts the force $k_{12}(x^2 - x^1)$ on the first mass and the opposite force $k_{12}(x^1 - x^2)$ on the second mass. Newton's second law therefore gives

$$\begin{aligned} m_1\ddot{x}^1 &= -k_1x^1 + k_{12}(x^2 - x^1), \\ m_2\ddot{x}^2 &= -k_2x^2 + k_{12}(x^1 - x^2). \end{aligned}$$

Equivalently,

$$\begin{aligned} m_1\ddot{x}^1 + (k_1 + k_{12})x^1 - k_{12}x^2 &= 0, \\ m_2\ddot{x}^2 - k_{12}x^1 + (k_2 + k_{12})x^2 &= 0. \end{aligned}$$

Introducing

$$\mathbf{x} = \begin{bmatrix} x^1 \\ x^2 \end{bmatrix},$$

the equations take the matrix form

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{0},$$

where

$$\mathbf{M} = \begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix}, \qquad \mathbf{K} = \begin{bmatrix} k_1 + k_{12} & -k_{12} \\ -k_{12} & k_2 + k_{12} \end{bmatrix}.$$

The mass matrix $\mathbf{M}$ and stiffness matrix $\mathbf{K}$ are symmetric. Since all masses and spring constants are positive, both matrices are positive definite.

Equal masses and equal springs Suppose now that

$$m_1 = m_2 = m, \qquad k_1 = k_2 = k_{12} = k.$$

Then

$$m\ddot{x}^1 = k(-2x^1 + x^2), \qquad m\ddot{x}^2 = k(x^1 - 2x^2),$$

or

$$m\begin{bmatrix}\ddot{x}^1\\ \ddot{x}^2\end{bmatrix} + k\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}\begin{bmatrix}x^1\\ x^2\end{bmatrix} = \mathbf{0}.$$

Thus the stiffness matrix is

$$\mathbf{K}_2 = k\begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}.$$

This form suggests the corresponding system for an arbitrary number of equal masses.

The general fixed-end chain Consider a chain of n equal masses, each of mass m, connected by springs of stiffness k. The two endpoints are fixed. If $x^j(t)$ denotes the displacement of the jth mass, the fixed-end conditions are represented by

$$x^0(t) = 0, \qquad x^{n+1}(t) = 0.$$

The jth mass is acted upon by the springs on its left and right. Newton's second law gives

$$m\ddot{x}^j = k(x^{j-1} - x^j) + k(x^{j+1} - x^j),$$

and hence

$$m\ddot{x}^j = k(x^{j-1} - 2x^j + x^{j+1}), \qquad j = 1, \ldots, n.$$

Introduce the displacement vector

$$\mathbf{x} = \begin{bmatrix}x^1\\ \vdots\\ x^n\end{bmatrix}.$$

The system may then be written as

$$m\ddot{\mathbf{x}} + kL_n^{(D)}\mathbf{x} = \mathbf{0},$$

where

$$L_n^{(D)} = \begin{bmatrix}2 & -1 & 0 & \cdots & 0\\ -1 & 2 & -1 & \ddots & \vdots\\ 0 & -1 & 2 & \ddots & 0\\ \vdots & \ddots & \ddots & \ddots & -1\\ 0 & \cdots & 0 & -1 & 2\end{bmatrix}.$$

Thus the stiffness matrix of the chain is

$$\mathbf{K}_n = kL_n^{(D)}.$$

The matrix $L_n^{(D)}$ is the one-dimensional discrete Laplacian with homogeneous Dirichlet boundary conditions. Its action is

$$\left(L_n^{(D)}\mathbf{x}\right)^j = 2x^j - x^{j-1} - x^{j+1},$$

where $x^0 = x^{n+1} = 0$.

Thus the spring force on the jth mass is proportional to the discrete second difference

$$x^{j-1} - 2x^j + x^{j+1}.$$

This nearest-neighbor coupling mechanism is fundamental in vibrating strings, elastic lattices, finite-difference methods, and many other discrete models arising in science and engineering.

Exercise 5.1 (A coupled chain with one free end)**.** Consider a chain of n equal masses, each of mass m, connected by springs of stiffness k. The left end is attached to a fixed wall, while the right end is free. Let $x^j(t)$ denote the displacement of the jth mass from equilibrium.

Derive the equations of motion and show that they can be written as

$$m\ddot{\mathbf{x}} + kL_n^{(DN)}\mathbf{x} = \mathbf{0},$$

where

$$L_n^{(DN)} = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & -1 & 2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix}.$$

Interpret the last equation mechanically and show that the free-end condition may be represented formally by

$$x^{n+1} = x^n.$$

Compare this matrix with the fixed-end discrete Laplacian $L_n^{(D)}$.

5.1.2 Coupled RLC circuits

In Section 3.1.2, we derived the equation for a single series RLC circuit. We now consider a chain of interconnected RLC loops. The new feature is that the equations depend not only on the individual components, but also on the way in which the loops are connected.

An electrical circuit can be represented by a graph

$$G = (V, E).$$

The vertices are the junctions of the circuit, and the edges are the connections between them. Each oriented edge e carries an edge current I_e and an oriented voltage drop V_e. If e is oriented from the vertex v_i to the vertex v_j, then

$$V_e = u_i - u_j,$$

where u_i and u_j are the potentials at the two vertices. Reversing the orientation of the edge changes the signs of both I_e and V_e.

The circuit equations are determined by Kirchhoff's laws and the relations governing the components:

- Kirchhoff's current law states that, at every vertex, the algebraic sum of the outgoing edge currents is zero.
- Kirchhoff's voltage law states that the algebraic sum of the oriented voltage drops around every closed loop is zero.
- For a resistor, capacitor, and inductor, respectively, the edge relations are

$$V_e = RI_e, \qquad I_e = C\dot{V}_e, \qquad V_e = L\dot{I}_e.$$

A capacitively coupled RLC ladder Consider the circuit shown in Figure 5.2. It consists of a chain of RLC loops coupled through their capacitor branches. The horizontal R–L paths carry the loop currents, while Kirchhoff's current law expresses the current through each capacitor as the difference of the currents in the neighboring loops. Consequently, the capacitors generate the nearest-neighbor coupling that leads to a discrete-Laplacian term in the equations for the loop currents.

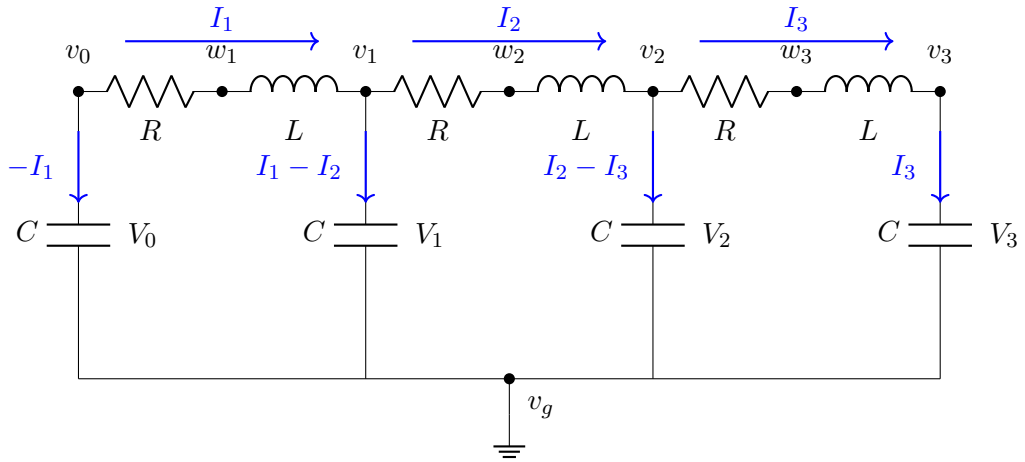


Figure 5.2: A chain of three coupled RLC loops. The variables V_0, V_1, V_2, V_3 are the downward voltage drops across the four capacitors. The currents I_1, I_2, I_3 are oriented from left to right along the horizontal R–L paths. The downward currents through the capacitors are determined by Kirchhoff's current law. The lower wire represents the single ground vertex v_g.

We now consider a chain of n adjacent loops. Let

$$V_0, V_1, \ldots, V_n$$

denote the downward voltage drops across the $n+1$ vertical capacitors, and let

$$I_1, I_2, \ldots, I_n$$

denote the currents through the horizontal R–L paths, oriented from left to right.

Introduce the conventions

$$I_0 = 0, \qquad I_{n+1} = 0.$$

Kirchhoff's current law at the upper vertex v_j shows that the downward current through the jth capacitor is

$$I_j - I_{j+1}, \qquad j = 0, \dots, n.$$

For example, in Figure 5.2, the four downward capacitor currents are

$$-I_1, \qquad I_1 - I_2, \qquad I_2 - I_3, \qquad I_3.$$

Since the current through a capacitor equals C times the derivative of its oriented voltage drop, we obtain

$$C\dot{V}_j = I_j - I_{j+1}, \qquad j = 0, \dots, n.$$

We next apply Kirchhoff's voltage law around the jth loop. Traversing the upper R–L path from left to right gives the voltage drop

$$RI_j + L\dot{I}_j.$$

The downward voltage drop along the right capacitor is V_j, while traversing the left capacitor upward gives the voltage drop $-V_{j-1}$. Therefore,

$$L\dot{I}_j + RI_j + V_j - V_{j-1} = 0, \qquad j = 1, \dots, n.$$

Differentiating this equation gives

$$L\ddot{I}_j + R\dot{I}_j + \dot{V}_j - \dot{V}_{j-1} = 0.$$

Using

$$\dot{V}_j = \frac{1}{C}(I_j - I_{j+1}), \qquad \dot{V}_{j-1} = \frac{1}{C}(I_{j-1} - I_j),$$

we find

$$L\ddot{I}_j + R\dot{I}_j + \frac{1}{C}\,(2I_j - I_{j-1} - I_{j+1}) = 0, \qquad j = 1, \dots, n.$$

Introducing

$$\mathbf{I} = \begin{bmatrix} I_1 \\ \vdots \\ I_n \end{bmatrix},$$

the system becomes

$$L\ddot{\mathbf{I}} + R\dot{\mathbf{I}} + \frac{1}{C}L_n^{(D)}\mathbf{I} = \mathbf{0},$$

where

$$L_n^{(D)} = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}.$$

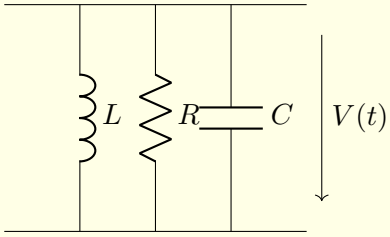
Thus the coupled RLC circuit has the same mathematical form as the coupled spring–mass system:

$$m\ddot{\mathbf{x}} + \gamma\dot{\mathbf{x}} + kL_n^{(D)}\mathbf{x} = \mathbf{0}.$$

The correspondence is

$$m \longleftrightarrow L, \qquad \gamma \longleftrightarrow R, \qquad k \longleftrightarrow \frac{1}{C}.$$

Exercise 5.2 (A source-free parallel RLC circuit)**.** An inductor L, a resistor R, and a capacitor C are connected in parallel with no external source.



Let $V(t)$ be the voltage drop from the upper node to the lower node, and orient all three branch currents downward. Using Kirchhoff's current law and

$$L\dot{I}_L = V, \qquad I_R = \frac{V}{R}, \qquad I_C = C\dot{V},$$

derive a second-order ODE for $V(t)$.

Exercise 5.3 (Energy in a capacitively coupled RLC circuit)**.** For the chain of n coupled RLC loops, suppose

$$C\dot{V}_j = I_j - I_{j+1}, \qquad j = 0, \dots, n, \qquad I_0 = I_{n+1} = 0,$$

and

$$L\dot{I}_j + RI_j + V_j - V_{j-1} = 0, \qquad j = 1, \dots, n.$$

Define

$$\mathcal{H}(t) = \frac{C}{2}\sum_{j=0}^{n} V_j(t)^2 + \frac{L}{2}\sum_{j=1}^{n} I_j(t)^2.$$

1. Show that
$$\frac{d\mathcal{H}}{dt} = -R\sum_{j=1}^{n} I_j(t)^2.$$
2. Interpret this identity when $R > 0$ and when $R = 0$.

Exercise 5.4 (An inductively coupled LC ladder)**.** The LC ladder in Figure 5.3 is obtained from the lossless version of the capacitively coupled ladder in the text by interchanging the

roles of capacitors and inductors.

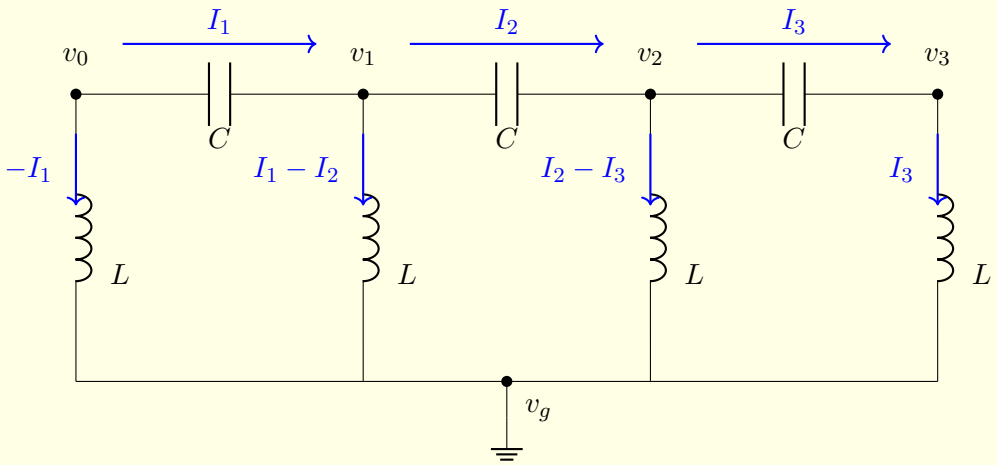


Figure 5.3: Three LC loops coupled through shared inductor branches. This ladder is the circuit-dual counterpart of the capacitively coupled ladder in Figure 5.2.

For a ladder with n loops, let Q_j be the charge on the jth horizontal capacitor and set

$$I_j = \dot{Q}_j, \qquad I_0 = I_{n+1} = 0.$$

Use Kirchhoff's laws to derive

$$L L_n^{(D)} \ddot{\mathbf{Q}} + \frac{1}{C}\mathbf{Q} = \mathbf{0}.$$

Compare this with the lossless capacitively coupled ladder,

$$L\ddot{\mathbf{I}} + \frac{1}{C} L_n^{(D)} \mathbf{I} = \mathbf{0}.$$

Explain how the interchange of capacitors and inductors moves the discrete Laplacian from the restoring term to the inertial term. This illustrates the duality between the two ladder networks.

5.1.3 Chemical reaction chains

Linear systems also arise naturally in chemical kinetics. In this subsection we study a chain of chemical species connected by reversible reactions. This provides another example of a serially connected system: each species interacts only with its nearest neighbors in the chain.

Consider a sequence of chemical species

$$A_1,\ A_2,\ \ldots,\ A_n,$$

connected by reversible reactions

$$A_1 \underset{k}{\overset{k}{\rightleftharpoons}} A_2 \underset{k}{\overset{k}{\rightleftharpoons}} A_3 \underset{k}{\overset{k}{\rightleftharpoons}} \cdots \underset{k}{\overset{k}{\rightleftharpoons}} A_n.$$

Let $c_j(t)$ denote the concentration of the species A_j. According to the law of mass action, the reaction rate is proportional to the concentration of the reactant. For the reversible reaction

$$A_j \underset{k_j^-}{\overset{k_j^+}{\rightleftharpoons}} A_{j+1},$$

the forward flux is $k_j^+ c_j$, and the backward flux is $k_j^- c_{j+1}$. Thus the net flux from A_j to A_{j+1} is

$$k_j^+ c_j - k_j^- c_{j+1}.$$

Three species We first consider the simple chain

$$A_1 \underset{k}{\overset{k}{\rightleftharpoons}} A_2 \underset{k}{\overset{k}{\rightleftharpoons}} A_3.$$

The concentration c_1 loses material to A_2 and gains material from A_2. Hence

$$c_1' = k(c_2 - c_1).$$

Similarly,

$$c_2' = k(c_1 - c_2) + k(c_3 - c_2) = k(c_1 - 2c_2 + c_3),$$

and

$$c_3' = k(c_2 - c_3).$$

Therefore,

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}' = -k \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}.$$

This matrix has the same nearest-neighbor structure as the matrices appearing in spring–mass chains and coupled circuit networks. The entries reflect the fact that each species exchanges material only with adjacent species in the reaction chain.

Remark 5.1 (General reaction rates). The assumption that all reaction rates are equal is made only to highlight the common structure of the models. In a more realistic reaction network, each reaction may have its own forward and backward rates. For example, consider

$$A_1 \underset{k_1^-}{\overset{k_1^+}{\rightleftharpoons}} A_2 \underset{k_2^-}{\overset{k_2^+}{\rightleftharpoons}} A_3.$$

By the law of mass action,

$$c_1' = -k_1^+ c_1 + k_1^- c_2,$$
$$c_2' = k_1^+ c_1 - (k_1^- + k_2^+) c_2 + k_2^- c_3,$$

and

$$c_3' = k_2^+ c_2 - k_2^- c_3.$$

In matrix form,

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}' = \begin{bmatrix} -k_1^+ & k_1^- & 0 \\ k_1^+ & -(k_1^- + k_2^+) & k_2^- \\ 0 & k_2^+ & -k_2^- \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}.$$

The coefficient matrix remains tridiagonal because the reaction chain has only nearest-neighbor interactions. When

$$k_1^+ = k_1^- = k_2^+ = k_2^- = k,$$

this matrix reduces to

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

General reaction chain For a chain of n species with equal reversible reaction rate k, the interior equations are

$$c_j' = k(c_{j-1} - 2c_j + c_{j+1}), \qquad j = 2, \dots, n-1,$$

while the boundary equations are

$$c_1' = k(c_2 - c_1), \qquad c_n' = k(c_{n-1} - c_n).$$

Introducing the concentration vector

$$\mathbf{c} = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix},$$

we obtain

$$\mathbf{c}' = -kL_n\mathbf{c},$$

where

$$L_n = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & -1 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}.$$

The matrix L_n is the one-dimensional discrete Laplacian with Neumann-type boundary conditions. It describes the transfer of material between neighboring species in the reaction chain. The boundary rows are different from the interior rows because the endpoint species have only one neighbor.

Thus chemical reaction chains provide another example of the same general modeling principle:

$$\text{local interaction} \quad \Longrightarrow \quad \text{discrete Laplacian} \quad \Longrightarrow \quad \mathbf{c}' = -kL_n\mathbf{c}.$$

5.1.4 Diffusion on a graph

The previous models all involve nearest-neighbor coupling along a one-dimensional chain. Such coupling leads to a one-dimensional discrete Laplacian. We now generalize this idea by allowing the components to interact according to an arbitrary graph.

Let $G = (V, E)$ be an undirected graph with vertices

$$V = \{1, \dots, n\}.$$

Let $u^j(t)$ denote the amount of a quantity located at vertex j. For example, vertex j may represent the jth city in a transportation network, and $u^j(t)$ may denote the population of that city at time t. People may move between neighboring cities along roads connecting them. If the population difference between two neighboring cities is large, then more people tend to move from the more populated city to the less populated one. This leads naturally to a diffusion process on the graph.

If two vertices i and j are connected by an edge, we assume that the flux from i to j is proportional to the population difference

$$u^i - u^j.$$

Summing all contributions from neighboring vertices, we obtain

$$\dot{u}^j = \kappa \sum_{i \sim j} (u^i - u^j),$$

where $i \sim j$ means that vertices i and j are connected by an edge, and $\kappa > 0$ is the diffusion coefficient.

Equivalently,

$$\dot{u}^j = -\kappa \left(d_j u^j - \sum_{i \sim j} u^i \right),$$

where d_j is the degree of vertex j.

Introducing

$$\mathbf{u} = \begin{bmatrix} u^1 \\ \vdots \\ u^n \end{bmatrix},$$

the system becomes

$$\dot{\mathbf{u}} = -\kappa L_G \mathbf{u},$$

where

$$L_G = D - A_G$$

is the *graph Laplacian*. Here D is the diagonal degree matrix and A_G is the adjacency matrix of the graph.

Thus the graph Laplacian is the matrix representation of local diffusion on a network. Its diagonal entries record how many neighbors each vertex has, while its off-diagonal entries record which vertices are connected by edges.

Example 5.2 (A three-node graph)**.** Consider the graph

$$1 \sim 2, \qquad 2 \sim 3, \qquad 1 \sim 3.$$

This is the triangle graph. Each vertex is connected to the other two. Hence

$$A_G = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}, \qquad D = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Therefore

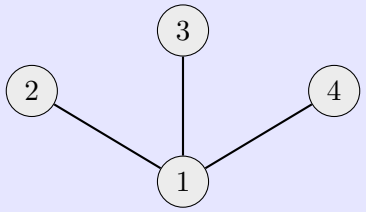
$$L_G = D - A_G = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}.$$

The diffusion equation on this graph is

$$\dot{\mathbf{u}} = -\kappa \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \mathbf{u}.$$

Example 5.3 (Diffusion on a star graph)**.** Consider the graph with four vertices in which vertex 1 is connected to vertices 2, 3, and 4, but the outer vertices are not connected to one another:

$$2 \sim 1, \qquad 3 \sim 1, \qquad 4 \sim 1.$$



The adjacency matrix is

$$A_G = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix},$$

and the degree matrix is

$$D = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Hence the graph Laplacian is

$$L_G = D - A_G = \begin{bmatrix} 3 & -1 & -1 & -1 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}.$$

The diffusion equation on this graph is

$$\dot{\mathbf{u}} = -\kappa L_G \mathbf{u}.$$

Componentwise, this means

$$\dot{u}^1 = \kappa\big(u^2 + u^3 + u^4 - 3u^1\big),$$

and

$$\dot{u}^2 = \kappa(u^1 - u^2), \qquad \dot{u}^3 = \kappa(u^1 - u^3), \qquad \dot{u}^4 = \kappa(u^1 - u^4).$$

This example is different from the one-dimensional chain. The central vertex interacts with three neighboring vertices at once, so the coupling matrix is no longer tridiagonal. Nevertheless, the same principle remains: each row of L_G records the local connections of one vertex.

Remark 5.4 (The path graph). The one-dimensional chain is a special case of graph diffusion. If G is the path graph

$$1 - 2 - \cdots - n,$$

then L_G becomes the one-dimensional discrete Laplacian with Neumann-type boundary conditions:

$$L_G = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & -1 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}.$$

Thus graph diffusion extends the idea of nearest-neighbor coupling from a line to an arbitrary network.

Exercise 5.5 (Migration on a transportation network)**.** Five cities are connected by

$$1 \sim 2, \qquad 1 \sim 3, \qquad 2 \sim 4, \qquad 3 \sim 4, \qquad 4 \sim 5.$$

Let $u^j(t)$ be the population of city j, and suppose migration along each road is proportional to the population difference between its endpoints.

1. Draw the graph, construct its Laplacian L_G, and write the model as

$$\dot{\mathbf{u}} = -\kappa L_G \mathbf{u}.$$

2. Write the five scalar differential equations.
3. Show that the total population $u^1 + \cdots + u^5$ is conserved.

5.1.5 State-space formulation

The abstract theory developed in this chapter applies to first-order systems of the form

$$\mathbf{y}' = \mathbf{A}\mathbf{y}.$$

Some of the models derived above already have this form. For example, the chemical reaction chain and graph diffusion are first-order systems. The spring–mass system and the coupled RLC circuit, however, are second-order systems. We now show how to rewrite them as first-order systems by enlarging the state variables.

Spring–mass system Recall that the coupled spring–mass system satisfies

$$m\ddot{\mathbf{x}} + K_n\mathbf{x} = \mathbf{0}.$$

Introduce the velocity vector

$$\mathbf{v} = \dot{\mathbf{x}},$$

and define the state vector

$$\mathbf{y} = \begin{bmatrix} \mathbf{x} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{x} \\ \dot{\mathbf{x}} \end{bmatrix}.$$

Then

$$\dot{\mathbf{y}} = \begin{bmatrix} \dot{\mathbf{x}} \\ \dot{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \mathbf{v} \\ -\frac{1}{m}K_n\mathbf{x} \end{bmatrix}.$$

Hence

$$\boxed{\dot{\mathbf{y}} = \begin{bmatrix} 0 & I \\ -\frac{1}{m}K_n & 0 \end{bmatrix}\mathbf{y}.}$$

Thus the second-order spring–mass system is equivalent to a first-order linear system

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y},$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & I \\ -\frac{1}{m}K_n & 0 \end{bmatrix}.$$

Coupled RLC circuit The coupled RLC circuit satisfies

$$L\ddot{\mathbf{I}} + R\dot{\mathbf{I}} + \frac{1}{C}L_n^{(D)}\mathbf{I} = \mathbf{0}.$$

Introduce

$$\mathbf{J} = \dot{\mathbf{I}},$$

and define the state vector

$$\mathbf{y} = \begin{bmatrix} \mathbf{I} \\ \mathbf{J} \end{bmatrix} = \begin{bmatrix} \mathbf{I} \\ \dot{\mathbf{I}} \end{bmatrix}.$$

Then

$$\dot{\mathbf{y}} = \begin{bmatrix} \mathbf{J} \\ -\frac{1}{L}\left(R\mathbf{J} + \frac{1}{C}L_n^{(D)}\mathbf{I}\right) \end{bmatrix}.$$

Therefore

$$\boxed{\dot{\mathbf{y}} = \begin{bmatrix} 0 & I \\ -\frac{1}{LC}L_n^{(D)} & -\frac{R}{L}I \end{bmatrix}\mathbf{y}.}$$

Again we obtain a first-order linear system of the form

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}.$$

Summary The models discussed in this section all lead to linear systems with constant coefficients. Some systems, such as chemical reaction chains and graph diffusion, are naturally first-order. Others, such as spring–mass systems and RLC circuits, become first-order systems after introducing additional state variables.

Thus a wide range of interconnected systems can be written in the common form

$$\mathbf{y}' = \mathbf{A}\mathbf{y}.$$

This observation is the starting point of the abstract linear theory developed in the next section.

Exercise 5.6 (State-space form and modal eigenvalues)**.** Consider

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{D}\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{0},$$

where $\mathbf{M}$ is invertible and $\mathbf{M}$, $\mathbf{D}$, and $\mathbf{K}$ are constant $n \times n$ matrices.

1. Write the equation as a first-order system

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}, \qquad \mathbf{y} = \begin{bmatrix} \mathbf{x} \\ \dot{\mathbf{x}} \end{bmatrix}.$$

2. Show that λ is an eigenvalue of $\mathbf{A}$ if and only if

$$\det\left(\lambda^2\mathbf{M} + \lambda\mathbf{D} + \mathbf{K}\right) = 0.$$

3. Compute the eigenvalues of the state-space matrix when

$$\mathbf{M} = \mathbf{I}, \qquad \mathbf{D} = \begin{bmatrix} 5 & -3 \\ -3 & 5 \end{bmatrix}, \qquad \mathbf{K} = \begin{bmatrix} 4 & -2 \\ -2 & 4 \end{bmatrix}.$$

   Identify the two modes and determine their damping types.

5.2 Abstract Linear Theory

The models in the previous section lead to constant-coefficient systems of the form

$$\mathbf{y}' = \mathbf{A}\mathbf{y}.$$

The basic linear theory, however, applies more generally to

$$\mathbf{y}' = \mathbf{A}(t)\mathbf{y},$$

where $\mathbf{A}(t)$ is a continuous $n \times n$ matrix-valued function on an interval I.

In this section, we study the structure of the set of solutions rather than methods for computing them explicitly. The existence–uniqueness theorem shows that a solution is determined by its value at one time. Linearity then implies that the solutions of the homogeneous system form an n-dimensional vector space. A basis of this space may be assembled into a fundamental matrix, which represents every solution of the homogeneous system.

We then consider the nonhomogeneous system

$$\mathbf{y}' = \mathbf{A}(t)\mathbf{y} + \mathbf{f}(t)$$

and derive the variation-of-constants formula. This formula separates the evolution of the initial data from the contribution of the forcing term and is the system version of Duhamel's principle.

5.2.1 Existence and uniqueness

We begin with the homogeneous linear system

$$\mathbf{y}'(t) = \mathbf{A}(t)\mathbf{y}(t), \tag{5.1}$$

where

$$\mathbf{y}(t) = \begin{bmatrix} y^1(t) \\ \vdots \\ y^n(t) \end{bmatrix},$$

and $\mathbf{A}(t)$ is a continuous $n \times n$ matrix-valued function on an interval $I \subset \mathbb{R}$.

Writing the system componentwise, we have

$$\frac{dy^i}{dt} = \sum_{j=1}^{n} a_{ij}(t) y^j, \qquad i = 1, \dots, n.$$

Since the right-hand side is continuous in t and linear in $\mathbf{y}$, it is locally Lipschitz with respect to $\mathbf{y}$. Therefore, the Cauchy–Lipschitz theorem gives a unique local solution.

For a linear system, this solution extends throughout the interval I. Indeed, let $J \subset I$ be a compact interval containing t_0. Since $\mathbf{A}$ is continuous, there exists $M > 0$ such that

$$\|\mathbf{A}(t)\| \le M, \qquad t \in J.$$

The integral equation

$$\mathbf{y}(t) = \mathbf{y}_0 + \int_{t_0}^{t} \mathbf{A}(s)\mathbf{y}(s)\, ds$$

and Gronwall's inequality give

$$\|\mathbf{y}(t)\| \le \|\mathbf{y}_0\| e^{M|t-t_0|}, \qquad t \in J.$$

Thus the solution remains bounded on every compact subinterval of I and can be continued throughout I.

Theorem 5.5 (Existence and uniqueness)**.** *Let $\mathbf{A} : I \to \mathbb{R}^{n\times n}$ be continuous on an interval I. For every $t_0 \in I$ and every $\mathbf{y}_0 \in \mathbb{R}^n$, the initial value problem*

$$\begin{cases} \mathbf{y}'(t) = \mathbf{A}(t)\mathbf{y}(t), \\ \mathbf{y}(t_0) = \mathbf{y}_0 \end{cases}$$

has a unique solution defined on the entire interval I.

5.2.2 The solution space

Let

$$\mathcal{S}_0 := \{\mathbf{y}(\cdot) : \mathbf{y}' = \mathbf{A}(t)\mathbf{y}\}$$

denote the set of all solutions of the homogeneous system.

Because the differential equation is linear, any linear combination of solutions is again a solution.

Proposition 5.6. *The set $\mathcal{S}_0$ is a vector space.*

Proof. Let $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{S}_0$ and let $c_1, c_2 \in \mathbb{R}$. Since

$$\mathbf{y}_1' = \mathbf{A}(t)\mathbf{y}_1, \qquad \mathbf{y}_2' = \mathbf{A}(t)\mathbf{y}_2,$$

we have

$$(c_1\mathbf{y}_1 + c_2\mathbf{y}_2)' = c_1\mathbf{y}_1' + c_2\mathbf{y}_2' = \mathbf{A}(t)(c_1\mathbf{y}_1 + c_2\mathbf{y}_2).$$

Hence $c_1\mathbf{y}_1 + c_2\mathbf{y}_2$ is again a solution. Therefore $\mathcal{S}_0$ is a vector space. □

The existence and uniqueness theorem determines the dimension of this vector space.

Theorem 5.7. *Let $\mathcal{S}_0$ be the solution space of*

$$\mathbf{y}' = \mathbf{A}(t)\mathbf{y},$$

where $\mathbf{A}(t)$ is continuous on an interval I. Then $\mathcal{S}_0$ is an n-dimensional vector space.

Proof. Fix $t_0 \in I$. For every initial value

$$\mathbf{y}(t_0) = \mathbf{y}_0,$$

there exists a unique solution of (5.1). Thus the mapping

$$\mathcal{S}_0 \longrightarrow \mathbb{R}^n, \qquad \mathbf{y}(\cdot) \longmapsto \mathbf{y}(t_0),$$

is a linear isomorphism. Hence $\mathcal{S}_0$ is isomorphic to $\mathbb{R}^n$, and therefore

$$\dim \mathcal{S}_0 = n.$$

□

The previous theorem generalizes a familiar fact from ordinary differential equations: the solution space of a homogeneous second-order linear equation is two-dimensional. More generally, an n-dimensional homogeneous linear system has exactly n linearly independent solutions.

The importance of the existence and uniqueness theorem extends beyond the existence of solutions. It identifies each solution uniquely by its initial value. Consequently, many questions about solutions can be translated into questions about vectors in $\mathbb{R}^n$. This observation will serve as the foundation for the remainder of this chapter.

5.2.3 Fundamental solutions and fundamental matrices

Let

$$\mathbf{y}_1(t), \dots, \mathbf{y}_m(t)$$

be solutions of

$$\mathbf{y}' = \mathbf{A}(t)\mathbf{y}.$$

They are said to be linearly independent if

$$c_1\mathbf{y}_1(t) + \cdots + c_m\mathbf{y}_m(t) = 0$$

for all $t \in I$ implies

$$c_1 = \cdots = c_m = 0.$$

Since the solution space $\mathcal{S}_0$ is n-dimensional, it has a basis consisting of n solutions,

$$\mathbf{y}_1(t), \dots, \mathbf{y}_n(t).$$

Such a basis is called a *fundamental set of solutions*. Every solution can then be written uniquely as

$$\mathbf{y}(t) = c_1\mathbf{y}_1(t) + \cdots + c_n\mathbf{y}_n(t).$$

The matrix obtained by placing the solutions of a fundamental set in its columns,

$$\mathbf{Y}(t) = \begin{bmatrix} \mathbf{y}_1(t) & \cdots & \mathbf{y}_n(t) \end{bmatrix},$$

is called a *fundamental matrix*. The general solution may then be written as

$$\mathbf{y}(t) = \mathbf{Y}(t)\mathbf{c}, \qquad \mathbf{c} = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}.$$

Since each column of $\mathbf{Y}(t)$ is a solution,

$$\mathbf{Y}'(t) = \mathbf{A}(t)\mathbf{Y}(t).$$

The determinant

$$W(t) = \det \mathbf{Y}(t)$$

is called the Wronskian of the solutions $\mathbf{y}_1, \dots, \mathbf{y}_n$. It detects whether these solutions are linearly independent at time t.

Proposition 5.8 (Liouville's formula)**.** *Let $\mathbf{Y}(t)$ be a matrix solution of*

$$\mathbf{Y}'(t) = \mathbf{A}(t)\mathbf{Y}(t),$$

and let

$$W(t) = \det \mathbf{Y}(t).$$

Then

$$\boxed{W(t) = W(t_0) \exp\left(\int_{t_0}^{t} \operatorname{tr} \mathbf{A}(s)\, ds \right).}$$

Consequently, if $W(t_0) \neq 0$ for one point t_0, then $W(t) \neq 0$ for all t in the interval of existence. In particular, the solutions $\mathbf{y}_1, \dots, \mathbf{y}_n$ are linearly independent at one point if and only if they are linearly independent at every point.

Proof. We first prove the differential equation for $W(t)$. Let

$$\mathbf{Y}(t) = \begin{bmatrix} \mathbf{y}_1(t) & \cdots & \mathbf{y}_n(t) \end{bmatrix},$$

where $\mathbf{y}_j(t)$ denotes the jth column. By the multilinearity of the determinant with respect to its columns,

$$\frac{d}{dt} \det \mathbf{Y}(t) = \sum_{j=1}^{n} \det \begin{bmatrix} \mathbf{y}_1(t) & \cdots & \mathbf{y}_j'(t) & \cdots & \mathbf{y}_n(t) \end{bmatrix}.$$

Since $\mathbf{Y}' = \mathbf{A}(t)\mathbf{Y}$, the jth column satisfies

$$\mathbf{y}_j'(t) = \mathbf{A}(t)\mathbf{y}_j(t).$$

Thus

$$W'(t) = \sum_{j=1}^{n} \det \begin{bmatrix} \mathbf{y}_1 & \cdots & \mathbf{A}(t)\mathbf{y}_j & \cdots & \mathbf{y}_n \end{bmatrix}.$$

We now use the following elementary identity: for any square matrix $\mathbf{B}$ and any matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$,

$$\sum_{j=1}^{n} \det \begin{bmatrix} \mathbf{y}_1 & \cdots & \mathbf{B}\mathbf{y}_j & \cdots & \mathbf{y}_n \end{bmatrix} = (\operatorname{tr} \mathbf{B}) \det \mathbf{Y}.$$

Applying this identity with $\mathbf{B} = \mathbf{A}(t)$ gives

$$W'(t) = \operatorname{tr} \mathbf{A}(t)\, W(t).$$

Solving this scalar linear equation yields

$$W(t) = W(t_0) \exp\left(\int_{t_0}^{t} \operatorname{tr} \mathbf{A}(s)\, ds \right).$$

It follows immediately that if $W(t_0) \neq 0$, then $W(t) \neq 0$ for all t in the interval. Since $W(t) = \det \mathbf{Y}(t)$, this means that the vectors $\mathbf{y}_1(t), \ldots, \mathbf{y}_n(t)$ are linearly independent at one time if and only if they are linearly independent at every time. □

Exercise 5.7 (Computing a Wronskian)**.** Let

$$\mathbf{y}_1(t) = e^{t} \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \qquad \mathbf{y}_2(t) = e^{3t} \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad \mathbf{Y}(t) = \begin{bmatrix} \mathbf{y}_1(t) & \mathbf{y}_2(t) \end{bmatrix}.$$

1. Compute $\det \mathbf{Y}(t)$ and determine whether $\mathbf{y}_1, \mathbf{y}_2$ are linearly independent.
2. Find the constant matrix $\mathbf{A}$ for which $\mathbf{Y}' = \mathbf{A}\mathbf{Y}$.

Exercise 5.8 (Using Liouville's formula)**.** Suppose $\mathbf{Y}(t)$ is a fundamental matrix for $\mathbf{y}' = \mathbf{A}(t)\mathbf{y}$, where

$$\mathbf{A}(t) = \begin{bmatrix} 1 & t \\ 0 & -2 \end{bmatrix}, \qquad \det \mathbf{Y}(0) = 3.$$

1. Use Liouville's formula to compute $\det \mathbf{Y}(t)$.

2. Deduce that the columns of $\mathbf{Y}(t)$ remain linearly independent for all t.

5.2.4 Representation of solutions

Let $\mathbf{Y}(t)$ be a fundamental matrix of the homogeneous system

$$\mathbf{y}' = \mathbf{A}(t)\mathbf{y}.$$

Since the columns of $\mathbf{Y}(t)$ form a basis of the solution space, every solution of the homogeneous system can be written uniquely in the form

$$\mathbf{y}(t) = \mathbf{Y}(t)\mathbf{c},$$

where $\mathbf{c} \in \mathbb{R}^n$ is a constant vector.

Now consider the initial value problem

$$\mathbf{y}' = \mathbf{A}(t)\mathbf{y}, \qquad \mathbf{y}(t_0) = \mathbf{y}_0.$$

Since every solution has the form

$$\mathbf{y}(t) = \mathbf{Y}(t)\mathbf{c},$$

substituting $t = t_0$ gives

$$\mathbf{y}_0 = \mathbf{Y}(t_0)\mathbf{c}.$$

Because $\mathbf{Y}(t_0)$ is invertible,

$$\mathbf{c} = \mathbf{Y}(t_0)^{-1}\mathbf{y}_0.$$

Therefore

$$\boxed{\mathbf{y}(t) = \mathbf{Y}(t)\mathbf{Y}(t_0)^{-1}\mathbf{y}_0.}$$

The matrix

$$\mathbf{\Phi}(t, t_0) = \mathbf{Y}(t)\mathbf{Y}(t_0)^{-1}$$

is called the *transition matrix* from time t_0 to time t. It maps the initial state $\mathbf{y}_0$ at time t_0 to the state $\mathbf{y}(t)$ at time t:

$$\mathbf{y}(t) = \mathbf{\Phi}(t, t_0)\mathbf{y}_0.$$

It satisfies

$$\mathbf{\Phi}(t_0, t_0) = \mathbf{I}, \qquad \frac{\partial}{\partial t}\mathbf{\Phi}(t, t_0) = \mathbf{A}(t)\mathbf{\Phi}(t, t_0).$$

Moreover, for any $r, s, t \in I$,

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

and hence

$$\mathbf{\Phi}(t, s)^{-1} = \mathbf{\Phi}(s, t).$$

5.2.5 Nonhomogeneous systems and Duhamel's principle

We now consider the nonhomogeneous linear system

$$\mathbf{y}' = \mathbf{A}(t)\mathbf{y} + \mathbf{f}(t),$$

where $\mathbf{A} : I \to \mathbb{R}^{n\times n}$ and $\mathbf{f} : I \to \mathbb{R}^n$ are continuous.

The guiding idea is that the homogeneous system already provides a basis of solutions through the fundamental matrix $\mathbf{Y}(t)$. For the nonhomogeneous equation, we allow the coefficient vector to vary with time. Thus we look for a solution of the form

$$\mathbf{y}(t) = \mathbf{Y}(t)\mathbf{c}(t).$$

Differentiating gives

$$\mathbf{y}'(t) = \mathbf{Y}'(t)\mathbf{c}(t) + \mathbf{Y}(t)\mathbf{c}'(t).$$

Since $\mathbf{Y}'(t) = \mathbf{A}(t)\mathbf{Y}(t)$, this becomes

$$\mathbf{y}'(t) = \mathbf{A}(t)\mathbf{Y}(t)\mathbf{c}(t) + \mathbf{Y}(t)\mathbf{c}'(t) = \mathbf{A}(t)\mathbf{y}(t) + \mathbf{Y}(t)\mathbf{c}'(t).$$

Comparing with

$$\mathbf{y}' = \mathbf{A}(t)\mathbf{y} + \mathbf{f}(t),$$

we obtain

$$\mathbf{Y}(t)\mathbf{c}'(t) = \mathbf{f}(t).$$

Since $\mathbf{Y}(t)$ is invertible,

$$\mathbf{c}'(t) = \mathbf{Y}(t)^{-1}\mathbf{f}(t).$$

Integrating from t_0 to t gives

$$\mathbf{c}(t) = \mathbf{c}(t_0) + \int_{t_0}^{t} \mathbf{Y}(s)^{-1}\mathbf{f}(s)\, ds.$$

If $\mathbf{y}(t_0) = \mathbf{y}_0$, then

$$\mathbf{y}_0 = \mathbf{Y}(t_0)\mathbf{c}(t_0),$$

so

$$\mathbf{c}(t_0) = \mathbf{Y}(t_0)^{-1}\mathbf{y}_0.$$

Hence

$$\mathbf{y}(t) = \mathbf{Y}(t)\mathbf{Y}(t_0)^{-1}\mathbf{y}_0 + \mathbf{Y}(t)\int_{t_0}^{t} \mathbf{Y}(s)^{-1}\mathbf{f}(s)\, ds.$$

Equivalently, using the transition matrix

$$\mathbf{\Phi}(t, t_0) = \mathbf{Y}(t)\mathbf{Y}(t_0)^{-1},$$

the solution can be written as

$$\boxed{\mathbf{y}(t) = \mathbf{\Phi}(t, t_0)\mathbf{y}_0 + \int_{t_0}^{t} \mathbf{\Phi}(t, s)\mathbf{f}(s)\, ds.}$$

The first term describes the evolution of the initial state under the homogeneous dynamics. The integral term is the accumulated contribution of the forcing: a contribution created at time s is transported from time s to time t by the transition matrix $\mathbf{\Phi}(t, s)$. This representation is the variation-of-constants formula and expresses *Duhamel's principle*.

Example 5.9 (Sinusoidal forcing and synchronized response)**.** Consider the forced linear system

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y} + \Re\left(e^{i\omega t}\mathbf{b}\right),$$

where $\omega \in \mathbb{R}$. We seek a particular solution oscillating with the same frequency as the forcing:

$$\mathbf{y}_{\text{sync}}(t) = \Re\left(e^{i\omega t}\mathbf{c}\right).$$

Substitution into the complexified equation gives

$$i\omega\mathbf{c} = \mathbf{A}\mathbf{c} + \mathbf{b}.$$

If $i\omega$ is not an eigenvalue of $\mathbf{A}$, then

$$\mathbf{c} = (i\omega\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}.$$

Therefore

$$\boxed{\mathbf{y}_{\text{sync}}(t) = \Re\left[e^{i\omega t}(i\omega\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}\right].}$$

The response has the same frequency as the forcing, but its amplitude and phase are changed by the matrix

$$(i\omega\mathbf{I} - \mathbf{A})^{-1}.$$

This matrix is the frequency-response operator of the system.

If every solution of the homogeneous system

$$\dot{\mathbf{z}} = \mathbf{A}\mathbf{z}$$

tends to $\mathbf{0}$, then the difference between any solution and the synchronized solution satisfies the homogeneous equation. Consequently,

$$\mathbf{y}(t) - \mathbf{y}_{\text{sync}}(t) \longrightarrow \mathbf{0}.$$

Thus the synchronized solution is the steady periodic response, while the homogeneous solution represents the transient response. A criterion for this decay in terms of the eigenvalues of $\mathbf{A}$ will be developed later in the chapter. This system-theoretic formulation should be compared with the damped forced oscillator studied in Section 3.5.2.

Exercise 5.9 (Duhamel's formula)**.** Use Duhamel's formula to solve

$$\mathbf{y}' = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}\mathbf{y} + \begin{bmatrix} e^{-t} \\ 1 \end{bmatrix}, \qquad \mathbf{y}(0) = \mathbf{0}.$$

Verify the resulting solution by direct substitution, and explain the appearance of the factor te^{-t} in its first component.

Exercise 5.10 (Resonance in a rotating system)**.** Consider

$$\mathbf{y}' = \begin{bmatrix} 0 & -\omega \\ \omega & 0 \end{bmatrix} \mathbf{y} + \begin{bmatrix} \cos \nu t \\ \sin \nu t \end{bmatrix}, \qquad \mathbf{y}(0) = \mathbf{0},$$

with principal fundamental matrix

$$\mathbf{Y}(t) = \begin{bmatrix} \cos \omega t & -\sin \omega t \\ \sin \omega t & \cos \omega t \end{bmatrix}.$$

Use Duhamel's formula to solve the system. Show that the solution is bounded when $\nu \neq \omega$, whereas for $\nu = \omega$,

$$\mathbf{y}(t) = t \begin{bmatrix} \cos \omega t \\ \sin \omega t \end{bmatrix}.$$

Explain the resonance geometrically.

5.3 Modal Representation

We now specialize to the constant-coefficient homogeneous system

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}, \qquad \mathbf{A} \in \mathbb{R}^{n \times n}.$$

The simplest solutions are obtained by separating the time dependence from the direction in the state space. This leads to solutions of the form

$$\mathbf{y}(t) = e^{\lambda t}\mathbf{v},$$

where λ is an eigenvalue of $\mathbf{A}$ and $\mathbf{v}$ is a corresponding eigenvector. We call such a solution an *ordinary mode*.

For a real matrix, nonreal eigenvalues occur in complex conjugate pairs. Although the associated eigenvectors and exponential modes are complex, their real and imaginary parts give real solutions.

When eigenvectors coalesce, ordinary modes no longer provide enough independent solutions. The missing solutions contain polynomial factors multiplying the exponential term. These are the *generalized modes* associated with Jordan chains.

The purpose of this section is to identify these elementary building blocks. The question of whether they form a complete basis of the solution space will be addressed in the next two sections.

5.3.1 Eigenvectors and ordinary modes

We begin with the exponential ansatz

$$\mathbf{y}(t) = e^{\lambda t}\mathbf{v},$$

where $\lambda \in \mathbb{C}$ and $\mathbf{v} \in \mathbb{C}^n$ is a constant nonzero vector. Substitution into the differential equation gives

$$\lambda e^{\lambda t}\mathbf{v} = e^{\lambda t}\mathbf{A}\mathbf{v}.$$

Since $e^{\lambda t} \neq 0$, this reduces to

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}.$$

Thus the construction of exponential solutions is equivalent to the eigenvalue problem for $\mathbf{A}$.

Proposition 5.10 (Ordinary modes)**.** *Let $\lambda \in \mathbb{C}$ and $\mathbf{v} \in \mathbb{C}^n$, with $\mathbf{v} \neq \mathbf{0}$. Then*

$$\mathbf{y}(t) = e^{\lambda t}\mathbf{v}$$

is a solution of

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}$$

if and only if

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}.$$

Proof. If $\mathbf{y}(t) = e^{\lambda t}\mathbf{v}$, then

$$\dot{\mathbf{y}}(t) = \lambda e^{\lambda t}\mathbf{v}.$$

Hence $\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}$ if and only if

$$\lambda e^{\lambda t}\mathbf{v} = e^{\lambda t}\mathbf{A}\mathbf{v},$$

which is equivalent to

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}.$$

□

The vector $\mathbf{v}$ describes the spatial or geometric pattern of the mode, while $e^{\lambda t}$ describes its time dependence. Thus the eigenvalue problem separates the state-space structure from the temporal evolution.

If $\mathbf{v}_1, \ldots, \mathbf{v}_m$ are eigenvectors corresponding to eigenvalues $\lambda_1, \ldots, \lambda_m$, then linearity gives the solution

$$\mathbf{y}(t) = c_1 e^{\lambda_1 t}\mathbf{v}_1 + \cdots + c_m e^{\lambda_m t}\mathbf{v}_m.$$

At this stage, however, we do not assume that the eigenvectors form a basis of $\mathbb{C}^n$. That structural question will be studied later.

Exercise 5.11 (Ordinary modes)**.** For

$$\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix},$$

find two independent exponential modes of $\mathbf{y}' = \mathbf{A}\mathbf{y}$. Use them to solve the initial-value problem

$$\mathbf{y}(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

and describe which mode dominates as $t \to \infty$.

5.3.2 Complex modes and real modal pairs

Suppose that $\mathbf{A} \in \mathbb{R}^{n\times n}$ has a nonreal eigenvalue

$$\lambda = \alpha + i\beta, \qquad \beta \neq 0,$$

with eigenvector

$$\mathbf{v} = \mathbf{p} + i\mathbf{q}, \qquad \mathbf{p}, \mathbf{q} \in \mathbb{R}^n.$$

Then

$$\begin{aligned} e^{\lambda t}\mathbf{v} &= e^{\alpha t}\big(\cos\beta t + i\sin\beta t\big)(\mathbf{p}+i\mathbf{q}) \\ &= e^{\alpha t}\big(\cos\beta t\,\mathbf{p} - \sin\beta t\,\mathbf{q}\big) \\ &\qquad + ie^{\alpha t}\big(\sin\beta t\,\mathbf{p} + \cos\beta t\,\mathbf{q}\big). \end{aligned}$$

Since $\mathbf{A}$ is real, the real and imaginary parts of a complex solution are real solutions.

Proposition 5.11 (Complex modes and real modal pairs)**.** *Let*

$$\lambda = \alpha + i\beta, \qquad \mathbf{v} = \mathbf{p} + i\mathbf{q}, \qquad \beta \neq 0.$$

Then

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$$

if and only if

$$\mathbf{y}_1(t) = e^{\alpha t}\big(\cos\beta t\,\mathbf{p} - \sin\beta t\,\mathbf{q}\big)$$

and

$$\mathbf{y}_2(t) = e^{\alpha t}\big(\sin\beta t\,\mathbf{p} + \cos\beta t\,\mathbf{q}\big)$$

are solutions of

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}.$$

Proof. If $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$, then $e^{\lambda t}\mathbf{v}$ is a complex solution, and $\mathbf{y}_1$ and $\mathbf{y}_2$ are its real and imaginary parts.

Conversely, suppose that $\mathbf{y}_1$ and $\mathbf{y}_2$ are solutions. Since

$$\mathbf{y}_1(0) = \mathbf{p}, \qquad \mathbf{y}_2(0) = \mathbf{q},$$

and

$$\dot{\mathbf{y}}_1(0) = \alpha\mathbf{p} - \beta\mathbf{q}, \qquad \dot{\mathbf{y}}_2(0) = \beta\mathbf{p} + \alpha\mathbf{q},$$

we obtain

$$\mathbf{A}\mathbf{p} = \alpha\mathbf{p} - \beta\mathbf{q}, \qquad \mathbf{A}\mathbf{q} = \beta\mathbf{p} + \alpha\mathbf{q}.$$

Therefore,

$$\mathbf{A}(\mathbf{p}+i\mathbf{q}) = (\alpha+i\beta)(\mathbf{p}+i\mathbf{q}).$$

□

The conjugate eigenvalue

$$\overline{\lambda} = \alpha - i\beta$$

has eigenvector

$$\overline{\mathbf{v}} = \mathbf{p} - i\mathbf{q}$$

and yields the same pair of real solutions. The vectors $\mathbf{p}$ and $\mathbf{q}$ are linearly independent over $\mathbb{R}$. Indeed, if they were linearly dependent, then the complex eigenvector $\mathbf{v} = \mathbf{p} + i\mathbf{q}$ could be written as

$$\mathbf{v} = z\mathbf{r}$$

for some nonzero $z \in \mathbb{C}$ and some nonzero $\mathbf{r} \in \mathbb{R}^n$. The eigenvalue equation would then imply

$$\mathbf{A}\mathbf{r} = \lambda\mathbf{r}.$$

This is impossible because $\mathbf{Ar}$ is real, whereas $\lambda\mathbf{r}$ is nonreal when $\beta \neq 0$.

Let

$$\mathbf{P} = \begin{bmatrix} \mathbf{p} & \mathbf{q} \end{bmatrix}.$$

Then the restriction of $\mathbf{A}$ to this invariant subspace is represented, with respect to the basis $\{\mathbf{p}, \mathbf{q}\}$, by

$$\mathbf{C} = \begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix};$$

equivalently,

$$\mathbf{AP} = \mathbf{PC}.$$

Example 5.12 (Rotation about a fixed axis)**.** Consider

$$\dot{\mathbf{y}} = \boldsymbol{\omega} \times \mathbf{y} = \widehat{\boldsymbol{\omega}}\, \mathbf{y},$$

where

$$\widehat{\boldsymbol{\omega}} = \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix}, \qquad \omega = |\boldsymbol{\omega}| > 0.$$

The characteristic polynomial is

$$\det(\lambda \mathbf{I} - \widehat{\boldsymbol{\omega}}) = \lambda(\lambda^2 + \omega^2),$$

so the eigenvalues are

$$0, \qquad i\omega, \qquad -i\omega.$$

Set

$$\mathbf{e} = \frac{\boldsymbol{\omega}}{\omega}.$$

Since

$$\widehat{\boldsymbol{\omega}}\mathbf{e} = \boldsymbol{\omega} \times \mathbf{e} = \mathbf{0},$$

the eigenvector associated with 0 is $\mathbf{e}$, which is parallel to $\boldsymbol{\omega}$.

Choose a unit vector $\mathbf{p} \perp \mathbf{e}$ and define

$$\mathbf{q} = \mathbf{p} \times \mathbf{e}.$$

Then $\mathbf{p}, \mathbf{q}, \mathbf{e}$ form an orthonormal basis, and

$$\widehat{\boldsymbol{\omega}}\mathbf{p} = -\omega\mathbf{q}, \qquad \widehat{\boldsymbol{\omega}}\mathbf{q} = \omega\mathbf{p}.$$

It follows that

$$\widehat{\boldsymbol{\omega}}(\mathbf{p} + i\mathbf{q}) = i\omega(\mathbf{p} + i\mathbf{q}),$$

while

$$\widehat{\boldsymbol{\omega}}(\mathbf{p} - i\mathbf{q}) = -i\omega(\mathbf{p} - i\mathbf{q}).$$

Thus the eigenvalues and corresponding eigenvectors are

$$\begin{array}{c|c} 0 & \mathbf{e} \\ i\omega & \mathbf{p}+i\mathbf{q} \\ -i\omega & \mathbf{p}-i\mathbf{q} \end{array}.$$

By Proposition 5.11, the mode

$$e^{i\omega t}(\mathbf{p}+i\mathbf{q})$$

gives the real modal pair

$$\mathbf{y}_1(t) = \cos\omega t\,\mathbf{p} - \sin\omega t\,\mathbf{q},$$
$$\mathbf{y}_2(t) = \sin\omega t\,\mathbf{p} + \cos\omega t\,\mathbf{q}.$$

Together with the zero mode $\mathbf{e}$, these form a basis of solutions. Hence every solution is

$$\mathbf{y}(t) = a\mathbf{y}_1(t) + b\mathbf{y}_2(t) + c\mathbf{e}.$$

The component parallel to $\boldsymbol{\omega}$ remains constant, while the perpendicular component rotates about the axis $\boldsymbol{\omega}$.

Exercise 5.12 (Fixed-axis rotation and Rodrigues' formula)**.** Consider

$$\dot{\mathbf{y}} = \boldsymbol{\omega} \times \mathbf{y}, \qquad \boldsymbol{\omega} \neq \mathbf{0},$$

and set

$$\omega = |\boldsymbol{\omega}|, \qquad \mathbf{e} = \frac{\boldsymbol{\omega}}{\omega}.$$

For an initial vector $\mathbf{y}_0$, write

$$\mathbf{y}_0 = (\mathbf{y}_0 \cdot \mathbf{e})\mathbf{e} + \mathbf{y}_\perp, \qquad \mathbf{y}_\perp \cdot \mathbf{e} = 0.$$

1. Derive Rodrigues' formula

$$\mathbf{y}(t) = \cos(\omega t)\mathbf{y}_0 + \sin(\omega t)(\mathbf{e} \times \mathbf{y}_0) + \big(1 - \cos(\omega t)\big)(\mathbf{y}_0 \cdot \mathbf{e})\mathbf{e}.$$

2. Show directly from the formula that

$$|\mathbf{y}(t)| = |\mathbf{y}_0| \qquad \text{and} \qquad \mathbf{y}(t) \cdot \boldsymbol{\omega} = \mathbf{y}_0 \cdot \boldsymbol{\omega}.$$

5.3.3 Mode coalescence and generalized eigenvectors

Ordinary modes may fail to provide enough independent solutions when two eigenvalues and their eigenvectors coalesce. The missing solution is then obtained by allowing a polynomial factor to multiply the exponential mode.

To see the mechanism, consider the family

$$\mathbf{A}_\varepsilon = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda+\varepsilon \end{bmatrix}, \qquad \varepsilon \neq 0.$$

The two eigenvalues are λ and $\lambda + \varepsilon$. Corresponding eigenvectors may be chosen as

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \mathbf{e}_1 + \varepsilon \mathbf{e}_2 = \begin{bmatrix} 1 \\ \varepsilon \end{bmatrix}.$$

Hence the system has the two ordinary modes

$$e^{\lambda t}\mathbf{e}_1$$

and

$$e^{(\lambda+\varepsilon)t}(\mathbf{e}_1 + \varepsilon \mathbf{e}_2).$$

As $\varepsilon \to 0$, the two eigenvalues approach λ, while the two eigenvectors both approach $\mathbf{e}_1$. Thus the two modes coalesce. To extract a second independent limiting solution, consider the difference quotient

$$\frac{e^{(\lambda+\varepsilon)t}(\mathbf{e}_1 + \varepsilon \mathbf{e}_2) - e^{\lambda t}\mathbf{e}_1}{\varepsilon}.$$

Using

$$e^{\varepsilon t} = 1 + \varepsilon t + O(\varepsilon^2),$$

we find

$$\begin{aligned} e^{(\lambda+\varepsilon)t}(\mathbf{e}_1 + \varepsilon \mathbf{e}_2) &= e^{\lambda t} e^{\varepsilon t}(\mathbf{e}_1 + \varepsilon \mathbf{e}_2) \\ &= e^{\lambda t}\left[\mathbf{e}_1 + \varepsilon(t\mathbf{e}_1 + \mathbf{e}_2) + O(\varepsilon^2)\right]. \end{aligned}$$

Therefore,

$$\lim_{\varepsilon \to 0} \frac{e^{(\lambda+\varepsilon)t}(\mathbf{e}_1 + \varepsilon \mathbf{e}_2) - e^{\lambda t}\mathbf{e}_1}{\varepsilon} = e^{\lambda t}(t\mathbf{e}_1 + \mathbf{e}_2).$$

At $\varepsilon = 0$, the limiting matrix is

$$\mathbf{A}_0 = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}.$$

It satisfies

$$(\mathbf{A}_0 - \lambda \mathbf{I})\mathbf{e}_1 = \mathbf{0}, \qquad (\mathbf{A}_0 - \lambda \mathbf{I})\mathbf{e}_2 = \mathbf{e}_1.$$

Thus $\mathbf{e}_1$ is an eigenvector, while $\mathbf{e}_2$ is mapped to the eigenvector $\mathbf{e}_1$ by $\mathbf{A}_0 - \lambda \mathbf{I}$. The second solution is no longer an ordinary exponential mode; it contains the polynomial factor t.

The same phenomenon occurs in the critically damped oscillator.

Example 5.13 (Critical damping and coalescing modes)**.** Recall from Section 3.3.2 that

$$m\ddot{y} + \gamma \dot{y} + ky = 0$$

has characteristic roots

$$\lambda_{\pm} = \frac{-\gamma \pm \sqrt{\gamma^2 - 4mk}}{2m}.$$

With

$$\mathbf{y} = \begin{bmatrix} y \\ \dot{y} \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} 0 & 1 \\ -k/m & -\gamma/m \end{bmatrix},$$

corresponding eigenvectors are

$$\mathbf{v}_\pm = \begin{bmatrix} 1 \\ \lambda_\pm \end{bmatrix}.$$

Thus the modal solutions are

$$e^{\lambda_\pm t}\mathbf{v}_\pm,$$

whose first components are the displacement modes $e^{\lambda_\pm t}$.

At critical damping, $\gamma^2 = 4mk$, the eigenvalues and eigenvectors coalesce:

$$\lambda_+ = \lambda_- = \lambda = -\frac{\gamma}{2m}, \qquad \mathbf{v}_\pm \longrightarrow \mathbf{v}_0 = \begin{bmatrix} 1 \\ \lambda \end{bmatrix}.$$

To recover a second direction, observe that

$$\frac{\mathbf{v}_+ - \mathbf{v}_-}{\lambda_+ - \lambda_-} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

This suggests introducing

$$\mathbf{v}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Then

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_0 = \mathbf{0}, \qquad (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_1 = \mathbf{v}_0.$$

The associated chain solutions are

$$e^{\lambda t}\mathbf{v}_0$$

and

$$e^{\lambda t}(t\mathbf{v}_0 + \mathbf{v}_1) = e^{\lambda t}\begin{bmatrix} t \\ 1 + \lambda t \end{bmatrix}.$$

Their first components are

$$y_1(t) = e^{\lambda t}, \qquad y_2(t) = te^{\lambda t},$$

which recover the critically damped modes from Section 3.3.2.

The preceding examples motivate a broader notion of eigenvector.

Definition 5.14 (Generalized eigenvector)**.** Let λ be an eigenvalue of $\mathbf{A}$. A nonzero vector $\mathbf{v}$ is called a generalized eigenvector associated with λ if

$$(\mathbf{A} - \lambda\mathbf{I})^m\mathbf{v} = \mathbf{0}$$

for some positive integer m.

An ordinary eigenvector is a generalized eigenvector with $m = 1$. If

$$(\mathbf{A} - \lambda\mathbf{I})^2\mathbf{v}_1 = \mathbf{0}, \qquad (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_1 \neq \mathbf{0},$$

then $\mathbf{v}_1$ is a generalized eigenvector of rank two. Setting

$$\mathbf{v}_0 = (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_1,$$

we have

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_0 = \mathbf{0}, \qquad (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_1 = \mathbf{v}_0.$$

These relations generate two solutions:

$$\mathbf{y}_0(t) = e^{\lambda t}\mathbf{v}_0$$

and

$$\mathbf{y}_1(t) = e^{\lambda t}(t\mathbf{v}_0 + \mathbf{v}_1).$$

Indeed,

$$\dot{\mathbf{y}}_1 = e^{\lambda t}\left[\lambda(t\mathbf{v}_0 + \mathbf{v}_1) + \mathbf{v}_0\right],$$

while

$$\begin{aligned}\mathbf{A}\mathbf{y}_1 &= e^{\lambda t}\left[t\mathbf{A}\mathbf{v}_0 + \mathbf{A}\mathbf{v}_1\right] \\ &= e^{\lambda t}\left[t\lambda\mathbf{v}_0 + \lambda\mathbf{v}_1 + \mathbf{v}_0\right].\end{aligned}$$

Thus $\dot{\mathbf{y}}_1 = \mathbf{A}\mathbf{y}_1$.

The polynomial factor in a generalized mode is therefore not an additional guess. It arises naturally when ordinary exponential modes coalesce. Longer sequences of generalized eigenvectors produce higher powers of t, as we now show.

5.3.4 Jordan chains and generalized modes

The preceding construction extends to longer sequences of generalized eigenvectors.

Definition 5.15 (Jordan chain)**.** A sequence of nonzero vectors

$$\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_m$$

is called a Jordan chain of length $m+1$ associated with the eigenvalue λ if

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_0 = \mathbf{0}$$

and

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_k = \mathbf{v}_{k-1}, \qquad k = 1, \ldots, m.$$

Equivalently,

$$\mathbf{A}\mathbf{v}_0 = \lambda\mathbf{v}_0,$$

and

$$\mathbf{A}\mathbf{v}_k = \lambda\mathbf{v}_k + \mathbf{v}_{k-1}, \qquad k = 1, \ldots, m.$$

The first vector $\mathbf{v}_0$ is an eigenvector. The remaining vectors are generalized eigenvectors of increasing rank.

A Jordan chain generates a corresponding chain of solutions.

Proposition 5.16 (Generalized modes and Jordan chains)**.** *Let $\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_m$ be nonzero vectors, and for $j = 0, \ldots, m$ define*

$$\mathbf{y}_j(t) = e^{\lambda t}\sum_{k=0}^{j}\frac{t^{j-k}}{(j-k)!}\mathbf{v}_k.$$

Then $\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_m$ form a Jordan chain associated with λ if and only if each $\mathbf{y}_j$ is a solution of

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}.$$

Explicitly,

$$\begin{aligned}
\mathbf{y}_0(t) &= e^{\lambda t}\mathbf{v}_0, \\
\mathbf{y}_1(t) &= e^{\lambda t}(t\mathbf{v}_0 + \mathbf{v}_1), \\
\mathbf{y}_2(t) &= e^{\lambda t}\left(\frac{t^2}{2!}\mathbf{v}_0 + t\mathbf{v}_1 + \mathbf{v}_2\right),
\end{aligned}$$

and so forth.

Proof. Suppose first that $\mathbf{v}_0, \dots, \mathbf{v}_m$ form a Jordan chain. For fixed j, set

$$\mathbf{p}_j(t) = \sum_{k=0}^{j} \frac{t^{j-k}}{(j-k)!}\mathbf{v}_k, \qquad \mathbf{y}_j(t) = e^{\lambda t}\mathbf{p}_j(t).$$

Using

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_0 = \mathbf{0}, \qquad (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_k = \mathbf{v}_{k-1},$$

we obtain

$$\begin{aligned}
(\mathbf{A} - \lambda\mathbf{I})\mathbf{p}_j &= \sum_{k=1}^{j} \frac{t^{j-k}}{(j-k)!}\mathbf{v}_{k-1} \\
&= \dot{\mathbf{p}}_j.
\end{aligned}$$

Hence

$$\mathbf{A}\mathbf{y}_j = e^{\lambda t}\big(\lambda\mathbf{p}_j + \dot{\mathbf{p}}_j\big) = \dot{\mathbf{y}}_j.$$

Conversely, suppose that every $\mathbf{y}_j$ is a solution. Then

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{p}_j = \dot{\mathbf{p}}_j.$$

For $j = 0$, this gives

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_0 = \mathbf{0}.$$

For $j \geq 1$, comparing coefficients of powers of t gives

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_k = \mathbf{v}_{k-1}, \qquad k = 1, \dots, j.$$

Taking $j = m$ yields all the Jordan-chain relations. □

Example 5.17 (A Jordan chain of length three)**.** Let

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

and let

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \qquad \mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Then

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{e}_1 = \mathbf{0},$$

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{e}_2 = \mathbf{e}_1,$$

and

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{e}_3 = \mathbf{e}_2.$$

Thus

$$\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$$

form a Jordan chain of length three. The associated generalized modes are

$$e^{\lambda t}\mathbf{e}_1,$$

$$e^{\lambda t}(t\mathbf{e}_1 + \mathbf{e}_2),$$

and

$$e^{\lambda t}\left(\frac{t^2}{2}\mathbf{e}_1 + t\mathbf{e}_2 + \mathbf{e}_3\right).$$

At $t = 0$, these solutions take the values

$$\mathbf{e}_1, \qquad \mathbf{e}_2, \qquad \mathbf{e}_3,$$

respectively, and are therefore linearly independent.

Complex Jordan chains are treated in the same way. If

$$\lambda = \alpha + i\beta$$

and

$$\mathbf{v}_0, \ldots, \mathbf{v}_m$$

is a complex Jordan chain, then the generalized complex modes are

$$e^{(\alpha + i\beta)t} \sum_{k=0}^{j} \frac{t^{j-k}}{(j-k)!}\mathbf{v}_k, \qquad j = 0, \ldots, m.$$

For a real matrix, their real and imaginary parts give real generalized modes. These combine polynomial factors with the sine and cosine factors associated with the complex eigenvalue.

Ordinary and generalized modes are therefore described by the common form

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

where $\mathbf{p}(t)$ is a vector-valued polynomial. For an ordinary mode, $\mathbf{p}(t)$ is constant. For a Jordan chain of length greater than one, nonconstant polynomial factors appear.

The remaining structural question is whether the state space can always be decomposed into subspaces generated by Jordan chains. We turn to this question in the next section.

Exercise 5.13 (A real generalized mode)**.** Let

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ -1 & 3 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

Show that $\lambda = 2$ has algebraic multiplicity two but only one independent eigenvector. Construct a Jordan chain for $\lambda = 2$ and an eigenvector for $\lambda = -1$. Use them to write three independent real solutions of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ and hence the general solution.

5.4 Decoupling and the Structure of a Matrix

The modal solutions constructed in the previous section arise from eigenvectors and Jordan chains. To obtain a complete description of the system, we must understand whether the whole state space can be assembled from such structures.

The key idea is to decompose the space into *invariant subspaces*. On each invariant subspace, the matrix acts independently of the others. Once a basis is chosen in each component, the matrix acquires a block form, and the original problem is reduced to several smaller *decoupled* ones.

We first formulate this notion of decoupling. We then show that the generalized eigenspaces associated with the distinct eigenvalues form an invariant direct-sum decomposition of $\mathbb{C}^n$. Finally, we examine the structure inside each generalized eigenspace and construct a basis of Jordan chains.

5.4.1 Invariant subspaces and decoupling

A subspace $\mathcal{V} \subseteq \mathbb{C}^n$ is called invariant under $\mathbf{A}$ if

$$\mathbf{A}\mathcal{V} \subseteq \mathcal{V}.$$

Equivalently, whenever $\mathbf{v} \in \mathcal{V}$, one also has $\mathbf{A}\mathbf{v} \in \mathcal{V}$.

If $\mathcal{V}$ is invariant, then the action of $\mathbf{A}$ may be restricted to $\mathcal{V}$. Thus $\mathbf{A}$ defines a linear map

$$\mathbf{A}|_{\mathcal{V}} : \mathcal{V} \longrightarrow \mathcal{V}.$$

The behavior of $\mathbf{A}$ on $\mathcal{V}$ can therefore be studied without reference to vectors outside $\mathcal{V}$.

The simplest invariant subspace is an eigenspace. If

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v},$$

then the line

$$\operatorname{span}\{\mathbf{v}\}$$

is invariant. More generally, the span of a Jordan chain is invariant. Indeed, if

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_0 = \mathbf{0}, \qquad (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_k = \mathbf{v}_{k-1},$$

then

$$\mathbf{A}\mathbf{v}_0 = \lambda\mathbf{v}_0, \qquad \mathbf{A}\mathbf{v}_k = \lambda\mathbf{v}_k + \mathbf{v}_{k-1},$$

so $\mathbf{A}$ maps every vector in

$$\operatorname{span}\{\mathbf{v}_0, \dots, \mathbf{v}_m\}$$

back into the same subspace.

The simplest decoupling occurs when $\mathbf{A}$ has a basis of eigenvectors. Suppose

$$\mathbf{A}\mathbf{v}_j = \lambda_j \mathbf{v}_j, \qquad j = 1, \dots, n,$$

and $\mathbf{v}_1, \dots, \mathbf{v}_n$ form a basis of $\mathbb{C}^n$. Then

$$\mathbb{C}^n = \operatorname{span}\{\mathbf{v}_1\} \oplus \cdots \oplus \operatorname{span}\{\mathbf{v}_n\},$$

and each one-dimensional subspace is invariant under $\mathbf{A}$.

Let

$$\mathbf{P} = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_n \end{bmatrix}, \qquad \mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}.$$

The eigenvalue relations can then be written together as

$$\mathbf{A}\mathbf{P} = \mathbf{P}\mathbf{D}.$$

Since $\mathbf{P}$ is invertible,

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D}.$$

Thus a matrix is diagonalizable precisely when the space is a direct sum of one-dimensional invariant subspaces.

A matrix need not have enough eigenvectors to yield a decomposition into one-dimensional invariant subspaces. In that case, the natural invariant components may have higher dimension, and the corresponding matrix representation is block diagonal rather than diagonal.

Suppose

$$\mathbb{C}^n = \mathcal{V}_1 \oplus \cdots \oplus \mathcal{V}_r,$$

where each $\mathcal{V}_j$ is invariant under $\mathbf{A}$. Then $\mathbf{A}$ acts independently on the components $\mathcal{V}_j$. This is the algebraic meaning of decoupling.

Proposition 5.18 (Invariant decomposition and block form)**.** *For each j, let $\mathbf{P}_j$ be a matrix whose columns form a basis of $\mathcal{V}_j$, and set*

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_1 & \cdots & \mathbf{P}_r \end{bmatrix}.$$

Then $\mathbf{P}$ is invertible and

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{A}_r \end{bmatrix},$$

where $\mathbf{A}_j$ is the matrix of the restriction $\mathbf{A}|_{\mathcal{V}_j}$ in the chosen basis of $\mathcal{V}_j$.

Proof. Since the sum is direct, the columns of $\mathbf{P}$ form a basis of $\mathbb{C}^n$, so $\mathbf{P}$ is invertible. The invariance of $\mathcal{V}_j$ implies

$$\mathbf{A}\mathbf{P}_j = \mathbf{P}_j\mathbf{A}_j.$$

Combining these identities gives the stated block-diagonal representation. □

The main question is therefore not how to manipulate a block-diagonal matrix once it is given, but whether such invariant components can be found systematically. Distinct eigenvalues provide the first level of this decomposition. For each eigenvalue λ, the appropriate component is not merely the eigenspace

$$\ker(\mathbf{A} - \lambda\mathbf{I}),$$

but the larger generalized eigenspace

$$\ker(\mathbf{A} - \lambda\mathbf{I})^m$$

for a sufficiently large exponent m. We now show that the generalized eigenspaces associated with distinct eigenvalues form a direct-sum decomposition of the whole space.

5.4.2 Generalized eigenspaces and primary decomposition

Let $\lambda_1, \ldots, \lambda_r$ be the distinct eigenvalues of $\mathbf{A}$, and let

$$p_{\mathbf{A}}(z) = \det(z\mathbf{I} - \mathbf{A}) = \prod_{j=1}^{r}(z - \lambda_j)^{m_j}.$$

Here m_j is called the algebraic multiplicity of λ_j.

For each eigenvalue λ_j, define the generalized eigenspace

$$\mathcal{G}_j = \ker(\mathbf{A} - \lambda_j\mathbf{I})^{m_j}.$$

The exponent m_j may be replaced by any larger integer. Once the kernels stabilize, increasing the exponent does not enlarge the generalized eigenspace.

Each $\mathcal{G}_j$ is invariant under $\mathbf{A}$. Indeed, if $\mathbf{v} \in \mathcal{G}_j$, then

$$(\mathbf{A} - \lambda_j\mathbf{I})^{m_j}\mathbf{v} = \mathbf{0}.$$

Since $\mathbf{A}$ commutes with every polynomial in $\mathbf{A}$,

$$(\mathbf{A} - \lambda_j\mathbf{I})^{m_j}\mathbf{A}\mathbf{v} = \mathbf{A}(\mathbf{A} - \lambda_j\mathbf{I})^{m_j}\mathbf{v} = \mathbf{0}.$$

Hence $\mathbf{A}\mathbf{v} \in \mathcal{G}_j$.

The main result of this subsection is that the generalized eigenspaces associated with the distinct eigenvalues give a complete invariant decomposition of the state space:

$$\mathbb{C}^n = \mathcal{G}_1 \oplus \cdots \oplus \mathcal{G}_r.$$

We first prove a two-factor splitting result.

Proposition 5.19 (Splitting by relatively prime factors)**.** *Suppose f and g are relatively prime polynomials and*

$$f(\mathbf{A})g(\mathbf{A}) = \mathbf{0}.$$

Then

$$\mathbb{C}^n = \ker f(\mathbf{A}) \oplus \ker g(\mathbf{A}).$$

Proof. Since f and g are relatively prime, Bézout's identity gives polynomials a and b such that

$$a(z)f(z) + b(z)g(z) = 1.$$

Replacing z by $\mathbf{A}$, we obtain

$$a(\mathbf{A})f(\mathbf{A}) + b(\mathbf{A})g(\mathbf{A}) = \mathbf{I}.$$

Hence every $\mathbf{v} \in \mathbb{C}^n$ can be written as

$$\mathbf{v} = a(\mathbf{A})f(\mathbf{A})\mathbf{v} + b(\mathbf{A})g(\mathbf{A})\mathbf{v}.$$

Because

$$g(\mathbf{A})a(\mathbf{A})f(\mathbf{A})\mathbf{v} = \mathbf{0},$$

the first term belongs to $\ker g(\mathbf{A})$. Similarly, the second term belongs to $\ker f(\mathbf{A})$.

If $\mathbf{v}$ lies in both kernels, then

$$\mathbf{v} = a(\mathbf{A})f(\mathbf{A})\mathbf{v} + b(\mathbf{A})g(\mathbf{A})\mathbf{v} = \mathbf{0}.$$

Thus the sum is direct. □

We also use the Cayley–Hamilton theorem.

Theorem 5.20 (Cayley–Hamilton theorem)**.** *Every square matrix satisfies its characteristic equation. Thus, if*

$$p_{\mathbf{A}}(z) = \det(z\mathbf{I} - \mathbf{A}),$$

then

$$p_{\mathbf{A}}(\mathbf{A}) = \mathbf{0}.$$

We now apply the splitting proposition successively to the relatively prime factors of the characteristic polynomial.

Theorem 5.21 (Primary decomposition)**.** *Let $\lambda_1, \dots, \lambda_r$ be the distinct eigenvalues of $\mathbf{A}$, and write*

$$p_{\mathbf{A}}(z) = \prod_{j=1}^{r} (z - \lambda_j)^{m_j}.$$

Then

$$\mathbb{C}^n = \mathcal{G}_1 \oplus \cdots \oplus \mathcal{G}_r, \qquad \mathcal{G}_j = \ker(\mathbf{A} - \lambda_j \mathbf{I})^{m_j}.$$

Each $\mathcal{G}_j$ is invariant under $\mathbf{A}$.

Proof. Set

$$p_j(z) = (z - \lambda_j)^{m_j}, \qquad R_j(z) = \prod_{k=j}^{r} p_k(z).$$

By the Cayley–Hamilton theorem,

$$R_1(\mathbf{A}) = p_{\mathbf{A}}(\mathbf{A}) = \mathbf{0}.$$

Since p_1 and R_2 are relatively prime, the splitting proposition gives

$$\mathbb{C}^n = \ker p_1(\mathbf{A}) \oplus \ker R_2(\mathbf{A}) = \mathcal{G}_1 \oplus \ker R_2(\mathbf{A}).$$

The subspace $\ker R_2(\mathbf{A})$ is invariant under $\mathbf{A}$. Restricting $\mathbf{A}$ to this subspace and applying the same argument to p_2 and R_3, we obtain

$$\ker R_2(\mathbf{A}) = \mathcal{G}_2 \oplus \ker R_3(\mathbf{A}).$$

Continuing in this way yields

$$\mathbb{C}^n = \mathcal{G}_1 \oplus \cdots \oplus \mathcal{G}_r.$$

□

The primary decomposition separates the contributions from distinct eigenvalues. Relative to a basis adapted to

$$\mathbb{C}^n = \mathcal{G}_1 \oplus \cdots \oplus \mathcal{G}_r,$$

the matrix $\mathbf{A}$ has the block-diagonal form

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{A}_r \end{bmatrix},$$

where $\mathbf{A}_j$ is the restriction of $\mathbf{A}$ to $\mathcal{G}_j$.

Each block $\mathbf{A}_j$ has only one eigenvalue, namely λ_j. Thus the remaining problem is local: we must understand the structure of a matrix with a single eigenvalue. This structure is encoded by the nested kernels

$$\ker(\mathbf{A} - \lambda_j \mathbf{I}) \subseteq \ker(\mathbf{A} - \lambda_j \mathbf{I})^2 \subseteq \cdots,$$

which we study next.

5.4.3 Kernel filtration and Jordan block sizes

Fix an eigenvalue λ of $\mathbf{A}$, and let

$$\mathcal{G} = \ker(\mathbf{A} - \lambda \mathbf{I})^m$$

be the corresponding generalized eigenspace, where m is the algebraic multiplicity of λ. On $\mathcal{G}$, define

$$\mathbf{N} = \mathbf{A} - \lambda \mathbf{I}.$$

Then $\mathbf{N}$ is nilpotent:

$$\mathbf{N}^m = \mathbf{0} \qquad \text{on } \mathcal{G}.$$

Thus the structure of $\mathbf{A}$ on $\mathcal{G}$ is determined by the nilpotent map $\mathbf{N}$.

Consider the nested sequence

$$\mathcal{K}_1 \subseteq \mathcal{K}_2 \subseteq \cdots \subseteq \mathcal{K}_m = \mathcal{G},$$

where

$$\mathcal{K}_k = \ker \mathbf{N}^k.$$

This sequence is called the *kernel filtration* associated with λ. Set

$$\mathcal{K}_0 = \{\mathbf{0}\}, \qquad d_k = \dim \mathcal{K}_k, \qquad k = 0, \ldots, m.$$

Since $\mathbf{N}^m = \mathbf{0}$ on $\mathcal{G}$, the filtration has stabilized at level m. For convenience, define

$$\mathcal{K}_k = \mathcal{G}, \qquad d_k = \dim \mathcal{G}, \qquad k \geq m.$$

The increments

$$d_k - d_{k-1}$$

measure how many new independent directions enter the kernel at level k.

A Jordan chain

$$\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_{\ell-1}$$

of length ℓ satisfies

$$\mathbf{N}\mathbf{v}_0 = \mathbf{0}, \qquad \mathbf{N}\mathbf{v}_j = \mathbf{v}_{j-1}, \qquad j = 1, \ldots, \ell - 1.$$

For this chain,

$$\mathbf{v}_0 \in \mathcal{K}_1, \qquad \mathbf{v}_1 \in \mathcal{K}_2, \qquad \ldots, \qquad \mathbf{v}_{\ell-1} \in \mathcal{K}_\ell.$$

The chain contributes one new independent direction to each of

$$\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_\ell.$$

This gives the following interpretation of the kernel dimensions.

Proposition 5.22 (Kernel increments and Jordan block sizes)**.** *Suppose $\mathcal{G}$ has a basis consisting of Jordan chains for the eigenvalue λ. Then*

$$d_k - d_{k-1}$$

is the number of Jordan blocks of size at least k.

Consequently,

$$(d_k - d_{k-1}) - (d_{k+1} - d_k)$$

is the number of Jordan blocks of size exactly k.

Proof. A Jordan chain of length ℓ contributes one independent vector to $\ker \mathbf{N}^k$ for each $k \leq \ell$, and contributes no new vector after $k > \ell$. Hence the increase from $\ker \mathbf{N}^{k-1}$ to $\ker \mathbf{N}^k$ counts exactly the chains whose lengths are at least k. Taking the difference of two successive increments gives the number of chains whose lengths are exactly k. □

Example 5.23 (Reading block sizes from kernel dimensions)**.** Suppose λ has algebraic multiplicity 9 and

$$d_1 = 4, \qquad d_2 = 7, \qquad d_3 = 8, \qquad d_4 = 9.$$

The increments are

$$d_1 - d_0 = 4, \qquad d_2 - d_1 = 3, \qquad d_3 - d_2 = 1, \qquad d_4 - d_3 = 1.$$

Thus

k	number of blocks of size at least k	number of blocks of size k
1	4	1
2	3	2
3	1	0
4	1	1

Therefore the block sizes are

$$4,\ 2,\ 2,\ 1.$$

The restriction of $\mathbf{A}$ to the generalized eigenspace associated with λ has Jordan form

$$\mathbf{J}_4(\lambda) \oplus \mathbf{J}_2(\lambda) \oplus \mathbf{J}_2(\lambda) \oplus \mathbf{J}_1(\lambda).$$

Thus the kernel filtration determines the number and sizes of the Jordan blocks associated with λ. It does not yet choose the vectors in the corresponding Jordan chains. We turn to that construction next.

5.4.4 Jordan chains, Jordan bases, and Jordan form

We now construct Jordan chains and assemble them into a basis of the whole space.

Fix an eigenvalue λ, let $\mathcal{G}$ be its generalized eigenspace, and set

$$\mathbf{N} = \mathbf{A} - \lambda \mathbf{I} \qquad \text{on } \mathcal{G}.$$

Suppose

$$\mathbf{w} \in \ker \mathbf{N}^{\ell} \setminus \ker \mathbf{N}^{\ell-1}.$$

Then

$$\mathbf{N}^{\ell}\mathbf{w} = \mathbf{0}, \qquad \mathbf{N}^{\ell-1}\mathbf{w} \neq \mathbf{0}.$$

Define

$$\mathbf{v}_j = \mathbf{N}^{\ell-1-j}\mathbf{w}, \qquad j = 0, \dots, \ell - 1.$$

Then

$$\mathbf{N}\mathbf{v}_0 = \mathbf{0}, \qquad \mathbf{N}\mathbf{v}_j = \mathbf{v}_{j-1}, \qquad j = 1, \dots, \ell - 1.$$

Thus

$$\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_{\ell-1}$$

form a Jordan chain of length ℓ. The vector $\mathbf{w} = \mathbf{v}_{\ell-1}$ is called the terminal vector of the chain.

To construct a full chain basis, one first determines the block sizes from the kernel filtration. The longest chains are then constructed first, followed by the shorter chains. At each stage, the terminal vectors are chosen so that all resulting chain vectors remain linearly independent.

Theorem 5.24 (Existence of a Jordan-chain basis)**.** *Let $\mathcal{G}$ be the generalized eigenspace associated with an eigenvalue λ. Then $\mathcal{G}$ has a basis consisting of Jordan chains for λ.*

Proof. Set

$$\mathbf{N} = \mathbf{A} - \lambda \mathbf{I} \qquad \text{on } \mathcal{G}.$$

Since $\mathbf{N}$ is nilpotent, we prove the result by induction on $\dim \mathcal{G}$.

If $\mathbf{N} = \mathbf{0}$, then every basis of $\mathcal{G}$ consists of Jordan chains of length one.

Suppose $\mathbf{N} \neq \mathbf{0}$ and set

$$\mathcal{W} = \operatorname{range} \mathbf{N}.$$

The subspace $\mathcal{W}$ is invariant under $\mathbf{N}$, and

$$\dim \mathcal{W} < \dim \mathcal{G}.$$

By the induction hypothesis, $\mathcal{W}$ has a basis consisting of Jordan chains. Write these chains as

$$\mathbf{v}_{q,0}, \dots, \mathbf{v}_{q,\ell_q - 1}, \qquad q = 1, \dots, r.$$

For each terminal vector $\mathbf{v}_{q,\ell_q-1} \in \mathcal{W}$, choose $\mathbf{w}_q \in \mathcal{G}$ such that

$$\mathbf{N}\mathbf{w}_q = \mathbf{v}_{q,\ell_q-1}.$$

Then

$$\mathbf{v}_{q,0}, \dots, \mathbf{v}_{q,\ell_q-1}, \mathbf{w}_q$$

is a Jordan chain in $\mathcal{G}$.

The number of chains in a Jordan-chain basis of $\mathcal{W}$ is

$$r = \dim \ker(\mathbf{N}|_{\mathcal{W}}) = \dim(\ker \mathbf{N} \cap \mathcal{W}).$$

Choose vectors

$$\mathbf{z}_1, \dots, \mathbf{z}_s$$

that complete a basis of $\ker \mathbf{N} \cap \mathcal{W}$ to a basis of $\ker \mathbf{N}$. Each $\mathbf{z}_j$ forms a Jordan chain of length one.

We claim that the extended chains together with $\mathbf{z}_1, \dots, \mathbf{z}_s$ form a basis of $\mathcal{G}$. To see their linear independence, suppose

$$\sum_{q=1}^{r} a_q \mathbf{w}_q + \mathbf{w} + \sum_{j=1}^{s} b_j \mathbf{z}_j = \mathbf{0}, \qquad \mathbf{w} \in \mathcal{W}.$$

Applying $\mathbf{N}$ gives

$$\sum_{q=1}^{r} a_q \mathbf{v}_{q,\ell_q-1} + \mathbf{N}\mathbf{w} = \mathbf{0}.$$

The terminal vectors of the chains in $\mathcal{W}$ are linearly independent modulo $\mathbf{N}\mathcal{W}$. Hence

$$a_1 = \cdots = a_r = 0.$$

The original relation then becomes

$$\mathbf{w} + \sum_{j=1}^{s} b_j \mathbf{z}_j = \mathbf{0}.$$

Since the vectors $\mathbf{z}_j$ are independent modulo $\ker \mathbf{N} \cap \mathcal{W}$, it follows that

$$b_1 = \cdots = b_s = 0, \qquad \mathbf{w} = \mathbf{0}.$$

Thus the constructed vectors are linearly independent.

Finally, the number of new vectors added to the basis of $\mathcal{W}$ is

$$\begin{aligned} r + s &= \dim(\ker \mathbf{N} \cap \mathcal{W}) + \dim \ker \mathbf{N} - \dim(\ker \mathbf{N} \cap \mathcal{W}) \\ &= \dim \ker \mathbf{N}. \end{aligned}$$

Therefore, the total number of constructed vectors is

$$\dim \mathcal{W} + \dim \ker \mathbf{N} = \dim \mathcal{G}$$

by the rank–nullity theorem. Hence they form a Jordan-chain basis of $\mathcal{G}$. □

Example 5.25 (Determining the Jordan form and constructing a Jordan basis)**.** Consider

$$\mathbf{A} = \begin{bmatrix} 3 & -1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix}.$$

Its characteristic polynomial is

$$\det(\mathbf{A} - \lambda \mathbf{I}) = (2 - \lambda)^3,$$

so the only eigenvalue is $\lambda = 2$, with algebraic multiplicity three.

Set

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

Since

$$\operatorname{rank} \mathbf{N} = 1,$$

we have

$$d_1 = \dim \ker \mathbf{N} = 2.$$

Moreover,

$$\mathbf{N}^2 = \mathbf{0},$$

so

$$d_2 = \dim \ker \mathbf{N}^2 = 3.$$

Thus

$$d_1 - d_0 = 2$$

shows that there are two Jordan blocks, while

$$d_2 - d_1 = 1$$

shows that exactly one block has size at least two. Hence the block sizes are

$$2,\ 1.$$

To construct the block of size two, choose a vector outside $\ker \mathbf{N}$. Taking

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

we obtain

$$\mathbf{v}_0 = \mathbf{N}\mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}.$$

Since $\mathbf{N}^2 = \mathbf{0}$,

$$\mathbf{N}\mathbf{v}_0 = \mathbf{0}, \qquad \mathbf{N}\mathbf{v}_1 = \mathbf{v}_0.$$

Thus $\mathbf{v}_0, \mathbf{v}_1$ form a Jordan chain of length two.

For the remaining block, choose an eigenvector independent of $\mathbf{v}_0$. Since

$$\ker \mathbf{N} = \left\{ \begin{bmatrix} x \\ y \\ z \end{bmatrix} : x - y + z = 0 \right\},$$

we may take

$$\mathbf{w}_0 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}.$$

The vectors

$$\mathbf{v}_0, \mathbf{v}_1, \mathbf{w}_0$$

are linearly independent. Therefore

$$\mathbf{P} = \begin{bmatrix} \mathbf{v}_0 & \mathbf{v}_1 & \mathbf{w}_0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

is invertible.

With

$$\mathbf{J} = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix},$$

the chain relations give

$$\mathbf{A}\mathbf{P} = \mathbf{P}\mathbf{J}.$$

Equivalently,

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J}.$$

We now assemble the chain bases of the individual generalized eigenspaces.

Let $\lambda_1, \ldots, \lambda_r$ be the distinct eigenvalues of $\mathbf{A}$. For each j, let

$$\mathbf{v}_{j,q,0}, \mathbf{v}_{j,q,1}, \ldots, \mathbf{v}_{j,q,\ell_{j,q}-1}, \qquad q = 1, \ldots, s_j,$$

be Jordan chains forming a basis of the generalized eigenspace $\mathcal{G}_j$. Thus

$$(\mathbf{A} - \lambda_j \mathbf{I})\mathbf{v}_{j,q,0} = \mathbf{0},$$

and

$$(\mathbf{A} - \lambda_j \mathbf{I})\mathbf{v}_{j,q,k} = \mathbf{v}_{j,q,k-1}, \qquad k = 1, \ldots, \ell_{j,q} - 1.$$

For each chain, set

$$\mathbf{P}_{j,q} = \begin{bmatrix} \mathbf{v}_{j,q,0} & \mathbf{v}_{j,q,1} & \cdots & \mathbf{v}_{j,q,\ell_{j,q}-1} \end{bmatrix}.$$

Since

$$\mathbb{C}^n = \mathcal{G}_1 \oplus \cdots \oplus \mathcal{G}_r,$$

the vectors from all these chains form a basis of $\mathbb{C}^n$. Such a basis is called a *Jordan basis* for $\mathbf{A}$.

For a single chain

$$\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_{\ell-1},$$

we have

$$\mathbf{A}\mathbf{v}_0 = \lambda \mathbf{v}_0,$$

and

$$\mathbf{A}\mathbf{v}_k = \mathbf{v}_{k-1} + \lambda \mathbf{v}_k, \qquad k = 1, \ldots, \ell - 1.$$

Hence, relative to the ordered basis

$$\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_{\ell-1},$$

the restriction of $\mathbf{A}$ is represented by the Jordan block

$$\mathbf{J}_\ell(\lambda) = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda & 1 \\ 0 & \cdots & \cdots & 0 & \lambda \end{bmatrix} = \lambda \mathbf{I} + \mathbf{N}_\ell,$$

where

$$\mathbf{N}_\ell = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 0 & 1 \\ 0 & \cdots & \cdots & 0 & 0 \end{bmatrix}, \qquad \mathbf{N}_\ell^\ell = \mathbf{0}.$$

Assemble the Jordan basis as

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{1,1} & \cdots & \mathbf{P}_{1,s_1} & \cdots & \mathbf{P}_{r,1} & \cdots & \mathbf{P}_{r,s_r} \end{bmatrix},$$

and let

$$\mathbf{J} = \bigoplus_{j=1}^{r} \bigoplus_{q=1}^{s_j} \mathbf{J}_{\ell_{j,q}}(\lambda_j).$$

Then the Jordan-chain relations combine into

$$\mathbf{AP} = \mathbf{PJ}.$$

The matrix $\mathbf{P}$ whose columns form a Jordan basis is called a *Jordan basis matrix* of $\mathbf{A}$.

We summarize the representation of $\mathbf{A}$ with respect to this basis in the following theorem.

Theorem 5.26 (Jordan canonical form)**.** *For every matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$, there exists a Jordan basis matrix $\mathbf{P}$ such that*

$$\mathbf{AP} = \mathbf{PJ},$$

where $\mathbf{J}$ is a block-diagonal matrix whose blocks are Jordan blocks. Equivalently,

$$\mathbf{P}^{-1}\mathbf{AP} = \mathbf{J}, \qquad \mathbf{A} = \mathbf{PJP}^{-1}.$$

The Jordan form is unique up to the order of its blocks. For matrices with several eigenvalues, the block sizes are determined separately on each generalized eigenspace, and the resulting blocks are then combined by direct sum.

For a fixed eigenvalue λ, the number of Jordan blocks is

$$\dim \ker(\mathbf{A} - \lambda\mathbf{I}),$$

and the size of the largest block is the smallest integer k for which

$$\ker(\mathbf{A} - \lambda\mathbf{I})^k = \ker(\mathbf{A} - \lambda\mathbf{I})^{k+1}.$$

A matrix is diagonalizable precisely when every Jordan block has size one. Equivalently, for each eigenvalue λ_j,

$$\mathcal{G}_j = \ker(\mathbf{A} - \lambda_j\mathbf{I}).$$

In that case, the Jordan form reduces to a diagonal matrix.

Remark 5.27 (Real form for complex eigenvalues). If $\mathbf{A}$ is real, nonreal eigenvalues occur in conjugate pairs. For

$$\lambda = \alpha + i\beta, \qquad \beta \neq 0,$$

the pair of complex Jordan blocks

$$\mathbf{J}_\ell(\lambda) \oplus \mathbf{J}_\ell(\overline{\lambda})$$

can be replaced by the real block

$$\begin{bmatrix} \mathbf{C} & \mathbf{I}_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C} & \mathbf{I}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{C} & \mathbf{I}_2 \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{C} \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix}.$$

Thus a real matrix can be reduced by a real change of basis to a block form containing ordinary Jordan blocks for real eigenvalues and real $2\ell \times 2\ell$ blocks for complex conjugate pairs.

Remark 5.28 (Jordan form and numerical computation). Jordan form is fundamental for structural analysis, but its direct numerical computation is sensitive to perturbations. A matrix with a nontrivial Jordan block may become diagonalizable after an arbitrarily small change in its entries. For numerical computation, the Schur decomposition is generally more stable. Jordan form should therefore be regarded primarily as a theoretical and structural tool.

The identity

$$\mathbf{AP} = \mathbf{PJ}$$

is the central conclusion of this section. It expresses the action of $\mathbf{A}$ in a basis adapted to its invariant Jordan-chain structure. In the next section, we apply this representation to

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}.$$

The change of variables

$$\mathbf{y} = \mathbf{P}\mathbf{z}$$

then reduces the system to

$$\dot{\mathbf{z}} = \mathbf{J}\mathbf{z}.$$

Exercise 5.14 (Constructing a Jordan basis)**.** Let

$$\mathbf{A} = \begin{bmatrix} 3 & 1 & -1 \\ 0 & 2 & 1 \\ 0 & -1 & 4 \end{bmatrix}, \qquad \mathbf{N} = \mathbf{A} - 3\mathbf{I}.$$

The only eigenvalue of $\mathbf{A}$ is 3.

1. Compute $\ker \mathbf{N}$ and $\ker \mathbf{N}^2$, and determine the Jordan-block sizes.
2. Construct a Jordan chain of length two and an independent eigenvector.
3. Assemble these vectors into a matrix $\mathbf{P}$ and verify

$$\mathbf{AP} = \mathbf{P} \begin{bmatrix} 3 & 1 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

5.5 Complete Modal Representation and Linear Stability

The previous section established the structural decomposition

$$\mathbf{AP} = \mathbf{PJ},$$

where the columns of $\mathbf{P}$ form a Jordan basis and $\mathbf{J}$ is the direct sum of the corresponding Jordan blocks. We now apply this representation to the linear system

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}.$$

The change of variables

$$\mathbf{y} = \mathbf{P}\mathbf{z}$$

transforms the original system into

$$\dot{\mathbf{z}} = \mathbf{J}\mathbf{z}.$$

Since $\mathbf{J}$ is block diagonal, the transformed system separates into independent subsystems, one for each Jordan block. The solutions of these subsystems are precisely the ordinary and generalized modes constructed in Section 5.3.

This leads to a complete modal representation of every solution. It also shows directly how eigenvalues and Jordan blocks determine the stability of the zero solution.

5.5.1 Jordan coordinates and decoupled subsystems

Suppose

$$\mathbf{A}\mathbf{P} = \mathbf{P}\mathbf{J},$$

where $\mathbf{P}$ is invertible. Introduce the Jordan coordinates $\mathbf{z}$ by

$$\mathbf{y} = \mathbf{P}\mathbf{z}.$$

Since $\mathbf{P}$ is constant,

$$\dot{\mathbf{y}} = \mathbf{P}\dot{\mathbf{z}}.$$

Substitution into $\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}$ gives

$$\mathbf{P}\dot{\mathbf{z}} = \mathbf{A}\mathbf{P}\mathbf{z} = \mathbf{P}\mathbf{J}\mathbf{z}.$$

Multiplying by $\mathbf{P}^{-1}$ yields

$$\dot{\mathbf{z}} = \mathbf{J}\mathbf{z}.$$

Thus similarity of matrices corresponds to a change of coordinates in the differential system. The matrix $\mathbf{A}$ describes the system in the original coordinates, while $\mathbf{J}$ describes the same system in a basis adapted to its invariant Jordan-chain structure.

Suppose

$$\mathbf{J} = \mathbf{J}_{\ell_1}(\lambda_1) \oplus \cdots \oplus \mathbf{J}_{\ell_s}(\lambda_s).$$

Decompose $\mathbf{z}$ accordingly:

$$\mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_s \end{bmatrix}, \qquad \mathbf{z}_q \in \mathbb{C}^{\ell_q}.$$

Then

$$\dot{\mathbf{z}}_q = \mathbf{J}_{\ell_q}(\lambda_q)\mathbf{z}_q, \qquad q = 1, \ldots, s.$$

These subsystems are completely decoupled.

Consider one Jordan block

$$\mathbf{J}_\ell(\lambda) = \begin{bmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda & 1 \\ 0 & \cdots & \cdots & 0 & \lambda \end{bmatrix}.$$

Writing

$$\mathbf{z} = \begin{bmatrix} z_0 \\ z_1 \\ \vdots \\ z_{\ell-1} \end{bmatrix},$$

the corresponding subsystem is

$$\begin{aligned} \dot{z}_0 &= \lambda z_0 + z_1, \\ \dot{z}_1 &= \lambda z_1 + z_2, \\ &\vdots \\ \dot{z}_{\ell-2} &= \lambda z_{\ell-2} + z_{\ell-1}, \\ \dot{z}_{\ell-1} &= \lambda z_{\ell-1}. \end{aligned}$$

It can be solved successively from the last equation upward. The final component is

$$z_{\ell-1}(t) = c_{\ell-1} e^{\lambda t}.$$

Substitution into the preceding equation gives

$$z_{\ell-2}(t) = e^{\lambda t}\big(c_{\ell-2} + c_{\ell-1} t\big),$$

and continuing in this way yields

$$z_j(t) = e^{\lambda t} \sum_{k=j}^{\ell-1} c_k \frac{t^{k-j}}{(k-j)!}.$$

Equivalently, the solution of the Jordan-block subsystem is a linear combination of

$$e^{\lambda t}\mathbf{e}_1,$$

$$e^{\lambda t}(t\mathbf{e}_1 + \mathbf{e}_2),$$

$$e^{\lambda t}\left(\frac{t^2}{2!}\mathbf{e}_1 + t\mathbf{e}_2 + \mathbf{e}_3\right),$$

and the analogous generalized modes up to degree $\ell - 1$.

Returning to the original coordinates, if

$$\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_{\ell-1}$$

are the columns of $\mathbf{P}$ corresponding to this block, then these coordinate modes become

$$e^{\lambda t}\mathbf{v}_0,$$

$$e^{\lambda t}(t\mathbf{v}_0 + \mathbf{v}_1),$$

$$e^{\lambda t}\left(\frac{t^2}{2!}\mathbf{v}_0 + t\mathbf{v}_1 + \mathbf{v}_2\right),$$

and so forth. Thus the generalized modes are not merely particular solutions: they are the coordinate solutions of the decoupled Jordan subsystems.

5.5.2 Complete modal representation

The Jordan coordinates provide a complete basis of solutions for the transformed system, and therefore for the original system.

Theorem 5.29 (Complete modal representation)**.** *Let*

$$\mathbf{AP} = \mathbf{PJ},$$

where the columns of $\mathbf{P}$ *form a Jordan basis of* $\mathbb{C}^n$*. For each Jordan chain*

$$\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_{\ell-1}$$

associated with an eigenvalue λ*, form the generalized modes*

$$\mathbf{y}_j(t) = e^{\lambda t} \sum_{k=0}^{j} \frac{t^{j-k}}{(j-k)!} \mathbf{v}_k, \qquad j = 0, \ldots, \ell - 1.$$

The generalized modes obtained from all Jordan chains form a basis of the solution space of

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}.$$

Consequently, every solution has a unique representation as a linear combination of these modes.

Proof. In Jordan coordinates, the system becomes

$$\dot{\mathbf{z}} = \mathbf{J}\mathbf{z}.$$

Each Jordan block of size ℓ contributes ℓ generalized modes. Their values at $t = 0$ are

$$\mathbf{e}_1, \ldots, \mathbf{e}_\ell,$$

so they are linearly independent. Taking all blocks together gives n linearly independent solutions of the transformed system.

Multiplication by the invertible matrix $\mathbf{P}$ maps these coordinate solutions to the generalized modes associated with the columns of $\mathbf{P}$. Their values at $t = 0$ are precisely the vectors of the Jordan basis, so they remain linearly independent. Since the solution space has dimension n, they form a basis. □

Thus every initial vector

$$\mathbf{y}_0 \in \mathbb{C}^n$$

has a unique expansion in the Jordan basis. If

$$\mathbf{y}_0 = \mathbf{P}\mathbf{z}_0,$$

then the coefficients in $\mathbf{z}_0$ determine the coefficients of the corresponding ordinary and generalized modes.

For one Jordan chain

$$\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_{\ell-1},$$

suppose the component of the initial value in the span of this chain is

$$c_0\mathbf{v}_0 + \cdots + c_{\ell-1}\mathbf{v}_{\ell-1}.$$

The associated part of the solution is

$$\begin{aligned}\mathbf{y}(t) = {} & c_0 e^{\lambda t}\mathbf{v}_0 \\ & + c_1 e^{\lambda t}(t\mathbf{v}_0 + \mathbf{v}_1) \\ & + c_2 e^{\lambda t}\left(\frac{t^2}{2!}\mathbf{v}_0 + t\mathbf{v}_1 + \mathbf{v}_2\right) + \cdots .\end{aligned}$$

At $t = 0$, this reduces to

$$c_0\mathbf{v}_0 + \cdots + c_{\ell-1}\mathbf{v}_{\ell-1},$$

as required.

If $\mathbf{A}$ is real, complex eigenvalues and Jordan chains occur in conjugate pairs. Taking real and imaginary parts of the corresponding complex generalized modes produces a real basis of solutions. Hence every real solution can be represented entirely in terms of real ordinary and generalized modes.

Example 5.30 (A defective system)**.** Suppose $\mathbf{A} \in \mathbb{C}^{4\times 4}$ has a Jordan chain

$$\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2$$

associated with λ, together with an eigenvector $\mathbf{w}$ associated with a distinct eigenvalue μ. Then

$$\mathbf{v}_0, \mathbf{v}_1, \mathbf{v}_2, \mathbf{w}$$

form a Jordan basis of $\mathbb{C}^4$. Every solution is of the form

$$\begin{aligned}\mathbf{y}(t) = {} & c_0 e^{\lambda t}\mathbf{v}_0 \\ & + c_1 e^{\lambda t}(t\mathbf{v}_0 + \mathbf{v}_1) \\ & + c_2 e^{\lambda t}\left(\frac{t^2}{2}\mathbf{v}_0 + t\mathbf{v}_1 + \mathbf{v}_2\right) \\ & + d e^{\mu t}\mathbf{w}.\end{aligned}$$

The four coefficients are determined uniquely by the initial condition.

The complete modal representation separates the solution into contributions from the individual Jordan chains. This makes the long-time behavior transparent: exponential factors are controlled by the real parts of the eigenvalues, while nontrivial Jordan blocks introduce polynomial factors in t.

5.5.3 Linear stability

We now apply the complete modal representation to the stability of the zero solution of

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}.$$

Recall that the zero solution is stable if, for every $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|\mathbf{y}(0)| < \delta$$

implies

$$|\mathbf{y}(t)| < \varepsilon \qquad \text{for all } t \geq 0.$$

It is asymptotically stable if it is stable and, in addition,

$$\mathbf{y}(t) \longrightarrow \mathbf{0} \qquad \text{as } t \to \infty.$$

For an ordinary mode

$$e^{\lambda t}\mathbf{v},$$

the exponential factor is

$$|e^{\lambda t}| = e^{(\Re\lambda)t}.$$

Thus the sign of $\Re\lambda$ determines whether the mode decays, remains bounded, or grows.

For a generalized mode associated with a Jordan block of size greater than one, polynomial factors also appear:

$$e^{\lambda t}\left(\mathbf{v}_j + t\mathbf{v}_{j-1} + \cdots + \frac{t^j}{j!}\mathbf{v}_0\right).$$

If $\Re\lambda < 0$, the exponential decay dominates every polynomial factor. If $\Re\lambda = 0$, however, a nonconstant polynomial factor is unbounded.

This yields the stability criterion.

Theorem 5.31 (Linear stability criterion)**.** *Consider*

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}.$$

1. *The zero solution is asymptotically stable if and only if*

$$\Re\lambda < 0$$

for every eigenvalue λ of $\mathbf{A}$.

2. *The zero solution is stable if and only if*

$$\Re\lambda \leq 0$$

for every eigenvalue λ, and every eigenvalue satisfying

$$\Re\lambda = 0$$

is semisimple; equivalently, all Jordan blocks associated with such an eigenvalue have size one.

Proof. Suppose first that every eigenvalue satisfies $\Re\lambda < 0$. Each modal term is a polynomial in t multiplied by a decaying exponential. More precisely, for every $\alpha > 0$ and every nonnegative integer k, the function

$$t^k e^{-\alpha t}$$

is uniformly bounded on $[0, \infty)$ and tends to zero as $t \to \infty$. Hence every generalized mode tends to zero and remains uniformly bounded for $t \geq 0$. Let $\boldsymbol{\phi}_1(t), \ldots, \boldsymbol{\phi}_n(t)$ be a basis of modal solutions. Since

$$\mathbf{y}(t) = \sum_{j=1}^{n} c_j \boldsymbol{\phi}_j(t)$$

and the coefficients c_j depend linearly on $\mathbf{y}(0)$, there is a constant $C_0 > 0$ such that

$$\sum_{j=1}^{n} |c_j| \leq C_0 |\mathbf{y}(0)|.$$

Since the modal solutions are uniformly bounded on $[0, \infty)$, there is a constant $C_1 > 0$ such that

$$|\boldsymbol{\phi}_j(t)| \leq C_1, \qquad t \geq 0, \quad j = 1, \ldots, n.$$

Therefore

$$|\mathbf{y}(t)| \leq C_0 C_1 |\mathbf{y}(0)|, \qquad t \geq 0.$$

This proves stability. Since every modal solution also tends to $\mathbf{0}$, every solution tends to $\mathbf{0}$, and hence the zero solution is asymptotically stable.

Conversely, if some eigenvalue satisfies $\Re\lambda > 0$, the corresponding ordinary mode grows exponentially. If $\Re\lambda = 0$, an ordinary mode does not tend to zero. Hence asymptotic stability requires $\Re\lambda < 0$ for all eigenvalues.

For stability, an eigenvalue with positive real part again gives an unbounded mode. If $\Re\lambda = 0$ and there is a Jordan block of size greater than one, the corresponding generalized mode contains a nonconstant polynomial factor and is unbounded. Thus stability requires all eigenvalues to satisfy $\Re\lambda \leq 0$ and all eigenvalues on the imaginary axis to be semisimple.

Under these conditions, modes with negative real part decay, while modes associated with semisimple imaginary-axis eigenvalues remain bounded. Thus every modal solution is uniformly bounded on $[0, \infty)$. Applying the same finite-basis estimate used above, there is a constant $M > 0$ such that

$$|\mathbf{y}(t)| \leq M |\mathbf{y}(0)|, \qquad t \geq 0.$$

Therefore the zero solution is stable. □

Remark 5.32. An eigenvalue is semisimple if its algebraic and geometric multiplicities are equal. Equivalently, every Jordan block associated with that eigenvalue has size one.

5.5.4 Modal analysis of important coupled systems

We now return to several coupled systems introduced in Section 5.1. Since the completeness of the modal representation has been established, it is enough to identify the eigenvectors of the coupling matrix. These eigenvectors give the spatial modes, while the corresponding scalar equations determine their temporal behavior.

Example 5.33 (Normal modes of a fixed-end spring–mass chain)**.** Consider a chain of n identical masses, each of mass m, connected by identical springs of stiffness $\kappa > 0$. Both endpoints are fixed. The displacement vector

$$\mathbf{x}(t) = \begin{bmatrix} x^1(t) \\ \vdots \\ x^n(t) \end{bmatrix}$$

satisfies

$$m\ddot{\mathbf{x}} + \kappa L_n^{(D)} \mathbf{x} = \mathbf{0},$$

where

$$L_n^{(D)} = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & -1 & 2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{bmatrix}$$

is the discrete Dirichlet Laplacian.

For $\ell = 1, \dots, n$, its eigenvalues and eigenvectors are

$$\mu_\ell = 2\left(1 - \cos\frac{\ell\pi}{n+1}\right) = 4\sin^2\left(\frac{\ell\pi}{2(n+1)}\right),$$

and

$$(\mathbf{v}_\ell)^j = \sin\left(\frac{j\ell\pi}{n+1}\right), \qquad j = 1, \dots, n.$$

Indeed, using

$$\sin((j-1)\theta) + \sin((j+1)\theta) = 2\sin(j\theta)\cos\theta,$$

with $\theta = \ell\pi/(n+1)$, one obtains

$$L_n^{(D)}\mathbf{v}_\ell = \mu_\ell \mathbf{v}_\ell.$$

For a motion of the form

$$\mathbf{x}(t) = q_\ell(t)\mathbf{v}_\ell,$$

the amplitude satisfies

$$m\ddot{q}_\ell + \kappa\mu_\ell q_\ell = 0.$$

Hence the natural frequency of the ℓth mode is

$$\omega_\ell = \sqrt{\frac{\kappa\mu_\ell}{m}} = 2\sqrt{\frac{\kappa}{m}}\sin\left(\frac{\ell\pi}{2(n+1)}\right).$$

Since $\mathbf{v}_1, \dots, \mathbf{v}_n$ form an orthogonal basis of $\mathbb{R}^n$, every motion has the complete modal representation

$$\mathbf{x}(t) = \sum_{\ell=1}^{n} \left[A_\ell \cos(\omega_\ell t) + B_\ell \sin(\omega_\ell t) \right] \mathbf{v}_\ell .$$

The lower modes vary slowly along the chain, whereas the higher modes have more rapid spatial oscillations and higher natural frequencies.

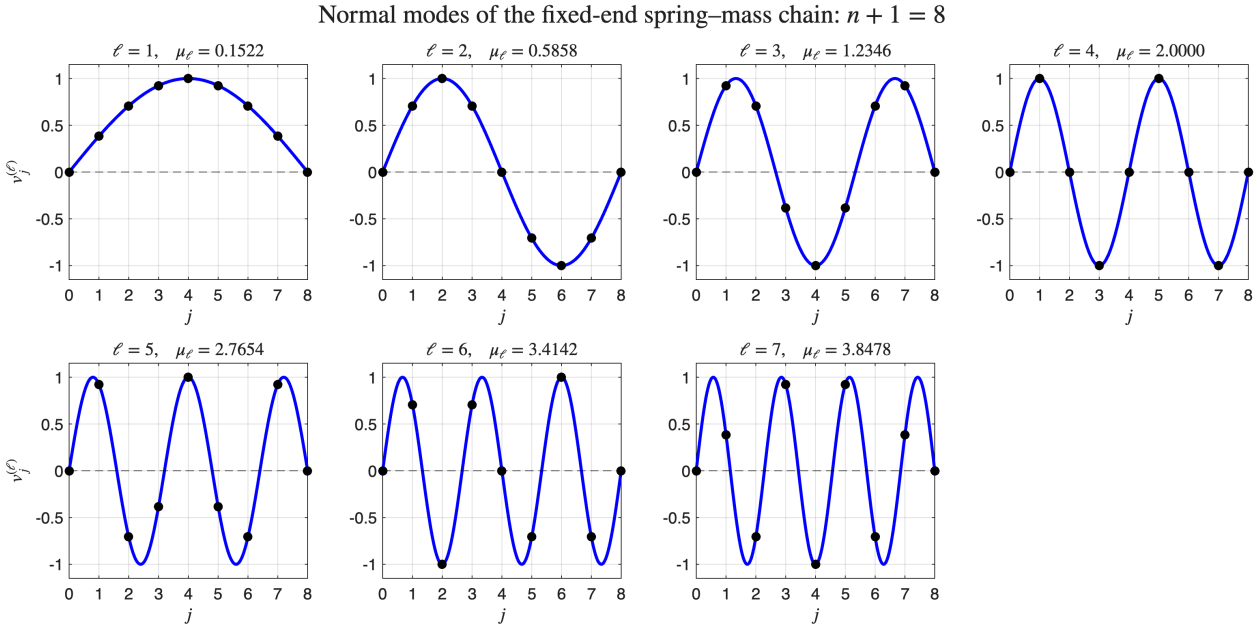


Figure 5.4: Normal modes of a fixed-end spring–mass chain with seven moving masses. The ℓth mode has spatial profile $(\mathbf{v}_\ell)^j = \sin(j\ell\pi/8)$, $j = 1, \dots, 7$. Higher modes have more rapid spatial oscillations and higher natural frequencies.

Remark 5.34 (The same spatial modes in a coupled RLC ladder). The coupled RLC ladder introduced in Section 5.1.2 satisfies

$$L\ddot{\mathbf{I}} + R\dot{\mathbf{I}} + \frac{1}{C} L_n^{(D)} \mathbf{I} = \mathbf{0}.$$

Because the same discrete Laplacian appears, the spatial modes are again $\mathbf{v}_1, \dots, \mathbf{v}_n$. Writing

$$\mathbf{I}(t) = \sum_{\ell=1}^{n} q_\ell(t) \mathbf{v}_\ell ,$$

the modal amplitudes satisfy

$$L\ddot{q}_\ell + R\dot{q}_\ell + \frac{\mu_\ell}{C} q_\ell = 0.$$

Thus the spring–mass chain and the RLC ladder have the same spatial modes, although their physical variables and temporal equations are different.

Example 5.35 (Graph Laplacian and consensus)**.** Let a network have n vertices, and let $y_j(t)$ be a scalar state associated with vertex j. Suppose neighboring vertices adjust their states toward one another:

$$\dot{y}_j = \kappa \sum_{k \sim j} (y_k - y_j), \qquad j = 1, \dots, n.$$

In matrix form,

$$\dot{\mathbf{y}} = -\kappa \mathbf{L}_G \mathbf{y},$$

where

$$\mathbf{L}_G = \mathbf{D} - \mathbf{W}$$

is the graph Laplacian, $\mathbf{W}$ is the adjacency matrix, and $\mathbf{D}$ is the diagonal matrix of vertex degrees.

For an undirected graph, $\mathbf{L}_G$ is symmetric and positive semidefinite. Indeed,

$$\mathbf{y}^T \mathbf{L}_G \mathbf{y} = \frac{1}{2} \sum_{j,k} w_{jk} (y_j - y_k)^2 \geq 0.$$

Moreover, the multiplicity of the eigenvalue 0 is equal to the number of connected components of the graph. Hence, if the graph is connected,

$$0 = \mu_1 < \mu_2 \leq \cdots \leq \mu_n,$$

and

$$\ker \mathbf{L}_G = \operatorname{span}\{\mathbf{1}\}.$$

We may therefore take

$$\mathbf{v}_1 = \frac{1}{\sqrt{n}} \mathbf{1}.$$

Expanding the initial state as

$$\mathbf{y}_0 = \sum_{j=1}^{n} c_j \mathbf{v}_j,$$

we obtain

$$\mathbf{y}(t) = c_1 \mathbf{v}_1 + \sum_{j=2}^{n} c_j e^{-\kappa \mu_j t} \mathbf{v}_j.$$

All nonconstant modes decay, while the constant mode remains. Therefore,

$$\mathbf{y}(t) \longrightarrow c_1 \mathbf{v}_1 = \frac{\mathbf{1}^T \mathbf{y}_0}{n} \mathbf{1}.$$

Every component approaches the initial average

$$\overline{y}_0 = \frac{1}{n} \sum_{j=1}^{n} y_j(0).$$

The zero solution is stable but not asymptotically stable, because of the zero eigenvalue. The consensus subspace

$$\operatorname{span}\{\mathbf{1}\}$$

is invariant and attracts every solution. The rate of convergence is controlled by the first positive eigenvalue μ_2, often called the spectral gap or algebraic connectivity of the graph.

Example 5.36 (Diffusion on a path graph)**.** For the path graph with n vertices, the graph Laplacian is

$$L_n^{(N)} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & -1 & 2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix}.$$

Its eigenvalues and eigenvectors are

$$\mu_k = 2\left(1 - \cos\frac{k\pi}{n}\right), \qquad k = 0, \dots, n-1,$$

and

$$(\mathbf{v}_k)^j = \cos\left[\frac{k\pi}{n}\left(j - \frac{1}{2}\right)\right], \qquad j = 1, \dots, n.$$

Hence

$$\mathbf{u}(t) = \sum_{k=0}^{n-1} c_k e^{-\kappa\mu_k t}\mathbf{v}_k.$$

The first mode is constant:

$$\mu_0 = 0, \qquad \mathbf{v}_0 = \mathbf{1}.$$

All remaining modes decay. Higher modes oscillate more rapidly in space and decay faster because their eigenvalues are larger.

Figures 5.4 and 5.5 show how boundary conditions affect the modal structure. Fixed endpoints lead to sine modes and strictly positive eigenvalues. The path graph has cosine modes and retains a constant zero mode. In both cases, completeness of the eigenvectors gives a complete representation of every motion.

Remark 5.37 (The same modes in a reversible reaction chain). The reversible reaction chain introduced in Section 5.1.3 satisfies

$$\dot{\mathbf{c}} = -k\mathbf{L}_n^{(N)}\mathbf{c},$$

where $\mathbf{L}_n^{(N)}$ is the same discrete Neumann Laplacian that appears for diffusion on the path graph. Therefore the concentration modes are the cosine modes

$$(\mathbf{v}_\ell)_j = \cos\left[\frac{\ell\pi}{n}\left(j - \frac{1}{2}\right)\right], \qquad \ell = 0, \dots, n-1.$$

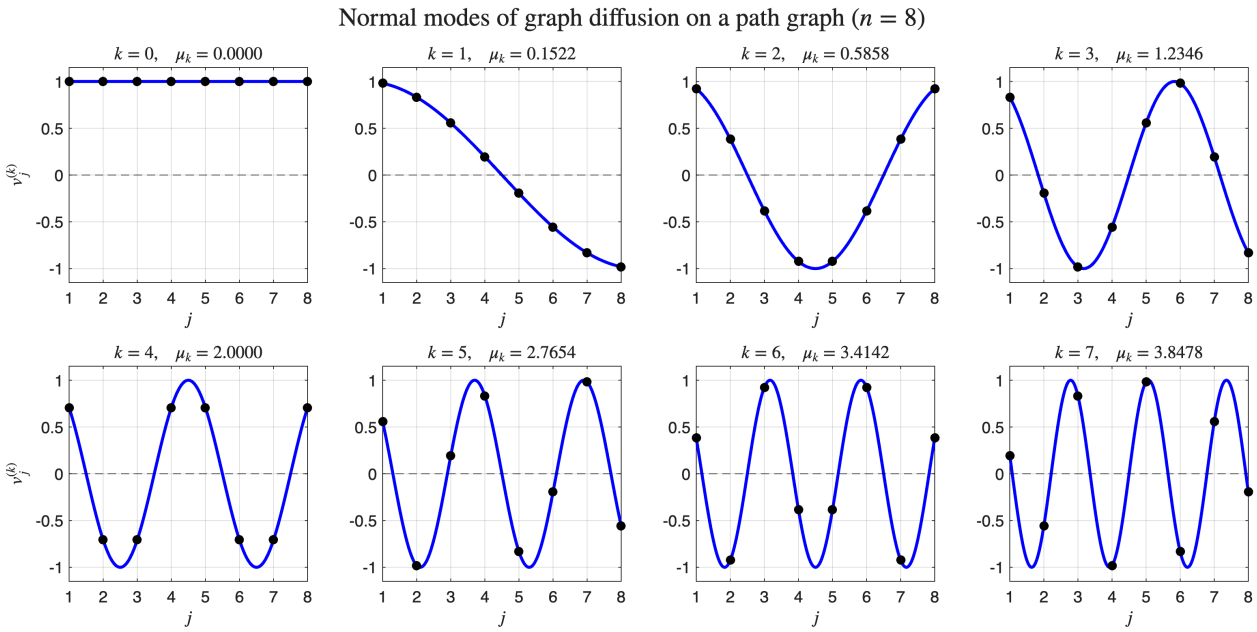


Figure 5.5: Normal modes of diffusion on the path graph with $n = 8$ vertices. The constant mode is the equilibrium state. Higher cosine modes oscillate more rapidly in space and decay faster.

Writing

$$\mathbf{c}(t) = \sum_{\ell=0}^{n-1} a_\ell(t)\mathbf{v}_\ell,$$

the modal amplitudes satisfy

$$\dot{a}_\ell = -k\mu_\ell a_\ell,$$

and hence

$$a_\ell(t) = a_\ell(0)e^{-k\mu_\ell t}.$$

The constant mode $\mathbf{v}_0 = \mathbf{1}$ represents conservation of the total concentration, while all nonconstant modes decay. Thus the reaction chain approaches the uniform concentration determined by its initial total mass.

Exercise 5.15 (A decaying Jordan mode and a constant mode)**.** For

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

Show that the eigenvalues are $-1, -1, 0$ and determine the associated Jordan structure. Use a Jordan chain to find the general solution of $\mathbf{y}' = \mathbf{A}\mathbf{y}$. Is the origin stable? Is it asymptotically stable?

Exercise 5.16 (Imaginary eigenvalues with Jordan blocks)**.** Let

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

Show that the eigenvalues are $\pm i$, with one Jordan block of size two for each eigenvalue over $\mathbb{C}$. Determine whether the origin of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ is stable.

Exercise 5.17 (Designing systems with prescribed stability)**.** Construct a real 3×3 matrix for each of the following cases:

1. the origin is asymptotically stable;
2. the origin is stable but not asymptotically stable;
3. all eigenvalues satisfy $\Re\lambda \leq 0$, but the origin is unstable.

Justify each answer using the eigenvalues and Jordan structure.

5.6 Matrix Exponential

The preceding sections showed that every matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ admits a Jordan decomposition

$$\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1},$$

where $\mathbf{P}$ is invertible and $\mathbf{J}$ is block diagonal. The corresponding change of variables

$$\mathbf{y} = \mathbf{P}\mathbf{z}$$

reduces

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}$$

to

$$\dot{\mathbf{z}} = \mathbf{J}\mathbf{z}.$$

The matrix exponential packages this evolution into a single operator. After defining it, we will show that

$$e^{t\mathbf{A}} = \mathbf{P}e^{t\mathbf{J}}\mathbf{P}^{-1}.$$

Thus its computation reduces to the exponential of a Jordan block. In Jordan coordinates, the columns of $e^{t\mathbf{J}}$ are precisely the ordinary and generalized modes obtained earlier. The columns of $e^{t\mathbf{A}}$ are their corresponding linear combinations in the original coordinates.

5.6.1 Definition and evolution operator

The scalar exponential is defined by the power series

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

The same series defines the exponential of a square matrix.

Definition 5.38 (Matrix exponential)**.** For an $n \times n$ matrix $\mathbf{A}$, define

$$e^{t\mathbf{A}} = \sum_{k=0}^{\infty} \frac{t^{k}}{k!}\mathbf{A}^{k}.$$

The series converges for every t. Indeed, for any submultiplicative matrix norm,

$$\left\| \frac{t^{k}}{k!}\mathbf{A}^{k} \right\| \le \frac{|t|^{k}}{k!}\|\mathbf{A}\|^{k},$$

and the scalar majorant sums to $e^{|t|\|\mathbf{A}\|}$.

Proposition 5.39 (Evolution equation)**.** *The matrix exponential satisfies*

$$\frac{d}{dt}e^{t\mathbf{A}} = \mathbf{A}e^{t\mathbf{A}} = e^{t\mathbf{A}}\mathbf{A}, \qquad e^{0\mathbf{A}} = \mathbf{I}.$$

Consequently, it is the unique solution of

$$\dot{\mathbf{\Phi}}(t) = \mathbf{A}\mathbf{\Phi}(t), \qquad \mathbf{\Phi}(0) = \mathbf{I}.$$

Proof. By the Weierstrass M-test, both the matrix exponential series and its term-by-term derivative converge uniformly on every bounded interval. Hence the series may be differentiated term by term:

$$\begin{aligned} \frac{d}{dt}e^{t\mathbf{A}} &= \sum_{k=1}^{\infty} \frac{t^{k-1}}{(k-1)!}\mathbf{A}^{k} \\ &= \mathbf{A}\sum_{j=0}^{\infty} \frac{t^{j}}{j!}\mathbf{A}^{j} \\ &= \mathbf{A}e^{t\mathbf{A}}. \end{aligned}$$

Since $\mathbf{A}$ commutes with every power of itself,

$$\mathbf{A}e^{t\mathbf{A}} = e^{t\mathbf{A}}\mathbf{A}.$$

The initial condition follows directly from the series. Uniqueness is a consequence of the existence–uniqueness theorem for linear systems. □

It follows that the initial-value problem

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}, \qquad \mathbf{y}(t_0) = \mathbf{y}_0,$$

has the solution

$$\mathbf{y}(t) = e^{(t-t_0)\mathbf{A}}\mathbf{y}_0.$$

Thus the transition matrix is

$$\mathbf{\Phi}(t, t_0) = e^{(t-t_0)\mathbf{A}}.$$

The matrix exponential therefore gives the evolution operator for the constant-coefficient system. If

$$\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1},$$

then

$$\mathbf{A}^k = \mathbf{P}\mathbf{J}^k\mathbf{P}^{-1} \qquad \text{for every } k \geq 0.$$

Substitution into the defining series gives

$$\begin{aligned} e^{t\mathbf{A}} &= \sum_{k=0}^{\infty} \frac{t^k}{k!}\mathbf{A}^k \\ &= \mathbf{P}\left(\sum_{k=0}^{\infty} \frac{t^k}{k!}\mathbf{J}^k\right)\mathbf{P}^{-1} \\ &= \mathbf{P}e^{t\mathbf{J}}\mathbf{P}^{-1}. \end{aligned}$$

Thus it remains to compute $e^{t\mathbf{J}}$ block by block.

5.6.2 Exponential of a Jordan block

Let

$$\mathbf{J}_\ell(\lambda) = \lambda\mathbf{I} + \mathbf{N},$$

where

$$\mathbf{N} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 0 & 1 \\ 0 & \cdots & \cdots & 0 & 0 \end{bmatrix}, \qquad \mathbf{N}^\ell = \mathbf{0}.$$

Since $\lambda\mathbf{I}$ commutes with $\mathbf{N}$,

$$e^{t\mathbf{J}_\ell(\lambda)} = e^{\lambda t\mathbf{I}}e^{t\mathbf{N}} = e^{\lambda t}e^{t\mathbf{N}}.$$

The nilpotency of $\mathbf{N}$ terminates the exponential series:

$$e^{t\mathbf{N}} = \mathbf{I} + t\mathbf{N} + \frac{t^2}{2!}\mathbf{N}^2 + \cdots + \frac{t^{\ell-1}}{(\ell-1)!}\mathbf{N}^{\ell-1}.$$

Therefore,

$$e^{t\mathbf{J}_\ell(\lambda)} = e^{\lambda t}\left(\mathbf{I} + t\mathbf{N} + \frac{t^2}{2!}\mathbf{N}^2 + \cdots + \frac{t^{\ell-1}}{(\ell-1)!}\mathbf{N}^{\ell-1}\right).$$

Equivalently,

$$e^{t\mathbf{J}_\ell(\lambda)} = e^{\lambda t}\begin{bmatrix} 1 & t & \frac{t^2}{2!} & \cdots & \frac{t^{\ell-1}}{(\ell-1)!} \\ 0 & 1 & t & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \frac{t^2}{2!} \\ 0 & \cdots & 0 & 1 & t \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix}.$$

The columns of this matrix are precisely the coordinate generalized modes associated with the standard Jordan chain

$$\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_\ell.$$

For example, the first three columns are

$$e^{\lambda t}\mathbf{e}_1,$$

$$e^{\lambda t}(t\mathbf{e}_1 + \mathbf{e}_2),$$

and

$$e^{\lambda t}\left(\frac{t^2}{2!}\mathbf{e}_1 + t\mathbf{e}_2 + \mathbf{e}_3\right).$$

Thus the polynomial factors in the generalized modes arise directly from the nilpotent part of the Jordan block.

If

$$\mathbf{J} = \mathbf{J}_{\ell_1}(\lambda_1) \oplus \cdots \oplus \mathbf{J}_{\ell_s}(\lambda_s),$$

then

$$e^{t\mathbf{J}} = e^{t\mathbf{J}_{\ell_1}(\lambda_1)} \oplus \cdots \oplus e^{t\mathbf{J}_{\ell_s}(\lambda_s)}.$$

Hence the exponential of a Jordan matrix is obtained by exponentiating each Jordan block separately.

Real blocks for complex eigenvalues. For a real matrix, a conjugate pair

$$\lambda = \alpha + i\beta, \qquad \overline{\lambda} = \alpha - i\beta, \qquad \beta \neq 0,$$

may be represented in real coordinates by the block

$$\mathbf{C} = \begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix} = \alpha\mathbf{I}_2 + \beta\mathbf{R}, \qquad \mathbf{R} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Since $\mathbf{R}^2 = -\mathbf{I}_2$,

$$e^{t\mathbf{C}} = e^{\alpha t}\begin{bmatrix} \cos\beta t & \sin\beta t \\ -\sin\beta t & \cos\beta t \end{bmatrix}.$$

Thus $e^{t\mathbf{C}}$ combines exponential growth or decay at rate α with rotation at angular speed $|\beta|$.

More generally, the real block corresponding to a complex Jordan chain of length ℓ is

$$\mathbf{J}_\ell^{\mathbb{R}}(\alpha,\beta) = \begin{bmatrix} \mathbf{C} & \mathbf{I}_2 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C} & \mathbf{I}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{C} & \mathbf{I}_2 \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{C} \end{bmatrix}.$$

Writing

$$\mathbf{J}_\ell^{\mathbb{R}}(\alpha,\beta) = \mathbf{I}_\ell \otimes \mathbf{C} + \mathbf{N}_\ell \otimes \mathbf{I}_2,$$

the two terms commute, and hence

$$e^{t\mathbf{J}_\ell^{\mathbb{R}}(\alpha,\beta)} = e^{t\mathbf{N}_\ell} \otimes e^{t\mathbf{C}}.$$

Equivalently,

$$e^{t\mathbf{J}_\ell^{\mathbb{R}}(\alpha,\beta)} = \begin{bmatrix} e^{t\mathbf{C}} & te^{t\mathbf{C}} & \frac{t^2}{2!}e^{t\mathbf{C}} & \cdots & \frac{t^{\ell-1}}{(\ell-1)!}e^{t\mathbf{C}} \\ \mathbf{0} & e^{t\mathbf{C}} & te^{t\mathbf{C}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \frac{t^2}{2!}e^{t\mathbf{C}} \\ \mathbf{0} & \cdots & \mathbf{0} & e^{t\mathbf{C}} & te^{t\mathbf{C}} \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & e^{t\mathbf{C}} \end{bmatrix}.$$

The diagonal blocks describe rotation combined with exponential growth or decay, while the blocks above the diagonal introduce the polynomial factors associated with generalized modes.

5.6.3 Computing the matrix exponential from Jordan form

Suppose

$$\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1},$$

where $\mathbf{J}$ is the Jordan form of $\mathbf{A}$. Since

$$\mathbf{A}^k = \mathbf{P}\mathbf{J}^k\mathbf{P}^{-1},$$

the power-series definition gives

$$\begin{aligned} e^{t\mathbf{A}} &= \sum_{k=0}^{\infty} \frac{t^k}{k!}\mathbf{A}^k \\ &= \mathbf{P}\left(\sum_{k=0}^{\infty} \frac{t^k}{k!}\mathbf{J}^k\right)\mathbf{P}^{-1}. \end{aligned}$$

Therefore,

$$\boxed{e^{t\mathbf{A}} = \mathbf{P}e^{t\mathbf{J}}\mathbf{P}^{-1}.}$$

This formula reduces the computation of $e^{t\mathbf{A}}$ to the Jordan-block formulas obtained in the preceding subsection. If

$$\mathbf{J} = \mathbf{J}_{\ell_1}(\lambda_1) \oplus \cdots \oplus \mathbf{J}_{\ell_s}(\lambda_s),$$

then

$$e^{t\mathbf{J}} = e^{t\mathbf{J}_{\ell_1}(\lambda_1)} \oplus \cdots \oplus e^{t\mathbf{J}_{\ell_s}(\lambda_s)}.$$

If $\mathbf{A}$ is diagonalizable, every Jordan block has size one. Thus

$$\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}, \qquad \mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n),$$

and

$$e^{t\mathbf{A}} = \mathbf{P} \operatorname{diag}\big(e^{\lambda_1 t}, \ldots, e^{\lambda_n t}\big)\mathbf{P}^{-1}.$$

This is the matrix form of the ordinary modal representation.

When nontrivial Jordan blocks are present, the same formula introduces the polynomial factors associated with generalized modes. Hence the Jordan decomposition unifies the diagonalizable and defective cases.

The matrix exponential therefore contains the complete modal representation in a single matrix. For every initial value $\mathbf{y}_0$,

$$\mathbf{y}(t) = e^{t\mathbf{A}}\mathbf{y}_0$$

is obtained by first resolving $\mathbf{y}_0$ into Jordan coordinates, evolving each Jordan component independently, and then returning to the original coordinates.

Exercise 5.18 (Computing a matrix exponential)**.** Let

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

1. Identify the Jordan blocks and compute $e^{t\mathbf{A}}$.
2. Use the result to solve $\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}$ with

$$\mathbf{y}(0) = \begin{bmatrix} a \\ b \\ c \end{bmatrix}.$$

5.6.4 Basic properties of the matrix exponential

The matrix exponential satisfies the familiar exponential laws associated with a one-parameter evolution. Some identities hold for every matrix, while others require commutativity.

Proposition 5.40. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be square matrices of the same size. Then:*

1.

$$e^{0\mathbf{A}} = \mathbf{I}.$$

2. For all $s, t \in \mathbb{R}$,

$$e^{(t+s)\mathbf{A}} = e^{t\mathbf{A}}e^{s\mathbf{A}}.$$

3.

$$\big(e^{t\mathbf{A}}\big)^{-1} = e^{-t\mathbf{A}}.$$

4. If $\mathbf{P}$ *is invertible, then*

$$e^{t\mathbf{P}^{-1}\mathbf{A}\mathbf{P}} = \mathbf{P}^{-1}e^{t\mathbf{A}}\mathbf{P}.$$

5. If

$$\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A},$$

then

$$e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{A}}e^{\mathbf{B}} = e^{\mathbf{B}}e^{\mathbf{A}}.$$

Proof. The first identity follows directly from the power-series definition.

For the semigroup property, fix s and define

$$\mathbf{F}(t) = e^{t\mathbf{A}}e^{s\mathbf{A}}.$$

Then

$$\dot{\mathbf{F}}(t) = \mathbf{A}\mathbf{F}(t), \qquad \mathbf{F}(0) = e^{s\mathbf{A}}.$$

The matrix

$$\mathbf{G}(t) = e^{(t+s)\mathbf{A}}$$

satisfies the same initial-value problem. Hence uniqueness gives

$$e^{(t+s)\mathbf{A}} = e^{t\mathbf{A}}e^{s\mathbf{A}}.$$

Taking $s = -t$ yields

$$e^{t\mathbf{A}}e^{-t\mathbf{A}} = \mathbf{I},$$

and therefore

$$\left(e^{t\mathbf{A}}\right)^{-1} = e^{-t\mathbf{A}}.$$

For the similarity property,

$$\left(\mathbf{P}^{-1}\mathbf{A}\mathbf{P}\right)^{k} = \mathbf{P}^{-1}\mathbf{A}^{k}\mathbf{P},$$

so the conclusion follows by substituting into the exponential series.

Finally, if $\mathbf{A}$ and $\mathbf{B}$ commute, then the binomial formula applies:

$$(\mathbf{A}+\mathbf{B})^{k} = \sum_{j=0}^{k} \binom{k}{j} \mathbf{A}^{j}\mathbf{B}^{k-j}.$$

Substitution into the power series gives

$$e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{A}}e^{\mathbf{B}}.$$

Since $\mathbf{A}$ and $\mathbf{B}$ commute, their exponentials also commute. □

The semigroup property expresses the composition of successive evolutions:

$$e^{(t-s)\mathbf{A}}e^{(s-r)\mathbf{A}} = e^{(t-r)\mathbf{A}}.$$

Thus the transition matrix

$$\mathbf{\Phi}(t,s) = e^{(t-s)\mathbf{A}}$$

satisfies

$$\mathbf{\Phi}(t,s)\mathbf{\Phi}(s,r) = \mathbf{\Phi}(t,r).$$

In general,

$$e^{\mathbf{A}+\mathbf{B}} \neq e^{\mathbf{A}}e^{\mathbf{B}}$$

when $\mathbf{A}$ and $\mathbf{B}$ do not commute. This obstruction becomes important for nonautonomous systems, where the matrices at different times need not commute.

5.6.5 Product approximation for nonautonomous systems

Consider the nonautonomous linear system

$$\dot{\mathbf{y}} = \mathbf{A}(t)\mathbf{y}, \qquad \mathbf{y}(0) = \mathbf{y}_0,$$

where $\mathbf{A}(t)$ is a smooth matrix-valued function. In general, its solution is not

$$\mathbf{y}(t) = \exp\left(\int_0^t \mathbf{A}(s)\,ds\right)\mathbf{y}_0,$$

because matrices evaluated at different times need not commute.

To see the role of noncommutativity, let $\mathbf{A}$ and $\mathbf{B}$ be two constant matrices. For a small time step Δt, expansion of the matrix exponentials gives

$$e^{\Delta t\mathbf{A}}e^{\Delta t\mathbf{B}} = \mathbf{I} + \Delta t(\mathbf{A}+\mathbf{B}) + (\Delta t)^2\left(\frac{1}{2}\mathbf{A}^2 + \mathbf{AB} + \frac{1}{2}\mathbf{B}^2\right) + O((\Delta t)^3),$$

whereas

$$e^{\Delta t\mathbf{B}}e^{\Delta t\mathbf{A}} = \mathbf{I} + \Delta t(\mathbf{A}+\mathbf{B}) + (\Delta t)^2\left(\frac{1}{2}\mathbf{A}^2 + \mathbf{BA} + \frac{1}{2}\mathbf{B}^2\right) + O((\Delta t)^3).$$

Therefore,

$$e^{\Delta t\mathbf{A}}e^{\Delta t\mathbf{B}} - e^{\Delta t\mathbf{B}}e^{\Delta t\mathbf{A}} = (\Delta t)^2[\mathbf{A},\mathbf{B}] + O((\Delta t)^3),$$

where

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

is the matrix commutator. Thus the order in which short-time evolutions are applied first becomes visible at order $(\Delta t)^2$.

We now choose a fixed step size $\Delta t > 0$ and define

$$t^k = k\Delta t, \qquad k = 0, 1, \ldots, N,$$

where $T = N\Delta t$. On the interval $[t^k, t^{k+1}]$, replace $\mathbf{A}(t)$ by the constant matrix

$$\mathbf{A}^{(k)} = \mathbf{A}(t^k).$$

The resulting first-order scheme is

$$\boxed{\mathbf{y}^{k+1} = e^{\Delta t\mathbf{A}^{(k)}}\mathbf{y}^k, \qquad \mathbf{y}^0 = \mathbf{y}_0.}$$

After N steps,

$$\mathbf{y}^N = e^{\Delta t\mathbf{A}^{(N-1)}}e^{\Delta t\mathbf{A}^{(N-2)}}\cdots e^{\Delta t\mathbf{A}^{(0)}}\mathbf{y}_0.$$

The earliest-time factor acts first and therefore appears on the right. This ordered product retains the effect of the noncommutativity of the matrices $\mathbf{A}(t^k)$.

If $\mathbf{A}$ is continuously differentiable on $[0, T]$, then the method is first-order convergent: there is a constant C_T, independent of Δt, such that

$$\max_{0 \le k \le N} \left| \mathbf{y}(t^k) - \mathbf{y}^k \right| \le C_T \Delta t$$

for all sufficiently small Δt. Equivalently, the ordered products converge to the transition matrix as $\Delta t \to 0$.

Since

$$e^{\Delta t \mathbf{A}^{(k)}} = \mathbf{I} + \Delta t \mathbf{A}^{(k)} + O((\Delta t)^2),$$

this method is closely related to the forward Euler scheme

$$\mathbf{y}^{k+1} = \mathbf{y}^k + \Delta t \, \mathbf{A}(t^k)\mathbf{y}^k.$$

The convergence analysis of first-order numerical methods of this type will be developed in Chapter 10.

Thus approximations connect the exact matrix exponential for autonomous systems with numerical time stepping for nonautonomous systems. The key point is that short-time evolution operators must be composed in chronological order when the coefficient matrices do not commute.

Exercise 5.19 (Exponentials of noncommuting matrices)**.** Let

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

1. Compute $\mathbf{AB}$, $\mathbf{BA}$, and $[\mathbf{A}, \mathbf{B}]$.
2. Use $\mathbf{A}^2 = \mathbf{B}^2 = \mathbf{0}$ and $(\mathbf{A} + \mathbf{B})^2 = \mathbf{I}$ to compute

   $$e^{t\mathbf{A}} e^{t\mathbf{B}} \qquad \text{and} \qquad e^{t(\mathbf{A}+\mathbf{B})}.$$

   Show that they are unequal for $t \neq 0$.
3. Expand both matrices for small t and identify the first term at which they differ.

Homework Problems

Modeling and modal analysis

Homework 5.1 (Diffusion on a star graph)**.** Let G be the star graph with one central vertex connected to $n - 1$ outer vertices, and consider

$$\mathbf{u}' = -\kappa L_G \mathbf{u}, \qquad \kappa > 0.$$

1. Construct L_G and determine all its eigenvalues and eigenspaces.

2. Write the modal representation of the solution.

3. Find the long-time limit of $\mathbf{u}(t)$ and interpret it physically.

Homework 5.2 (Critical damping in a coupled system)**.** Consider

$$m\ddot{\mathbf{x}} + \gamma\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{0}, \qquad \mathbf{K} = \begin{bmatrix} k+\kappa & -\kappa \\ -\kappa & k+\kappa \end{bmatrix},$$

where $m, k, \kappa > 0$.

1. Show that
$$\mathbf{p} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad \mathbf{q} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$
are eigenvectors of $\mathbf{K}$, and find their eigenvalues.

2. Writing $\mathbf{x} = u\mathbf{p} + v\mathbf{q}$, derive the scalar equations for u and v.

3. Choose γ so that the $\mathbf{p}$-mode is critically damped. Give its two independent solutions and determine the damping type of the $\mathbf{q}$-mode.

Homework 5.3 (Resonant transfer in a nonreciprocally coupled oscillator)**.** Consider

$$\begin{aligned} \ddot{x}_1 + \omega^2 x_1 + \kappa(x_1 - x_2) &= 0, \\ \ddot{x}_2 + \omega^2 x_2 + \kappa(x_1 - x_2) &= 0, \end{aligned} \qquad \omega > 0, \quad \kappa \neq 0.$$

Equivalently, $\ddot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{0}$, where

$$\mathbf{K} = \begin{bmatrix} \omega^2 + \kappa & -\kappa \\ \kappa & \omega^2 - \kappa \end{bmatrix}.$$

1. Show that $\mathbf{K} = \omega^2\mathbf{I} + \mathbf{N}$, where $\mathbf{N}^2 = \mathbf{0}$ but $\mathbf{N} \neq \mathbf{0}$. Construct a Jordan chain for ω^2.

2. Set $u = x_1 - x_2$ and $v = x_1 + x_2$. Derive the equations for u, v and find the general solution. Explain the terms $t\sin\omega t$ and $t\cos\omega t$.

3. Compare this internal resonant transfer with
$$\ddot{y} + \omega^2 y = F_0 \cos\omega t.$$
Identify the common mechanism and the difference in interpretation.

Generalized modes

Homework 5.4 (Complex generalized modes)**.** Suppose $\mathbf{A} \in \mathbb{R}^{4\times 4}$ has a complex Jordan chain

$\mathbf{v}_0, \mathbf{v}_1$ for

$$\lambda = \alpha + i\beta, \qquad \beta \neq 0,$$

so that

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{v}_0 = \mathbf{0}, \qquad (\mathbf{A} - \lambda \mathbf{I})\mathbf{v}_1 = \mathbf{v}_0.$$

Write $\mathbf{v}_j = \mathbf{p}_j + i\mathbf{q}_j$.

1. Take the real and imaginary parts of

$$e^{\lambda t}\mathbf{v}_0, \qquad e^{\lambda t}(t\mathbf{v}_0 + \mathbf{v}_1),$$

   and write the resulting four real solutions explicitly.

2. Show that their values at $t = 0$ are linearly independent whenever $\mathbf{p}_0, \mathbf{q}_0, \mathbf{p}_1, \mathbf{q}_1$ are linearly independent.

Linear stability

Homework 5.5 (Uniform exponential decay)**.** Suppose every eigenvalue of $\mathbf{A}$ satisfies

$$\Re\lambda \leq -\gamma$$

for some $\gamma > 0$. Show that, for every $0 < \beta < \gamma$, there is a constant $C > 0$ such that every solution of $\mathbf{y}' = \mathbf{A}\mathbf{y}$ satisfies

$$\|\mathbf{y}(t)\| \leq Ce^{-\beta t}\|\mathbf{y}(0)\|, \qquad t \geq 0.$$

Explain why nontrivial Jordan blocks do not prevent asymptotic stability.

Matrix exponential

Homework 5.6 (A closed formula for a 2×2 exponential)**.** Let $\mathbf{A}$ be a real 2×2 matrix, and set

$$\tau = \operatorname{tr}\mathbf{A}, \qquad \delta = \det\mathbf{A}, \qquad \mathbf{B} = \mathbf{A} - \frac{\tau}{2}\mathbf{I}, \qquad \Delta = \frac{\tau^2}{4} - \delta.$$

1. Use the Cayley–Hamilton theorem to show that

$$\mathbf{B}^2 = \Delta\mathbf{I}, \qquad e^{t\mathbf{A}} = e^{\tau t/2}e^{t\mathbf{B}}.$$

2. Derive the following three formulas:

$$\begin{aligned}
&\Delta > 0,\ \mu = \sqrt{\Delta}: && e^{t\mathbf{A}} = e^{\tau t/2}\left(\cosh(\mu t)\mathbf{I} + \frac{\sinh(\mu t)}{\mu}\mathbf{B}\right),\\
&\Delta = 0: && e^{t\mathbf{A}} = e^{\tau t/2}(\mathbf{I} + t\mathbf{B}),\\
&\Delta < 0,\ \nu = \sqrt{-\Delta}: && e^{t\mathbf{A}} = e^{\tau t/2}\left(\cos(\nu t)\mathbf{I} + \frac{\sin(\nu t)}{\nu}\mathbf{B}\right).
\end{aligned}$$

3. Relate the three cases to the eigenvalue types of $\mathbf{A}$.

Product approximation

Homework 5.7 (Strang splitting)**.** Let $\mathbf{A}$ and $\mathbf{B}$ be square matrices. By expanding through terms of order h^2, show that

$$e^{\frac{h}{2}\mathbf{B}}e^{h\mathbf{A}}e^{\frac{h}{2}\mathbf{B}} = \mathbf{I} + h(\mathbf{A}+\mathbf{B}) + \frac{h^2}{2}(\mathbf{A}+\mathbf{B})^2 + O(h^3).$$

Deduce that

$$e^{h(\mathbf{A}+\mathbf{B})} = e^{\frac{h}{2}\mathbf{B}}e^{h\mathbf{A}}e^{\frac{h}{2}\mathbf{B}} + O(h^3),$$

even when $\mathbf{A}$ and $\mathbf{B}$ do not commute.

Projects

Project 5.1 (Evolution for a nonautonomous linear system)**.** This project examines how a time-dependent, noncommuting coefficient matrix changes the evolution of a linear system. Consider

$$\dot{\mathbf{y}} = \mathbf{A}(t)\mathbf{y}, \qquad \mathbf{A}(t) = \begin{bmatrix} 1 & e^{2t} \\ 0 & -1 \end{bmatrix}.$$

1. Show that $\mathbf{A}(t_1)$ and $\mathbf{A}(t_2)$ need not commute. Solve the system, construct the principal fundamental matrix $\mathbf{X}(t)$, and compute

$$\mathbf{\Phi}(t,s) = \mathbf{X}(t)\mathbf{X}(s)^{-1}.$$

2. Verify the composition law for $\mathbf{\Phi}$, use Liouville's formula to check $\det \mathbf{X}(t)$, and explain why generally

$$\mathbf{X}(t) \neq \exp\left(\int_0^t \mathbf{A}(s)\,ds\right).$$

3. Use variation of constants to solve

$$\dot{\mathbf{y}} = \mathbf{A}(t)\mathbf{y} + \begin{bmatrix} e^{t} \\ e^{-t} \end{bmatrix}, \qquad \mathbf{y}(0) = \mathbf{0},$$

and verify the result by substitution.

Historical Notes and Further Reading

Historical notes. The study of coupled linear systems was strongly influenced by problems of small oscillation in mechanics and celestial dynamics. In his *Mécanique analytique*, Lagrange (1788) showed how the motion of a system near equilibrium could be reduced to a characteristic algebraic equation. The resulting independent patterns of oscillation are now called normal modes. Each mode has its own frequency and spatial pattern, and a general small oscillation is obtained by superposing these modes. This physical idea is the origin of the modal viewpoint developed in this chapter.

During the nineteenth century, the same algebraic questions arose in the study of linear substitutions and systems of differential equations. If a matrix has enough independent eigenvectors, a change of coordinates decouples the system into independent scalar equations. When eigenvectors coalesce, generalized eigenvectors are required. The canonical reduction associated with such chains was developed by Jordan (1870). In differential equations, a Jordan chain produces generalized modes of the form

$$e^{\lambda t}\left(\mathbf{v}_k + t\mathbf{v}_{k-1} + \cdots + \frac{t^k}{k!}\mathbf{v}_0\right),$$

so the algebraic structure of the matrix determines the time dependence of the solutions.

Matrices arising from interconnected systems often also reflect the geometry of the underlying network. In his analysis of electrical networks, Kirchhoff (1847) introduced a matrix closely related to what is now called the graph Laplacian. Its nullspace records conserved or constant modes, while its nonzero eigenvalues describe differences across the network. In mechanical systems these eigenvectors represent normal modes; in diffusion and consensus models they represent spatial patterns that decay at rates determined by the corresponding eigenvalues.

The matrix exponential combines all these modes into the evolution operator

$$\mathbf{y}(t) = e^{t\mathbf{A}}\mathbf{y}_0.$$

For a diagonalizable matrix, this formula superposes ordinary exponential modes. For a matrix with Jordan blocks, it automatically incorporates the polynomial factors associated with generalized modes. Chapter 6 gives the complementary transform-domain description, in which the same eigenvalues and Jordan structure appear as poles of the resolvent $(s\mathbf{I} - \mathbf{A})^{-1}$.

Further reading. For a systematic treatment of linear systems, fundamental matrices, matrix exponentials, and stability, see Hirsch et al. (2013). The algebraic theory of eigenvalues, invariant subspaces, Jordan form, and matrix functions is developed further in Horn and Johnson (2013). For normal modes and their applications to coupled mechanical systems, see Meirovitch (2001). Graph Laplacians, spectral modes, and their connections with diffusion and network structure are treated in Chung (1997).

Chapter 6

Laplace Transform Methods

The Laplace transform converts differentiation in time into multiplication by a complex variable s. Its basic mechanism comes from the exponential kernel e^{-st}, for which

$$De^{-st} = -se^{-st}.$$

After integration by parts, differentiation of a function becomes multiplication by s, together with a term determined by its initial value:

$$\mathcal{L}\{Df\}(s) = s\mathcal{L}\{f\}(s) - f(0).$$

Thus, for functions with zero initial value, the formal correspondence is

$$D \longleftrightarrow s.$$

For a scalar constant-coefficient equation

$$P(D)y = f,$$

the Laplace transform gives

$$P(s)Y(s) = I(s) + F(s),$$

where $I(s)$ contains the initial data. With zero initial data,

$$Y(s) = \frac{1}{P(s)}F(s).$$

Thus the causal inverse of the differential operator is represented in the transform domain by

$$P(D)^{-1} \longleftrightarrow \frac{1}{P(s)}.$$

The same principle applies to a first-order linear system

$$\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{f}.$$

Taking the Laplace transform gives

$$(s\mathbf{I} - \mathbf{A})\mathbf{Y}(s) = \mathbf{y}_0 + \mathbf{F}(s).$$

With zero initial data,

$$\mathbf{Y}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{F}(s),$$

so the corresponding inverse-operator relation is

$$(D\mathbf{I} - \mathbf{A})^{-1} \longleftrightarrow (s\mathbf{I} - \mathbf{A})^{-1}.$$

The matrix-valued function

$$(s\mathbf{I} - \mathbf{A})^{-1}$$

is the resolvent of $\mathbf{A}$.

These two correspondences organize the chapter:

$$P(D)^{-1} \longleftrightarrow \frac{1}{P(s)}, \qquad (D\mathbf{I} - \mathbf{A})^{-1} \longleftrightarrow (s\mathbf{I} - \mathbf{A})^{-1}.$$

In the scalar case, the inverse transform of $1/P(s)$ is the causal Green's function. Its action on a general forcing term is expressed by convolution, and its response to the delta source is the impulse response. In the system case, the inverse transform of the resolvent is the evolution operator:

$$\mathcal{L}^{-1}\left\{(s\mathbf{I} - \mathbf{A})^{-1}\right\} = e^{t\mathbf{A}}.$$

The structure developed in Chapter 5 appears naturally in the resolvent. If

$$\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1},$$

then

$$(s\mathbf{I} - \mathbf{A})^{-1} = \mathbf{P}(s\mathbf{I} - \mathbf{J})^{-1}\mathbf{P}^{-1}.$$

The resolvent may therefore be analyzed block by block in Jordan coordinates. Its poles are the eigenvalues of $\mathbf{A}$, while their orders are determined by the sizes of the Jordan blocks. Under inverse Laplace transformation, simple poles produce ordinary exponential modes and higher-order poles produce the polynomial factors occurring in generalized modes.

Because the Laplace transform is taken over $t \geq 0$, it is naturally adapted to initial-value problems, delayed inputs, discontinuous forcing, and impulses. It is therefore not merely a table of transform formulas. It provides an operator-theoretic framework connecting differential equations, Green's functions, convolution, resolvents, and the modal structure of linear systems.

6.1 Laplace Transform

The Laplace transform associates a function of time with a function of the complex variable s. Its principal use in this chapter is to replace differentiation in time by multiplication by s, with the initial data appearing as boundary terms. Constant-coefficient differential equations can then be treated algebraically in the s-domain.

We first define the transform and compute a small collection of basic examples. We then develop only the operational rules needed for differential equations. Convolution will be introduced later, when multiplication in the s-domain leads us to ask for the corresponding operation in the time domain.

6.1.1 Definition and basic examples

Let f be a function defined on $[0, \infty)$. Its *Laplace transform* is defined by

$$\mathcal{L}\{f\}(s) = F(s) := \int_0^\infty e^{-st} f(t)\, dt,$$

whenever the improper integral converges.

The variable s is complex. We write

$$s = \sigma + i\omega, \qquad \sigma = \Re(s).$$

Since

$$|e^{-st}| = e^{-\sigma t},$$

the convergence of the transform is controlled by the real part of s.

For the purposes of this chapter, we shall mainly consider functions that are piecewise continuous on every bounded interval and are of exponential order. A function f is said to be of *exponential order* α if there exist constants $C > 0$ and $T \geq 0$ such that

$$|f(t)| \leq Ce^{\alpha t}, \qquad t \geq T.$$

For such a function, the Laplace integral converges absolutely whenever $\Re(s) > \alpha$. The set of values of s for which the transform converges is called its *region of convergence*.

The constant function. For $\Re(s) > 0$,

$$\begin{aligned} \mathcal{L}\{1\}(s) &= \int_0^\infty e^{-st}\, dt \\ &= \frac{1}{s}. \end{aligned}$$

Thus

$$\boxed{\mathcal{L}\{1\}(s) = \frac{1}{s}.}$$

Exponential functions. Let $\lambda \in \mathbb{C}$. If $\Re(s) > \Re(\lambda)$, then

$$\begin{aligned} \mathcal{L}\{e^{\lambda t}\}(s) &= \int_0^\infty e^{-(s-\lambda)t}\, dt \\ &= \frac{1}{s-\lambda}. \end{aligned}$$

Therefore

$$\boxed{\mathcal{L}\{e^{\lambda t}\}(s) = \frac{1}{s-\lambda}.}$$

This gives the basic correspondence

$$e^{\lambda t} \quad \longleftrightarrow \quad \frac{1}{s-\lambda}.$$

An exponential mode in the time domain is represented by a pole in the s-domain.

Power functions. For $n = 0, 1, 2, \ldots$, set

$$I_n(s) = \int_0^\infty t^n e^{-st}\, dt.$$

For $\Re(s) > 0$, integration by parts gives

$$I_n(s) = \frac{n}{s} I_{n-1}(s), \qquad I_0(s) = \frac{1}{s}.$$

Hence

$$\boxed{\mathcal{L}\{t^n\}(s) = \frac{n!}{s^{n+1}}.}$$

In particular,

$$\mathcal{L}\{t\}(s) = \frac{1}{s^2}, \qquad \mathcal{L}\{t^2\}(s) = \frac{2}{s^3}.$$

More generally,

$$\begin{aligned}\mathcal{L}\{t^m e^{\lambda t}\}(s) &= \int_0^\infty t^m e^{-(s-\lambda)t}\, dt \\ &= \frac{m!}{(s-\lambda)^{m+1}},\end{aligned}$$

provided $\Re(s) > \Re(\lambda)$. Thus

$$\boxed{\mathcal{L}\{t^m e^{\lambda t}\}(s) = \frac{m!}{(s-\lambda)^{m+1}}.}$$

The generalized mode $t^m e^{\lambda t}$ corresponds to a pole of order $m+1$ at $s = \lambda$.

Sine and cosine. Using

$$\cos \omega t = \frac{e^{i\omega t} + e^{-i\omega t}}{2}, \qquad \sin \omega t = \frac{e^{i\omega t} - e^{-i\omega t}}{2i},$$

we obtain

$$\boxed{\mathcal{L}\{\cos \omega t\}(s) = \frac{s}{s^2 + \omega^2}, \qquad \mathcal{L}\{\sin \omega t\}(s) = \frac{\omega}{s^2 + \omega^2}.}$$

The poles at $s = \pm i\omega$ correspond to the oscillatory modes $e^{\pm i\omega t}$.

The basic transforms used most often in this chapter are therefore

time-domain function	Laplace transform
1	$\dfrac{1}{s}$
$e^{\lambda t}$	$\dfrac{1}{s-\lambda}$
t^n	$\dfrac{n!}{s^{n+1}}$
$t^n e^{\lambda t}$	$\dfrac{n!}{(s-\lambda)^{n+1}}$
$\cos \omega t$	$\dfrac{s}{s^2+\omega^2}$
$\sin \omega t$	$\dfrac{\omega}{s^2+\omega^2}$

These formulas form the beginning of a dictionary between time-domain behavior and algebraic structure in the s-domain. In particular, exponential modes correspond to poles, and polynomial factors in time correspond to higher-order poles.

Exercise 6.1 (Time scaling and transform poles)**.** Let $F(s) = \mathcal{L}\{f\}(s)$ and let $a > 0$. Prove directly from the definition that

$$\mathcal{L}\{f(at)\}(s) = \frac{1}{a} F\Big(\frac{s}{a}\Big).$$

Use this rule to obtain the transforms of $\cos \omega t$ and $\sin \omega t$ from the unit-frequency formulas, and explain how time scaling moves their poles.

6.1.2 Operational rules and operator algebra

We now develop the operational rules needed to transform differential equations. The central idea is that operations in the time domain become algebraic operations in the s-domain. Throughout this subsection, let

$$F(s) = \mathcal{L}\{f\}(s), \qquad G(s) = \mathcal{L}\{g\}(s),$$

whenever the transforms exist.

Linearity. The Laplace transform is linear:

$$\mathcal{L}\{af + bg\}(s) = aF(s) + bG(s).$$

This follows directly from the linearity of the integral. Thus addition and scalar multiplication are preserved by the transform:

$$af + bg \quad \longleftrightarrow \quad aF + bG.$$

Uniqueness. The Laplace transform is also one-to-one on the usual class of admissible functions.

Theorem 6.1 (Uniqueness of the Laplace transform)**.** *Let f and g be piecewise continuous on every bounded interval and of exponential order. If*

$$\mathcal{L}\{f\}(s) = \mathcal{L}\{g\}(s)$$

for all sufficiently large real s, then

$$f(t) = g(t)$$

at every point where both functions are continuous.

The proof of this uniqueness theorem requires an approximation argument and will not be given here. It allows us to define the inverse Laplace transform on the range of $\mathcal{L}$ and to write

$$f = \mathcal{L}^{-1}\{F\}.$$

Inverse Laplace transform. The inverse transform is also linear:

$$\mathcal{L}^{-1}\{aF + bG\} = a\mathcal{L}^{-1}\{F\} + b\mathcal{L}^{-1}\{G\}.$$

Differentiation and initial data. Let

$$D = \frac{d}{dt}.$$

Assume that f is continuous, piecewise continuously differentiable on every bounded interval, and that both f and f' are of exponential order. Then, for $\Re(s)$ sufficiently large, integration by parts gives

$$\begin{aligned}\mathcal{L}\{Df\}(s) &= \int_0^\infty e^{-st} f'(t)\, dt \\ &= \left[e^{-st} f(t)\right]_0^\infty + s\int_0^\infty e^{-st} f(t)\, dt.\end{aligned}$$

Since the boundary term at infinity vanishes for $\Re(s)$ sufficiently large, we obtain

$$\boxed{\mathcal{L}\{Df\}(s) = sF(s) - f(0).}$$

Thus differentiation becomes multiplication by s, together with a boundary term containing the initial value. If $f(0) = 0$, then

$$\mathcal{L}\{Df\}(s) = sF(s),$$

and hence

$$D \longleftrightarrow s.$$

More generally, suppose that $f, f', \ldots, f^{(k-1)}$ are continuous, that $f^{(k)}$ is piecewise continuous on every bounded interval, and that these functions are of exponential order. Applying the differentiation formula repeatedly gives

$$\mathcal{L}\{D^2 f\}(s) = s^2 F(s) - sf(0) - f'(0),$$

and, more generally,

$$\boxed{\mathcal{L}\{D^k f\}(s) = s^k F(s) - \sum_{j=0}^{k-1} s^{k-1-j} f^{(j)}(0).}$$

If

$$f(0) = f'(0) = \cdots = f^{(k-1)}(0) = 0,$$

then

$$\mathcal{L}\{D^k f\}(s) = s^k F(s).$$

This is the basic reason that the Laplace transform is adapted to constant-coefficient differential equations. Powers of the differential operator become powers of the algebraic variable s, while the initial data appear explicitly as boundary terms.

Causal integration. Define the causal integration operator $\mathcal{J}$ by

$$\mathcal{J}(f)(t) = \int_0^t f(\tau)\, d\tau.$$

Then

$$(\mathcal{J}f)(0) = 0, \qquad D(\mathcal{J}f) = f.$$

Applying the differentiation rule,

$$s\mathcal{L}\{\mathcal{J}f\}(s) = F(s),$$

and therefore

$$\boxed{\mathcal{L}\{\mathcal{J}f\}(s) = \frac{F(s)}{s}.}$$

Thus

$$\mathcal{J} \longleftrightarrow \frac{1}{s}.$$

In the causal setting, $\mathcal{J}$ is the inverse of differentiation with zero initial value:

$$D\mathcal{J}f = f, \qquad \mathcal{J}Df = f - f(0).$$

The transform-domain relations

$$D \longleftrightarrow s, \qquad \mathcal{J} \longleftrightarrow \frac{1}{s}$$

are therefore inverse to one another when the appropriate initial value vanishes.

Multiplication by an exponential. Multiplication by an exponential in the time domain translates the transform variable. Indeed,

$$\begin{aligned}\mathcal{L}\{e^{\lambda t}f(t)\}(s) &= \int_0^\infty e^{-st}e^{\lambda t}f(t)\,dt \\ &= \int_0^\infty e^{-(s-\lambda)t}f(t)\,dt \\ &= F(s-\lambda).\end{aligned}$$

Hence

$$\boxed{\mathcal{L}\{e^{\lambda t}f(t)\}(s) = F(s-\lambda).}$$

Equivalently,

$$\mathcal{L}^{-1}\{F(s-\lambda)\}(t) = e^{\lambda t}f(t).$$

This rule explains how exponential factors shift poles. If $F(s)$ has a pole at $s = \mu$, then $F(s-\lambda)$ has a pole at

$$s = \mu + \lambda.$$

Thus multiplication by $e^{\lambda t}$ shifts every exponential mode by λ.

A secondary rule: multiplication by time. Differentiating the transform with respect to s gives

$$\begin{aligned}F'(s) &= \frac{d}{ds}\int_0^\infty e^{-st}f(t)\,dt \\ &= -\int_0^\infty te^{-st}f(t)\,dt.\end{aligned}$$

Therefore

$$\mathcal{L}\{tf(t)\}(s) = -F'(s).$$

More generally,

$$\boxed{\mathcal{L}\{t^m f(t)\}(s) = (-1)^m F^{(m)}(s).}$$

This identity is useful for generating additional transform pairs, but it plays a secondary role in the operator method developed below.

The basic operator dictionary. The rules needed most often in this chapter are

time-domain operation	s-domain operation
$af + bg$	$aF + bG$
Df	$sF - f(0)$
$D^k f$	$s^k F - \sum_{j=0}^{k-1} s^{k-1-j} f^{(j)}(0)$
$Jf = \int_0^t f(\tau)\, d\tau$	$\dfrac{F}{s}$
$e^{\lambda t} f(t)$	$F(s-\lambda)$
$t^m f(t)$	$(-1)^m F^{(m)}(s)$

The essential correspondences are

$$D \longleftrightarrow s, \qquad J \longleftrightarrow \frac{1}{s},$$

with the understanding that J is the causal inverse of D under the zero-initial-value condition.

Multiplication of two transforms will lead to a different time-domain operation. Rather than introduce that operation formally here, we shall encounter it later when solving forced equations. It will lead naturally to convolution and Green functions.

Exercise 6.2 (A mixed operational identity)**.** Let $F(s) = \mathcal{L}\{f\}(s)$, where f is sufficiently smooth and of exponential order. Show that

$$\mathcal{L}\{tf'(t)\}(s) = -sF'(s) - F(s).$$

Interpret the formula using the commutator of multiplication by t and differentiation.

Exercise 6.3 (Iterated causal integration)**.** Let

$$(\mathcal{J}f)(t) = \int_0^t f(\tau)\, d\tau.$$

Show that $(\mathcal{J}D - D\mathcal{J})f = -f(0)\mathbf{1}$. Then prove, for $m \geq 1$, that

$$\mathcal{J}^m f(t) = \frac{1}{(m-1)!} \int_0^t (t-\tau)^{m-1} f(\tau)\, d\tau,$$

and use the Laplace transform to verify that

$$\mathcal{L}\{\mathcal{J}^m f\}(s) = \frac{F(s)}{s^m}.$$

Exercise 6.4 (Time factors and pole structure)**.** Let $F(s) = \mathcal{L}\{f\}(s)$. For $m \geq 0$, prove

$$\mathcal{L}\{t^m e^{\lambda t} f(t)\}(s) = (-1)^m \frac{d^m}{ds^m} F(s - \lambda).$$

Apply the formula to $t^m e^{\lambda t} \cos \omega t$ and $t^m e^{\lambda t} \sin \omega t$. Locate the poles, determine their orders, and relate them to the exponential, oscillatory, and polynomial factors.

6.2 Scalar Equations and Polynomial Differential Operators

We now apply the operational rules of the Laplace transform to scalar constant-coefficient differential equations. This is the transform-domain counterpart of the characteristic-polynomial method developed in Chapter 3. There, one substitutes an exponential mode $y = e^{\lambda t}$ into the homogeneous equation

$$P(D)y = 0$$

and obtains the characteristic equation

$$P(\lambda) = 0.$$

The roots of P determine the elementary time modes.

The Laplace transform acts instead on the entire initial value problem. The differential operator $P(D)$ becomes the polynomial $P(s)$, while the initial data produce an additional polynomial $I(s)$. Thus

$$P(D)y = f$$

is transformed into

$$P(s)Y(s) = I(s) + F(s).$$

In this section we concentrate on cases for which the transformed solution $Y(s)$ is a rational function of s. The initial-data term $I(s)$ is always a polynomial, while many standard forcing terms have rational Laplace transforms. In such cases, partial fractions reduce the inversion problem to the basic transform pairs established in the preceding section. The poles of $Y(s)$ then reveal the exponential, oscillatory, and generalized modes of the solution.

6.2.1 From $P(D)y = f$ to $P(s)Y = I(s) + F(s)$

Consider the scalar constant-coefficient equation

$$P(D)y = f(t),$$

where

$$D = \frac{d}{dt}$$

and

$$P(D) = a_n D^n + a_{n-1} D^{n-1} + \cdots + a_1 D + a_0, \qquad a_n \neq 0.$$

The corresponding polynomial in the transform variable is

$$P(s) = a_n s^n + a_{n-1} s^{n-1} + \cdots + a_1 s + a_0.$$

Let

$$Y(s) = \mathcal{L}\{y\}(s), \qquad F(s) = \mathcal{L}\{f\}(s).$$

Using

$$\mathcal{L}\{D^k y\}(s) = s^k Y(s) - \sum_{j=0}^{k-1} s^{k-1-j} y^{(j)}(0),$$

we obtain

$$\begin{aligned} \mathcal{L}\{P(D)y\}(s) &= \sum_{k=0}^{n} a_k \mathcal{L}\{D^k y\}(s) \\ &= P(s)Y(s) - I(s), \end{aligned}$$

where

$$\boxed{I(s) = \sum_{k=1}^{n} a_k \sum_{j=0}^{k-1} s^{k-1-j} y^{(j)}(0).}$$

The polynomial $I(s)$ is determined entirely by the initial data

$$y(0),\ \dot{y}(0),\ \ldots,\ y^{(n-1)}(0).$$

Therefore the differential equation

$$P(D)y = f$$

is transformed into the algebraic equation

$$\boxed{P(s)Y(s) = I(s) + F(s).}$$

Solving for $Y(s)$ gives

$$\boxed{Y(s) = \frac{I(s)}{P(s)} + \frac{F(s)}{P(s)}.}$$

The first term is generated by the initial data, while the second is generated by the forcing. Thus the Laplace transform separates the free and forced responses at the algebraic level.

If the initial data vanish, then $I(s) = 0$, and

$$Y(s) = \frac{F(s)}{P(s)}.$$

In this case, the inverse operator $P(D)^{-1}$ is represented in the transform domain by multiplication by $1/P(s)$. Its time-domain interpretation through Green's functions and convolution will be developed in the next section.

6.2.2 Partial fractions, poles, and scalar modes

We now consider cases in which the transformed solution

$$Y(s) = \frac{I(s) + F(s)}{P(s)}$$

is a rational function of s. This occurs automatically for the initial-data contribution $I(s)/P(s)$, and also for many standard forcing terms whose Laplace transforms are rational. When $Y(s)$ is a proper rational function,* it can be inverted by partial fractions. If $Y(s)$ is improper, polynomial division also produces a polynomial part whose inverse involves impulses and their derivatives at $t = 0$. Such terms will be considered in Section 6.3.3 after impulsive forcing has been introduced. The factorization of the denominator determines the possible time-domain modes, while the numerator determines their coefficients.

Simple real poles. Suppose

$$P(s) = a_n(s - \lambda_1) \cdots (s - \lambda_n),$$

where the roots $\lambda_1, \ldots, \lambda_n$ are distinct and real, and let $\deg Q < \deg P$. Then

$$\frac{Q(s)}{P(s)} = \sum_{j=1}^{n} \frac{c_j}{s - \lambda_j}.$$

For a simple pole, the coefficient is

$$c_j = \lim_{s \to \lambda_j} (s - \lambda_j) \frac{Q(s)}{P(s)} = \frac{Q(\lambda_j)}{P'(\lambda_j)}.$$

Taking inverse Laplace transforms gives

$$\mathcal{L}^{-1} \left\{ \frac{Q(s)}{P(s)} \right\} (t) = \sum_{j=1}^{n} c_j e^{\lambda_j t}.$$

Thus a simple pole at $s = \lambda$ produces the exponential mode

$$e^{\lambda t}.$$

Repeated poles. Suppose that $Y(s)$ has a pole of order m at $s = \lambda$. Its partial-fraction expansion then contains terms of the form

$$\frac{c_1}{s - \lambda} + \frac{c_2}{(s - \lambda)^2} + \cdots + \frac{c_m}{(s - \lambda)^m}.$$

Since

$$\mathcal{L}^{-1} \left\{ \frac{1}{(s - \lambda)^k} \right\} (t) = \frac{t^{k-1}}{(k-1)!} e^{\lambda t},$$

the corresponding time-domain contribution is

$$e^{\lambda t} \left(c_1 + c_2 t + \cdots + c_m \frac{t^{m-1}}{(m-1)!} \right).$$

Thus a pole of order m at $s = \lambda$ produces the generalized modes

$$e^{\lambda t}, \quad t e^{\lambda t}, \quad \ldots, \quad t^{m-1} e^{\lambda t}.$$

*A rational function $Q(s)/P(s)$ is called *proper* if $\deg Q < \deg P$.

Example 6.2 (A repeated pole)**.** Find the inverse Laplace transform of

$$Y(s) = \frac{2s+1}{(s+1)^2}.$$

Since

$$2s + 1 = 2(s+1) - 1,$$

we have

$$Y(s) = \frac{2}{s+1} - \frac{1}{(s+1)^2}.$$

Therefore

$$y(t) = 2e^{-t} - te^{-t}.$$

The factor t is produced by the second-order pole at $s = -1$.

Complex conjugate poles. For equations with real coefficients, nonreal poles occur in conjugate pairs. Let

$$\lambda = \alpha + i\beta, \qquad \overline{\lambda} = \alpha - i\beta, \qquad \beta \neq 0.$$

The corresponding real quadratic factor is

$$(s-\lambda)(s-\overline{\lambda}) = (s-\alpha)^2 + \beta^2.$$

For a real numerator $As + B$, where $A, B \in \mathbb{R}$, write

$$As + B = A(s-\alpha) + (A\alpha + B).$$

It follows that

$$\mathcal{L}^{-1}\left\{\frac{As+B}{(s-\alpha)^2+\beta^2}\right\}(t) = e^{\alpha t}\left(A\cos\beta t + \frac{A\alpha+B}{\beta}\sin\beta t\right).$$

Thus the conjugate poles

$$\alpha \pm i\beta$$

produce the real modes

$$e^{\alpha t}\cos\beta t, \qquad e^{\alpha t}\sin\beta t.$$

The real part α determines exponential growth or decay, while $|\beta|$ determines the angular frequency.

Pole–mode correspondence. The preceding cases may be summarized as follows:

pole structure in the s-domain	corresponding time-domain mode
$\dfrac{1}{s-\lambda}$	$e^{\lambda t}$
$\dfrac{1}{(s-\lambda)^m}$	$\dfrac{t^{m-1}}{(m-1)!}e^{\lambda t}$
$\dfrac{s-\alpha}{(s-\alpha)^2+\beta^2}$	$e^{\alpha t}\cos\beta t$
$\dfrac{\beta}{(s-\alpha)^2+\beta^2}$	$e^{\alpha t}\sin\beta t$

The locations of the poles determine growth, decay, and oscillation, while their orders determine the polynomial factors multiplying the exponential modes.

6.2.3 Examples from second-order equations

We now revisit several second-order equations from the transform-domain point of view. The purpose is not to repeat the methods of Chapter 3, but to show how the initial data and forcing determine the numerator of $Y(s)$, while the denominator determines the possible modes.

Example 6.3 (The harmonic oscillator)**.** Let $\omega > 0$ and consider

$$\ddot{y} + \omega^2 y = 0, \qquad y(0) = y_0, \qquad \dot{y}(0) = v_0.$$

Taking the Laplace transform gives

$$s^2 Y(s) - s y_0 - v_0 + \omega^2 Y(s) = 0.$$

Hence

$$(s^2 + \omega^2) Y(s) = s y_0 + v_0,$$

and therefore

$$Y(s) = y_0 \frac{s}{s^2 + \omega^2} + \frac{v_0}{\omega} \frac{\omega}{s^2 + \omega^2}.$$

Taking inverse transforms yields

$$\boxed{y(t) = y_0 \cos \omega t + \frac{v_0}{\omega} \sin \omega t.}$$

The denominator has poles at

$$s = \pm i\omega,$$

which determine the oscillatory modes. The initial-data polynomial

$$I(s) = s y_0 + v_0$$

determines their coefficients.

Example 6.4 (Resonant forcing)**.** Let $\omega > 0$ and consider

$$\ddot{y} + \omega^2 y = \cos \omega t, \qquad y(0) = 0, \qquad \dot{y}(0) = 0.$$

Since

$$\mathcal{L}\{\cos \omega t\}(s) = \frac{s}{s^2 + \omega^2},$$

the transformed equation is

$$(s^2 + \omega^2) Y(s) = \frac{s}{s^2 + \omega^2}.$$

Hence

$$Y(s) = \frac{s}{(s^2+\omega^2)^2}.$$

The forcing poles coincide with the poles of the inverse differential operator, producing second-order poles at

$$s = \pm i\omega.$$

Using

$$\mathcal{L}\{t\sin\omega t\}(s) = \frac{2\omega s}{(s^2+\omega^2)^2},$$

we obtain

$$\boxed{y(t) = \frac{t}{2\omega}\sin\omega t.}$$

The factor t is the transform-domain signature of resonance: the coincidence of the forcing poles with the natural poles increases their order and produces a secular term.

The examples in this section have rational transforms and can therefore be handled by algebraic manipulation and partial fractions. For a general forcing term, however, $F(s)$ need not be rational. The next section isolates the inverse operator represented by $1/P(s)$ and describes its action through Green's functions and convolution.

Exercise 6.5 (Free and forced response)**.** Use the Laplace transform to solve

$$y'' - 3y' + 2y = e^t, \qquad y(0) = y_0, \qquad y'(0) = y_1.$$

Separate the free and forced responses and determine the long-time behavior.

Exercise 6.6 (A repeated scalar mode)**.** Let

$$P(D)y = 0, \qquad P(s) = (s-\lambda)^2(s-\mu), \qquad \lambda \neq \mu.$$

Assume

$$y(0) = y_0, \qquad \dot{y}(0) = y_1, \qquad \ddot{y}(0) = y_2.$$

Use the Laplace transform to find the initial-data polynomial $I(s)$ and write

$$Y(s) = \frac{I(s)}{(s-\lambda)^2(s-\mu)}.$$

Explain from the pole structure why

$$y(t) = (c_0 + c_1 t)e^{\lambda t} + c_2 e^{\mu t}.$$

and determine the coefficients from the initial data.

Exercise 6.7 (Coalescing poles and resonance)**.** For $\omega > 0$ and $\nu \geq 0$, $\nu \neq \omega$, use the Laplace transform to solve

$$\ddot{y} + \omega^2 y = \cos \nu t, \qquad y(0) = 0, \qquad \dot{y}(0) = 0.$$

Take the limit $\nu \to \omega$ and recover

$$y(t) = \frac{t}{2\omega} \sin \omega t.$$

Explain how coalescing simple poles produce the second-order poles of the resonant problem.

6.3 Green's Functions, Convolution, and Impulses

We now turn to equations with a general forcing term. For zero initial data,

$$P(D)y = f$$

is transformed into

$$P(s)Y(s) = F(s),$$

and hence

$$Y(s) = \frac{1}{P(s)} F(s).$$

When $F(s)$ is rational, this expression may sometimes be inverted by partial fractions. For a general forcing term, however, such an algebraic inversion is not available.

The factor $1/P(s)$ depends only on the differential operator, whereas $F(s)$ depends on the forcing. We therefore first study the inverse transform of $1/P(s)$. This leads to the Green's function of the operator. Convolution then describes how this kernel acts on an arbitrary forcing term. Finally, the Heaviside function and the delta distribution allow us to treat delayed, discontinuous, and impulsive inputs.

6.3.1 Green's functions as inverse operators

Consider the constant-coefficient differential operator

$$P(D) = a_n D^n + a_{n-1} D^{n-1} + \cdots + a_1 D + a_0, \qquad a_n \neq 0,$$

and the initial value problem

$$P(D)y = f(t),$$

with zero initial data

$$y(0) = \dot{y}(0) = \cdots = y^{(n-1)}(0) = 0.$$

Taking the Laplace transform gives

$$P(s)Y(s) = F(s),$$

so that

$$Y(s) = \frac{1}{P(s)} F(s).$$

The multiplier $1/P(s)$ represents the causal inverse of $P(D)$ in the transform domain.

Definition 6.5 (Green's function)**.** The causal Green's function associated with $P(D)$ is

$$G(t) = \mathcal{L}^{-1}\left\{\frac{1}{P(s)}\right\}(t).$$

Thus

$$\mathcal{L}\{G\}(s) = \frac{1}{P(s)}.$$

The function G depends only on the operator $P(D)$, not on the particular forcing term. Once G is known, the response to a general forcing will be obtained by applying the inverse operator represented by this kernel.

Formally,

$$P(s)\mathcal{L}\{G\}(s) = 1.$$

Since the inverse Laplace transform of 1 is the delta distribution, this corresponds to

$$P(D)G = \delta.$$

This identity will be justified after the delta distribution is introduced later in this section. It shows that the Green's function is the response of the operator to a unit impulse.

For example,

$P(D)$	$G(t)$
$D+\alpha$	$e^{-\alpha t}$
$D^2+\omega^2$	$\dfrac{1}{\omega}\sin\omega t$

Example 6.6 (A fourth-order Green's function)**.** Consider the operator

$$P(D) = (D^2+\omega_1^2)(D^2+\omega_2^2), \qquad 0 < \omega_1 < \omega_2.$$

Its transform-domain inverse is

$$\frac{1}{P(s)} = \frac{1}{(s^2+\omega_1^2)(s^2+\omega_2^2)}.$$

Using partial fractions,

$$\frac{1}{(s^2+\omega_1^2)(s^2+\omega_2^2)} = \frac{1}{\omega_2^2-\omega_1^2}\left(\frac{1}{s^2+\omega_1^2} - \frac{1}{s^2+\omega_2^2}\right).$$

Therefore the causal Green's function is

$$\boxed{G(t) = \frac{1}{\omega_2^2-\omega_1^2}\left(\frac{\sin\omega_1 t}{\omega_1} - \frac{\sin\omega_2 t}{\omega_2}\right).}$$

The four poles

$$s = \pm i\omega_1, \qquad s = \pm i\omega_2$$

produce two oscillatory components. Thus the Green's function displays the two natural frequencies of the fourth-order operator.

Although each individual term is of order t near $t = 0$, their leading terms cancel. Expanding at $t = 0$ gives

$$G(t) = \frac{t^3}{6} + O(t^5).$$

Consequently,

$$G(0) = \dot{G}(0) = \ddot{G}(0) = 0, \qquad G^{(3)}(0+) = 1.$$

These are precisely the causal initial conditions associated with the impulse equation

$$P(D)G = \delta.$$

The pole structure of $1/P(s)$ determines the qualitative behavior of G. Poles in the left half-plane produce decaying modes, poles on the imaginary axis produce undamped oscillations, and poles in the right half-plane produce exponentially growing modes. Higher-order poles produce additional polynomial factors in time.

Exercise 6.8 (Four causal Green's functions)**.** For $\alpha > 0$, $\omega > 0$, and $\lambda \in \mathbb{C}$, find the causal Green's functions of

$$D + \alpha, \qquad D^2 - \alpha^2, \qquad D^2 + \omega^2, \qquad (D - \lambda)^2.$$

6.3.2 Convolution and Duhamel's principle

Let f_1 and f_2 be piecewise continuous and of exponential order in $[0, \infty)$, and write

$$F_1(s) = \mathcal{L}\{f_1\}(s), \qquad F_2(s) = \mathcal{L}\{f_2\}(s).$$

Their causal convolution is defined by

$$(f_1 * f_2)(t) = \int_0^t f_1(t - \tau) f_2(\tau)\, d\tau.$$

Equivalently,

$$(f_1 * f_2)(t) = \int_0^t f_1(\tau) f_2(t - \tau)\, d\tau.$$

Thus

$$f_1 * f_2 = f_2 * f_1.$$

The convolution theorem states that

$$\mathcal{L}\{f_1 * f_2\}(s) = F_1(s) F_2(s).$$

Now consider

$$P(D)y = f$$

with zero initial data. If G is the causal Green's function of $P(D)$, then

$$\mathcal{L}\{G\}(s) = \frac{1}{P(s)}.$$

Since

$$Y(s) = \frac{1}{P(s)} F(s),$$

the convolution theorem gives

$$Y(s) = \mathcal{L}\{G * f\}(s).$$

By uniqueness of the Laplace transform,

$$\boxed{y(t) = (G * f)(t) = \int_0^t G(t-\tau) f(\tau)\, d\tau.}$$

This formula is Duhamel's principle. It expresses the response to a general forcing as a superposition of responses generated at earlier times. An input applied at time τ is propagated from τ to t by the kernel $G(t-\tau)$, and the total response is obtained by integrating over all $0 \le \tau \le t$.

The convolution formula also shows directly that the Green's function is the kernel of the causal inverse:

$$P(D)^{-1} f = G * f.$$

Example 6.7 (Response to ground acceleration)**.** Consider a mass attached to a moving support by a spring and a damper. Let $u(t)$ denote the displacement of the support and let $x(t)$ denote the absolute displacement of the mass. The spring and damping forces depend on the relative displacement

$$y(t) = x(t) - u(t).$$

Newton's law gives

$$m\ddot{x} + \gamma(\dot{x} - \dot{u}) + k(x - u) = 0,$$

where $\gamma > 0$ is the damping coefficient. Since $x = y + u$, the relative displacement satisfies

$$m\ddot{y} + \gamma\dot{y} + ky = -m\ddot{u}(t).$$

Dividing by m and setting

$$2\alpha = \frac{\gamma}{m}, \qquad \omega_0^2 = \frac{k}{m},$$

we obtain

$$\ddot{y} + 2\alpha\dot{y} + \omega_0^2 y = -\ddot{u}(t).$$

Assume that the structure is initially at rest relative to the support:

$$y(0) = 0, \qquad \dot{y}(0) = 0.$$

In the underdamped case, let

$$\omega_d = \sqrt{\omega_0^2 - \alpha^2}.$$

The Green's function is

$$G(t) = \frac{1}{\omega_d} e^{-\alpha t} \sin \omega_d t.$$

Duhamel's principle therefore gives

$$\boxed{y(t) = -\int_0^t \frac{1}{\omega_d} e^{-\alpha(t-\tau)} \sin\big(\omega_d(t-\tau)\big)\ddot{u}(\tau)\,d\tau.}$$

Thus the relative displacement is determined by the entire history of the ground acceleration. An acceleration applied at time τ excites a damped free oscillation, and its contribution at time t is weighted by $G(t-\tau)$.

This formula is fundamental in vibration theory and earthquake engineering. The forcing $\ddot{u}(t)$ may be an experimentally measured acceleration record and need not have a simple Laplace transform.

For nonzero initial data, the solution is the sum of the free response and the forced response:

$$y(t) = y_{\text{free}}(t) + (G * f)(t),$$

where y_{free} is determined by the initial-data term $I(s)/P(s)$.

Exercise 6.9 (Delayed Duhamel principle)**.** Let G be the causal Green's function of $P(D)$ and assume zero initial data. If

$$P(D)y = H(t-a)q(t-a),$$

show that

$$y(t) = H(t-a)\int_0^{t-a} G(t-a-\tau)q(\tau)\,d\tau.$$

6.3.3 Discontinuous and impulsive forcing

The Laplace transform is especially useful for forcing terms that are switched on at a prescribed time or concentrated at a single instant. Such inputs are described by the Heaviside function and the delta distribution.

For $a \geq 0$, define

$$H(t-a) = \begin{cases} 0, & 0 \leq t < a, \\ 1, & t \geq a. \end{cases}$$

A delayed input with profile q may be written as

$$f(t) = H(t-a)q(t-a).$$

If

$$Q(s) = \mathcal{L}\{q\}(s),$$

then the translation rule gives

$$\boxed{\mathcal{L}\{H(t-a)q(t-a)\}(s) = e^{-as}Q(s).}$$

Thus a delay in time is represented by multiplication by e^{-as} in the transform domain.

Example 6.8 (A delayed step input)**.** Consider

$$\dot{y} + \alpha y = H(t-a), \qquad y(0) = 0, \qquad \alpha > 0.$$

Taking the Laplace transform gives

$$(s+\alpha)Y(s) = \frac{e^{-as}}{s},$$

and hence

$$Y(s) = e^{-as}\frac{1}{s(s+\alpha)}.$$

Since

$$\frac{1}{s(s+\alpha)} = \frac{1}{\alpha}\left(\frac{1}{s} - \frac{1}{s+\alpha}\right),$$

we obtain

$$\boxed{y(t) = \frac{1}{\alpha}H(t-a)\left(1 - e^{-\alpha(t-a)}\right).}$$

The solution vanishes before the input is switched on and then evolves as the response to a constant forcing, with the time origin shifted to $t = a$.

We next introduce the delta distribution. It is characterized by the sampling property

$$\int_{-\infty}^{\infty} \delta(t-a)\varphi(t)\,dt = \varphi(a)$$

for every smooth test function φ. In the causal setting, for $a > 0$,

$$\int_{0}^{\infty} \delta(t-a)\varphi(t)\,dt = \varphi(a).$$

The expression $\delta(t-a)$ represents a unit impulse applied at time a. Its Laplace transform is

$$\boxed{\mathcal{L}\{\delta(t-a)\}(s) = e^{-as}.}$$

In particular,

$$\mathcal{L}\{\delta(t)\}(s) = 1.$$

Moreover,

$$\frac{d}{dt}H(t-a) = \delta(t-a)$$

in the sense of distributions.

Consider now

$$P(D)y = \delta(t-a),$$

with zero initial data before the impulse. Taking the Laplace transform gives

$$P(s)Y(s) = e^{-as},$$

and hence

$$Y(s) = e^{-as}\frac{1}{P(s)}.$$

If

$$G(t) = \mathcal{L}^{-1}\left\{\frac{1}{P(s)}\right\}(t),$$

then

$$\boxed{y(t) = H(t-a)G(t-a).}$$

Thus the response to an impulse applied at time a is the delayed Green's function. For an impulse applied at the initial time,

$$P(D)y = \delta(t),$$

the response is precisely

$$y(t) = G(t).$$

This justifies the interpretation of the Green's function as the impulse response of the operator.

Example 6.9 (Impulse applied to an oscillator)**.** Let $\omega > 0$ and consider

$$\ddot{y} + \omega^2 y = \delta(t-a),$$

with zero initial data before $t = a$. Since

$$G(t) = \frac{1}{\omega}\sin\omega t,$$

the solution is

$$\boxed{y(t) = H(t-a)\frac{1}{\omega}\sin\big(\omega(t-a)\big).}$$

The impulse initiates a free oscillation at time a.

The explicit solution above is continuous at $t = a$. The jump in its derivative can also be obtained directly from the differential equation. Integrating

$$\ddot{y} + \omega^2 y = \delta(t-a)$$

from $a-\varepsilon$ to $a+\varepsilon$ gives

$$\dot{y}(a+\varepsilon) - \dot{y}(a-\varepsilon) + \omega^2\int_{a-\varepsilon}^{a+\varepsilon} y(t)\,dt = 1.$$

If y remains bounded, then letting $\varepsilon \to 0$ yields

$$\boxed{\dot{y}(a+) - \dot{y}(a-) = 1.}$$

Thus the velocity has a unit jump, while the explicit formula confirms that the displacement remains continuous.

More generally, if

$$a_n y^{(n)} + \cdots + a_0 y = \delta(t-a),$$

and the lower derivatives remain bounded, then

$$a_n \left[y^{(n-1)}(a+) - y^{(n-1)}(a-) \right] = 1.$$

Thus an impulse produces a jump in the derivative immediately below the highest derivative appearing in the equation.

The main correspondences of this section are

time domain	Laplace domain
$G(t)$	$\dfrac{1}{P(s)}$
$(G * f)(t)$	$\dfrac{1}{P(s)}F(s)$
$\delta(t-a)$	e^{-as}
$H(t-a)G(t-a)$	$e^{-as}\dfrac{1}{P(s)}$

Green's functions, convolution, and impulse response are therefore three forms of the same causal inverse-operator structure.

Exercises

Exercise 6.10 (Two impulses)**.** Solve

$$y'' + \omega^2 y = \delta(t-a) - \delta(t-b), \qquad 0 < a < b,$$

with zero initial data before $t = a$. Determine the jumps of y' and find all values of $b - a$ for which the second impulse cancels the motion completely.

Exercise 6.11 (Response to a rectangular pulse)**.** Let

$$P(D) = D^2 + 2\gamma D + \omega_0^2, \qquad 0 < \gamma < \omega_0,$$

and let G be its causal Green's function. For $f(t) = H(t-a) - H(t-b)$, $0 < a < b$, show that

$$y(t) = H(t-a)S(t-a) - H(t-b)S(t-b),$$

where

$$S(t) = \int_0^t G(\tau)\, d\tau.$$

Describe the motion before, during, and after the pulse.

6.4 Linear Systems, Resolvents, and Modes

We now extend the Laplace transform method from scalar equations to linear systems. Consider

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y} + \mathbf{f}(t), \qquad \mathbf{y}(0) = \mathbf{y}_0,$$

where $\mathbf{A}$ is a constant matrix. The scalar polynomial $P(s)$ is replaced by the matrix

$$s\mathbf{I} - \mathbf{A},$$

and its reciprocal is replaced by the resolvent

$$(s\mathbf{I} - \mathbf{A})^{-1}.$$

Thus the correspondence

$$P(D)^{-1} \longleftrightarrow \frac{1}{P(s)}$$

for scalar equations becomes

$$(D\mathbf{I} - \mathbf{A})^{-1} \longleftrightarrow (s\mathbf{I} - \mathbf{A})^{-1}$$

for systems. The poles of the resolvent are the eigenvalues of $\mathbf{A}$, and their orders reflect the generalized eigenvector structure studied in Chapter 5.

6.4.1 Laplace transform of first-order systems

Let

$$\mathbf{Y}(s) = \mathcal{L}\{\mathbf{y}\}(s), \qquad \mathbf{F}(s) = \mathcal{L}\{\mathbf{f}\}(s),$$

where the Laplace transform is applied componentwise. Taking the transform of

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y} + \mathbf{f}(t)$$

gives

$$s\mathbf{Y}(s) - \mathbf{y}_0 = \mathbf{A}\mathbf{Y}(s) + \mathbf{F}(s).$$

Therefore

$$\boxed{(s\mathbf{I} - \mathbf{A})\mathbf{Y}(s) = \mathbf{y}_0 + \mathbf{F}(s).}$$

Whenever $s\mathbf{I} - \mathbf{A}$ is invertible,

$$\boxed{\mathbf{Y}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{y}_0 + (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{F}(s).}$$

The first term is generated by the initial data, while the second is generated by the forcing. Thus the transformed solution again separates into

$$\text{free response} \;+\; \text{forced response}.$$

The matrix-valued function

$$(s\mathbf{I} - \mathbf{A})^{-1}$$

is called the resolvent of $\mathbf{A}$.

For the homogeneous system

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}, \qquad \mathbf{y}(0) = \mathbf{y}_0,$$

we have

$$\mathbf{Y}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{y}_0.$$

On the other hand, Chapter 5 gives

$$\mathbf{y}(t) = e^{t\mathbf{A}}\mathbf{y}_0.$$

Since this identity holds for every initial vector $\mathbf{y}_0$, we obtain the matrix transform formula

$$\boxed{\mathcal{L}\{e^{t\mathbf{A}}\}(s) = (s\mathbf{I} - \mathbf{A})^{-1}.}$$

Equivalently,

$$\boxed{e^{t\mathbf{A}} = \mathcal{L}^{-1}\left\{(s\mathbf{I} - \mathbf{A})^{-1}\right\}(t).}$$

Thus the matrix exponential is the inverse Laplace transform of the resolvent.

The forced system. For the full equation,

$$\mathbf{Y}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{y}_0 + (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{F}(s).$$

Using

$$\mathcal{L}^{-1}\left\{(s\mathbf{I} - \mathbf{A})^{-1}\right\} = e^{t\mathbf{A}},$$

and applying the convolution theorem componentwise, we obtain

$$\boxed{\mathbf{y}(t) = e^{t\mathbf{A}}\mathbf{y}_0 + \int_0^t e^{(t-\tau)\mathbf{A}}\mathbf{f}(\tau)\,d\tau.}$$

This is the variation-of-constants formula for a constant-coefficient linear system.

The formula separates the solution into

$$\underbrace{e^{t\mathbf{A}}\mathbf{y}_0}_{\text{free response}} + \underbrace{\int_0^t e^{(t-\tau)\mathbf{A}}\mathbf{f}(\tau)\,d\tau}_{\text{forced response}}.$$

The matrix $e^{(t-\tau)\mathbf{A}}$ propagates the input applied at time τ to the observation time t.

Example 6.10 (A diagonal system)**.** Let

$$\mathbf{A} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}.$$

Then

$$s\mathbf{I} - \mathbf{A} = \begin{bmatrix} s - \lambda_1 & 0 \\ 0 & s - \lambda_2 \end{bmatrix},$$

and hence

$$(s\mathbf{I} - \mathbf{A})^{-1} = \begin{bmatrix} \dfrac{1}{s - \lambda_1} & 0 \\ 0 & \dfrac{1}{s - \lambda_2} \end{bmatrix}.$$

Taking the inverse Laplace transform entrywise gives

$$e^{t\mathbf{A}} = \begin{bmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{bmatrix}.$$

Thus each pole of the resolvent produces the corresponding exponential mode.

Example 6.11 (A Jordan block)**.** Let

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

Then

$$s\mathbf{I} - \mathbf{A} = \begin{bmatrix} s-\lambda & -1 \\ 0 & s-\lambda \end{bmatrix},$$

and

$$(s\mathbf{I} - \mathbf{A})^{-1} = \begin{bmatrix} \dfrac{1}{s-\lambda} & \dfrac{1}{(s-\lambda)^2} \\ 0 & \dfrac{1}{s-\lambda} \end{bmatrix}.$$

Therefore

$$e^{t\mathbf{A}} = \begin{bmatrix} e^{\lambda t} & te^{\lambda t} \\ 0 & e^{\lambda t} \end{bmatrix}.$$

The second-order pole in the upper-right entry produces the generalized mode

$$te^{\lambda t}.$$

This is the resolvent counterpart of the Jordan-block structure.

Exercise 6.12 (The planar rotation resolvent)**.** For

$$\mathbf{A} = \begin{bmatrix} 0 & -\omega \\ \omega & 0 \end{bmatrix},$$

compute $(s\mathbf{I} - \mathbf{A})^{-1}$ and recover

$$e^{t\mathbf{A}} = \begin{bmatrix} \cos\omega t & -\sin\omega t \\ \sin\omega t & \cos\omega t \end{bmatrix}$$

by inverse Laplace transformation.

6.4.2 Properties of the resolvent

The identity

$$\mathcal{L}\{e^{t\mathbf{A}}\}(s) = (s\mathbf{I} - \mathbf{A})^{-1}$$

shows that the resolvent is the transform-domain representation of the evolution operator. We now examine this relation more closely.

Resolvent expansion for large s. For $|s|$ sufficiently large,

$$s\mathbf{I} - \mathbf{A} = s\left(\mathbf{I} - \frac{\mathbf{A}}{s}\right).$$

Hence

$$\begin{aligned}(s\mathbf{I} - \mathbf{A})^{-1} &= \frac{1}{s}\left(\mathbf{I} - \frac{\mathbf{A}}{s}\right)^{-1} \\ &= \frac{1}{s}\left(\mathbf{I} + \frac{\mathbf{A}}{s} + \frac{\mathbf{A}^2}{s^2} + \cdots\right).\end{aligned}$$

Therefore

$$(s\mathbf{I} - \mathbf{A})^{-1} = \frac{\mathbf{I}}{s} + \frac{\mathbf{A}}{s^2} + \frac{\mathbf{A}^2}{s^3} + \cdots.$$

Taking the inverse Laplace transform term by term gives

$$e^{t\mathbf{A}} = \mathbf{I} + t\mathbf{A} + \frac{t^2}{2!}\mathbf{A}^2 + \cdots.$$

Thus the usual power-series definition

$$e^{t\mathbf{A}} = \sum_{k=0}^{\infty} \frac{t^k}{k!}\mathbf{A}^k$$

is recovered directly from the large-s expansion of the resolvent.

This correspondence may be written term by term as

$$\frac{\mathbf{A}^k}{s^{k+1}} \quad \longleftrightarrow \quad \frac{t^k}{k!}\mathbf{A}^k.$$

Differentiation of the resolvent. Let us denote the resolvent $(s\mathbf{I} - \mathbf{A})^{-1}$ by $\mathbf{R}(s; \mathbf{A})$. Differentiating

$$(s\mathbf{I} - \mathbf{A})\mathbf{R}(s; \mathbf{A}) = \mathbf{I}$$

with respect to s gives

$$\mathbf{R}(s; \mathbf{A}) + (s\mathbf{I} - \mathbf{A})\frac{d}{ds}\mathbf{R}(s; \mathbf{A}) = 0.$$

Multiplying by $\mathbf{R}(s; \mathbf{A})$ yields

$$\frac{d}{ds}\mathbf{R}(s; \mathbf{A}) = -\mathbf{R}(s; \mathbf{A})^2.$$

Repeated differentiation gives

$$\frac{d^k}{ds^k}\mathbf{R}(s; \mathbf{A}) = (-1)^k k!\, \mathbf{R}(s; \mathbf{A})^{k+1}.$$

On the other hand,

$$\mathcal{L}\{t^k e^{t\mathbf{A}}\}(s) = (-1)^k \frac{d^k}{ds^k}\mathbf{R}(s; \mathbf{A}).$$

Therefore

$$\boxed{\mathcal{L}\{t^k e^{t\mathbf{A}}\}(s) = k!\,(s\mathbf{I} - \mathbf{A})^{-(k+1)}.}$$

This is the matrix analogue of

$$\mathcal{L}\{t^k e^{\lambda t}\}(s) = \frac{k!}{(s-\lambda)^{k+1}}.$$

6.4.3 Resolvent poles, Jordan structure, and modal decomposition

In Chapter 5, the Jordan decomposition

$$\mathbf{A} = \mathbf{P}\mathbf{J}\mathbf{P}^{-1}$$

reduced the evolution of the system to the exponential of the Jordan matrix:

$$e^{t\mathbf{A}} = \mathbf{P}e^{t\mathbf{J}}\mathbf{P}^{-1}.$$

The resolvent admits the analogous representation. Since

$$s\mathbf{I} - \mathbf{A} = \mathbf{P}(s\mathbf{I} - \mathbf{J})\mathbf{P}^{-1},$$

we have

$$(s\mathbf{I} - \mathbf{A})^{-1} = \mathbf{P}(s\mathbf{I} - \mathbf{J})^{-1}\mathbf{P}^{-1}.$$

Thus the pole structure of the resolvent is determined entirely by the resolvent of the Jordan matrix. Since $\mathbf{J}$ is block diagonal, the calculation reduces to a single Jordan block.

Let

$$\mathbf{J}_\ell(\lambda) = \lambda\mathbf{I} + \mathbf{N}, \qquad \mathbf{N}^\ell = 0.$$

Then

$$\begin{aligned} s\mathbf{I} - \mathbf{J}_\ell(\lambda) &= (s-\lambda)\mathbf{I} - \mathbf{N} \\ &= (s-\lambda)\left(\mathbf{I} - \frac{\mathbf{N}}{s-\lambda}\right). \end{aligned}$$

Therefore,

$$(s\mathbf{I} - \mathbf{J}_\ell(\lambda))^{-1} = \frac{1}{s-\lambda}\left(\mathbf{I} - \frac{\mathbf{N}}{s-\lambda}\right)^{-1}.$$

Since $\mathbf{N}$ is nilpotent, the geometric series terminates:

$$\left(\mathbf{I} - \frac{\mathbf{N}}{s-\lambda}\right)^{-1} = \mathbf{I} + \frac{\mathbf{N}}{s-\lambda} + \cdots + \frac{\mathbf{N}^{\ell-1}}{(s-\lambda)^{\ell-1}}.$$

Hence

$$(s\mathbf{I} - \mathbf{J}_\ell(\lambda))^{-1} = \frac{\mathbf{I}}{s-\lambda} + \frac{\mathbf{N}}{(s-\lambda)^2} + \cdots + \frac{\mathbf{N}^{\ell-1}}{(s-\lambda)^\ell}.$$

Equivalently,

$$(s\mathbf{I} - \mathbf{J}_\ell(\lambda))^{-1} = \begin{bmatrix} \frac{1}{s-\lambda} & \frac{1}{(s-\lambda)^2} & \cdots & \frac{1}{(s-\lambda)^\ell} \\ 0 & \frac{1}{s-\lambda} & \cdots & \frac{1}{(s-\lambda)^{\ell-1}} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{1}{s-\lambda} \end{bmatrix}.$$

Thus a Jordan block of size ℓ produces poles of orders up to ℓ at $s = \lambda$. In particular, the order of the pole of the matrix resolvent at λ is the size of the largest Jordan block associated with λ; it need not equal the algebraic multiplicity of λ.

Taking inverse Laplace transforms term by term gives

$$\begin{aligned}\mathcal{L}^{-1}\left\{(s\mathbf{I}-\mathbf{J}_\ell(\lambda))^{-1}\right\} &= e^{\lambda t}\left(\mathbf{I}+t\mathbf{N}+\frac{t^2}{2!}\mathbf{N}^2+\cdots+\frac{t^{\ell-1}}{(\ell-1)!}\mathbf{N}^{\ell-1}\right)\\ &= e^{t\mathbf{J}_\ell(\lambda)}.\end{aligned}$$

The columns of this matrix are

$$e^{\lambda t}\mathbf{e}_1,$$
$$e^{\lambda t}(t\mathbf{e}_1+\mathbf{e}_2),$$
$$e^{\lambda t}\left(\frac{t^2}{2!}\mathbf{e}_1+t\mathbf{e}_2+\mathbf{e}_3\right),$$

and the analogous generalized modes up to degree $\ell-1$.

Now suppose

$$\mathbf{J}=\mathbf{J}_{\ell_1}(\lambda_1)\oplus\cdots\oplus\mathbf{J}_{\ell_r}(\lambda_r).$$

Then

$$(s\mathbf{I}-\mathbf{J})^{-1}=\bigoplus_{q=1}^{r}(s\mathbf{I}-\mathbf{J}_{\ell_q}(\lambda_q))^{-1},$$

and therefore

$$\mathcal{L}^{-1}\left\{(s\mathbf{I}-\mathbf{J})^{-1}\right\}=e^{t\mathbf{J}}.$$

Returning to the original coordinates,

$$\begin{aligned}\mathcal{L}^{-1}\left\{(s\mathbf{I}-\mathbf{A})^{-1}\right\} &= \mathbf{P}\mathcal{L}^{-1}\left\{(s\mathbf{I}-\mathbf{J})^{-1}\right\}\mathbf{P}^{-1}\\ &= \mathbf{P}e^{t\mathbf{J}}\mathbf{P}^{-1}\\ &= e^{t\mathbf{A}}.\end{aligned}$$

Thus the same structure appears in three equivalent forms:

$$\begin{array}{c}\text{Jordan block of size } \ell,\\ \Updownarrow\\ \text{pole of order } \ell \text{ in the resolvent},\\ \Updownarrow\\ t^{\ell-1}e^{\lambda t}\text{-type generalized modes.}\end{array}$$

The resolvent therefore gives the transform-domain representation of the complete modal structure developed in Chapter 5.

Example 6.12 (Resolvent of a matrix with one nontrivial Jordan block)**.** Suppose

$$\mathbf{A}=\mathbf{P}\mathbf{J}\mathbf{P}^{-1},$$

where

$$\mathbf{J}=\begin{bmatrix}2&1&0\\0&2&0\\0&0&-1\end{bmatrix}=\mathbf{J}_2(2)\oplus\mathbf{J}_1(-1).$$

Then

$$(s\mathbf{I} - \mathbf{A})^{-1} = \mathbf{P}(s\mathbf{I} - \mathbf{J})^{-1}\mathbf{P}^{-1},$$

where

$$(s\mathbf{I} - \mathbf{J})^{-1} = \begin{bmatrix} \dfrac{1}{s-2} & \dfrac{1}{(s-2)^2} & 0 \\ 0 & \dfrac{1}{s-2} & 0 \\ 0 & 0 & \dfrac{1}{s+1} \end{bmatrix}.$$

Thus the resolvent has a second-order pole at $s = 2$ and a simple pole at $s = -1$. Taking inverse Laplace transforms gives

$$e^{t\mathbf{J}} = \begin{bmatrix} e^{2t} & te^{2t} & 0 \\ 0 & e^{2t} & 0 \\ 0 & 0 & e^{-t} \end{bmatrix},$$

and hence

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

The second-order pole produces the generalized mode te^{2t}, while the simple pole at $s = -1$ produces the ordinary decaying mode e^{-t}.

For a diagonalizable matrix,

$$\mathbf{J} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n),$$

and hence

$$(s\mathbf{I} - \mathbf{J})^{-1} = \operatorname{diag}\left(\frac{1}{s-\lambda_1}, \ldots, \frac{1}{s-\lambda_n}\right).$$

Thus only simple poles occur, even when an eigenvalue is repeated, provided that all Jordan blocks associated with it have size one.

If $\mathbf{A}$ is real, nonreal eigenvalues and resolvent poles occur in complex-conjugate pairs. A pair

$$\lambda = \alpha + i\beta, \qquad \overline{\lambda} = \alpha - i\beta, \qquad \beta \neq 0,$$

produces the real modes

$$e^{\alpha t}\cos\beta t, \qquad e^{\alpha t}\sin\beta t.$$

If the corresponding complex Jordan blocks have size ℓ, then poles of orders up to ℓ occur, and the associated real generalized modes contain the factors

$$t^k e^{\alpha t}\cos\beta t, \qquad t^k e^{\alpha t}\sin\beta t, \qquad k = 0, \ldots, \ell - 1.$$

Eigenvalues determine the locations of the resolvent poles, Jordan block sizes determine their orders, and inverse Laplace transformation converts these poles into the ordinary and generalized modes of the system.

Exercise 6.13 (A rank-one nilpotent resolvent)**.** Let

$$\mathbf{A} = \begin{bmatrix} 3 & -1 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 2 \end{bmatrix}, \qquad \mathbf{N} = \mathbf{A} - 2\mathbf{I}.$$

Verify that $\mathbf{N}^2 = \mathbf{0}$. Use this fact to compute the resolvent, identify the order of its pole at $s = 2$, and recover $e^{t\mathbf{A}}$ without first finding a Jordan basis.

Exercise 6.14 (Same eigenvalue, different pole orders)**.** Let

$$\mathbf{A}_1 = \lambda \mathbf{I}_3, \qquad \mathbf{A}_2 = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}.$$

Both matrices have characteristic polynomial $(s - \lambda)^3$. Compute their resolvents and compare the orders of their poles at $s = \lambda$. Explain how the difference is reflected in the solutions of $\mathbf{y}' = \mathbf{A}_j \mathbf{y}$.

Exercise 6.15 (A delayed constant input)**.** Consider

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y} + \mathbf{b}\, H(t - a), \qquad \mathbf{y}(0) = \mathbf{0},$$

where $a > 0$ and $\mathbf{b}$ is a constant vector.

Use the Laplace transform to show that

$$\mathbf{y}(t) = H(t - a) \int_0^{t-a} e^{(t-a-\tau)\mathbf{A}} \mathbf{b}\, d\tau.$$

If $\mathbf{A}$ is invertible, show further that

$$\mathbf{y}(t) = H(t - a)\mathbf{A}^{-1} \left(e^{(t-a)\mathbf{A}} - \mathbf{I} \right) \mathbf{b}.$$

Homework Problems

Homework 6.1 (Resolvent from the minimal polynomial)**.** Suppose that $\mathbf{A}$ has minimal polynomial

$$m_{\mathbf{A}}(z) = (z - \lambda)^2 (z - \mu), \qquad \lambda \neq \mu.$$

Show that its resolvent has the form

$$(s\mathbf{I} - \mathbf{A})^{-1} = \frac{\mathbf{P}_\lambda}{s - \lambda} + \frac{\mathbf{N}_\lambda}{(s - \lambda)^2} + \frac{\mathbf{P}_\mu}{s - \mu},$$

where $\mathbf{P}_\lambda$, $\mathbf{P}_\mu$, and $\mathbf{N}_\lambda$ are polynomials in $\mathbf{A}$. Determine them explicitly as polynomials in $\mathbf{A}$, and use the result to obtain a formula for $e^{t\mathbf{A}}$.

Homework 6.2 (Repeated poles and resonant response)**.** Consider

$$P(D)y = f, \qquad P(D) = (D-\lambda)^2(D-\mu), \qquad \lambda \neq \mu,$$

with zero initial data.

1. Find the causal Green's function G of $P(D)$.
2. Let $f(t) = e^{\lambda t}$. Determine the pole structure of
$$Y(s) = \frac{1}{P(s)}\frac{1}{s-\lambda}$$
and find the corresponding solution.
3. Explain why the forcing produces a term of the form
$$t^2 e^{\lambda t}.$$
4. Compare this resonance mechanism with the appearance of polynomial factors produced by a Jordan block.

Homework 6.3 (A switched and impulsively driven RLC circuit)**.** A series RLC circuit has charge $q(t)$, current $i(t) = q'(t)$, inductance $L > 0$, resistance $R > 0$, and capacitance $C > 0$. It is driven by
$$V(t) = V_0 H(t-a) + J\delta(t-b), \qquad 0 < a < b,$$
where the step voltage is switched on at $t = a$ and the later voltage impulse has strength J. Assume $q(0) = i(0) = 0$.

1. Write $\mathbf{x} = (q, i)^{\mathsf{T}}$ and derive a system of the form
$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}V(t).$$
Compute $(s\mathbf{I} - \mathbf{A})^{-1}$ and identify its poles.
2. Use the resolvent and the shift rule to show that
$$\mathbf{x}(t) = V_0 H(t-a)\int_0^{t-a} e^{(t-a-\tau)\mathbf{A}}\mathbf{b}\,d\tau + JH(t-b)e^{(t-b)\mathbf{A}}\mathbf{b}.$$
Determine the jumps of q and i at $t = b$.

3. In the underdamped case $R^2 < 4L/C$, describe the transient after each input and determine the limiting charge and current.

Historical Notes and Further Reading

Historical notes. The integral transform now bearing Laplace's name developed from his work on probability, generating functions, and differential and difference equations. In his *Théorie analytique des probabilités*, Laplace (1812) used integrals with exponential kernels as part of a broader method for converting analytic problems into algebraic ones. The modern one-sided Laplace transform,

$$\mathcal{L}\{f\}(s) = \int_0^\infty e^{-st} f(t)\, dt,$$

retains this central idea: differentiation in time is replaced by multiplication by the transform variable, together with terms determined by the initial data.

Transform methods became especially important in electrical engineering through the operational calculus of Oliver Heaviside. In his treatment of electromagnetic and circuit problems, Heaviside (1899) manipulated the differentiation operator algebraically and used inverse operators to describe the response of a system to an applied source. Although his methods were initially formal, Laplace-transform theory later supplied a systematic analytic justification. The correspondence

$$P(D)y = f \qquad \longleftrightarrow \qquad P(s)Y(s) = I(s) + F(s)$$

is the modern form of this operational viewpoint.

The ideas of Green functions and Duhamel's principle were introduced in Chapter 3. In the present chapter they are unified by the transform: multiplication by $1/P(s)$ in the transform domain becomes convolution with the causal Green function in the time domain. Discontinuous inputs and concentrated impulses can therefore be treated within the same framework as ordinary forcing terms.

The delta function was used systematically by Dirac (1930) to represent a quantity concentrated at a single point. The later theory of distributions developed by Schwartz (1950–1951) gave such generalized functions a rigorous mathematical foundation. In an initial-value problem, the delta source represents an instantaneous impulse, and the corresponding solution is the impulse response of the system.

For a linear system, the transform-domain inverse is the resolvent

$$(s\mathbf{I} - \mathbf{A})^{-1}.$$

Its poles are the eigenvalues of $\mathbf{A}$, and the order of each pole records the size of the largest Jordan block associated with that eigenvalue. Inverse transformation therefore recovers the ordinary and generalized modes of Chapter 5, while

$$\mathcal{L}^{-1}\left\{(s\mathbf{I} - \mathbf{A})^{-1}\right\} = e^{t\mathbf{A}}$$

connects the resolvent directly with the evolution operator.

Further reading. For a systematic treatment of the Laplace transform, including convergence, inversion, differentiation, and convolution, see Doetsch (1974). A classical analytic account of the transform and its inversion theory is given by Widder (1946). For an accessible introduction to generalized functions and delta sources, see Lighthill (1958). The relation among evolution operators, semigroups, generators, and resolvents is developed at a more advanced level by Engel and Nagel (2000).

Orientation to Part III

Part III develops the variational and Hamiltonian structures underlying classical mechanics. A variational problem assigns a scalar quantity to an entire curve and asks which curves make it stationary. The resulting Euler–Lagrange equation converts this global condition into a differential equation.

Chapter 7 introduces functionals, variations, and constrained variational problems. First integrals reduce the order of an Euler–Lagrange equation and lead to quadrature formulas for periods, turning points, separatrices, and connecting orbits. Hamilton's principle then extends this viewpoint to mechanical systems. Generalized coordinates incorporate constraints, while conjugate momentum, cyclic coordinates, and Noether's theorem connect continuous symmetries with conservation laws. In the Kepler problem, rotational symmetry gives conservation of angular momentum, reduces the motion to a plane, and helps determine the geometry of the orbit.

Chapter 8 passes from configuration and velocity variables $(\mathbf{q}, \dot{\mathbf{q}})$ to phase-space variables $(\mathbf{q}, \mathbf{p})$. The Legendre transform converts the Euler–Lagrange equations into Hamilton's canonical equations and reveals the geometry preserved by their flow. Symplectic structure identifies the transformations that preserve Hamiltonian form, while Liouville's theorem shows that phase-space volume is conserved. Normal modes, invariant tori, quasiperiodic motion, and Poincaré recurrence illustrate the consequences of this conservative geometry.

Poisson brackets provide an algebraic language for observables, symmetry, and conservation laws. Commuting first integrals lead to Liouville integrability and the decomposition of certain coupled systems into independent motions. Point-vortex dynamics shows that this framework extends beyond particle mechanics to an interacting fluid system with a Hamiltonian, a weighted symplectic structure, and conservation laws generated by Euclidean symmetries.

The central progression is

$$\begin{aligned}\text{variational principles} &\longrightarrow \text{Euler–Lagrange equations} \longrightarrow \text{first integrals and symmetry} \\ &\longrightarrow \text{Hamiltonian phase space} \longrightarrow \text{symplectic geometry and integrability.}\end{aligned}$$

This part also prepares a central contrast in the study of long-time dynamics. Hamiltonian systems preserve geometric structure and phase-space volume, whereas the dissipative systems of Part V may contract volume and approach attractors.

Chapter 7

Variational Principles and Euler–Lagrange Equations

Many differential equations arise from local laws of change. In mechanics, for example, Newton's law prescribes the acceleration from the applied force. There is another route. One begins with a quantity assigned to an entire curve—length, travel time, energy, or action—and asks which curves make this quantity stationary. The resulting condition is the Euler–Lagrange equation.

The calculus of variations grew from classical problems in geometry and mechanics: the shortest path, the catenary, the isoperimetric problem, and the brachistochrone. Euler and Lagrange developed a general method for such problems, and Hamilton later placed mechanics itself within the same variational framework.

Our first emphasis is the relation between variational principles and ordinary differential equations:

$$\text{variational problem} \longrightarrow \text{Euler–Lagrange equation} \longrightarrow \text{first integral} \longrightarrow \text{quadrature}.$$

For a system with one degree of freedom, a first integral gives an algebraic relation for an orbit in the phase plane. It reveals turning points, closed orbits, and connecting orbits. Quadrature then restores the parametrization along the orbit, giving quantitative information such as the period of a closed orbit or the profile of a homoclinic or heteroclinic connection.

The final section develops Lagrangian mechanics for systems with several degrees of freedom and geometric constraints. Generalized coordinates and the mass matrix provide a systematic description of coupled mechanical systems, while invariance under time translation and configuration-space symmetries leads to conservation of energy and momentum. Noether's theorem makes the relation between symmetry and conservation precise, and the Kepler problem illustrates how these conserved quantities reduce and ultimately solve a nontrivial two-body problem.

This formulation also prepares the transition to the next chapter. There the conjugate momenta replace the generalized velocities, and the Euler–Lagrange equations are rewritten as Hamilton's equations on phase space.

7.1 Variational Problems and the Euler–Lagrange Equation

The calculus of variations began with geometric and mechanical problems in which the unknown is a curve rather than a number. Among the classical examples are the shortest path between two

points, the shape of a hanging chain, the curve enclosing the largest area for a prescribed length, and the curve of fastest descent under gravity. These problems differ in origin, but they have a common form: a quantity assigned to a curve is required to be stationary.

7.1.1 Functionals and variations

A functional assigns a number to a function. We consider functionals of the form

$$J[y] = \int_a^b L(x, y(x), y'(x))\, dx,$$

where

$$L \in C^2([a,b] \times \mathbb{R} \times \mathbb{R})$$

is called the Lagrangian. Suppose that the endpoint values are prescribed:

$$y(a) = y_a, \qquad y(b) = y_b.$$

For the classical derivation below, we consider stationary curves $y \in C^2([a,b])$ and smooth variations that preserve these endpoint conditions. Thus an admissible variation of y is a one-parameter family

$$y_\varepsilon(x) = y(x) + \varepsilon h(x),$$

where

$$h(a) = h(b) = 0.$$

Here h determines the direction of variation, while ε measures the size of the perturbation.

The first variation of J at y in the direction h is

$$\delta J[y](h) = \frac{d}{d\varepsilon} J[y + \varepsilon h]\bigg|_{\varepsilon=0}.$$

A curve y is stationary if

$$\delta J[y](h) = 0$$

for every admissible variation h.

Stationarity is only a necessary condition for an extremum. A stationary curve may be a minimum, a maximum, or neither.

Exercise 7.1 (A stationary curve that is not an extremum)**.** On functions satisfying $y(0) = y(\pi) = 0$, consider

$$J[y] = \int_0^\pi \left(\frac{1}{2}y'^2 - y^2\right) dx.$$

Show that $y = 0$ is stationary. Then evaluate $J[\varepsilon \sin x]$ and $J[\varepsilon \sin 2x]$ to prove that $y = 0$ is neither a local minimum nor a local maximum.

7.1.2 The Euler–Lagrange equation

For

$$J[y] = \int_a^b L(x, y, y')\,dx,$$

we have

$$\begin{aligned}\delta J[y](h) &= \frac{d}{d\varepsilon}\int_a^b L(x, y+\varepsilon h, y'+\varepsilon h')\,dx\bigg|_{\varepsilon=0} \\ &= \int_a^b \left(L_y h + L_{y'} h'\right)\,dx.\end{aligned}$$

Integrating the second term by parts gives

$$\int_a^b L_{y'} h'\,dx = \left[L_{y'} h\right]_a^b - \int_a^b \frac{d}{dx} L_{y'}\,h\,dx.$$

Since $h(a) = h(b) = 0$,

$$\delta J[y](h) = \int_a^b \left(L_y - \frac{d}{dx} L_{y'}\right) h\,dx.$$

We use the following elementary lemma.

Lemma 7.1 (Fundamental lemma of the calculus of variations)**.** *Let g be continuous on $[a, b]$. If*

$$\int_a^b g(x)h(x)\,dx = 0$$

for every smooth function h satisfying $h(a) = h(b) = 0$, then

$$g(x) = 0 \qquad \text{for all } x \in [a, b].$$

Proof. Suppose that $g(x_0) > 0$ at some $x_0 \in (a, b)$. By continuity, $g > 0$ on a small interval about x_0. Choose a nonnegative smooth function h, supported in that interval and not identically zero. Then

$$\int_a^b g(x)h(x)\,dx > 0,$$

a contradiction. The case $g(x_0) < 0$ is similar. Thus $g = 0$ on (a, b), and continuity gives $g(a) = g(b) = 0$ as well. □

Applying the lemma yields the Euler–Lagrange equation.

Theorem 7.2 (Euler–Lagrange equation)**.** *Let*

$$y \in C^2([a, b])$$

be stationary for

$$J[y] = \int_a^b L(x, y, y')\,dx$$

among all smooth curves with the same endpoint values. Then y satisfies

$$\boxed{\frac{d}{dx} L_{y'}(x, y(x), y'(x)) = L_y(x, y(x), y'(x)), \qquad a < x < b.}$$

Thus a global stationarity condition has been converted into a local differential equation.

Example 7.3 (The shortest curve)**.** The length of a graph $y = y(x)$ is

$$J[y] = \int_a^b \sqrt{1 + y'^2}\, dx.$$

Here

$$L = \sqrt{1 + y'^2}, \qquad L_y = 0.$$

The Euler–Lagrange equation gives

$$\frac{d}{dx}\left(\frac{y'}{\sqrt{1 + y'^2}} \right) = 0.$$

Hence y' is constant, and the stationary curve is the straight line joining the prescribed endpoints.

Exercise 7.2 (A nonlinear boundary-value problem)**.** Consider

$$J[y] = \int_0^1 \left[\frac{1}{2}(1 + x^2)y'^2 + \frac{1}{4}y^4 - f(x)y \right] dx,$$

subject to

$$y(0) = 0, \qquad y(1) = 1.$$

1. Derive the nonlinear boundary-value problem satisfied by every stationary curve.
2. Show that two stationary curves with the same boundary values must coincide.

   Hint. If y_1 and y_2 are stationary curves, subtract their equations, multiply by $y_1 - y_2$, and integrate by parts.
3. Explain which properties of the Lagrangian are responsible for uniqueness.

Exercise 7.3 (Equivalent Lagrangians)**.** Let $F = F(x, y)$ be C^2 and define

$$\widetilde{L}(x, y, y') = L(x, y, y') + \frac{d}{dx}F(x, y) = L + F_x + F_y y'.$$

Show that the actions defined by L and $\widetilde{L}$ differ only by an endpoint term. Conclude that, for fixed endpoints, they have the same stationary curves, and verify this directly from their Euler–Lagrange equations.

Exercise 7.4 (A free endpoint)**.** Let

$$J[y] = \int_a^b L(x, y, y')\, dx,$$

where $y(a)$ is fixed and $y(b)$ is free.

1. Compute the first variation without imposing $h(b) = 0$.
2. Derive the Euler–Lagrange equation on (a, b) and the natural boundary condition

$$L_{y'}(b, y(b), y'(b)) = 0.$$

7.1.3 Constrained variations and Lagrange multipliers

In many variational problems, the admissible curves are required to satisfy an additional condition. We therefore consider

$$J[y] = \int_a^b L(x, y, y')\, dx$$

subject to the constraint

$$K[y] = \int_a^b G(x, y, y')\, dx = K_0,$$

together with the prescribed endpoint values.

A variation

$$y_\varepsilon = y + \varepsilon h$$

is admissible only when it preserves the constraint. Differentiating

$$K[y_\varepsilon] = K_0$$

at $\varepsilon = 0$ gives the linearized condition

$$\delta K[y](h) = 0.$$

Thus constrained stationarity means

$$\delta J[y](h) = 0$$

for every variation h satisfying

$$\delta K[y](h) = 0.$$

The Lagrange multiplier principle replaces this restricted stationarity condition by an unrestricted one: under a nondegeneracy assumption, there is a constant λ such that

$$\delta J[y](h) + \lambda\, \delta K[y](h) = 0$$

for every variation h. Equivalently, y is stationary for the augmented functional

$$J[y] + \lambda K[y] = \int_a^b \big(L + \lambda G\big)(x, y, y')\, dx.$$

Thus the Euler–Lagrange equation for y can be derived from the augmented Lagrangian $L + \lambda G$.

Theorem 7.4 (Lagrange multiplier principle)**.** *Suppose that y is stationary for J subject to*

$$K[y] = K_0,$$

and assume that

$$\delta K[y] \neq 0.$$

Then there exists a constant λ such that

$$\delta J[y] + \lambda\, \delta K[y] = 0.$$

Equivalently, y is stationary for the augmented functional

$$J + \lambda K.$$

Consequently, y satisfies the Euler–Lagrange equation for the augmented Lagrangian

$$\widetilde{L} = L + \lambda G :$$

$$\frac{d}{dx}\left(L_{y'} + \lambda G_{y'}\right) = L_y + \lambda G_y.$$

Proof. Fix a variation h and consider the two-parameter family

$$y_{\varepsilon,\eta} = y + \varepsilon h + \eta h_0.$$

Since

$$\frac{\partial}{\partial \eta} K[y_{\varepsilon,\eta}]\bigg|_{(0,0)} = \delta K[y](h_0) \neq 0,$$

the implicit function theorem gives a function $\eta = \eta(\varepsilon)$ such that

$$K[y_{\varepsilon,\eta(\varepsilon)}] = K_0.$$

Stationarity of J along this constrained family gives

$$0 = \delta J[y](h) + \eta'(0)\delta J[y](h_0).$$

Differentiating the constraint gives

$$0 = \delta K[y](h) + \eta'(0)\delta K[y](h_0).$$

Eliminating $\eta'(0)$ yields

$$\delta J[y](h) + \lambda \delta K[y](h) = 0,$$

where

$$\lambda = -\frac{\delta J[y](h_0)}{\delta K[y](h_0)}.$$

□

The multiplier λ is determined together with the stationary curve from the Euler–Lagrange equation, the endpoint conditions, and the original constraint.

Example 7.5 (The catenary)**.** A flexible chain of uniform density hangs between two fixed points. If y is measured upward, its potential energy is proportional to

$$J[y] = \int_a^b y\sqrt{1 + y'^2}\, dx.$$

Its length is prescribed:

$$\int_a^b \sqrt{1 + y'^2}\, dx = \ell.$$

Introducing a Lagrange multiplier, we obtain the augmented Lagrangian

$$\widetilde{L} = (y + \lambda)\sqrt{1 + y'^2}.$$

The Euler–Lagrange equation is

$$\frac{d}{dx}\left(\frac{(y + \lambda)y'}{\sqrt{1 + y'^2}}\right) = \sqrt{1 + y'^2}.$$

Thus the constrained variational problem has been reduced to a differential equation. In the next section, a first integral will reduce this equation to first order and lead to the catenary.

Example 7.6 (The isoperimetric problem)**.** The classical isoperimetric problem asks for the closed plane curve of fixed length that encloses the greatest area.

Let

$$\boldsymbol{\gamma}(s) = \begin{bmatrix} x(s) \\ y(s) \end{bmatrix}, \qquad 0 \le s \le L,$$

be a closed curve. Its signed area is

$$A[\boldsymbol{\gamma}] = \frac{1}{2}\int_0^L \left(xy' - yx'\right)\, ds,$$

while its length is

$$\mathcal{L}[\boldsymbol{\gamma}] = \int_0^L \sqrt{x'^2 + y'^2}\, ds.$$

Introducing a multiplier for the length constraint and applying the Euler–Lagrange equations gives the condition that the curvature is constant. A closed plane curve of constant nonzero curvature is a circle.

The variational calculation identifies the stationary curve. It does not, by itself, prove the global isoperimetric inequality

$$4\pi A \le \mathcal{L}^2,$$

with equality only for circles.

The isoperimetric problem is among the oldest problems in optimization. Its study influenced the development of both the calculus of variations and geometric analysis.

Exercise 7.5 (An elastic string with prescribed mean displacement)**.** A stretched elastic string on a nonlinear foundation has energy

$$J[y] = \int_a^b \left(\frac{1}{2} y'^2 + V(y) \right) \, dx.$$

With fixed endpoints, find the stationary equation subject to

$$\int_a^b y(x) \, dx = M.$$

Interpret the Lagrange multiplier physically.

Exercise 7.6 (A capillary interface with prescribed area)**.** A two-dimensional liquid interface joining two fixed points has surface energy

$$J[y] = \sigma \int_a^b \sqrt{1 + y'^2} \, dx, \qquad \sigma > 0.$$

Assume that the enclosed area is prescribed:

$$\int_a^b y(x) \, dx = A.$$

Derive the constrained Euler–Lagrange equation and show that every stationary interface has constant curvature. Interpret the Lagrange multiplier physically.

7.2 First Integrals and Quadrature

The Euler–Lagrange equation is usually a second-order differential equation. When the Lagrangian has a symmetry, it may possess a first integral, reducing the equation to first order. The resulting algebraic relation describes an orbit in the phase plane; quadrature then determines how that orbit is traversed. We first consider two basic cases: independence of the dependent variable yields conservation of conjugate momentum, while independence of the independent variable yields the Beltrami identity.

7.2.1 Cyclic variables and momentum

Proposition 7.7 (Cyclic variable)**.** *If*

$$L_y = 0,$$

then every solution of the Euler–Lagrange equation satisfies

$$L_{y'} = C.$$

Proof. The Euler–Lagrange equation reduces to

$$\frac{d}{dx} L_{y'} = 0.$$

□

The quantity $L_{y'}$ is called the momentum conjugate to y. Thus independence of the Lagrangian from a dependent variable gives a conserved momentum.

Example 7.8 (Inertial motion)**.** For a free particle of mass m,

$$L(q, \dot{q}) = \frac{1}{2} m \dot{q}^2.$$

Since L is independent of q, the conjugate momentum

$$p = L_{\dot{q}} = m\dot{q}$$

is constant. Hence

$$\dot{q} = v_0, \qquad q(t) = q_0 + v_0 t.$$

Thus conservation of linear momentum gives uniform motion.

Example 7.9 (A weighted path)**.** Consider

$$J[y] = \int_a^b n(x)\sqrt{1 + y'^2}\, dx, \qquad n(x) > 0.$$

Since L is independent of y, the conjugate momentum is conserved:

$$\frac{n(x)y'}{\sqrt{1 + y'^2}} = C.$$

Solving for y' gives

$$y' = \frac{C}{\sqrt{n(x)^2 - C^2}},$$

where the sign is absorbed into the constant C. The path is restricted to regions in which

$$n(x) > |C|.$$

Therefore

$$y(x) = y(x_0) + \int_{x_0}^{x} \frac{C}{\sqrt{n(\xi)^2 - C^2}}\, d\xi.$$

The constant C is determined by the endpoint conditions. If the path joins (x_0, y_0) to (x_1, y_1), then

$$y_1 - y_0 = \int_{x_0}^{x_1} \frac{C}{\sqrt{n(x)^2 - C^2}}\, dx.$$

Thus C is obtained from a scalar equation, usually by numerical solution.

Thus the conservation law reduces the Euler–Lagrange equation directly to quadrature.

Exercise 7.7 (A cyclic coordinate)**.** Let

$$J[y] = \int_a^b \left(\frac{1}{2} a(x) y'^2 + b(x) y' \right) dx, \qquad a(x) > 0.$$

Find the conserved conjugate momentum and reduce the Euler–Lagrange equation to one quadrature.

7.2.2 The Beltrami identity

A second first integral appears when the Lagrangian has no explicit dependence on the independent variable.

Proposition 7.10 (Beltrami identity)**.** *If*

$$L_x = 0,$$

then every solution of the Euler–Lagrange equation satisfies

$$L - y' L_{y'} = C.$$

Proof. Along a solution,

$$\begin{aligned} \frac{d}{dx}\left(L - y' L_{y'} \right) &= L_x + L_y y' + L_{y'} y'' - y'' L_{y'} - y' \frac{d}{dx} L_{y'} \\ &= L_x + y' \left(L_y - \frac{d}{dx} L_{y'} \right) = 0. \end{aligned}$$

□

The Beltrami identity is the variational origin of the energy integral in autonomous mechanical systems. More generally, the cyclic-variable identity and the Beltrami identity are two elementary manifestations of the same principle: invariance of the Lagrangian under a continuous symmetry gives rise to a conserved quantity. This principle will be developed systematically as Noether's theorem in Section 7.3.4.

Example 7.11 (The catenary)**.** For the constrained catenary problem in Example 7.5, which describes a hanging chain of fixed length, the augmented Lagrangian is

$$\widetilde{L} = (y + \lambda)\sqrt{1 + y'^2}.$$

Since $\widetilde{L}$ is independent of x, the Beltrami identity gives

$$\widetilde{L} - y' \widetilde{L}_{y'} = \frac{y + \lambda}{\sqrt{1 + y'^2}} = C.$$

Therefore

$$y'^2 = \frac{(y + \lambda)^2}{C^2} - 1.$$

Separating variables yields

$$y + \lambda = C \cosh\left(\frac{x - x_0}{C}\right).$$

Thus the stationary shape is a catenary,

$$y(x) = C \cosh\left(\frac{x - x_0}{C}\right) - \lambda.$$

Suppose that the endpoints are

$$y(a) = y_a, \qquad y(b) = y_b,$$

and that the prescribed length is ℓ. Then the constants $C > 0$, x_0, and λ are determined by

$$y_a = C \cosh\left(\frac{a - x_0}{C}\right) - \lambda,$$

$$y_b = C \cosh\left(\frac{b - x_0}{C}\right) - \lambda,$$

together with

$$\ell = \int_a^b \sqrt{1 + y'^2}\, dx.$$

Since

$$y'(x) = \sinh\left(\frac{x - x_0}{C}\right),$$

we have

$$\sqrt{1 + y'^2} = \cosh\left(\frac{x - x_0}{C}\right),$$

and therefore

$$\ell = C\left[\sinh\left(\frac{x - x_0}{C}\right)\right]_a^b.$$

Hence C, x_0, and λ are obtained from the three equations

$$y_a = C \cosh\left(\frac{a - x_0}{C}\right) - \lambda,$$

$$y_b = C \cosh\left(\frac{b - x_0}{C}\right) - \lambda,$$

and

$$\ell = C\left[\sinh\left(\frac{b - x_0}{C}\right) - \sinh\left(\frac{a - x_0}{C}\right)\right].$$

In general, this nonlinear system is solved numerically.

The catenary problem was posed in the seventeenth century. Galileo believed that a hanging chain formed a parabola, but the correct curve was found in 1691 by Johann Bernoulli, Jacob Bernoulli, and Huygens. Its derivation was one of the early successes of the calculus of variations.

Example 7.12 (The brachistochrone)**.** The brachistochrone problem asks for the curve along which a particle, starting from rest, descends between two points in the shortest time. Johann Bernoulli posed the problem in 1696; its solutions by Newton, Leibniz, Jacob Bernoulli, and L'Hospital helped stimulate the development of the calculus of variations.

Let y measure vertical distance downward from the initial point. By conservation of mechanical energy, the speed is

$$v = \sqrt{2gy}.$$

For a graph $y = y(x)$, the travel time is

$$T[y] = \frac{1}{\sqrt{2g}} \int \sqrt{\frac{1 + y'^2}{y}}\, dx.$$

Ignoring the constant factor, the Lagrangian is

$$L = \sqrt{\frac{1 + y'^2}{y}}.$$

Since $L_x = 0$, the Beltrami identity gives

$$\frac{1}{\sqrt{y}\sqrt{1 + y'^2}} = C.$$

Equivalently,

$$y(1 + y'^2) = a$$

for some $a > 0$.

Set

$$y = \frac{a}{2}(1 - \cos\theta).$$

Then the first integral gives

$$\frac{dx}{d\theta} = \frac{a}{2}(1 - \cos\theta),$$

and hence

$$x - x_0 = \frac{a}{2}(\theta - \sin\theta), \qquad y = \frac{a}{2}(1 - \cos\theta).$$

The brachistochrone is therefore an arc of a cycloid.

7.2.3 Periods of closed orbits

Consider the autonomous equation

$$m\ddot{q} + V'(q) = 0, \qquad m > 0.$$

It is the Euler–Lagrange equation associated with

$$L(q, \dot{q}) = \frac{1}{2}m\dot{q}^2 - V(q).$$

Since the Lagrangian is independent of t, the Beltrami identity gives

$$L - \dot{q}L_{\dot{q}} = C.$$

Equivalently,

$$\frac{1}{2}m\dot{q}^2 + V(q) = E,$$

where $E = -C$ is the total energy.

In the phase plane, the energy relation

$$\frac{1}{2}mv^2 + V(q) = E, \qquad v = \dot{q},$$

defines an orbit. Motion is possible only where

$$V(q) \le E,$$

and the turning points are the points at which

$$V(q) = E.$$

The points at which

$$V(q) = E$$

are the possible turning points of the motion. At a simple turning point, the velocity vanishes and the direction of motion reverses.

Suppose that the equation

$$V(q) = E$$

has two consecutive simple roots

$$q_-(E) < q_+(E),$$

and that

$$V(q) < E, \qquad q_-(E) < q < q_+(E).$$

Then the corresponding energy orbit is closed and the solution oscillates between the two turning points. Solving the energy relation for $\dot{q}$ gives

$$\dot{q} = \pm\sqrt{\frac{2}{m}\big(E - V(q)\big)}.$$

The time required to travel from q_- to q_+ is therefore

$$\sqrt{\frac{m}{2}} \int_{q_-}^{q_+} \frac{dq}{\sqrt{E - V(q)}}.$$

This is one half of a complete oscillation. Hence the period is

$$T(E) = \sqrt{2m} \int_{q_-(E)}^{q_+(E)} \frac{dq}{\sqrt{E - V(q)}}.$$

Thus the closed energy orbit determines the geometry of the oscillation, while quadrature determines its period.

Example 7.13 (The harmonic oscillator)**.** For

$$V(q) = \frac{1}{2}kq^2, \qquad k > 0,$$

the energy equation is

$$\frac{1}{2}m\dot{q}^2 + \frac{1}{2}kq^2 = E.$$

The turning points are

$$q_\pm = \pm A, \qquad A = \sqrt{\frac{2E}{k}}.$$

Therefore

$$T(E) = \sqrt{2m} \int_{-A}^{A} \frac{dq}{\sqrt{E - \frac{1}{2}kq^2}}.$$

With $q = A\sin\theta$, this becomes

$$T(E) = 2\pi\sqrt{\frac{m}{k}}.$$

The period is independent of the energy and hence of the amplitude.

The harmonic oscillator is exceptional. For a general potential, the period depends on the energy and therefore on the amplitude of oscillation.

Example 7.14 (The simple pendulum)**.** Let θ denote the angle of a pendulum of length ℓ. Its equation is

$$\ddot{\theta} + \frac{g}{\ell}\sin\theta = 0,$$

and its energy is

$$\frac{1}{2}\dot{\theta}^2 + \frac{g}{\ell}(1 - \cos\theta) = E.$$

Suppose that the pendulum is released from rest at

$$\theta = \theta_0, \qquad 0 < \theta_0 < \pi.$$

Then

$$E = \frac{g}{\ell}(1 - \cos\theta_0),$$

and the turning points are $\pm\theta_0$. By symmetry,

$$T = 4\sqrt{\frac{\ell}{2g}} \int_0^{\theta_0} \frac{d\theta}{\sqrt{\cos\theta - \cos\theta_0}}.$$

Introduce

$$k = \sin\frac{\theta_0}{2}, \qquad \sin\frac{\theta}{2} = k\sin\phi.$$

Then

$$T = 4\sqrt{\frac{\ell}{g}}\, K(k),$$

where

$$K(k) = \int_0^{\pi/2} \frac{d\phi}{\sqrt{1 - k^2\sin^2\phi}}$$

is the complete elliptic integral of the first kind.

For small oscillations, $k \to 0$ and

$$K(k) \to \frac{\pi}{2},$$

so that

$$T \sim 2\pi\sqrt{\frac{\ell}{g}}.$$

This is the familiar period obtained from the linearized equation. As $\theta_0 \to \pi$, however, $k \to 1$ and

$$K(k) \to \infty.$$

Thus the period tends to infinity as the oscillatory orbit approaches the separatrix.

The dependence of the pendulum period on its amplitude led Huygens to seek a pendulum whose oscillations would be exactly isochronous.

Example 7.15 (The cycloidal pendulum)**.** Consider the cycloid

$$x = a(\theta + \sin\theta), \qquad y = a(1 - \cos\theta),$$

with its lowest point at $\theta = 0$. Let s be signed arc length measured from the lowest point. Since

$$ds = 2a\cos\frac{\theta}{2}\, d\theta,$$

we obtain

$$s = 4a\sin\frac{\theta}{2}.$$

Consequently,

$$y = a(1 - \cos\theta) = 2a\sin^2\frac{\theta}{2} = \frac{s^2}{8a}.$$

For a bead of mass m constrained to move along the cycloid, the Lagrangian in the arc-length coordinate is

$$L(s, \dot{s}) = \frac{1}{2}m\dot{s}^2 - \frac{mg}{8a}s^2.$$

The Euler–Lagrange equation is therefore

$$\ddot{s} + \frac{g}{4a}s = 0.$$

Hence every oscillation has period

$$T = 4\pi\sqrt{\frac{a}{g}},$$

independent of its amplitude. Thus the cycloidal pendulum is exactly isochronous, whereas the circular pendulum is only approximately isochronous for small oscillations.

Exercise 7.8 (A nonlinear period)**.** Consider

$$\ddot{q} + q^3 = 0, \qquad q(0) = A > 0, \qquad \dot{q}(0) = 0.$$

Use the energy integral to show that the solution is periodic and that

$$T(A) = \frac{C}{A},$$

where C is independent of A. Express C as a definite integral.

7.2.4 Separatrices and connecting orbits

Closed energy curves describe periodic motion. A different geometry occurs when an energy level passes through an unstable equilibrium. Such a critical energy level typically contains separatrices: distinguished orbits that separate regions of the phase plane with qualitatively different motions.

Consider again

$$m\ddot{q} + V'(q) = 0$$

with energy

$$\frac{1}{2}m\dot{q}^2 + V(q) = E.$$

An equilibrium q_* satisfies

$$V'(q_*) = 0.$$

If q_* is a local maximum of V, then $(q_*, 0)$ is an unstable equilibrium in the phase plane. The distinguished energy level

$$E_* = V(q_*)$$

may contain orbits that approach this equilibrium as $t \to \pm\infty$.

Solving the energy relation for $\dot q$ gives

$$\dot q = \pm\sqrt{\frac{2}{m}\big(E_* - V(q)\big)},$$

and hence

$$t - t_0 = \pm\sqrt{\frac{m}{2}}\int_{q_0}^{q(t)} \frac{du}{\sqrt{E_* - V(u)}}.$$

If q_* is a nondegenerate maximum, then

$$V(q) = E_* - \frac{\alpha}{2}(q-q_*)^2 + o\big((q-q_*)^2\big), \qquad \alpha > 0.$$

Therefore

$$\frac{1}{\sqrt{E_* - V(q)}} \sim \sqrt{\frac{2}{\alpha}}\,\frac{1}{|q-q_*|},$$

so the quadrature diverges logarithmically as $q \to q_*$. Thus the orbit approaches the unstable equilibrium only as $|t| \to \infty$.

An orbit that approaches the same equilibrium as $t \to -\infty$ and $t \to \infty$ is called *homoclinic*. An orbit that connects two different equilibria is called *heteroclinic*. Both are connecting orbits lying on separatrices.

Example 7.16 (The pendulum separatrix)**.** For the pendulum,

$$\ddot\theta + \frac{g}{\ell}\sin\theta = 0,$$

the energy is

$$\frac{1}{2}\dot\theta^2 + \frac{g}{\ell}(1-\cos\theta) = E.$$

The unstable equilibria occur at

$$\theta = (2k+1)\pi, \qquad \dot\theta = 0,$$

and their energy is

$$E_* = \frac{2g}{\ell}.$$

On this energy level,

$$\frac{1}{2}\dot\theta^2 = \frac{g}{\ell}(1+\cos\theta) = \frac{2g}{\ell}\cos^2\frac{\theta}{2}.$$

For the increasing orbit from $-\pi$ to π,

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

Separating variables and integrating gives

$$\theta(t) = 4\arctan\left(\exp\left(\sqrt{\frac{g}{\ell}}(t-t_0)\right)\right) - \pi.$$

Hence

$$\theta(t) \to -\pi \quad \text{as} \quad t \to -\infty, \qquad \theta(t) \to \pi \quad \text{as} \quad t \to \infty.$$

On the covering plane this is a heteroclinic orbit joining two adjacent unstable equilibria. Since the angles $-\pi$ and π represent the same point on the phase cylinder, the same orbit is homoclinic when the angular variable is regarded modulo 2π. See Figure 4.7 in Chapter 4.

Connecting orbits also arise naturally in variational models of transition layers.

Example 7.17 (A phase-field transition)**.** Consider the functional

$$J[u] = \int_{-\infty}^{\infty} \left[\frac{1}{2}u'^2 + \frac{1}{4}(1-u^2)^2 \right] dx$$

with the boundary conditions

$$u(x) \to -1 \quad \text{as} \quad x \to -\infty, \qquad u(x) \to 1 \quad \text{as} \quad x \to \infty.$$

Its Euler–Lagrange equation is

$$u'' = u(u^2 - 1).$$

Since the Lagrangian is independent of x, the Beltrami identity gives

$$\frac{1}{4}(1-u^2)^2 - \frac{1}{2}u'^2 = C.$$

The boundary conditions imply

$$u' \to 0, \qquad u \to \pm 1,$$

and hence $C = 0$. Therefore

$$\frac{1}{2}u'^2 = \frac{1}{4}(1-u^2)^2.$$

For the increasing connecting orbit,

$$u' = \frac{1}{\sqrt{2}}(1-u^2).$$

Separating variables gives

$$\int \frac{du}{1-u^2} = \frac{x - x_0}{\sqrt{2}},$$

and consequently

$$u(x) = \tanh\left(\frac{x - x_0}{\sqrt{2}} \right).$$

This solution is a heteroclinic orbit joining the equilibria $u = -1$ and $u = 1$. The constant x_0 reflects translation invariance and determines the location of the transition layer.

Thus the same sequence of ideas applies throughout this section: a first integral determines the distinguished orbit in the phase plane, and quadrature determines its temporal or spatial profile.

Exercise 7.9 (Periodic and homoclinic motion in a double well)**.** Consider the action

$$\mathcal{A}[q] = \int_{t_0}^{t_1} \left[\frac{1}{2}\dot{q}^2 - \frac{1}{4}(q^2-1)^2 \right] dt.$$

1. Derive the equation of motion and its conserved energy.
2. For $0 < E < 1/4$, find the turning points of the periodic motion in the right-hand well and express its period as a quadrature. Show that
$$T(E) \to \sqrt{2}\,\pi \quad \text{as } E \to 0, \qquad T(E) \to \infty \quad \text{as } E \to \frac{1}{4}.$$
3. At $E = 1/4$, use quadrature to obtain the homoclinic orbits
$$q(t) = \pm\sqrt{2}\,\operatorname{sech}(t - t_*).$$

7.3 Lagrangian Mechanics

The preceding examples, including the simple pendulum and the double-well oscillator, applied the Euler–Lagrange equation to systems with one degree of freedom. We now develop the variational formulation of mechanics for systems with several degrees of freedom and geometric constraints.

The essential step is to describe the configuration by *generalized coordinates* adapted to the constraints. The equations of motion then follow from a single scalar function, the Lagrangian, rather than from separate force balances in Cartesian coordinates. This formulation also reveals the connection between symmetries, cyclic coordinates, and conserved momenta.

7.3.1 Hamilton's principle

Let

$$\mathbf{q} = (q_1, \dots, q_n)$$

denote generalized coordinates describing the configuration of a mechanical system. For the moment, they may be regarded simply as coordinates specifying the configuration of the system; their construction from geometric constraints will be illustrated in the next subsection. A motion is then a curve

$$\mathbf{q} = \mathbf{q}(t)$$

in configuration space. The Lagrangian is a function

$$L = L(t, \mathbf{q}, \dot{\mathbf{q}}),$$

and the action on the time interval $[t_0, t_1]$ is

$$\mathcal{S}[\mathbf{q}] = \int_{t_0}^{t_1} L(t, \mathbf{q}, \dot{\mathbf{q}})\, dt.$$

Principle 7.1 (Hamilton's principle)**.** *The actual motion between two fixed configurations is stationary for the action:*

$$\delta \mathcal{S}[\mathbf{q}] = 0,$$

with respect to variations satisfying

$$\delta \mathbf{q}(t_0) = \delta \mathbf{q}(t_1) = 0.$$

Let

$$\mathbf{q}_\varepsilon(t) = \mathbf{q}(t) + \varepsilon \mathbf{h}(t), \qquad \mathbf{h}(t_0) = \mathbf{h}(t_1) = 0.$$

Then

$$\begin{aligned} \frac{d}{d\varepsilon}\mathcal{S}[\mathbf{q}_\varepsilon]\bigg|_{\varepsilon=0} &= \int_{t_0}^{t_1} \left(L_{\mathbf{q}} \cdot \mathbf{h} + L_{\dot{\mathbf{q}}} \cdot \dot{\mathbf{h}} \right) dt \\ &= \int_{t_0}^{t_1} \left(L_{\mathbf{q}} - \frac{d}{dt} L_{\dot{\mathbf{q}}} \right) \cdot \mathbf{h} \, dt. \end{aligned}$$

Hence Hamilton's principle gives the Euler–Lagrange system

$$\frac{d}{dt} L_{\dot{\mathbf{q}}} = L_{\mathbf{q}}.$$

In components,

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_j} = \frac{\partial L}{\partial q_j}, \qquad j = 1, \dots, n.$$

For a natural mechanical system,

$$L(\mathbf{q}, \dot{\mathbf{q}}) = T(\mathbf{q}, \dot{\mathbf{q}}) - V(\mathbf{q}),$$

where T is the kinetic energy and V is the potential energy. In Cartesian coordinates for a particle of mass m,

$$T = \frac{1}{2} m |\dot{\mathbf{q}}|^2,$$

and the Euler–Lagrange equation becomes

$$m\ddot{\mathbf{q}} = -\nabla V(\mathbf{q}).$$

Thus Newton's equation is recovered from Hamilton's principle.

7.3.2 Generalized coordinates and the mass matrix

For a mechanical system with generalized coordinates

$$\mathbf{q} = (q_1, \dots, q_n),$$

the kinetic energy is often a quadratic form in the generalized velocities:

$$T(\mathbf{q}, \dot{\mathbf{q}}) = \frac{1}{2} \dot{\mathbf{q}}^{\mathsf{T}} \mathbf{M}(\mathbf{q}) \dot{\mathbf{q}}.$$

The symmetric positive definite matrix $\mathbf{M}(\mathbf{q})$ is called the *mass matrix*. The Lagrangian of a natural mechanical system is then

$$L(\mathbf{q}, \dot{\mathbf{q}}) = \frac{1}{2}\dot{\mathbf{q}}^{\mathsf{T}}\mathbf{M}(\mathbf{q})\dot{\mathbf{q}} - V(\mathbf{q}).$$

When $\mathbf{M}$ depends on $\mathbf{q}$, the Euler–Lagrange equations contain terms quadratic in the velocities. These terms arise from the geometry of the coordinates and from the coupling among different degrees of freedom.

The simple pendulum revisited. The simple pendulum provides the basic example of a generalized coordinate. In Cartesian coordinates, the position of the mass is constrained to a circle. Using the angle θ instead, we write

$$x = \ell \sin\theta, \qquad y = -\ell\cos\theta,$$

so the constraint is satisfied automatically. The kinetic and potential energies are

$$T = \frac{1}{2}m\ell^2\dot{\theta}^2, \qquad V = mg\ell(1 - \cos\theta).$$

Hence

$$L(\theta, \dot{\theta}) = \frac{1}{2}m\ell^2\dot{\theta}^2 - mg\ell(1 - \cos\theta),$$

and the Euler–Lagrange equation gives

$$\ddot{\theta} + \frac{g}{\ell}\sin\theta = 0.$$

Thus one generalized coordinate replaces two Cartesian coordinates together with the circular constraint. The double pendulum extends this idea to two coupled angular coordinates.

Example 7.18 (Double pendulum)**.** Consider two point masses m_1 and m_2 connected by massless rigid links of lengths ℓ_1 and ℓ_2. The first link is attached to a fixed pivot, and the second link is attached to the end of the first. Let θ_1 and θ_2 denote the angles made by the two links with the downward vertical.

The positions of the two masses are

$$\begin{aligned} x_1 &= \ell_1\sin\theta_1, & y_1 &= -\ell_1\cos\theta_1, \\ x_2 &= \ell_1\sin\theta_1 + \ell_2\sin\theta_2, & y_2 &= -\ell_1\cos\theta_1 - \ell_2\cos\theta_2. \end{aligned}$$

Thus the four Cartesian coordinates of the two masses are determined by the two generalized coordinates θ_1 and θ_2.

Differentiating gives

$$\dot{x}_1^2 + \dot{y}_1^2 = \ell_1^2\dot{\theta}_1^2$$

and

$$\begin{aligned} \dot{x}_2^2 + \dot{y}_2^2 = \ell_1^2\dot{\theta}_1^2 + \ell_2^2\dot{\theta}_2^2 \\ + 2\ell_1\ell_2\cos(\theta_1 - \theta_2)\dot{\theta}_1\dot{\theta}_2. \end{aligned}$$

Hence

$$T = \frac{1}{2}(m_1 + m_2)\ell_1^2\dot{\theta}_1^2 + \frac{1}{2}m_2\ell_2^2\dot{\theta}_2^2 + m_2\ell_1\ell_2\cos(\theta_1 - \theta_2)\dot{\theta}_1\dot{\theta}_2.$$

The potential energy is

$$V = -(m_1 + m_2)g\ell_1\cos\theta_1 - m_2g\ell_2\cos\theta_2.$$

Therefore

$$L = T - V.$$

Writing

$$\mathbf{q} = \begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix},$$

the kinetic energy takes the form

$$T = \frac{1}{2}\dot{\mathbf{q}}^{\mathsf{T}}\mathbf{M}(\mathbf{q})\dot{\mathbf{q}},$$

where

$$\mathbf{M}(\mathbf{q}) = \begin{bmatrix}(m_1 + m_2)\ell_1^2 & m_2\ell_1\ell_2\cos(\theta_1 - \theta_2) \\ m_2\ell_1\ell_2\cos(\theta_1 - \theta_2) & m_2\ell_2^2\end{bmatrix}.$$

The off-diagonal entries describe inertial coupling: the kinetic energy cannot be separated into independent terms for θ_1 and θ_2.

The Euler–Lagrange equations are

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{\theta}_j} = \frac{\partial L}{\partial \theta_j}, \qquad j = 1, 2.$$

They form a coupled nonlinear system for θ_1 and θ_2. The same coordinate description applies to a planar two-link robotic arm; in that setting, the angles are joint variables and external torques may be added to the right-hand sides of the equations.

The mass-matrix form separates the basic ingredients of a mechanical system: the configuration-dependent inertia is encoded in $\mathbf{M}(\mathbf{q})$, while the forces derived from position are encoded in $V(\mathbf{q})$.

Exercise 7.10 (Motion constrained to a parabola)**.** A particle of mass m moves on $y = ax^2$ under a potential $V(x, y)$. Use x as a generalized coordinate.

1. Find the effective mass $M(x)$ and the reduced potential

$$U(x) = V(x, ax^2)$$

such that

$$T = \frac{1}{2}M(x)\dot{x}^2.$$

2. Derive
$$M(x)\ddot{x} + \frac{1}{2}M'(x)\dot{x}^2 + U'(x) = 0.$$

7.3.3 Conservation of energy, conjugate momentum, and cyclic coordinates

The Euler–Lagrange equations immediately yield two fundamental conservation laws. If the Lagrangian has no explicit time dependence, then the energy is conserved. If it is independent of one of the generalized coordinates, then the corresponding conjugate momentum is conserved.

Conservation of energy For a Lagrangian $L(t, \mathbf{q}, \dot{\mathbf{q}})$, define the energy function by
$$E = \dot{\mathbf{q}}^{\mathsf{T}} L_{\dot{\mathbf{q}}} - L.$$

Along a solution of the Euler–Lagrange equations,
$$\begin{aligned}\frac{dE}{dt} &= \sum_{j=1}^{n}\left(\ddot{q}_j\frac{\partial L}{\partial \dot{q}_j} + \dot{q}_j\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_j}\right) - \frac{\partial L}{\partial t} - \sum_{j=1}^{n}\frac{\partial L}{\partial q_j}\dot{q}_j - \sum_{j=1}^{n}\frac{\partial L}{\partial \dot{q}_j}\ddot{q}_j \\ &= \sum_{j=1}^{n}\dot{q}_j\left(\frac{d}{dt}\frac{\partial L}{\partial \dot{q}_j} - \frac{\partial L}{\partial q_j}\right) - \frac{\partial L}{\partial t}.\end{aligned}$$

The expression in parentheses vanishes by the Euler–Lagrange equations. Therefore,
$$\boxed{\frac{dE}{dt} = -\frac{\partial L}{\partial t}.}$$

Consequently, if L has no explicit time dependence, then E is conserved.

For a natural mechanical system,
$$L(\mathbf{q}, \dot{\mathbf{q}}) = \frac{1}{2}\dot{\mathbf{q}}^{\mathsf{T}}\mathbf{M}(\mathbf{q})\dot{\mathbf{q}} - V(\mathbf{q}),$$

the conserved energy is
$$E = \frac{1}{2}\dot{\mathbf{q}}^{\mathsf{T}}\mathbf{M}(\mathbf{q})\dot{\mathbf{q}} + V(\mathbf{q}) = T + V.$$

Conjugate momentum and cyclic coordinates The momentum conjugate to q_j is
$$p_j = \frac{\partial L}{\partial \dot{q}_j},$$

or, in vector form,
$$\mathbf{p} = L_{\dot{\mathbf{q}}}.$$

For a natural mechanical system,
$$L(\mathbf{q}, \dot{\mathbf{q}}) = \frac{1}{2}\dot{\mathbf{q}}^{\mathsf{T}}\mathbf{M}(\mathbf{q})\dot{\mathbf{q}} - V(\mathbf{q}),$$

this gives
$$\mathbf{p} = \mathbf{M}(\mathbf{q})\dot{\mathbf{q}}.$$

A generalized coordinate q_j is called *cyclic* if

$$\frac{\partial L}{\partial q_j} = 0.$$

The Euler–Lagrange equation then gives

$$\dot{p}_j = \frac{\partial L}{\partial q_j} = 0.$$

Thus the momentum conjugate to a cyclic coordinate is conserved.

Example 7.19 (A cyclic coordinate in polar coordinates)**.** Consider a particle of mass m moving in the plane under a central potential $V(r)$. In polar coordinates,

$$L(r, \theta, \dot{r}, \dot{\theta}) = \frac{1}{2} m \left(\dot{r}^2 + r^2 \dot{\theta}^2 \right) - V(r).$$

Since θ is cyclic,

$$p_\theta = \frac{\partial L}{\partial \dot{\theta}} = m r^2 \dot{\theta}$$

is constant. This is the angular momentum about the origin.

Writing

$$p_\theta = \ell,$$

we obtain

$$\dot{\theta} = \frac{\ell}{m r^2}.$$

The energy is

$$E = \frac{1}{2} m \dot{r}^2 + \frac{\ell^2}{2 m r^2} + V(r).$$

Thus the two-dimensional motion reduces to the one-dimensional radial equation

$$\frac{1}{2} m \dot{r}^2 + V_{\text{eff}}(r) = E,$$

where

$$V_{\text{eff}}(r) = V(r) + \frac{\ell^2}{2 m r^2}.$$

The cyclic coordinate therefore produces both a conserved quantity and a reduction of the dynamics.

Exercise 7.11 (The spherical pendulum)**.** A mass m is attached to a rigid rod of length ℓ and moves on a sphere under gravity. Let θ be the angle measured from the downward vertical and let ϕ be the azimuthal angle. Its Lagrangian is

$$L = \frac{1}{2} m \ell^2 \left(\dot{\theta}^2 + \sin^2 \theta \, \dot{\phi}^2 \right) - m g \ell (1 - \cos \theta).$$

1. Identify the cyclic coordinate and derive the corresponding conserved momentum
$$J = m\ell^2 \sin^2\theta\, \dot{\phi}.$$
2. Use conservation of J and energy to reduce the motion of θ to
$$\frac{1}{2} m\ell^2 \dot{\theta}^2 + V_{\text{eff}}(\theta) = E,$$
where
$$V_{\text{eff}}(\theta) = mg\ell(1 - \cos\theta) + \frac{J^2}{2m\ell^2 \sin^2\theta}.$$
3. Explain why a solution with $J \neq 0$ cannot pass through either vertical position. Determine the condition for a motion with constant θ, corresponding to steady rotation about the vertical axis.

7.3.4 Symmetry and Noether's theorem

The conservation law associated with a cyclic coordinate is a special case of a more general principle. Continuous symmetries of the Lagrangian give rise to conserved quantities.

Let
$$g^s : M \to M$$
be a one-parameter group of transformations of the configuration space. Its infinitesimal generator is the vector field
$$\mathbf{w}(\mathbf{q}) = \left.\frac{d}{ds}\right|_{s=0} g^s(\mathbf{q}).$$
The induced transformation of velocity is
$$\dot{\mathbf{q}} \longmapsto Dg^s(\mathbf{q})\dot{\mathbf{q}}.$$

Theorem 7.20 (Noether's theorem)**.** *Suppose that the Lagrangian is invariant under the one-parameter group g^s, so that*
$$L\big(t, g^s(\mathbf{q}), Dg^s(\mathbf{q})\dot{\mathbf{q}}\big) = L(t, \mathbf{q}, \dot{\mathbf{q}})$$
for s near 0. Then every solution of the Euler–Lagrange equations satisfies
$$\frac{d}{dt}\left(L_{\dot{\mathbf{q}}}(t, \mathbf{q}, \dot{\mathbf{q}}) \cdot \mathbf{w}(\mathbf{q})\right) = 0.$$
Thus the quantity
$$L_{\dot{\mathbf{q}}}(t, \mathbf{q}, \dot{\mathbf{q}}) \cdot \mathbf{w}(\mathbf{q})$$
is conserved.

Proof. Let
$$\mathbf{\Phi}(s, t) = g^s(\mathbf{q}(t)).$$
Differentiating
$$L\big(t, \mathbf{\Phi}(s,t), \mathbf{\Phi}_t(s,t)\big)$$

with respect to s at $s = 0$ gives

$$L_{\mathbf{q}} \cdot \mathbf{\Phi}_s(0,t) + L_{\dot{\mathbf{q}}} \cdot \frac{d}{dt}\mathbf{\Phi}_s(0,t) = 0.$$

Along a solution of the Euler–Lagrange equations,

$$L_{\mathbf{q}} = \frac{d}{dt}L_{\dot{\mathbf{q}}}.$$

Therefore

$$\begin{aligned} 0 &= \frac{d}{dt}L_{\dot{\mathbf{q}}} \cdot \mathbf{\Phi}_s(0,t) + L_{\dot{\mathbf{q}}} \cdot \frac{d}{dt}\mathbf{\Phi}_s(0,t) \\ &= \frac{d}{dt}\left(L_{\dot{\mathbf{q}}} \cdot \mathbf{\Phi}_s(0,t)\right). \end{aligned}$$

Since

$$\mathbf{\Phi}_s(0,t) = \mathbf{w}(\mathbf{q}(t)),$$

the result follows. □

Example 7.21 (Translation symmetry and linear momentum)**.** Consider a particle in $\mathbb{R}^n$ with Lagrangian

$$L(\mathbf{x}, \dot{\mathbf{x}}) = \frac{1}{2}m|\dot{\mathbf{x}}|^2 - V(\mathbf{x}).$$

Suppose that L is invariant under translation in the direction of a constant vector $\mathbf{a}$:

$$g^s(\mathbf{x}) = \mathbf{x} + s\mathbf{a}.$$

The infinitesimal generator is

$$\mathbf{w}(\mathbf{x}) = \mathbf{a}.$$

Noether's theorem gives

$$L_{\dot{\mathbf{x}}} \cdot \mathbf{a} = m\dot{\mathbf{x}} \cdot \mathbf{a} = \text{constant}.$$

Thus translation invariance in the direction $\mathbf{a}$ gives conservation of the corresponding component of linear momentum.

If the Lagrangian is invariant under all translations, then

$$m\dot{\mathbf{x}}$$

is conserved.

Example 7.22 (Rotation symmetry and angular momentum)**.** Consider a particle of mass m moving in $\mathbb{R}^3$ under a central potential $V(|\mathbf{x}|)$. Its Lagrangian is

$$L(\mathbf{x}, \dot{\mathbf{x}}) = \frac{1}{2}m|\dot{\mathbf{x}}|^2 - V(|\mathbf{x}|).$$

Since both $|\mathbf{x}|$ and $|\dot{\mathbf{x}}|$ are preserved by rotations, the Lagrangian is invariant under every

rotation about the origin.

Fix a unit vector $\mathbf{e}$, and let g^s be rotation through angle s about the axis $\mathbf{e}$. Its infinitesimal generator is

$$\mathbf{w}(\mathbf{x}) = \mathbf{e} \times \mathbf{x}.$$

Noether's theorem gives

$$\begin{aligned} L_{\dot{\mathbf{x}}} \cdot \mathbf{w}(\mathbf{x}) &= m\dot{\mathbf{x}} \cdot (\mathbf{e} \times \mathbf{x}) \\ &= \mathbf{e} \cdot \left(\mathbf{x} \times m\dot{\mathbf{x}}\right) = \text{constant}. \end{aligned}$$

Thus the component of angular momentum along every axis $\mathbf{e}$ is conserved. Consequently, the full angular momentum vector

$$\mathbf{L} = \mathbf{x} \times m\dot{\mathbf{x}}$$

is conserved.

Remark 7.23. A cyclic coordinate corresponds to a translation symmetry in that coordinate. Thus conservation of conjugate momentum for cyclic coordinates is already contained in Noether's theorem.

Remark 7.24. Time-translation symmetry of an autonomous Lagrangian gives conservation of

$$E = \dot{\mathbf{q}} \cdot L_{\dot{\mathbf{q}}} - L.$$

This is the system analogue of the Beltrami identity.

Exercise 7.12 (Linear and angular momentum of an interacting system)**.** Consider N particles in $\mathbb{R}^3$ with Lagrangian

$$L = \sum_{j=1}^{N} \frac{1}{2} m_j |\dot{\mathbf{x}}_j|^2 - \sum_{1 \le i < j \le N} W_{ij}(|\mathbf{x}_i - \mathbf{x}_j|).$$

Use translation and rotation invariance together with Noether's theorem to prove conservation of the total linear and angular momenta

$$\mathbf{P} = \sum_{j=1}^{N} m_j \dot{\mathbf{x}}_j, \qquad \mathbf{L} = \sum_{j=1}^{N} \mathbf{x}_j \times m_j \dot{\mathbf{x}}_j.$$

Deduce that the center of mass moves with constant velocity.

Exercise 7.13 (Motion on the hyperbolic plane)**.** In coordinates (q_1, q_2), consider the metric

$$ds^2 = dq_1^2 + e^{2q_1} dq_2^2.$$

A particle moving in this geometry under a potential $V(q_1)$ has Lagrangian

$$L = \frac{1}{2}\left(\dot{q}_1^2 + e^{2q_1}\dot{q}_2^2\right) - V(q_1).$$

1. Identify the cyclic coordinate and its conserved conjugate momentum.
2. Derive the Euler–Lagrange equation for q_1.
3. Use conservation of momentum and energy to reduce the motion of q_1 to a one-dimensional problem with an effective potential.

7.3.5 The Kepler problem

The Kepler problem brings together the main ideas developed in this section. Translation symmetry separates the center-of-mass motion, rotational symmetry gives conservation of angular momentum, and a cyclic angular coordinate reduces the relative motion to a one-dimensional radial problem.

Consider two particles of masses m_1 and m_2 with positions $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^3$. Suppose that they interact through the attractive potential

$$V(\mathbf{x}_1, \mathbf{x}_2) = -\frac{k}{|\mathbf{x}_1 - \mathbf{x}_2|}, \qquad k > 0.$$

Their Lagrangian is

$$L = \frac{1}{2} m_1 |\dot{\mathbf{x}}_1|^2 + \frac{1}{2} m_2 |\dot{\mathbf{x}}_2|^2 + \frac{k}{|\mathbf{x}_1 - \mathbf{x}_2|}.$$

Center-of-mass and relative coordinates Introduce

$$M = m_1 + m_2, \qquad \mu = \frac{m_1 m_2}{m_1 + m_2},$$

and

$$\mathbf{R} = \frac{m_1 \mathbf{x}_1 + m_2 \mathbf{x}_2}{M}, \qquad \mathbf{r} = \mathbf{x}_1 - \mathbf{x}_2.$$

The inverse relations are

$$\mathbf{x}_1 = \mathbf{R} + \frac{m_2}{M}\mathbf{r}, \qquad \mathbf{x}_2 = \mathbf{R} - \frac{m_1}{M}\mathbf{r}.$$

Differentiating and substituting into the kinetic energy gives

$$\frac{1}{2} m_1 |\dot{\mathbf{x}}_1|^2 + \frac{1}{2} m_2 |\dot{\mathbf{x}}_2|^2 = \frac{1}{2} M |\dot{\mathbf{R}}|^2 + \frac{1}{2} \mu |\dot{\mathbf{r}}|^2.$$

Hence

$$L = \frac{1}{2} M |\dot{\mathbf{R}}|^2 + L_{\mathrm{rel}}(\mathbf{r}, \dot{\mathbf{r}}),$$

where

$$L_{\mathrm{rel}} = \frac{1}{2} \mu |\dot{\mathbf{r}}|^2 + \frac{k}{|\mathbf{r}|}.$$

Since $\mathbf{R}$ is cyclic,

$$M\dot{\mathbf{R}}$$

is conserved, and the center of mass moves uniformly. The remaining relative motion is that of a particle of mass μ in the central potential $-k/r$.

Reduction by rotational symmetry The reduced Lagrangian

$$L_{\mathrm{rel}}(\mathbf{r}, \dot{\mathbf{r}}) = \frac{1}{2}\mu|\dot{\mathbf{r}}|^2 + \frac{k}{|\mathbf{r}|}$$

is invariant under rotations. By Noether's theorem, the angular momentum

$$\mathbf{L} = \mu \mathbf{r} \times \dot{\mathbf{r}}$$

is conserved.

Since $\mathbf{r}(t)$ and $\dot{\mathbf{r}}(t)$ are perpendicular to the fixed vector $\mathbf{L}$, the relative motion lies in a fixed plane whenever $\mathbf{L} \neq \mathbf{0}$. If $\mathbf{L} = \mathbf{0}$, the relative motion is purely radial. In what follows, we assume $\ell = |\mathbf{L}| > 0$. We may therefore introduce polar coordinates (r, θ) in that plane. The reduced Lagrangian becomes

$$L_{\mathrm{rel}}(r, \theta, \dot{r}, \dot{\theta}) = \frac{1}{2}\mu\left(\dot{r}^2 + r^2\dot{\theta}^2\right) + \frac{k}{r}.$$

The angle θ is cyclic, so its conjugate momentum

$$\ell = \frac{\partial L_{\mathrm{rel}}}{\partial \dot{\theta}} = \mu r^2 \dot{\theta}$$

is conserved. Here $\ell = |\mathbf{L}|$ is the magnitude of the angular momentum. The conserved energy is

$$E = \frac{1}{2}\mu\dot{r}^2 + \frac{1}{2}\mu r^2\dot{\theta}^2 - \frac{k}{r}.$$

Using

$$\dot{\theta} = \frac{\ell}{\mu r^2},$$

we obtain the one-dimensional energy equation

$$\frac{1}{2}\mu\dot{r}^2 + V_{\mathrm{eff}}(r) = E,$$

where

$$V_{\mathrm{eff}}(r) = \frac{\ell^2}{2\mu r^2} - \frac{k}{r}$$

is the effective potential.

The first term in V_{eff} is the centrifugal barrier, while the second is the attractive gravitational potential. The effective potential has a unique minimum at

$$r_* = \frac{\ell^2}{\mu k},$$

corresponding to a circular orbit. Its energy is

$$E_* = V_{\mathrm{eff}}(r_*) = -\frac{\mu k^2}{2\ell^2}.$$

More generally, the radial motion is restricted to the region in which

$$E \geq V_{\mathrm{eff}}(r).$$

The orbit equation To determine the shape of the orbit, set

$$u(\theta) = \frac{1}{r(\theta)}.$$

The radial Euler–Lagrange equation is

$$\mu\left(\ddot{r} - r\dot{\theta}^2\right) = -\frac{k}{r^2}.$$

Since

$$\dot{\theta} = \frac{\ell}{\mu r^2} = \frac{\ell}{\mu}u^2,$$

differentiation with respect to time gives

$$\dot{r} = -\frac{\ell}{\mu}u', \qquad \ddot{r} = -\frac{\ell^2}{\mu^2}u^2u'',$$

where the prime now denotes differentiation with respect to θ. Substitution into the radial equation yields

$$u'' + u = \frac{\mu k}{\ell^2}.$$

Hence

$$u(\theta) = \frac{\mu k}{\ell^2}\left[1 + e\cos(\theta - \theta_0)\right],$$

where $e \geq 0$ and θ_0 are constants. Therefore

$$r(\theta) = \frac{p}{1 + e\cos(\theta - \theta_0)}, \qquad p = \frac{\ell^2}{\mu k}.$$

Thus every nonradial Kepler orbit is a conic section with a focus at the center of attraction.

The eccentricity is related to the energy by

$$e^2 = 1 + \frac{2E\ell^2}{\mu k^2}.$$

Indeed, the energy may be written as

$$E = \frac{\ell^2}{2\mu}\left[(u')^2 + u^2\right] - ku.$$

Substituting the preceding expression for $u(\theta)$ gives

$$E = \frac{\mu k^2}{2\ell^2}(e^2 - 1),$$

which proves the formula.

Consequently,

E	e	orbit
$E < 0$	$0 \leq e < 1$	ellipse
$E = 0$	$e = 1$	parabola
$E > 0$	$e > 1$	hyperbola

The circular orbit is the special case $e = 0$.

Kepler's laws For $E < 0$, the relative orbit is an ellipse with the center of attraction at one focus. This is Kepler's first law.

The area swept out by the relative position vector satisfies

$$\frac{dA}{dt} = \frac{1}{2} r^2 \dot{\theta} = \frac{\ell}{2\mu}.$$

Since ℓ is constant, equal areas are swept out in equal times. This is Kepler's second law.

Let a and b denote the semimajor and semiminor axes of an elliptical relative orbit. Since the area of the ellipse is πab, its period is

$$T = \frac{\pi ab}{\ell/(2\mu)} = \frac{2\pi\mu ab}{\ell}.$$

Using

$$b^2 = a^2(1 - e^2)$$

and

$$\ell^2 = \mu k a(1 - e^2),$$

we obtain

$$T^2 = \frac{4\pi^2 \mu}{k} a^3.$$

This is Kepler's third law for the two-body potential $-k/r$.

For Newtonian gravitation,

$$k = G m_1 m_2.$$

Since

$$\frac{\mu}{k} = \frac{1}{G(m_1 + m_2)},$$

Kepler's third law becomes

$$\boxed{T^2 = \frac{4\pi^2}{G(m_1 + m_2)} a^3.}$$

The solution of the two-body problem is therefore organized by two symmetries:

translation symmetry $\longrightarrow$ uniform center-of-mass motion,

rotation symmetry $\longrightarrow$ conservation of angular momentum and planar relative motion.

After these reductions, the remaining radial equation can be analyzed by the one-dimensional effective potential. The complete orbit is then recovered from the conserved angular momentum.

Homework Problems

Homework 7.1 (A bead on a rotating hoop)**.** A bead of mass m moves without friction on a circular hoop of radius R in a uniform gravitational field. The hoop rotates with constant angular velocity Ω about its vertical diameter. Let θ denote the angle measured from the downward vertical.

1. Express the bead's Cartesian position and find its kinetic and potential energies.

2. Derive the equation of motion from the Euler–Lagrange equation.

3. Find the equilibria and determine when additional equilibria bifurcate from the downward position.

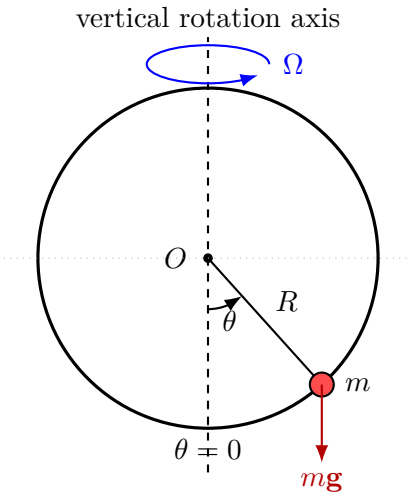


Figure 7.1: A bead moving on a vertical circular hoop that rotates with constant angular velocity Ω about its vertical diameter. The angle θ is measured from the downward vertical. The hoop is shown at one instant during its rotation.

Homework 7.2 (Small oscillations of the double pendulum)**.** Two point masses are joined by massless rods and move in a vertical plane under gravity. Starting from the double-pendulum Lagrangian derived in Example 7.18, with angles θ_1, θ_2 measured from the downward vertical, complete the following tasks.

1. Derive the nonlinear Euler–Lagrange equations. Then linearize them about the downward equilibrium $\theta_1 = \theta_2 = 0$ and write the result as

$$\mathbf{M}_0\ddot{\mathbf{q}} + \mathbf{K}_0\mathbf{q} = \mathbf{0}, \qquad \mathbf{q} = (\theta_1, \theta_2)^{\mathsf{T}}.$$

2. Show that the normal frequencies satisfy

$$\det(\mathbf{K}_0 - \omega^2\mathbf{M}_0) = 0.$$

3. When $m_1 = m_2 = m$ and $\ell_1 = \ell_2 = \ell$, find both frequencies and mode shapes. Describe the motion in each mode and explain why the higher-frequency mode has the masses moving more strongly against one another.

Homework 7.3 (Symmetry reduction of the double pendulum)**.** Consider the same planar

double pendulum with gravity removed. Write

$$A = (m_1 + m_2)\ell_1^2, \qquad B = m_2\ell_1\ell_2, \qquad C = m_2\ell_2^2.$$

Its kinetic Lagrangian is

$$L = \frac{1}{2}A\dot{\theta}_1^2 + B\cos(\theta_1 - \theta_2)\dot{\theta}_1\dot{\theta}_2 + \frac{1}{2}C\dot{\theta}_2^2.$$

Here the absence of gravity makes simultaneous rotation of both rods a continuous symmetry.

1. Use Noether's theorem to identify the conserved total angular momentum P.

2. Set
$$\phi = \theta_1 - \theta_2, \qquad \psi = \theta_2, \qquad D(\phi) = A + 2B\cos\phi + C.$$
Eliminate $\dot{\psi}$ using P and show that conservation of energy becomes
$$\frac{1}{2}\frac{AC - B^2\cos^2\phi}{D(\phi)}\dot{\phi}^2 + \frac{P^2}{2D(\phi)} = E.$$

3. Write the resulting quadrature for $\phi(t)$ and describe how its turning points depend on E and P.

Homework 7.4 (Amplitude dependence of the Duffing period)**.** Consider

$$\ddot{q} + \omega_0^2 q + \alpha q^3 = 0, \qquad q(0) = A > 0, \qquad \dot{q}(0) = 0,$$

where $\omega_0 > 0$ and $\alpha > 0$.

1. Use the energy integral and $q = Au$ to derive
$$T(A) = \frac{4}{\omega_0}\int_0^1 \frac{du}{\sqrt{(1-u^2)\left[1 + \dfrac{\alpha A^2}{2\omega_0^2}(1+u^2)\right]}}.$$

2. Expand the integrand for small A and show that
$$T(A) = \frac{2\pi}{\omega_0}\left(1 - \frac{3\alpha A^2}{8\omega_0^2} + O(A^4)\right).$$

3. Interpret the sign of the correction and explain why the oscillator is not isochronous.

Homework 7.5 (Radial collision in the Kepler problem)**.** Consider the reduced Kepler La-

grangian

$$L_{\mathrm{rel}} = \frac{1}{2}\mu|\dot{\mathbf{r}}|^2 + \frac{k}{|\mathbf{r}|}.$$

Suppose the angular momentum is zero.

1. Show that the motion remains on a fixed line and reduce the problem to an equation for $r(t) = |\mathbf{r}(t)|$.
2. For zero energy, use quadrature to show that a collision solution has the form

$$r(t) = \left(\frac{9k}{2\mu}\right)^{1/3} |t - t_*|^{2/3}.$$

3. Explain why the collision occurs in finite time and why $|\dot{r}(t)| \to \infty$ as $t \to t_*$.

Projects

Project 7.1 (A cantilever beam and natural boundary conditions)**.** A slender elastic beam occupies $0 \le x \le L$. Its small transverse deflection is $y(x)$, its bending stiffness is the constant $EI > 0$, and $f(x)$ is a transverse load per unit length. The total potential energy is modeled by

$$J[y] = \int_0^L \left(\frac{EI}{2}(y'')^2 - f(x)y\right) dx.$$

The left end is clamped, so $y(0) = y'(0) = 0$, while the right end is free.

1. Compute the first variation and integrate by parts twice, keeping all boundary terms.
2. Derive the beam equation
$$EI\, y^{(4)} = f(x)$$
and the natural conditions $y''(L) = y'''(L) = 0$ at the free end.
3. For a uniform load $f(x) = f_0$, solve the boundary-value problem and find the tip deflection $y(L)$.
4. Interpret EIy'' and EIy''' as bending moment and shear force, and explain the two free-end conditions physically.

Historical Notes and Further Reading

Historical notes. The calculus of variations grew from problems in which the unknown is a curve rather than a number. A celebrated early example is the brachistochrone problem posed by Johann Bernoulli in 1696: among all curves joining two prescribed points, determine the one along which a particle descends under gravity in the shortest time. The problem helped establish the close connection between variational questions, geometry, and mechanics.

A systematic theory was developed by Euler (1744) in his *Methodus inveniendi*. Lagrange subsequently introduced a more analytic method based on variations of the unknown curve. The differential condition now called the Euler–Lagrange equation emerged from this development.

Independence of the Lagrangian from a variable or from the independent variable leads to first integrals, represented in this chapter by conservation of conjugate momentum and the Beltrami identity.

Variational principles became a central formulation of mechanics during the eighteenth and nineteenth centuries. In the work of Hamilton (1834, 1835), the action principle was connected with a general formulation of dynamics. For mechanical systems, Hamilton's principle gives the equations of motion in generalized coordinates and makes the roles of kinetic energy, potential energy, momentum, and symmetry explicit.

The Kepler problem provides one of the earliest and most influential examples of the interaction between observation, differential equations, and conservation laws. Kepler formulated his three laws from astronomical data, while Newton showed that they follow from an inverse-square gravitational force. In the variational formulation developed here, translation and rotation symmetries explain the conservation of total linear momentum and angular momentum, and these conserved quantities reduce the two-body problem to an effectively one-dimensional radial equation.

The relation between continuous symmetry and conservation law was placed in its general form by Noether (1918). Translation symmetry gives linear momentum, rotational symmetry gives angular momentum, and time translation gives energy. In this way, the conserved quantities used to reduce differential equations are consequences of the underlying symmetry of the model.

Further reading. For a concise introduction to the calculus of variations and the Euler–Lagrange equation, see Gelfand and Fomin (2000). The connection between variational principles and classical mechanics is developed especially clearly by Lanczos (1986). A broader treatment of Lagrangian mechanics, generalized coordinates, constraints, and conservation laws is given by Goldstein et al. (2002). For the history and mathematical significance of Noether's theorem, see Kosmann-Schwarzbach (2011). The Hamiltonian formulation and phase-space geometry are continued in Chapter 8.

Chapter 8

Hamiltonian Systems and Phase-Space Geometry

The preceding chapter developed Lagrangian mechanics in terms of generalized coordinates and velocities. For a regular Lagrangian, the relation

$$\mathbf{p} = L_{\dot{\mathbf{q}}}(t, \mathbf{q}, \dot{\mathbf{q}})$$

can be inverted locally, allowing the variables $(\mathbf{q}, \dot{\mathbf{q}})$ to be replaced by the position–momentum variables $(\mathbf{q}, \mathbf{p})$. Hamiltonian mechanics begins with this change of variables and reformulates the Euler–Lagrange equations as a first-order system on phase space.

The resulting equations have the canonical form

$$\dot{\mathbf{q}} = H_{\mathbf{p}}, \qquad \dot{\mathbf{p}} = -H_{\mathbf{q}}.$$

Writing

$$\mathbf{z} = \begin{bmatrix} \mathbf{q} \\ \mathbf{p} \end{bmatrix}, \qquad \mathbf{J} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} \end{bmatrix},$$

they become

$$\dot{\mathbf{z}} = \mathbf{J}\nabla_{\mathbf{z}} H(t, \mathbf{z}).$$

The fixed matrix $\mathbf{J}$ encodes the canonical pairing of position and momentum variables.

The Hamiltonian formulation is more than a change of notation. It reveals geometric properties shared by a broad class of conservative systems. Hamiltonian flows preserve the canonical structure of phase space and, as a consequence, preserve phase-space volume. In particular, an equilibrium cannot attract an open neighborhood of initial conditions. When the motion is confined to a region of finite volume, volume preservation also leads to the recurrence phenomenon described by the Poincaré recurrence theorem.

The Hamiltonian determines the evolution not only of the phase-space variables but also of every observable through the Poisson bracket. Symmetries and first integrals can therefore be expressed directly in Hamiltonian form. When sufficiently many independent first integrals commute under the Poisson bracket, the system is integrable and its motion can often be decomposed into simpler oscillatory modes.

The examples in this chapter emphasize these structural ideas. Coupled linear oscillators illustrate normal modes, canonical transformations, and integrable motion on invariant tori. A charged

particle shows that canonical momentum need not equal mass times velocity. Point-vortex dynamics provides a nonmechanical first-order Hamiltonian model arising from fluid motion. Together these examples show how Hamiltonian structure organizes dynamics even when explicit solution formulas are unavailable.

8.1 From Lagrangian to Hamiltonian Form

8.1.1 From velocity to momentum variables

The Lagrangian formulation uses position and velocity as its basic variables:

$$(\mathbf{q}, \dot{\mathbf{q}}).$$

The conjugate momentum is defined by

$$\mathbf{p} = L_{\dot{\mathbf{q}}}(t, \mathbf{q}, \dot{\mathbf{q}}).$$

For each fixed t, consider the momentum–velocity map

$$(\mathbf{q}, \dot{\mathbf{q}}) \longmapsto (\mathbf{q}, \mathbf{p}).$$

Its Jacobian with respect to $(\mathbf{q}, \dot{\mathbf{q}})$ is the block matrix

$$\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ L_{\dot{\mathbf{q}}\mathbf{q}} & L_{\dot{\mathbf{q}}\dot{\mathbf{q}}} \end{bmatrix}.$$

Since this matrix is block triangular, it is nonsingular precisely when the velocity Hessian $L_{\dot{\mathbf{q}}\dot{\mathbf{q}}}$ is nonsingular.

Definition 8.1. A Lagrangian is called *regular* if its velocity Hessian

$$L_{\dot{\mathbf{q}}\dot{\mathbf{q}}} = \left[\frac{\partial^2 L}{\partial \dot{q}_i \, \partial \dot{q}_j} \right]$$

is nonsingular.

For a regular Lagrangian, the inverse function theorem shows that the momentum–velocity relation can be solved locally in the form

$$\dot{\mathbf{q}} = \mathbf{v}(t, \mathbf{q}, \mathbf{p}).$$

Consequently, $(\mathbf{q}, \mathbf{p})$ may be used as local coordinates on phase space, and we have the local change of variables

$$(\mathbf{q}, \dot{\mathbf{q}}) \longleftrightarrow (\mathbf{q}, \mathbf{p}).$$

For a natural mechanical system,

$$L(\mathbf{q}, \dot{\mathbf{q}}) = \frac{1}{2} \dot{\mathbf{q}}^{\mathsf{T}} \mathbf{M}(\mathbf{q}) \dot{\mathbf{q}} - V(\mathbf{q}),$$

the conjugate momentum is

$$\mathbf{p} = \mathbf{M}(\mathbf{q}) \dot{\mathbf{q}}.$$

If the mass matrix is positive definite, then it is invertible and

$$\dot{\mathbf{q}} = \mathbf{M}(\mathbf{q})^{-1}\mathbf{p}.$$

Thus every natural mechanical system with a positive definite mass matrix has a regular Lagrangian. This includes the double pendulum studied in Example 7.18; we return to that example after defining the Hamiltonian.

The next question is how to rewrite the dynamics in the variables $(\mathbf{q}, \mathbf{p})$. We seek a function whose derivatives recover both the velocity and the Euler–Lagrange equation. This leads to the Legendre transform.

8.1.2 The Legendre transform

Definition of the Hamiltonian Let

$$L = L(t, \mathbf{q}, \dot{\mathbf{q}})$$

be a regular Lagrangian. Its Hamiltonian is defined by the Legendre transform

$$H(t, \mathbf{q}, \mathbf{p}) = \mathbf{p} \cdot \dot{\mathbf{q}} - L(t, \mathbf{q}, \dot{\mathbf{q}}),$$

where the velocity on the right is replaced by

$$\dot{\mathbf{q}} = \mathbf{v}(t, \mathbf{q}, \mathbf{p}).$$

Thus H is a function of $(t, \mathbf{q}, \mathbf{p})$ rather than $(t, \mathbf{q}, \dot{\mathbf{q}})$.

The usefulness of this transformation is seen by differentiating:

$$dH = \dot{\mathbf{q}} \cdot d\mathbf{p} + \mathbf{p} \cdot d\dot{\mathbf{q}} - L_t\, dt - L_{\mathbf{q}} \cdot d\mathbf{q} - L_{\dot{\mathbf{q}}} \cdot d\dot{\mathbf{q}}.$$

Since

$$\mathbf{p} = L_{\dot{\mathbf{q}}},$$

the terms containing $d\dot{\mathbf{q}}$ cancel. Therefore

$$dH = -L_t\, dt - L_{\mathbf{q}} \cdot d\mathbf{q} + \dot{\mathbf{q}} \cdot d\mathbf{p},$$

and hence

$$H_t = -L_t, \qquad H_{\mathbf{q}} = -L_{\mathbf{q}}, \qquad H_{\mathbf{p}} = \dot{\mathbf{q}}.$$

These identities allow the Euler–Lagrange equations to be rewritten as a first-order system in $(\mathbf{q}, \mathbf{p})$.

Natural mechanical systems For a natural mechanical system,

$$L(\mathbf{q}, \dot{\mathbf{q}}) = \frac{1}{2}\dot{\mathbf{q}}^{\mathsf{T}}\mathbf{M}(\mathbf{q})\dot{\mathbf{q}} - V(\mathbf{q}),$$

where $\mathbf{M}(\mathbf{q})$ is symmetric and positive definite, we have

$$\mathbf{p} = \mathbf{M}(\mathbf{q})\dot{\mathbf{q}}, \qquad \dot{\mathbf{q}} = \mathbf{M}(\mathbf{q})^{-1}\mathbf{p}.$$

Substitution into the Legendre transform gives

$$\begin{aligned} H(\mathbf{q},\mathbf{p}) &= \mathbf{p}^{\mathsf{T}}\dot{\mathbf{q}} - \left(\frac{1}{2}\dot{\mathbf{q}}^{\mathsf{T}}\mathbf{M}(\mathbf{q})\dot{\mathbf{q}} - V(\mathbf{q})\right) \\ &= \frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{M}(\mathbf{q})^{-1}\mathbf{p} + V(\mathbf{q}). \end{aligned}$$

Thus

$$\boxed{H(\mathbf{q},\mathbf{p}) = \frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{M}(\mathbf{q})^{-1}\mathbf{p} + V(\mathbf{q}).}$$

The Hamiltonian is the total mechanical energy expressed in position and momentum variables.

Example 8.2 (Hamiltonian of the double pendulum)**.** Consider the double pendulum introduced in Example 7.18. Write

$$\mathbf{q} = \begin{bmatrix}\theta_1\\ \theta_2\end{bmatrix}, \qquad \mathbf{p} = \begin{bmatrix}p_1\\ p_2\end{bmatrix}, \qquad \delta = \theta_1 - \theta_2.$$

Its mass matrix is

$$\mathbf{M}(\mathbf{q}) = \begin{bmatrix}(m_1+m_2)\ell_1^2 & m_2\ell_1\ell_2\cos\delta \\ m_2\ell_1\ell_2\cos\delta & m_2\ell_2^2\end{bmatrix}.$$

Adding an irrelevant constant to the potential energy, we write

$$V(\theta_1,\theta_2) = (m_1+m_2)g\ell_1(1-\cos\theta_1) + m_2g\ell_2(1-\cos\theta_2).$$

The determinant of the mass matrix is

$$\det\mathbf{M}(\mathbf{q}) = m_2\ell_1^2\ell_2^2\left(m_1 + m_2\sin^2\delta\right) > 0.$$

Thus the momentum–velocity relation is invertible, with

$$\mathbf{M}(\mathbf{q})^{-1} = \frac{1}{m_2\ell_1^2\ell_2^2\left(m_1+m_2\sin^2\delta\right)}\begin{bmatrix}m_2\ell_2^2 & -m_2\ell_1\ell_2\cos\delta \\ -m_2\ell_1\ell_2\cos\delta & (m_1+m_2)\ell_1^2\end{bmatrix}.$$

Consequently,

$$H(\mathbf{q},\mathbf{p}) = \frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{M}(\mathbf{q})^{-1}\mathbf{p} + V(\mathbf{q}).$$

Explicitly,

$$\boxed{\begin{aligned} H(\theta_1,\theta_2,p_1,p_2) = {} & \frac{m_2\ell_2^2p_1^2 - 2m_2\ell_1\ell_2\cos(\theta_1-\theta_2)p_1p_2 + (m_1+m_2)\ell_1^2p_2^2}{2m_2\ell_1^2\ell_2^2\left[m_1+m_2\sin^2(\theta_1-\theta_2)\right]} \\ & + (m_1+m_2)g\ell_1(1-\cos\theta_1) + m_2g\ell_2(1-\cos\theta_2). \end{aligned}}$$

The mixed term p_1p_2 and the dependence on $\theta_1-\theta_2$ show that the pendula remain coupled in Hamiltonian variables. The Hamiltonian formulation does not remove the coupling; it expresses it through the position-dependent kinetic energy.

8.1.3 Hamilton's equations

We now use the Legendre transform to rewrite the Euler–Lagrange equations as a first-order system in the phase-space variables $(\mathbf{q}, \mathbf{p})$.

From the differential identity

$$dH = -L_t\, dt - L_{\mathbf{q}} \cdot d\mathbf{q} + \dot{\mathbf{q}} \cdot d\mathbf{p},$$

we have

$$H_{\mathbf{q}} = -L_{\mathbf{q}}, \qquad H_{\mathbf{p}} = \dot{\mathbf{q}}.$$

Since

$$\mathbf{p} = L_{\dot{\mathbf{q}}},$$

the Euler–Lagrange equations give

$$\dot{\mathbf{p}} = \frac{d}{dt} L_{\dot{\mathbf{q}}} = L_{\mathbf{q}} = -H_{\mathbf{q}}.$$

We have therefore proved the following equivalence.

Principle 8.1 (Hamilton's equations)**.** *Let $L(t, \mathbf{q}, \dot{\mathbf{q}})$ be a regular Lagrangian, and let*

$$H(t, \mathbf{q}, \mathbf{p}) = \mathbf{p} \cdot \dot{\mathbf{q}} - L(t, \mathbf{q}, \dot{\mathbf{q}})$$

be its Legendre transform. Under the change of variables

$$\mathbf{p} = L_{\dot{\mathbf{q}}}(t, \mathbf{q}, \dot{\mathbf{q}}),$$

the Euler–Lagrange equations are equivalent to

$$\boxed{\dot{\mathbf{q}} = H_{\mathbf{p}}(t, \mathbf{q}, \mathbf{p}), \qquad \dot{\mathbf{p}} = -H_{\mathbf{q}}(t, \mathbf{q}, \mathbf{p}).}$$

Thus the Lagrangian and Hamiltonian formulations describe the same dynamics but organize it differently. The Euler–Lagrange equations are second-order equations in configuration space, whereas Hamilton's equations form a first-order system on phase space.

Remark 8.3 (Time reversibility). For an autonomous mechanical Hamiltonian of the form

$$H(\mathbf{q}, \mathbf{p}) = \frac{1}{2} \mathbf{p}^{\mathsf{T}} \mathbf{M}(\mathbf{q})^{-1} \mathbf{p} + V(\mathbf{q}),$$

the equations are invariant under reversal of time together with reversal of momentum. More precisely, if

$$(\mathbf{q}(t), \mathbf{p}(t))$$

is a solution, then

$$\big(\mathbf{q}(-t), -\mathbf{p}(-t)\big)$$

is also a solution. This follows from the symmetry

$$H(\mathbf{q}, -\mathbf{p}) = H(\mathbf{q}, \mathbf{p}).$$

Note that time reversibility is not a defining property of Hamiltonian systems: a general Hamiltonian need not have this momentum-reversal symmetry, and a time-reversible differential equation need not be Hamiltonian.

8.1.4 Canonical form of Hamilton's equations

Introduce the phase-space variable

$$\mathbf{z} = \begin{bmatrix} \mathbf{q} \\ \mathbf{p} \end{bmatrix} \in \mathbb{R}^{2n}$$

and the canonical matrix

$$\mathbf{J} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} \end{bmatrix}.$$

Since

$$\nabla_{\mathbf{z}} H = \begin{bmatrix} H_{\mathbf{q}} \\ H_{\mathbf{p}} \end{bmatrix},$$

Hamilton's equations take the compact form

$$\boxed{\dot{\mathbf{z}} = \mathbf{J} \nabla_{\mathbf{z}} H(t, \mathbf{z}).}$$

The matrix $\mathbf{J}$ satisfies

$$\mathbf{J}^{\mathsf{T}} = -\mathbf{J}, \qquad \mathbf{J}^2 = -\mathbf{I}, \qquad \mathbf{J}^{-1} = -\mathbf{J}.$$

It is called the *canonical symplectic matrix*. Its skew symmetry is responsible for several basic properties of Hamiltonian systems.

For example, along a solution,

$$\begin{aligned} \frac{d}{dt} H(t, \mathbf{z}(t)) &= H_t + \nabla_{\mathbf{z}} H \cdot \dot{\mathbf{z}} \\ &= H_t + \nabla_{\mathbf{z}} H^{\mathsf{T}} \mathbf{J} \nabla_{\mathbf{z}} H \\ &= H_t, \end{aligned}$$

because

$$\mathbf{u}^{\mathsf{T}} \mathbf{J} \mathbf{u} = 0$$

for every $\mathbf{u} \in \mathbb{R}^{2n}$. Thus an autonomous Hamiltonian is conserved, whereas a time-dependent Hamiltonian need not be.

The following examples illustrate two features of the canonical formulation. The first shows how explicit time dependence permits an exchange of energy with an external forcing. The second shows that the canonical momentum need not equal the mechanical momentum when the Lagrangian depends linearly on velocity.

Example 8.4 (A time-dependent force)**.** Consider

$$H(t, q, p) = \frac{p^2}{2m} + V(q) - f(t) q.$$

Hamilton's equations are

$$\dot{q} = \frac{p}{m}, \qquad \dot{p} = -V'(q) + f(t),$$

and therefore

$$m\ddot{q} = -V'(q) + f(t).$$

Moreover,

$$\frac{dH}{dt} = H_t = -f'(t)q.$$

The prescribed force may therefore exchange energy with the mechanical system.

Example 8.5 (A charged particle in an electromagnetic field)**.** Let a particle of mass m and charge e move in electromagnetic fields described by a scalar potential $\phi(t, \mathbf{x})$ and a vector potential $\mathbf{A}(t, \mathbf{x})$:

$$\mathbf{E} = -\nabla\phi - \mathbf{A}_t, \qquad \mathbf{B} = \nabla \times \mathbf{A}.$$

The Lagrangian is

$$L(t, \mathbf{x}, \dot{\mathbf{x}}) = \frac{1}{2}m|\dot{\mathbf{x}}|^2 + e\mathbf{A}(t, \mathbf{x}) \cdot \dot{\mathbf{x}} - e\phi(t, \mathbf{x}).$$

Its conjugate momentum is

$$\boxed{\mathbf{p} = m\dot{\mathbf{x}} + e\mathbf{A}(t, \mathbf{x}).}$$

Thus the canonical momentum $\mathbf{p}$ differs from the mechanical momentum $m\dot{\mathbf{x}}$. Solving for the velocity gives

$$\dot{\mathbf{x}} = \frac{1}{m}\left(\mathbf{p} - e\mathbf{A}(t, \mathbf{x})\right).$$

The Legendre transform gives

$$\boxed{H(t, \mathbf{x}, \mathbf{p}) = \frac{1}{2m}\left|\mathbf{p} - e\mathbf{A}(t, \mathbf{x})\right|^2 + e\phi(t, \mathbf{x}).}$$

Hamilton's first equation recovers the velocity:

$$\dot{\mathbf{x}} = H_{\mathbf{p}} = \frac{1}{m}(\mathbf{p} - e\mathbf{A}).$$

The second equation gives

$$\dot{\mathbf{p}} = e(D_{\mathbf{x}}\mathbf{A})^{\mathsf{T}}\dot{\mathbf{x}} - e\nabla\phi.$$

Differentiating

$$\mathbf{p} = m\dot{\mathbf{x}} + e\mathbf{A}(t, \mathbf{x})$$

and combining the result with the preceding equation yields

$$\begin{aligned} m\ddot{\mathbf{x}} &= e\left[-\nabla\phi - \mathbf{A}_t + \left((D_{\mathbf{x}}\mathbf{A})^{\mathsf{T}} - D_{\mathbf{x}}\mathbf{A}\right)\dot{\mathbf{x}}\right] \\ &= e\left(\mathbf{E} + \dot{\mathbf{x}} \times \mathbf{B}\right). \end{aligned}$$

Thus Hamilton's equations recover the Lorentz force law.

The magnetic part of the force does no work because

$$\dot{\mathbf{x}} \cdot (\dot{\mathbf{x}} \times \mathbf{B}) = 0.$$

For static fields, the Hamiltonian is conserved and may be written as

$$H = \frac{1}{2}m|\dot{\mathbf{x}}|^2 + e\phi(\mathbf{x}).$$

For time-dependent fields, the system is still Hamiltonian, but H need not be conserved.

Remark 8.6 (Gauge freedom). The potentials are not unique. If

$$\mathbf{A}' = \mathbf{A} + \nabla\chi, \qquad \phi' = \phi - \chi_t,$$

then $\mathbf{E}$ and $\mathbf{B}$ are unchanged. The canonical momentum changes to

$$\mathbf{p}' = \mathbf{p} + e\nabla\chi,$$

but the mechanical momentum remains invariant:

$$\mathbf{p}' - e\mathbf{A}' = \mathbf{p} - e\mathbf{A} = m\dot{\mathbf{x}}.$$

Thus canonical momentum depends on the choice of potentials, whereas the mechanical momentum and the physical trajectory do not.

8.1.5 Two coupled linear oscillators

Coupled spring–mass systems and their normal modes were studied in Section 5.1.1. We now revisit the two-degree-of-freedom system to show that normal-mode coordinates also preserve the canonical position–momentum structure.

Consider

$$\mathbf{M}\ddot{\mathbf{q}} + \mathbf{K}\mathbf{q} = \mathbf{0}, \qquad \mathbf{q} \in \mathbb{R}^2,$$

where $\mathbf{M}$ and $\mathbf{K}$ are symmetric positive definite matrices. For two masses connected to fixed walls and coupled by a spring, these matrices have the form

$$\mathbf{M} = \begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix}, \qquad \mathbf{K} = \begin{bmatrix} k_1 + k_{12} & -k_{12} \\ -k_{12} & k_2 + k_{12} \end{bmatrix}.$$

The Lagrangian is

$$L(\mathbf{q}, \dot{\mathbf{q}}) = \frac{1}{2}\dot{\mathbf{q}}^{\mathsf{T}}\mathbf{M}\dot{\mathbf{q}} - \frac{1}{2}\mathbf{q}^{\mathsf{T}}\mathbf{K}\mathbf{q}.$$

Hence

$$\mathbf{p} = \mathbf{M}\dot{\mathbf{q}},$$

and the Hamiltonian is

$$H(\mathbf{q}, \mathbf{p}) = \frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{M}^{-1}\mathbf{p} + \frac{1}{2}\mathbf{q}^{\mathsf{T}}\mathbf{K}\mathbf{q}.$$

Normal modes The normal modes satisfy the generalized eigenvalue problem

$$\mathbf{K}\mathbf{v}_j = \omega_j^2 \mathbf{M}\mathbf{v}_j, \qquad j = 1, 2.$$

Since $\mathbf{M}$ and $\mathbf{K}$ are symmetric positive definite, the frequencies ω_1, ω_2 are positive, and the eigenvectors may be chosen $\mathbf{M}$-orthonormal:

$$\mathbf{v}_i^{\mathsf{T}} \mathbf{M} \mathbf{v}_j = \delta_{ij}.$$

Let

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 \end{bmatrix}.$$

Then

$$\mathbf{V}^{\mathsf{T}} \mathbf{M} \mathbf{V} = \mathbf{I}, \qquad \mathbf{V}^{\mathsf{T}} \mathbf{K} \mathbf{V} = \mathbf{\Omega}^2,$$

where

$$\mathbf{\Omega}^2 = \begin{bmatrix} \omega_1^2 & 0 \\ 0 & \omega_2^2 \end{bmatrix}.$$

Canonical modal coordinates Introduce the modal variables by

$$\mathbf{q} = \mathbf{V}\mathbf{Q}, \qquad \mathbf{p} = \mathbf{V}^{-\mathsf{T}} \mathbf{P}.$$

They preserve the position–momentum pairing because

$$\mathbf{p}^{\mathsf{T}} d\mathbf{q} = \mathbf{P}^{\mathsf{T}} d\mathbf{Q}.$$

Thus the transformation is canonical. In the new variables,

$$\begin{aligned} H &= \frac{1}{2} \mathbf{P}^{\mathsf{T}} \mathbf{P} + \frac{1}{2} \mathbf{Q}^{\mathsf{T}} \mathbf{\Omega}^2 \mathbf{Q} \\ &= \frac{1}{2} \left(P_1^2 + \omega_1^2 Q_1^2 \right) + \frac{1}{2} \left(P_2^2 + \omega_2^2 Q_2^2 \right). \end{aligned}$$

Hamilton's equations therefore separate into

$$\dot{Q}_j = P_j, \qquad \dot{P}_j = -\omega_j^2 Q_j, \qquad j = 1, 2.$$

Equivalently,

$$\ddot{Q}_j + \omega_j^2 Q_j = 0.$$

For initial data

$$\mathbf{q}(0) = \mathbf{q}_0, \qquad \mathbf{p}(0) = \mathbf{p}_0,$$

the initial modal variables are

$$Q_j(0) = \mathbf{v}_j^{\mathsf{T}} \mathbf{M} \mathbf{q}_0, \qquad P_j(0) = \mathbf{v}_j^{\mathsf{T}} \mathbf{p}_0.$$

Thus

$$Q_j(t) = Q_j(0) \cos(\omega_j t) + \frac{P_j(0)}{\omega_j} \sin(\omega_j t),$$

and

$$P_j(t) = -\omega_j Q_j(0) \sin(\omega_j t) + P_j(0) \cos(\omega_j t).$$

The original variables are recovered from

$$\mathbf{q} = \mathbf{V}\mathbf{Q}, \qquad \mathbf{p} = \mathbf{V}^{-\mathsf{T}} \mathbf{P}.$$

Modal energies Each modal energy

$$E_j = \frac{1}{2}\left(P_j^2 + \omega_j^2 Q_j^2\right), \qquad j = 1, 2,$$

is conserved, and

$$H = E_1 + E_2.$$

Thus the canonical transformation separates both the dynamics and the energy into two independent oscillatory modes. The simultaneous level sets of E_1 and E_2 are invariant tori when both modal energies are positive. We return to their dynamics and to the notion of integrability later in this chapter.

8.2 Canonical Structure of Hamiltonian Flows

We now turn from the construction of Hamilton's equations to the structure of the flow they generate. Using the phase-space variable $\mathbf{z}$ and the canonical matrix $\mathbf{J}$ introduced above, define the Hamiltonian vector field by

$$\mathbf{X}_H(t, \mathbf{z}) = \mathbf{J}\nabla_{\mathbf{z}} H(t, \mathbf{z}).$$

Hamilton's equations are therefore

$$\dot{\mathbf{z}} = \mathbf{X}_H(t, \mathbf{z}).$$

Hamiltonian dynamics has three fundamental structural properties:

- the flow preserves the canonical pairing of position and momentum;
- an autonomous Hamiltonian is constant along each trajectory;
- the flow preserves phase-space volume.

The first property is the strongest. It governs how infinitesimal variations in position and momentum are transported by the flow. Conservation of H constrains individual trajectories, while volume preservation describes the evolution of families of initial conditions.

8.2.1 Canonical pairing and canonical transformations

For

$$\mathbf{u} = \begin{bmatrix} \mathbf{u}_q \\ \mathbf{u}_p \end{bmatrix}, \qquad \mathbf{v} = \begin{bmatrix} \mathbf{v}_q \\ \mathbf{v}_p \end{bmatrix} \in \mathbb{R}^{2n},$$

define

$$\omega(\mathbf{u}, \mathbf{v}) = \mathbf{u}^{\mathsf{T}} \mathbf{J} \mathbf{v} = \mathbf{u}_q \cdot \mathbf{v}_p - \mathbf{u}_p \cdot \mathbf{v}_q.$$

This skew-symmetric bilinear pairing is called the *canonical symplectic form*. Unlike an inner product, it does not measure lengths or angles. Instead, it records oriented area in each conjugate coordinate plane (q_j, p_j).

Preservation by Hamiltonian flows Consider the possibly time-dependent Hamiltonian system

$$\dot{\mathbf{z}} = \mathbf{J}\nabla_{\mathbf{z}} H(t, \mathbf{z}),$$

and let

$$\varphi_{t,t_0}(\mathbf{z}_0)$$

denote its flow from time t_0 to time t. The derivative

$$\mathbf{F}(t) = D_{\mathbf{z}_0}\varphi_{t,t_0}(\mathbf{z}_0)$$

satisfies the variational equation

$$\dot{\mathbf{F}} = \mathbf{A}(t)\mathbf{F}, \qquad \mathbf{F}(t_0) = \mathbf{I},$$

where

$$\mathbf{A}(t) = \mathbf{J}D_{\mathbf{z}}^2 H\big(t, \varphi_{t,t_0}(\mathbf{z}_0)\big).$$

Because $D_{\mathbf{z}}^2 H$ is symmetric,

$$\mathbf{A}(t)^{\mathsf{T}}\mathbf{J} + \mathbf{J}\mathbf{A}(t) = \mathbf{0}.$$

It follows that

$$\begin{aligned}\frac{d}{dt}\left(\mathbf{F}^{\mathsf{T}}\mathbf{J}\mathbf{F}\right) &= \mathbf{F}^{\mathsf{T}}\left(\mathbf{A}^{\mathsf{T}}\mathbf{J} + \mathbf{J}\mathbf{A}\right)\mathbf{F} \\ &= \mathbf{0}.\end{aligned}$$

Since $\mathbf{F}(t_0) = \mathbf{I}$, we obtain

$$\boxed{\mathbf{F}(t)^{\mathsf{T}}\mathbf{J}\mathbf{F}(t) = \mathbf{J}.}$$

Principle 8.2 (Canonical structure of Hamiltonian flows)**.** *The flow of a Hamiltonian system preserves the canonical symplectic form:*

$$\omega\left(D\varphi_{t,t_0}(\mathbf{z})\mathbf{u}, D\varphi_{t,t_0}(\mathbf{z})\mathbf{v}\right) = \omega(\mathbf{u}, \mathbf{v}).$$

Equivalently,

$$D\varphi_{t,t_0}(\mathbf{z})^{\mathsf{T}}\mathbf{J}D\varphi_{t,t_0}(\mathbf{z}) = \mathbf{J}.$$

A flow with this property is called *symplectic*. Symplectic preservation implies phase-space volume preservation, but the converse is false. We develop the volume-preservation statement separately in Liouville's theorem.

Example 8.7 (Shear and rotation in the phase plane)**.** For one degree of freedom,

$$\omega(\mathbf{u}, \mathbf{v}) = u_q v_p - u_p v_q$$

is the signed area of the parallelogram spanned by $\mathbf{u}$ and $\mathbf{v}$ in the (q, p)-plane. Thus a symplectic flow preserves oriented phase-plane area.

For a free particle,

$$H(q, p) = \frac{p^2}{2m},$$

Hamilton's equations give

$$\dot{q} = \frac{p}{m}, \qquad \dot{p} = 0.$$

Hence

$$\begin{bmatrix} q(t) \\ p(t) \end{bmatrix} = \begin{bmatrix} 1 & t/m \\ 0 & 1 \end{bmatrix} \begin{bmatrix} q(0) \\ p(0) \end{bmatrix}.$$

The flow shears a region in the phase plane but does not change its area. Indeed,

$$\begin{bmatrix} 1 & 0 \\ t/m & 1 \end{bmatrix} \mathbf{J} \begin{bmatrix} 1 & t/m \\ 0 & 1 \end{bmatrix} = \mathbf{J}.$$

For the harmonic oscillator,

$$H(q,p) = \frac{1}{2}p^2 + \frac{1}{2}\omega^2 q^2,$$

the flow is

$$\begin{bmatrix} q(t) \\ p(t) \end{bmatrix} = \begin{bmatrix} \cos\omega t & \dfrac{\sin\omega t}{\omega} \\ -\omega \sin\omega t & \cos\omega t \end{bmatrix} \begin{bmatrix} q(0) \\ p(0) \end{bmatrix}.$$

In the scaled variables

$$Q = \sqrt{\omega}\, q, \qquad P = \frac{p}{\sqrt{\omega}},$$

this flow is an ordinary rotation. Thus phase-plane regions rotate around the origin without changing their oriented area.

These examples illustrate two basic kinds of symplectic motion: shear and rotation. A general Hamiltonian flow may stretch, bend, and mix different coordinate directions, but it preserves the canonical symplectic pairing. In a $2n$-dimensional phase space, this means that

$$\omega(\mathbf{u}, \mathbf{v})$$

is unchanged when both tangent vectors are transported by the flow. The flow need not preserve the ordinary area of each fixed coordinate plane, because different position and momentum directions may be mixed.

For one degree of freedom, symplectic preservation is equivalent to orientation-preserving area preservation. In higher dimensions it is stronger than preservation of total volume.

Canonical transformations A smooth change of phase-space variables

$$\mathbf{z} = \mathbf{\Phi}(\mathbf{Z})$$

is called a *canonical transformation* if

$$D\mathbf{\Phi}(\mathbf{Z})^{\mathsf{T}} \mathbf{J} D\mathbf{\Phi}(\mathbf{Z}) = \mathbf{J}.$$

If the transformation is linear,

$$\mathbf{z} = \mathbf{C}\mathbf{Z},$$

this condition becomes

$$\mathbf{C}^{\mathsf{T}} \mathbf{J} \mathbf{C} = \mathbf{J}.$$

A matrix satisfying this identity is called a *symplectic matrix*.

Canonical transformations preserve the form of Hamilton's equations. If

$$K(\mathbf{Z}) = H(\mathbf{\Phi}(\mathbf{Z})),$$

then

$$\dot{\mathbf{Z}} = \mathbf{J}\nabla_{\mathbf{Z}} K(\mathbf{Z}).$$

Thus a canonical transformation may simplify or decouple a Hamiltonian without destroying its position–momentum structure. The modal transformation in Section 8.1.5 is an important example.

Example 8.8 (A canonical change of configuration coordinates)**.** Let $\mathbf{V}$ be an invertible $n \times n$ matrix and introduce new configuration coordinates by

$$\mathbf{q} = \mathbf{V}\mathbf{Q}.$$

To preserve the position–momentum pairing, transform the momentum by

$$\mathbf{p} = \mathbf{V}^{-\mathsf{T}}\mathbf{P}.$$

Indeed,

$$\mathbf{p}^{\mathsf{T}} d\mathbf{q} = \mathbf{P}^{\mathsf{T}}\mathbf{V}^{-1}\mathbf{V}\, d\mathbf{Q} = \mathbf{P}^{\mathsf{T}} d\mathbf{Q}.$$

Equivalently, the phase-space transformation matrix is

$$\mathbf{C} = \begin{bmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}^{-\mathsf{T}} \end{bmatrix},$$

and it satisfies

$$\mathbf{C}^{\mathsf{T}}\mathbf{J}\mathbf{C} = \mathbf{J}.$$

Thus every invertible linear change of configuration coordinates becomes canonical when the momentum is transformed by the inverse transpose.

This is precisely the transformation used in Section 8.1.5 to pass from physical coordinates to normal-mode coordinates.

Exercise 8.1 (A harmonic oscillator in nonlinear coordinates)**.** Consider

$$L(q, \dot{q}) = \frac{\dot{q}^2}{2(1+q^2)} - \frac{\omega^2}{2}\big(\operatorname{arsinh} q\big)^2, \qquad \omega > 0.$$

Here,

$$\operatorname{arsinh} q = \log\left(q + \sqrt{1+q^2}\right)$$

is the inverse hyperbolic sine.

1. Find the conjugate momentum p and the Hamiltonian $H(q, p)$.

2. Introduce the new configuration coordinate

$$Q = \operatorname{arsinh} q.$$

Determine the momentum P for which

$$p\,dq = P\,dQ,$$

and verify that $(q,p) \mapsto (Q,P)$ is canonical.

3. Show that, in the new variables,

$$H(Q,P) = \frac{1}{2}P^2 + \frac{1}{2}\omega^2 Q^2.$$

Use this transformation to find the solution of the original nonlinear equation for arbitrary initial data.

Example 8.9 (Interchanging position and momentum)**.** Consider

$$\mathbf{q} = -\mathbf{P}, \qquad \mathbf{p} = \mathbf{Q}.$$

Equivalently,

$$\begin{bmatrix} \mathbf{q} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{Q} \\ \mathbf{P} \end{bmatrix}.$$

The transformation matrix is $-\mathbf{J}$ and satisfies

$$(-\mathbf{J})^{\mathsf{T}}\mathbf{J}(-\mathbf{J}) = \mathbf{J}.$$

Hence the transformation is canonical. The sign is essential: directly interchanging position and momentum without a sign change is not canonical.

The free-particle flow in Example 8.7 above is a *position shear*:

$$\mathbf{Q} = \mathbf{q} + \mathbf{B}\mathbf{p}, \qquad \mathbf{P} = \mathbf{p},$$

with $\mathbf{B} = (t/m)\mathbf{I}$. The following exercise studies the complementary shear in the momentum variable. It arises, for example, from the Hamiltonian flow generated by a quadratic potential.

Exercise 8.2 (A momentum shear)**.** Let $\mathbf{C}$ be a real $n \times n$ matrix and consider

$$\mathbf{Q} = \mathbf{q}, \qquad \mathbf{P} = \mathbf{p} + \mathbf{C}\mathbf{q}.$$

Show that this transformation is symplectic if and only if $\mathbf{C}$ is symmetric. Identify it as the time-t flow of a Hamiltonian

$$H(\mathbf{q}) = \frac{1}{2}\mathbf{q}^{\mathsf{T}}\mathbf{K}\mathbf{q}$$

for a suitable relation between $\mathbf{C}$, $\mathbf{K}$, and t.

8.2.2 Invariant energy surfaces

For an autonomous Hamiltonian, we have already shown that

$$H(\mathbf{z}(t)) = H(\mathbf{z}(0))$$

along every trajectory. Therefore each level set

$$\Sigma_E = \{\mathbf{z} \in \mathbb{R}^{2n} : H(\mathbf{z}) = E\}$$

is invariant under the Hamiltonian flow.

This invariance may also be seen geometrically. Since

$$\mathbf{X}_H(\mathbf{z}) = \mathbf{J}\nabla H(\mathbf{z}),$$

the skew symmetry of $\mathbf{J}$ gives

$$\nabla H(\mathbf{z}) \cdot \mathbf{X}_H(\mathbf{z}) = 0.$$

Thus the Hamiltonian vector field is tangent to the energy level through $\mathbf{z}$.

If

$$\nabla H(\mathbf{z}) \neq \mathbf{0} \qquad \text{for every } \mathbf{z} \in \Sigma_E,$$

then Σ_E is a smooth $(2n-1)$-dimensional hypersurface. When $n = 1$, it is a curve in the (q, p)-plane and often determines the trajectory geometrically. When $n > 1$, energy conservation removes only one dimension and generally does not determine the motion.

Energy conservation alone does not imply that trajectories are bounded, periodic, or explicitly solvable. Such conclusions require additional information about the energy surface or the existence of further first integrals.

8.2.3 Equilibria and linear Hamiltonian systems

Equilibria A point $\mathbf{z}_*$ is an equilibrium if

$$\mathbf{J}\nabla H(\mathbf{z}_*) = \mathbf{0}.$$

Since $\mathbf{J}$ is invertible, this is equivalent to

$$\nabla H(\mathbf{z}_*) = \mathbf{0}.$$

Thus the equilibria of a Hamiltonian system are precisely the critical points of its Hamiltonian.

Linearization Let

$$\mathbf{z} = \mathbf{z}_* + \boldsymbol{\xi},$$

where $\mathbf{z}_*$ is an equilibrium. The linearized system is

$$\dot{\boldsymbol{\xi}} = \mathbf{J}\mathbf{S}\boldsymbol{\xi}, \qquad \mathbf{S} = D^2 H(\mathbf{z}_*).$$

It is generated by the quadratic Hamiltonian

$$H_0(\boldsymbol{\xi}) = \frac{1}{2}\boldsymbol{\xi}^{\mathsf{T}}\mathbf{S}\boldsymbol{\xi}.$$

Indeed,

$$H(\mathbf{z}_* + \boldsymbol{\xi}) = H(\mathbf{z}_*) + H_0(\boldsymbol{\xi}) + O(|\boldsymbol{\xi}|^3).$$

Spectral symmetry The linearization matrix

$$\mathbf{A} = \mathbf{JS}$$

satisfies

$$\mathbf{A}^{\mathsf{T}}\mathbf{J} + \mathbf{JA} = \mathbf{0}.$$

A matrix with this property is called a *Hamiltonian matrix*. Since

$$\mathbf{J}^{-1}\mathbf{A}^{\mathsf{T}}\mathbf{J} = -\mathbf{A},$$

the matrices $\mathbf{A}^{\mathsf{T}}$ and $-\mathbf{A}$ are similar. It follows that if λ is an eigenvalue, then $-\lambda$ is also an eigenvalue. Because $\mathbf{A}$ is real, complex eigenvalues also occur in conjugate pairs. Thus the spectrum has the symmetry

$$\lambda, \qquad -\lambda, \qquad \overline{\lambda}, \qquad -\overline{\lambda}.$$

For one degree of freedom, the eigenvalues are either real or purely imaginary, including the possibility of zero eigenvalues. These cases correspond to the familiar *saddle* and *center* types encountered in the simple pendulum.

In higher dimensions, complex quartets with nonzero real and imaginary parts may occur, and oscillatory and hyperbolic directions may coexist.

For example, consider

$$H(q_1, q_2, p_1, p_2) = \frac{1}{2}(p_1^2 + p_2^2) + \frac{1}{2}\left(\omega^2 q_1^2 - \mu^2 q_2^2\right), \qquad \omega, \mu > 0.$$

Hamilton's equations separate into

$$\ddot{q}_1 + \omega^2 q_1 = 0, \qquad \ddot{q}_2 - \mu^2 q_2 = 0.$$

The eigenvalues are

$$\pm i\omega, \qquad \pm\mu.$$

Thus the origin has one oscillatory direction and one hyperbolic direction. It is an equilibrium of *center–saddle* type, a possibility that does not occur in a two-dimensional Hamiltonian system.

Another possibility in four-dimensional phase space is a genuinely complex quartet. Consider the quadratic Hamiltonian

$$H(q_1, q_2, p_1, p_2) = \alpha(q_1 p_1 + q_2 p_2) + \beta(q_1 p_2 - q_2 p_1), \qquad \alpha, \beta > 0.$$

Hamilton's equations are

$$\begin{aligned} \dot{q}_1 &= \alpha q_1 - \beta q_2, \qquad \dot{q}_2 = \beta q_1 + \alpha q_2, \\ \dot{p}_1 &= -\alpha p_1 - \beta p_2, \quad \dot{p}_2 = \beta p_1 - \alpha p_2. \end{aligned}$$

The eigenvalues are

$$\alpha \pm i\beta, \qquad -\alpha \pm i\beta.$$

They are distinct and therefore have independent eigenvectors over $\mathbb{C}$. The (q_1, q_2) components rotate while growing like $e^{\alpha t}$, whereas the (p_1, p_2) components rotate while decaying like $e^{-\alpha t}$. This example displays the full quartet symmetry of a real Hamiltonian matrix.

Positive-definite quadratic Hamiltonians Suppose that $\mathbf{S}$ is positive definite. Then

$$\mathbf{S}^{1/2}\mathbf{A}\mathbf{S}^{-1/2} = \mathbf{S}^{1/2}\mathbf{J}\mathbf{S}^{1/2}.$$

The matrix on the right is skew-symmetric:

$$\left(\mathbf{S}^{1/2}\mathbf{J}\mathbf{S}^{1/2}\right)^{\mathsf{T}} = -\mathbf{S}^{1/2}\mathbf{J}\mathbf{S}^{1/2}.$$

Therefore its eigenvalues, and hence those of $\mathbf{A}$, are purely imaginary. Moreover, the linearized flow is bounded. If $\mathbf{S}$ is only positive semidefinite, zero eigenvalues may occur.

Energy criterion for stability A strict local extremum of the Hamiltonian gives nonlinear stability. Suppose, for example, that $\mathbf{z}_*$ is a strict local minimum of H. Given a sufficiently small ball $B_r(\mathbf{z}_*)$, continuity and strict minimality imply

$$\delta = \min_{\partial B_r(\mathbf{z}_*)} \big(H(\mathbf{z}) - H(\mathbf{z}_*)\big) > 0.$$

If $\mathbf{z}(0)$ is sufficiently close to $\mathbf{z}_*$, then

$$H(\mathbf{z}(0)) - H(\mathbf{z}_*) < \delta.$$

Since H is conserved, the trajectory cannot reach $\partial B_r(\mathbf{z}_*)$. It therefore remains in $B_r(\mathbf{z}_*)$ for all times for which it is defined. The argument for a strict local maximum is analogous.

Principle 8.3 (Energy criterion for stability)**.** *If an equilibrium $\mathbf{z}_*$ is a strict local minimum or strict local maximum of an autonomous Hamiltonian, then $\mathbf{z}_*$ is stable.*

The conclusion is stability, not asymptotic stability. Since the Hamiltonian remains constant, nearby nonstationary trajectories cannot converge to an equilibrium lying on a different energy level.

Exercise 8.3 (Stability without an energy extremum)**.** Consider

$$H(q_1, q_2, p_1, p_2) = \frac{1}{2}(p_1^2 + q_1^2) - \frac{1}{2}(p_2^2 + q_2^2).$$

Show that the origin is stable even though H has neither a local minimum nor a local maximum there. Explain why this does not contradict the energy criterion for stability.

8.2.4 Liouville's theorem

Symplecticity already implies phase-space volume preservation. We now derive the same conclusion directly from the divergence of the Hamiltonian vector field. This proof also applies to time-dependent Hamiltonians.

Consider

$$\dot{\mathbf{z}} = \mathbf{X}_H(t, \mathbf{z}) = \mathbf{J}\nabla_{\mathbf{z}} H(t, \mathbf{z}).$$

Since $\mathbf{J}$ is constant,

$$D_{\mathbf{z}}\mathbf{X}_H = \mathbf{J}D_{\mathbf{z}}^2 H.$$

Therefore

$$\nabla_{\mathbf{z}} \cdot \mathbf{X}_H = \operatorname{tr}\left(\mathbf{J} D_{\mathbf{z}}^2 H\right).$$

The matrix $\mathbf{J}$ is skew-symmetric, while $D_{\mathbf{z}}^2 H$ is symmetric. The trace of their product is zero, so

$$\boxed{\nabla_{\mathbf{z}} \cdot \mathbf{X}_H = 0.}$$

Indeed, if $\mathbf{S}$ is symmetric, then

$$\operatorname{tr}(\mathbf{JS}) = \operatorname{tr}\left((\mathbf{JS})^{\mathsf{T}}\right) = \operatorname{tr}\left(\mathbf{S}^{\mathsf{T}}\mathbf{J}^{\mathsf{T}}\right) = -\operatorname{tr}(\mathbf{SJ}) = -\operatorname{tr}(\mathbf{JS}).$$

Therefore

$$\operatorname{tr}(\mathbf{JS}) = 0.$$

Taking $\mathbf{S} = D_{\mathbf{z}}^2 H$ gives

$$\nabla_{\mathbf{z}} \cdot \mathbf{X}_H = 0.$$

Let

$$\varphi_{t,t_0}(\mathbf{z}_0)$$

denote the flow from time t_0 to time t, and set

$$\mathbf{F}(t) = D_{\mathbf{z}_0}\varphi_{t,t_0}(\mathbf{z}_0).$$

The variational equation is

$$\dot{\mathbf{F}} = D_{\mathbf{z}}\mathbf{X}_H\big(t, \varphi_{t,t_0}(\mathbf{z}_0)\big)\mathbf{F}, \qquad \mathbf{F}(t_0) = \mathbf{I}.$$

We recall Jacobi's formula for the derivative of a determinant. If $\mathbf{F}(t)$ is a differentiable family of invertible matrices, then

$$\frac{d}{dt}\det \mathbf{F}(t) = \det \mathbf{F}(t)\ \operatorname{tr}\left(\mathbf{F}(t)^{-1}\dot{\mathbf{F}}(t)\right).$$

Applying Jacobi's formula to the derivative of the flow gives

$$\begin{aligned}\frac{d}{dt}\det \mathbf{F} &= \det \mathbf{F}\ \operatorname{tr}\left(\mathbf{F}^{-1}\dot{\mathbf{F}}\right)\\ &= \det \mathbf{F}\, \nabla_{\mathbf{z}} \cdot \mathbf{X}_H\big(t, \varphi_{t,t_0}(\mathbf{z}_0)\big)\\ &= 0.\end{aligned}$$

Since $\det \mathbf{F}(t_0) = 1$, we obtain

$$\boxed{\det D_{\mathbf{z}_0}\varphi_{t,t_0}(\mathbf{z}_0) = 1.}$$

Principle 8.4 (Liouville's theorem)**.** *The flow of a Hamiltonian system preserves phase-space volume. If $\Omega \subset \mathbb{R}^{2n}$ is a suitable region, then*

$$\operatorname{vol}\left(\varphi_{t,t_0}(\Omega)\right) = \operatorname{vol}(\Omega).$$

This remains true for a time-dependent Hamiltonian, even though the Hamiltonian itself need not be conserved.

Exercise 8.4 (Volume preserving but not Hamiltonian)**.** In the coordinates (q_1, q_2, p_1, p_2), consider

$$\dot{q}_1 = q_2, \qquad \dot{q}_2 = 0, \qquad \dot{p}_1 = 0, \qquad \dot{p}_2 = 0.$$

1. Show that the vector field has zero divergence and hence preserves four-dimensional volume.
2. Show that there is no function H for which these equations take the canonical Hamiltonian form
$$\dot{\mathbf{q}} = H_{\mathbf{p}}, \qquad \dot{\mathbf{p}} = -H_{\mathbf{q}}.$$
3. Explain why this example shows that the converse of Liouville's theorem is false.

No asymptotically stable equilibria Liouville's theorem implies that a Hamiltonian system cannot have an asymptotically stable equilibrium. Indeed, asymptotic stability would force a neighborhood of positive phase-space volume to enter and remain in arbitrarily small neighborhoods of the equilibrium. This contradicts preservation of phase-space volume.

This conclusion concerns attraction, not stability. An equilibrium may be stable, as occurs at a strict local extremum of the Hamiltonian, but it cannot attract a full neighborhood.

The more general notion of an attractor will be introduced in Chapter 11. There we will see that the same volume-preservation argument rules out attracting periodic orbits and other lower-dimensional attractors with open basins in Hamiltonian systems.

Contrast with dissipation Consider the damped mechanical system

$$\dot{\mathbf{q}} = \mathbf{p}, \qquad \dot{\mathbf{p}} = -\nabla V(\mathbf{q}) - \gamma \mathbf{p}, \qquad \gamma > 0.$$

Its phase-space divergence is

$$\nabla_{\mathbf{q}} \cdot \mathbf{p} + \nabla_{\mathbf{p}} \cdot (-\nabla V(\mathbf{q}) - \gamma \mathbf{p}) = -n\gamma.$$

Hence

$$\det D\varphi_t = e^{-n\gamma t},$$

so phase-space volume contracts exponentially. This contraction permits asymptotically stable equilibria and other attracting sets.

Stretching balanced by contraction Volume preservation does not prevent substantial deformation. A Hamiltonian flow may stretch in some directions while contracting in others, provided that the total phase-space volume is unchanged.

Example 8.10 (Hyperbolic stretching near an unstable equilibrium)**.** Consider

$$H(Q, P) = \frac{1}{2}P^2 - \frac{1}{2}Q^2.$$

Hamilton's equations are

$$\dot{Q} = P, \qquad \dot{P} = Q.$$

Introduce the canonical coordinates

$$q = \frac{Q+P}{\sqrt{2}}, \qquad p = \frac{P-Q}{\sqrt{2}}.$$

Then

$$H(q,p) = qp,$$

and Hamilton's equations become

$$\dot{q} = q, \qquad \dot{p} = -p.$$

Thus

$$q(t) = e^t q(0), \qquad p(t) = e^{-t} p(0).$$

The flow expands in the unstable direction and contracts by the reciprocal factor in the stable direction:

$$e^t e^{-t} = 1.$$

It therefore preserves phase-plane area despite the exponential stretching.

8.2.5 Normal modes, invariant tori, and recurrence

The two-oscillator calculation in Section 8.1.5 extends to a linear mechanical system with n degrees of freedom. After a canonical modal transformation, its Hamiltonian has the form

$$H = \sum_{j=1}^{n} E_j, \qquad E_j = \frac{1}{2}\left(P_j^2 + \omega_j^2 Q_j^2\right),$$

and the equations separate into

$$\dot{Q}_j = P_j, \qquad \dot{P}_j = -\omega_j^2 Q_j, \qquad j = 1, \ldots, n.$$

Each modal energy E_j is conserved. These functions will later be shown to commute under the Poisson bracket, providing the basic example of a Liouville-integrable system.

Invariant tori Fix positive modal energies

$$E_j = c_j > 0, \qquad j = 1, \ldots, n.$$

For each j,

$$P_j^2 + \omega_j^2 Q_j^2 = 2c_j$$

defines an ellipse, topologically equivalent to a circle. Their common level set is therefore

$$\mathcal{T}_{\mathbf{c}} = \{E_1 = c_1, \ldots, E_n = c_n\} \simeq \underbrace{S^1 \times \cdots \times S^1}_{n \text{ factors}},$$

an invariant n-dimensional torus.

Writing

$$Q_j = \frac{\sqrt{2c_j}}{\omega_j}\cos\theta_j, \qquad P_j = -\sqrt{2c_j}\sin\theta_j,$$

we obtain

$$\theta_j(t) = \theta_j(0) + \omega_j t \pmod{2\pi}.$$

The nature of the motion depends on the arithmetic relations among the frequencies. If they are commensurate, so that

$$\omega_j = m_j \omega_*, \qquad m_j \in \mathbb{N},$$

then every trajectory is periodic, with a common period

$$T = \frac{2\pi}{\omega_*}.$$

Otherwise, the motion is called *quasiperiodic*. A collection of frequencies is said to be rationally independent if

$$k_1\omega_1 + \cdots + k_n\omega_n = 0, \qquad k_j \in \mathbb{Z},$$

implies

$$k_1 = \cdots = k_n = 0.$$

If $\omega_1, \ldots, \omega_n$ are rationally independent, every trajectory is dense on the full invariant torus. It never returns exactly to its initial point, but it returns arbitrarily close infinitely often.

Intermediate cases are also possible. If the frequencies satisfy some integer relations but are not all commensurate, the motion is quasiperiodic on a lower-dimensional subtorus and is dense on that subtorus.

Poincaré recurrence Motion on an invariant torus is recurrent. A periodic trajectory returns exactly to its initial point, while a quasiperiodic trajectory returns arbitrarily close to it. A much more general recurrence principle follows from volume preservation.

Theorem 8.11 (Poincaré recurrence)**.** *Let $\Omega \subset \mathbb{R}^{2n}$ be an invariant measurable set with*

$$\mathrm{vol}(\Omega) < \infty.$$

Suppose that the Hamiltonian flow φ_t is defined on Ω for all $t \in \mathbb{R}$. Then, for almost every $\mathbf{z} \in \Omega$ and every neighborhood U of $\mathbf{z}$, there exist arbitrarily large times t such that

$$\varphi_t(\mathbf{z}) \in U.$$

Thus almost every point of Ω returns arbitrarily close to its initial position infinitely many times.

Proof. Fix a time $\tau > 0$ and consider the time-τ map

$$T = \varphi_\tau.$$

By Liouville's theorem, T preserves volume. Since the flow is defined for all positive and negative times, T is invertible.

Let $A \subset \Omega$ be measurable, and let

$$W = \left\{ \mathbf{z} \in A : T^k(\mathbf{z}) \notin A \text{ for every } k \geq 1 \right\}.$$

The sets

$$W,\ T(W),\ T^2(W), \ldots$$

are pairwise disjoint. Since T preserves volume, they all have the same volume. Because they are contained in the finite-volume set Ω, we must have

$$\operatorname{vol}(W) = 0.$$

Thus almost every point of A returns to A. The set of points of A that return only finitely many times is contained in a countable union of inverse images of W. Since T is invertible and volume preserving, each of these inverse images has volume zero. Hence almost every point of A returns to A infinitely many times.

Now choose a countable basis $\{B_j\}_{j=1}^{\infty}$ for $\mathbb{R}^{2n}$. Applying the preceding argument to each $A = B_j \cap \Omega$ and taking the union of the exceptional null sets, we find a set $N \subset \Omega$ of volume zero such that every $\mathbf{z} \in \Omega \setminus N$ returns infinitely often to every basis neighborhood containing it.

Therefore, for almost every $\mathbf{z} \in \Omega$ and every neighborhood U of $\mathbf{z}$, there exist arbitrarily large integers k such that

$$\varphi_{k\tau}(\mathbf{z}) \in U.$$

This proves the recurrence statement. □

The theorem does not say that a trajectory is periodic or that it returns exactly to its initial point. It says that, except for a set of initial conditions of volume zero, the trajectory returns arbitrarily close to its initial point infinitely many times. The theorem gives no estimate of the return time, which may be extremely large. Recurrence is therefore a consequence of finite available phase-space volume together with volume preservation, not a consequence of integrability. Periodic, quasiperiodic, and chaotic Hamiltonian trajectories may all be recurrent.

Exercise 8.5 (Why the recurrence hypotheses matter)**.** Consider the following systems:

$$\dot{q} = p, \qquad \dot{p} = 0,$$

and

$$\dot{q} = p, \qquad \dot{p} = -q - \gamma p, \qquad \gamma > 0.$$

For each system, determine whether phase-space volume is preserved and whether typical trajectories are recurrent. Explain which hypothesis of the Poincaré recurrence theorem fails.

8.3 Poisson Brackets, Symmetry, and First Integrals

Hamilton's equations describe the evolution of the phase-space point $\mathbf{z}(t)$. In applications, however, one is often interested in functions of the state, such as energy, momentum, or angular momentum. Such functions are called *observables*.

The Poisson bracket describes the evolution of observables and expresses the relation between conserved quantities and continuous symmetries. It also provides the natural language for collections of compatible first integrals and Hamiltonian integrability.

8.3.1 Evolution of observables and the Poisson bracket

Let

$$F = F(t, \mathbf{q}, \mathbf{p})$$

be a smooth observable. Along a solution of Hamilton's equations,

$$\begin{aligned}\frac{dF}{dt} &= F_t + \nabla_{\mathbf{q}} F \cdot \dot{\mathbf{q}} + \nabla_{\mathbf{p}} F \cdot \dot{\mathbf{p}} \\ &= F_t + \nabla_{\mathbf{q}} F \cdot \nabla_{\mathbf{p}} H - \nabla_{\mathbf{p}} F \cdot \nabla_{\mathbf{q}} H.\end{aligned}$$

Definition 8.12 (Poisson bracket)**.** For smooth observables F and G, define

$$\{F, G\} = \nabla_{\mathbf{q}} F \cdot \nabla_{\mathbf{p}} G - \nabla_{\mathbf{p}} F \cdot \nabla_{\mathbf{q}} G.$$

Equivalently,

$$\{F, G\} = \nabla_{\mathbf{z}} F^{\mathsf{T}} \mathbf{J} \nabla_{\mathbf{z}} G.$$

The evolution of an observable is therefore

$$\boxed{\frac{dF}{dt} = F_t + \{F, H\}.}$$

Thus Hamilton's equations describe the evolution of the phase-space point, whereas the Poisson-bracket formula describes the evolution of every observable.

Properties of the Poisson bracket The Poisson bracket is bilinear and antisymmetric:

$$\{F, G\} = -\{G, F\}.$$

It satisfies the Leibniz rule

$$\{FG, K\} = F\{G, K\} + G\{F, K\},$$

and the Jacobi identity

$$\{F, \{G, K\}\} + \{G, \{K, F\}\} + \{K, \{F, G\}\} = 0.$$

Thus, with either argument fixed, the Poisson bracket acts as a derivation with respect to multiplication.

The canonical coordinates satisfy

$$\{q_i, q_j\} = 0, \qquad \{p_i, p_j\} = 0, \qquad \{q_i, p_j\} = \delta_{ij}.$$

Consequently,

$$\dot{q}_i = \{q_i, H\}, \qquad \dot{p}_i = \{p_i, H\}.$$

If F has no explicit time dependence, then

$$\frac{dF}{dt} = \{F, H\}.$$

We therefore obtain the following characterization.

Principle 8.5 (First integrals)**.** *A time-independent observable $F(\mathbf{q}, \mathbf{p})$ is a first integral of the Hamiltonian system if and only if*

$$\{F, H\} = 0.$$

In particular, antisymmetry gives

$$\{H, H\} = 0,$$

and hence

$$\frac{dH}{dt} = H_t.$$

Thus the conservation of an autonomous Hamiltonian is the simplest example of Poisson commutation.

8.3.2 Generators and the Hamiltonian form of Noether's theorem

The Hamiltonian H generates physical time evolution through

$$\dot{\mathbf{z}} = \mathbf{X}_H(\mathbf{z}), \qquad \mathbf{X}_H = \mathbf{J}\nabla H.$$

More generally, every smooth function $G(\mathbf{q}, \mathbf{p})$ defines a Hamiltonian vector field

$$\mathbf{X}_G = \mathbf{J}\nabla G.$$

Let $\{\mathbf{\Phi}_s\}$ denote the local flow generated by $\mathbf{X}_G$:

$$\frac{d}{ds}\mathbf{\Phi}_s(\mathbf{z}) = \mathbf{X}_G\big(\mathbf{\Phi}_s(\mathbf{z})\big), \qquad \mathbf{\Phi}_0(\mathbf{z}) = \mathbf{z}.$$

The function G is called the *generator* of this one-parameter family of transformations.

For an observable F, its infinitesimal change under this transformation is

$$\begin{aligned}\delta_G F &= \left.\frac{d}{ds} F\big(\mathbf{\Phi}_s(\mathbf{z})\big)\right|_{s=0} \\ &= \nabla F^{\mathsf{T}} \mathbf{X}_G \\ &= \nabla F^{\mathsf{T}} \mathbf{J} \nabla G \\ &= \{F, G\}.\end{aligned}$$

Thus

$$\boxed{\delta_G F = \{F, G\}.}$$

The physical evolution generated by H is the special case

$$\dot{F} = \{F, H\}$$

for an observable with no explicit time dependence.

Hamiltonian form of Noether's theorem The transformations generated by G form a symmetry of the Hamiltonian if

$$H \circ \mathbf{\Phi}_s = H.$$

Differentiating with respect to s at $s = 0$ gives

$$\delta_G H = \{H, G\} = 0.$$

By antisymmetry,

$$\dot{G} = \{G, H\} = -\{H, G\} = 0.$$

Principle 8.6 (Hamiltonian form of Noether's theorem)**.** *Suppose that the transformations generated by*

$$\mathbf{X}_G = \mathbf{J}\nabla G$$

leave an autonomous Hamiltonian H invariant. Then G is conserved by the Hamiltonian flow.

The same function G therefore has two related roles: it generates a continuous transformation of phase space, and it is conserved whenever that transformation is a symmetry of the Hamiltonian.

Example 8.13 (Translations and conjugate momentum)**.** Let

$$G = p_j.$$

Then

$$\delta_G q_i = \{q_i, p_j\} = \delta_{ij}, \qquad \delta_G p_i = \{p_i, p_j\} = 0.$$

Thus p_j generates translations in the q_j direction:

$$\mathbf{\Phi}_s(\mathbf{q}, \mathbf{p}) = (\mathbf{q} + s\mathbf{e}_j, \mathbf{p}).$$

If H is independent of q_j, this translation leaves H invariant. Noether's theorem then gives

$$\dot{p}_j = 0.$$

Hence the momentum conjugate to a cyclic coordinate is both the generator of translations in that coordinate and the corresponding conserved quantity.

Example 8.14 (Rotations and angular momentum)**.** For a particle in $\mathbb{R}^3$, let

$$\mathbf{L} = \mathbf{x} \times \mathbf{p}.$$

Given a constant vector $\mathbf{a}$, define

$$G = \mathbf{a} \cdot \mathbf{L}.$$

A direct calculation gives

$$\delta_G \mathbf{x} = \{\mathbf{x}, G\} = \mathbf{a} \times \mathbf{x}, \qquad \delta_G \mathbf{p} = \{\mathbf{p}, G\} = \mathbf{a} \times \mathbf{p}.$$

Thus G generates simultaneous rotations of position and momentum about the axis $\mathbf{a}$.

For a central-force Hamiltonian

$$H(\mathbf{x}, \mathbf{p}) = \frac{1}{2m}|\mathbf{p}|^2 + V(|\mathbf{x}|),$$

these rotations leave H invariant. Therefore

$$\{H, \mathbf{a} \cdot \mathbf{L}\} = 0.$$

Since $\mathbf{a}$ is arbitrary, every component of $\mathbf{L}$ is conserved:

$$\dot{\mathbf{L}} = \mathbf{0}.$$

Thus angular momentum is both the generator of rotations and the conserved quantity associated with rotational symmetry.

Exercise 8.6 (Poisson brackets of angular momentum)**.** For a particle in $\mathbb{R}^3$, let

$$\mathbf{L} = \mathbf{x} \times \mathbf{p}.$$

Show that

$$\{L_1, L_2\} = L_3, \qquad \{L_2, L_3\} = L_1, \qquad \{L_3, L_1\} = L_2.$$

Use these identities to explain why rotations about different axes do not commute. Verify that

$$\{|\mathbf{L}|^2, L_j\} = 0, \qquad j = 1, 2, 3.$$

8.3.3 Commuting first integrals and integrability

A first integral restricts the motion to one of its level sets. If a Hamiltonian system has several functionally independent first integrals

$$F_1, \dots, F_r,$$

then every trajectory is confined to a common level set

$$F_1 = c_1, \qquad \dots, \qquad F_r = c_r.$$

At points where

$$\nabla F_1, \dots, \nabla F_r$$

are linearly independent, this common level set has codimension r.

Each F_j also generates a Hamiltonian vector field $\mathbf{X}_{F_j}$. The change of F_i along this vector field is

$$\mathbf{X}_{F_j}(F_i) = \{F_i, F_j\}.$$

Thus the condition

$$\{F_i, F_j\} = 0$$

means that the flow generated by F_j preserves the level sets of F_i. When this holds for every pair, the corresponding local Hamiltonian flows commute.

Definition 8.15. The functions $F_1, \dots, F_r$ are said to be *in involution* if

$$\{F_i, F_j\} = 0 \qquad \text{for all } i, j.$$

Closure under the Poisson bracket The Jacobi identity gives another useful consequence. If F and G are first integrals of H, then

$$\{F, H\} = 0, \qquad \{G, H\} = 0.$$

Consequently,

$$\begin{aligned} \{\{F, G\}, H\} &= \{F, \{G, H\}\} - \{G, \{F, H\}\} \\ &= 0. \end{aligned}$$

Thus $\{F, G\}$ is also a first integral. The first integrals are therefore closed under the Poisson bracket, although they need not commute with one another.

Liouville integrability A Hamiltonian system with n degrees of freedom is called *Liouville integrable* if it possesses n functionally independent first integrals

$$F_1, \dots, F_n$$

that are pairwise in involution:

$$\{F_i, F_j\} = 0, \qquad i, j = 1, \dots, n.$$

Functional independence means that

$$\nabla F_1, \dots, \nabla F_n$$

are linearly independent on the region under consideration.

The geometric significance of this definition is summarized by the Liouville–Arnold theorem. In an elementary form, it states that if a connected regular common level set

$$F_1 = c_1, \dots, F_n = c_n$$

is compact, then it is an n-dimensional torus. In suitable angular coordinates, the Hamiltonian motion on this torus becomes a linear rotation. A proof requires additional geometric machinery and will not be given here.

Example 8.16 (Decoupled oscillators)**.** For

$$H = \sum_{j=1}^{n} H_j, \qquad H_j = \frac{1}{2}\left(P_j^2 + \omega_j^2 Q_j^2\right),$$

each H_j depends only on the pair (Q_j, P_j). Hence

$$\{H_i, H_j\} = 0 \qquad \text{for all } i, j.$$

The modal energies are functionally independent wherever every mode is active. The system is therefore Liouville integrable.

For positive values

$$H_j = c_j > 0,$$

the common level set is the invariant torus described earlier, and the motion is periodic or quasiperiodic according to the arithmetic relations among the frequencies.

Example 8.17 (The Hénon–Heiles system)**.** Consider

$$H(q_1, q_2, p_1, p_2) = \frac{1}{2}(p_1^2 + p_2^2) + \frac{1}{2}(q_1^2 + q_2^2) + q_1^2 q_2 - \frac{1}{3} q_2^3.$$

This system has two degrees of freedom. Conservation of H supplies one first integral, but Liouville integrability would require a second independent first integral in involution with H.

The classical Hénon–Heiles system is nonintegrable in this sense: there is no second independent first integral of class C^1 that makes the system Liouville integrable (Llibre and Jiménez-Lara, 2011).

A Hamiltonian system may therefore preserve energy, symplectic structure, and phase-space volume while still exhibiting complicated dynamics, including chaotic motion. Nonintegrability does not mean that every trajectory is chaotic; periodic, quasiperiodic, and chaotic trajectories may coexist. The distinction between chaotic motion in conservative systems and strange attractors in dissipative systems is discussed further in Section 11.9.4.

Concluding perspective The symplectic and Poisson formulations describe the same canonical Hamiltonian structure from two complementary viewpoints. Symplecticity describes the geometry of the phase-space flow, while the Poisson bracket describes the evolution and interaction of observables. Commuting first integrals connect these two viewpoints and, when sufficiently numerous, lead to integrable motion on invariant tori.

The next section gives a nonmechanical example: point-vortex dynamics in an incompressible fluid. It retains Hamiltonian structure but uses a weighted Poisson bracket rather than the canonical matrix $\mathbf{J}$.

8.4 Point-Vortex Dynamics

Hamiltonian structure also arises in two-dimensional incompressible fluid flow. We begin with the motion of a passive tracer in a prescribed flow and then pass to the interacting dynamics of point vortices.

8.4.1 Passive tracers and stream functions

Let

$$\mathbf{u} = (u, v)$$

be a divergence-free velocity field in the plane. Locally, it can be represented by a stream function ψ:

$$u = \psi_y, \qquad v = -\psi_x.$$

Indeed,

$$\nabla \cdot \mathbf{u} = u_x + v_y = \psi_{yx} - \psi_{xy} = 0.$$

A passive tracer at position $(x(t), y(t))$ moves with the fluid:

$$\dot{x} = u(x, y), \qquad \dot{y} = v(x, y).$$

Consequently,

$$\dot{x} = \psi_y, \qquad \dot{y} = -\psi_x.$$

These are Hamilton's equations with Hamiltonian

$$H(x, y) = \psi(x, y),$$

where x and y form a conjugate pair. In particular,

$$\frac{d}{dt}\psi(x(t), y(t)) = \psi_x \dot{x} + \psi_y \dot{y} = 0.$$

Thus a passive tracer moves along a level curve of the stream function.

The scalar vorticity of the flow is

$$\omega = v_x - u_y = -\Delta\psi.$$

Exercise 8.7 (Passive tracers)**.** Let

$$\psi(x, y) = \sin x \sin y.$$

Write the equations for a passive tracer, classify the equilibria as centers or saddles, and sketch the principal level curves of ψ.

8.4.2 Hamiltonian equations for point vortices

A point vortex of strength Γ located at $\mathbf{x}_0$ is the idealized vorticity distribution

$$\omega(\mathbf{x}) = \Gamma\delta(\mathbf{x} - \mathbf{x}_0).$$

Its stream function satisfies

$$-\Delta\psi = \Gamma\delta(\mathbf{x} - \mathbf{x}_0).$$

Since the fundamental solution of $-\Delta$ in two dimensions is

$$-\frac{1}{2\pi}\log|\mathbf{x}|,$$

we obtain

$$\psi(\mathbf{x}) = -\frac{\Gamma}{2\pi}\log|\mathbf{x} - \mathbf{x}_0|.$$

The induced velocity is tangent to circles centered at $\mathbf{x}_0$, and its magnitude is

$$|\mathbf{u}(\mathbf{x})| = \frac{|\Gamma|}{2\pi|\mathbf{x} - \mathbf{x}_0|}.$$

Thus a single vortex generates a circular velocity field whose speed decreases inversely with the distance from the vortex.

Now consider N point vortices with nonzero strengths

$$\Gamma_1, \ldots, \Gamma_N$$

and positions

$$\mathbf{x}_i = (x_i, y_i), \qquad i = 1, \ldots, N.$$

The total vorticity is

$$\omega(\mathbf{x}) = \sum_{j=1}^{N} \Gamma_j \delta(\mathbf{x} - \mathbf{x}_j),$$

and the corresponding stream function is

$$\psi(\mathbf{x}) = -\frac{1}{2\pi}\sum_{j=1}^{N} \Gamma_j \log|\mathbf{x} - \mathbf{x}_j|.$$

Each vortex is transported by the velocity generated by all the other vortices. Its own singular velocity is omitted. Define

$$r_{ij} = |\mathbf{x}_i - \mathbf{x}_j| = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}.$$

The stream function acting on the ith vortex is

$$\psi_i^{\text{ext}} = -\frac{1}{2\pi} \sum_{j \neq i} \Gamma_j \log r_{ij}.$$

Therefore

$$\dot{x}_i = \frac{\partial \psi_i^{\text{ext}}}{\partial y_i}, \qquad \dot{y}_i = -\frac{\partial \psi_i^{\text{ext}}}{\partial x_i},$$

or explicitly,

$$\dot{x}_i = -\frac{1}{2\pi} \sum_{j \neq i} \Gamma_j \frac{y_i - y_j}{r_{ij}^2},$$

and

$$\dot{y}_i = \frac{1}{2\pi} \sum_{j \neq i} \Gamma_j \frac{x_i - x_j}{r_{ij}^2}.$$

The interaction Hamiltonian is

$$H = \frac{1}{2} \sum_{i=1}^{N} \Gamma_i \psi_i^{\text{ext}}.$$

The factor $1/2$ prevents each pair from being counted twice. Hence

$$\boxed{H = -\frac{1}{2\pi} \sum_{1 \leq i < j \leq N} \Gamma_i \Gamma_j \log r_{ij}.}$$

The singular self-interaction terms have been omitted because they do not contribute to the motion of the vortex centers.

Differentiating H gives

$$\frac{\partial H}{\partial y_i} = -\frac{\Gamma_i}{2\pi} \sum_{j \neq i} \Gamma_j \frac{y_i - y_j}{r_{ij}^2},$$

and

$$\frac{\partial H}{\partial x_i} = -\frac{\Gamma_i}{2\pi} \sum_{j \neq i} \Gamma_j \frac{x_i - x_j}{r_{ij}^2}.$$

Thus the vortex equations take the Hamiltonian form

$$\boxed{\Gamma_i \dot{x}_i = \frac{\partial H}{\partial y_i}, \qquad \Gamma_i \dot{y}_i = -\frac{\partial H}{\partial x_i}.}$$

The factors Γ_i show that (x_i, y_i) are not standard canonical coordinates. The corresponding weighted Poisson bracket is

$$\boxed{\{F, G\}_\Gamma = \sum_{i=1}^{N} \frac{1}{\Gamma_i} \left(F_{x_i} G_{y_i} - F_{y_i} G_{x_i} \right).}$$

With this bracket,

$$\dot{F} = \{F, H\}_\Gamma.$$

Standard canonical coordinates can be introduced by setting

$$q_i = x_i, \qquad p_i = \Gamma_i y_i.$$

Indeed,

$$\{q_i, p_j\}_\Gamma = \delta_{ij},$$

and the vortex equations become the usual canonical Hamiltonian equations in the variables $(\mathbf{q}, \mathbf{p})$.

8.4.3 Symmetries and conserved quantities

The point-vortex Hamiltonian depends only on the relative distances r_{ij}. It is therefore invariant under simultaneous translations and rotations of the entire vortex configuration.

The first moments of the vorticity are

$$X_\Gamma = \int_{\mathbb{R}^2} x\omega(\mathbf{x})\, d\mathbf{x} = \sum_{j=1}^{N} \Gamma_j x_j, \qquad Y_\Gamma = \int_{\mathbb{R}^2} y\omega(\mathbf{x})\, d\mathbf{x} = \sum_{j=1}^{N} \Gamma_j y_j.$$

Equivalently, the linear impulse vector is

$$\mathbf{P} = \begin{bmatrix} Y_\Gamma \\ -X_\Gamma \end{bmatrix}.$$

The factors Γ_j appear because each point vortex contributes the vorticity measure

$$\Gamma_j \delta(\mathbf{x} - \mathbf{x}_j).$$

Thus X_Γ and Y_Γ are the first spatial moments of the vorticity distribution.

Since the vortex Hamiltonian depends only on relative distances, it is invariant under translations. Consequently,

$$\dot{X}_\Gamma = \dot{Y}_\Gamma = 0.$$

Moreover,

$$\{x_i, Y_\Gamma\}_\Gamma = 1, \qquad \{y_i, -X_\Gamma\}_\Gamma = 1,$$

so Y_Γ generates translation in the x direction and $-X_\Gamma$ generates translation in the y direction.

The angular impulse is

$$I = \frac{1}{2} \sum_{j=1}^{N} \Gamma_j (x_j^2 + y_j^2).$$

This quantity is the second radial moment of the vorticity distribution:

$$I = \frac{1}{2} \int_{\mathbb{R}^2} |\mathbf{x}|^2 \omega(\mathbf{x})\, d\mathbf{x}.$$

Its form is also determined by the weighted Hamiltonian structure. Indeed,

$$\{x_i, I\}_\Gamma = y_i, \qquad \{y_i, I\}_\Gamma = -x_i.$$

Thus I generates simultaneous rotation of all vortex positions. The factor Γ_i reflects the strength of the ith vortex, while the factor $1/2$ removes the factor 2 produced by differentiating $x_i^2 + y_i^2$.

Rotational invariance gives

$$\{I, H\}_\Gamma = 0,$$

and therefore

$$\dot{I} = 0.$$

The two translation generators do not always commute:

$$\{X_\Gamma, Y_\Gamma\}_\Gamma = \sum_{j=1}^{N} \Gamma_j = \Gamma_T,$$

where

$$\Gamma_T = \sum_{j=1}^{N} \Gamma_j$$

is the total circulation. Thus X_Γ and Y_Γ are in involution precisely when the total circulation vanishes.

The Hamiltonian itself is also conserved:

$$\dot{H} = \{H, H\}_\Gamma = 0.$$

Thus the vortex motion preserves its interaction energy, linear impulses, and angular impulse.

8.4.4 Two-vortex motion

Consider two vortices at

$$\mathbf{x}_1 = (x_1, y_1), \qquad \mathbf{x}_2 = (x_2, y_2),$$

with nonzero strengths Γ_1 and Γ_2. Set

$$\mathbf{r} = \mathbf{x}_1 - \mathbf{x}_2, \qquad r = |\mathbf{r}|,$$

and let

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

The equations of motion are

$$\dot{\mathbf{x}}_1 = \frac{\Gamma_2}{2\pi r^2}\mathbf{R}\mathbf{r}, \qquad \dot{\mathbf{x}}_2 = -\frac{\Gamma_1}{2\pi r^2}\mathbf{R}\mathbf{r}.$$

Subtracting gives

$$\dot{\mathbf{r}} = \frac{\Gamma_1 + \Gamma_2}{2\pi r^2}\mathbf{R}\mathbf{r}.$$

Since

$$\mathbf{r} \cdot \mathbf{R}\mathbf{r} = 0,$$

we obtain

$$\frac{d}{dt} r^2 = 2\mathbf{r} \cdot \dot{\mathbf{r}} = 0.$$

Thus the distance between the two vortices remains constant.

Let

$$\Gamma_T = \Gamma_1 + \Gamma_2.$$

The form of the motion depends on whether Γ_T vanishes.

Nonzero total circulation. Suppose that

$$\Gamma_T \neq 0.$$

The center of circulation

$$\mathbf{C} = \frac{\Gamma_1 \mathbf{x}_1 + \Gamma_2 \mathbf{x}_2}{\Gamma_T}$$

is constant because

$$\Gamma_1 \dot{\mathbf{x}}_1 + \Gamma_2 \dot{\mathbf{x}}_2 = \mathbf{0}.$$

Since r is constant, the relative equation becomes

$$\dot{\mathbf{r}} = \Omega \mathbf{R} \mathbf{r}, \qquad \Omega = \frac{\Gamma_T}{2\pi r^2}.$$

Consequently,

$$\mathbf{r}(t) = e^{\Omega t \mathbf{R}} \mathbf{r}(0).$$

The vortex positions are recovered from

$$\mathbf{x}_1 = \mathbf{C} + \frac{\Gamma_2}{\Gamma_T}\mathbf{r}, \qquad \mathbf{x}_2 = \mathbf{C} - \frac{\Gamma_1}{\Gamma_T}\mathbf{r}.$$

Therefore

$$\boxed{\begin{aligned} \mathbf{x}_1(t) &= \mathbf{C} + \frac{\Gamma_2}{\Gamma_T} e^{\Omega t \mathbf{R}} \mathbf{r}(0), \\ \mathbf{x}_2(t) &= \mathbf{C} - \frac{\Gamma_1}{\Gamma_T} e^{\Omega t \mathbf{R}} \mathbf{r}(0). \end{aligned}}$$

Thus the vortices rotate rigidly about their center of circulation. The center may lie between the vortices or outside the segment joining them, depending on the signs and magnitudes of their strengths.

In particular, if

$$\Gamma_1 = \Gamma_2 = \Gamma,$$

then $\mathbf{C}$ is the midpoint and

$$\Omega = \frac{\Gamma}{\pi r^2}.$$

Zero total circulation: a vortex dipole. Suppose that

$$\Gamma_1 = \Gamma, \qquad \Gamma_2 = -\Gamma.$$

Then $\Gamma_T = 0$, and the relative equation gives

$$\dot{\mathbf{r}} = \mathbf{0}.$$

Thus the relative position remains fixed. Moreover,

$$\dot{\mathbf{x}}_1 = \dot{\mathbf{x}}_2 = -\frac{\Gamma}{2\pi r^2}\mathbf{R}\mathbf{r}.$$

Both vortices therefore translate with the same constant velocity

$$\mathbf{V} = -\frac{\Gamma}{2\pi r^2}\mathbf{R}\mathbf{r}.$$

Hence

$$\boxed{\mathbf{x}_i(t) = \mathbf{x}_i(0) + t\mathbf{V}, \qquad i = 1, 2.}$$

The velocity is perpendicular to the line joining the vortices and has magnitude

$$|\mathbf{V}| = \frac{|\Gamma|}{2\pi r}.$$

Such a pair is called a *vortex dipole*.

Integrability of the two-vortex problem. The phase space of two vortices is four-dimensional, so the system has two degrees of freedom. Besides the Hamiltonian

$$H = -\frac{\Gamma_1\Gamma_2}{2\pi}\log r,$$

the angular impulse

$$I = \frac{1}{2}\left(\Gamma_1|\mathbf{x}_1|^2 + \Gamma_2|\mathbf{x}_2|^2\right)$$

is conserved. Moreover,

$$\{H, I\}_\Gamma = 0.$$

On the regular part of phase space, H and I are functionally independent. The two-vortex problem is therefore Liouville integrable.

The explicit solution gives a direct description of this integrability. When the total circulation is nonzero, the relative configuration rotates uniformly about a fixed center of circulation. When the total circulation vanishes, the configuration translates uniformly without changing shape.

Remark 8.18 (Three and four vortices). The planar three-vortex problem is also integrable. Translation and rotation symmetries reduce its relative dynamics to an autonomous Hamiltonian system with one degree of freedom, which can be treated by quadrature.

For four vortices, the same reductions generally leave two effective degrees of freedom. The standard conserved quantities are then insufficient to establish Liouville integrability, and non-integrable or chaotic motion can occur. Special choices of vortex strengths and special invariant configurations may nevertheless produce integrable subsystems.

Homework Problems

Homework 8.1 (Canonical normal modes and invariant tori)**.** Consider the Hamiltonian

$$H(\mathbf{q}, \mathbf{p}) = \frac{1}{2}\mathbf{p}^{\mathsf{T}}\mathbf{M}^{-1}\mathbf{p} + \frac{1}{2}\mathbf{q}^{\mathsf{T}}\mathbf{K}\mathbf{q},$$

where

$$\mathbf{M} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{K} = \begin{bmatrix} 3 & -1 \\ -1 & 2 \end{bmatrix}.$$

1. Find $\mathbf{M}$-orthonormal generalized eigenvectors satisfying

$$\mathbf{K}\mathbf{v}_j = \omega_j^2 \mathbf{M}\mathbf{v}_j.$$

   Use them to construct a canonical transformation that decouples the Hamiltonian.

2. Find the two modal energies and verify that their Poisson bracket vanishes.

3. Determine which motions are periodic and which are quasiperiodic. Describe the corresponding invariant sets in phase space.

Homework 8.2 (The Runge–Lenz vector)**.** For the Kepler Hamiltonian

$$H(\mathbf{x}, \mathbf{p}) = \frac{|\mathbf{p}|^2}{2m} - \frac{k}{|\mathbf{x}|},$$

assume that the angular momentum

$$\mathbf{L} = \mathbf{x} \times \mathbf{p}$$

is nonzero, and define the Runge–Lenz vector

$$\mathbf{A} = \mathbf{p} \times \mathbf{L} - mk\frac{\mathbf{x}}{|\mathbf{x}|}.$$

1. Show directly from Hamilton's equations that

$$\frac{d\mathbf{A}}{dt} = \mathbf{0}.$$

2. Prove that

$$\mathbf{A} \cdot \mathbf{L} = 0.$$

3. Show that

$$|\mathbf{A}|^2 = m^2k^2 + 2mH|\mathbf{L}|^2.$$

4. Let $\ell = |\mathbf{L}|$. Taking the scalar product of $\mathbf{A}$ with $\mathbf{x}$, derive the orbit equation

$$r(\theta) = \frac{\ell^2/(mk)}{1 + e\cos(\theta - \theta_0)},$$

   and identify

$$e = \frac{|\mathbf{A}|}{mk}.$$

5. Explain why the Runge–Lenz vector determines the orientation of the orbit within its orbital plane.

Homework 8.3 (Gauge transformations as canonical transformations)**.** For the charged-particle Lagrangian

$$L = \frac{1}{2}m|\dot{\mathbf{x}}|^2 + e\mathbf{A}(t,\mathbf{x})\cdot\dot{\mathbf{x}} - e\phi(t,\mathbf{x}),$$

consider the gauge transformation

$$\mathbf{A}' = \mathbf{A} + \nabla\chi, \qquad \phi' = \phi - \chi_t.$$

A type-2 generating function $F_2(t,\mathbf{x},\mathbf{p}')$ defines a time-dependent canonical transformation by

$$\mathbf{p} = \frac{\partial F_2}{\partial \mathbf{x}}, \qquad \mathbf{x}' = \frac{\partial F_2}{\partial \mathbf{p}'}, \qquad H' = H + \frac{\partial F_2}{\partial t},$$

with the old variables expressed in terms of the new ones.

1. Show that
$$L' = L + e\frac{d\chi}{dt}, \qquad \mathbf{p}' = \mathbf{p} + e\nabla\chi.$$
2. Use
$$F_2(t,\mathbf{x},\mathbf{p}') = \mathbf{x}\cdot\mathbf{p}' - e\chi(t,\mathbf{x})$$
to recover the gauge transformation of the canonical variables and Hamiltonian.
3. Explain why the canonical description changes while the mechanical trajectory remains unchanged.

Homework 8.4 (A rotating polygon of point vortices)**.** Let $N \geq 2$ point vortices of equal strength Γ be located at the vertices

$$\mathbf{x}_j = R\begin{bmatrix}\cos(2\pi j/N)\\ \sin(2\pi j/N)\end{bmatrix}, \qquad j = 0,\dots,N-1.$$

1. Use symmetry to show that the velocity of each vortex is tangent to the circle $|\mathbf{x}| = R$.
2. Show that the configuration rotates rigidly and find its angular velocity:
$$\Omega = \frac{\Gamma(N-1)}{4\pi R^2}.$$
3. Verify that the Hamiltonian, the first moments of vorticity, and the angular impulse remain constant during this motion. Interpret the rotating polygon as a relative equilibrium.

Historical Notes and Further Reading

Historical notes. The Hamiltonian formulation of mechanics grew from the work of Hamilton (1834, 1835), who sought a unified method connecting dynamics with ideas he had previously

developed in geometrical optics. The resulting equations describe motion in terms of position and conjugate momentum and place the evolution in phase space. The Legendre transform provides the passage from the Lagrangian variables $(\mathbf{q}, \dot{\mathbf{q}})$ to the Hamiltonian variables $(\mathbf{q}, \mathbf{p})$.

The geometric and algebraic structures underlying Hamiltonian mechanics developed gradually. The bracket now named after Poisson arose in his work on the variation of constants in celestial mechanics (Poisson, 1809). Liouville (1838) established the differential-equation result underlying what is now called Liouville's theorem: Hamiltonian evolution preserves phase-space volume. Canonical transformations preserve the symplectic structure, while Poisson brackets describe both the evolution of observables and the relations among conserved quantities.

The relation between symmetry and conservation developed in Chapter 7 takes a particularly natural form in Hamiltonian variables: a conserved quantity generates a one-parameter family of canonical transformations, and a Hamiltonian symmetry produces a conserved generator. The commuting first integrals of a Liouville-integrable system lead, under suitable compactness and regularity conditions, to invariant tori carrying periodic or quasiperiodic motion.

Poincaré showed that volume preservation has important consequences for long-time dynamics. His recurrence principle states that, in an invariant region of finite phase-space volume, almost every trajectory returns arbitrarily close to its initial point. Recurrence does not imply integrability: a recurrent Hamiltonian system may still exhibit highly complicated dynamics.

The Hénon–Heiles system, introduced as a model in galactic dynamics, became a standard example in the study of nonintegrability and Hamiltonian chaos. It shows that conservation of energy and phase-space volume does not exclude the coexistence of periodic, quasiperiodic, and chaotic motion. For a rigorous nonintegrability result under suitable regularity assumptions, see Llibre and Jiménez-Lara (2011).

Point-vortex dynamics provides a different Hamiltonian model arising from ideal fluid motion. Helmholtz (1858) developed the dynamics of vortex filaments, and Kirchhoff (1876) formulated finite-dimensional equations for interacting point vortices. Their Poisson structure is weighted by the vortex strengths rather than initially written in standard position–momentum form. Nevertheless, the equations possess a Hamiltonian, conserved impulses, symmetry reductions, and integrable special cases. They illustrate that Hamiltonian structure extends beyond ordinary particle mechanics.

Further reading. For a geometric treatment of Lagrangian and Hamiltonian mechanics, canonical transformations, symplectic structure, and integrable systems, see Arnold (1989). A broad treatment of classical mechanics and canonical Hamiltonian methods is given by Goldstein et al. (2002). For a more advanced account of symplectic and Poisson geometry, symmetry, reduction, and momentum maps, see Marsden and Ratiu (1999). Point-vortex dynamics and its Hamiltonian structure are developed systematically in Newton (2001).

Orientation to Part IV

Part IV develops the analytic and numerical foundations of differential equations. Chapter 9 asks when an initial value problem actually determines a solution. By transforming the differential equation into an integral equation, fixed-point arguments establish local existence and uniqueness. A priori estimates then yield continuous dependence on initial data and parameters, continuation criteria, and maximal intervals of existence.

Chapter 10 asks how this continuous evolution can be approximated by discrete time steps. Consistency measures the error made in one step, stability controls the propagation of perturbations, and convergence describes the accumulated error over a finite time interval. Stiff equations and oscillatory systems show that accuracy alone is not enough: the stability and geometric properties of a numerical method must also be taken into account.

The two chapters are connected by a common principle. Estimates used to prove uniqueness and continuous dependence for the differential equation also control the growth of numerical errors. The central progression is

$$\begin{aligned}\text{integral formulation} &\longrightarrow \text{existence and uniqueness} \longrightarrow \text{continuous dependence}\\ &\longrightarrow \text{stable numerical approximation} \longrightarrow \text{convergence.}\end{aligned}$$

These foundations also prepare the dynamical viewpoint of Part V. Existence and uniqueness allow an autonomous equation to generate a flow, while numerical analysis clarifies how reliably that flow can be explored in computation.

Chapter 9

Existence, Uniqueness, and Well-Posedness

In the preceding chapters, differential equations were used to define trajectories, phase portraits, flows, invariant regions, and long-time dynamics. These constructions rest on several basic analytical questions. Given an initial value problem

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}), \qquad \mathbf{y}(t_0) = \mathbf{y}_0,$$

does a solution exist? Is it uniquely determined by the initial data? How does it change when the initial data or parameters are perturbed? How long can the solution be continued?

This chapter develops the fundamental theory needed to answer these questions. The two main tools are the contraction mapping theorem and Gronwall's inequality. Picard iteration and the contraction principle give local existence and uniqueness. Gronwall's inequality gives quantitative estimates for uniqueness and continuous dependence. When $\mathbf{f}$ is differentiable in $\mathbf{y}$, the solution also depends differentiably on the initial value, and its derivative satisfies the variational equation.

Gronwall's inequality also yields Hadamard well-posedness. For autonomous equations, uniqueness organizes the solutions into flow maps; differentiating these maps leads to the variational equation and transport along flows.

The main logical path is

integral formulation $\longrightarrow$ fixed point $\longrightarrow$ local existence and uniqueness
$\longrightarrow$ continuous dependence and well-posedness $\longrightarrow$ flows and variational equations
$\longrightarrow$ continuation and global existence.

9.1 Local Existence and Uniqueness

The starting point of the local theory is to regard a solution not as a single point in $\mathbb{R}^n$, but as an element of a space of functions. The differential equation will be transformed into an equivalent integral equation and then treated as a fixed-point problem in a suitable function space.

9.1.1 The integral formulation

From the differential equation to an integral equation. Consider

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}), \qquad \mathbf{y}(t_0) = \mathbf{y}_0,$$

where

$$\mathbf{f} : \Omega \subset \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n.$$

Proposition 9.1. *Let I be an interval containing t_0, suppose that $\mathbf{f}$ is continuous on Ω, and let*

$$\mathbf{y} \in C^1(I; \mathbb{R}^n)$$

have its graph in Ω. Then $\mathbf{y}$ solves

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}), \qquad \mathbf{y}(t_0) = \mathbf{y}_0,$$

on I if and only if

$$\mathbf{y}(t) = \mathbf{y}_0 + \int_{t_0}^{t} \mathbf{f}(s, \mathbf{y}(s))\, ds, \qquad t \in I.$$

Proof. If $\mathbf{y}$ solves the initial value problem, then integration from t_0 to t gives

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

which yields the integral equation.

Conversely, if $\mathbf{y}$ satisfies the integral equation, then continuity of $\mathbf{f}$ and the fundamental theorem of calculus give

$$\dot{\mathbf{y}}(t) = \mathbf{f}(t, \mathbf{y}(t)).$$

Setting $t = t_0$ gives $\mathbf{y}(t_0) = \mathbf{y}_0$. □

Remark 9.2 (Why the integral formulation?). The differential equation involves the derivative of the unknown function, whereas the integral equation involves only the function itself. Moreover, for $t_0 < t$, integration permits the basic estimate

$$\left| \int_{t_0}^{t} \mathbf{g}(s)\, ds \right| \leq \int_{t_0}^{t} |\mathbf{g}(s)|\, ds.$$

Thus an integral operator can be estimated directly in terms of the size of its input. By contrast, the derivative of a function cannot be controlled by the supremum norm of the function itself.

We may therefore define the integral operator

$$(\mathcal{T}\mathbf{y})(t) = \mathbf{y}_0 + \int_{t_0}^{t} \mathbf{f}(s, \mathbf{y}(s))\, ds.$$

A solution of the initial value problem is precisely a fixed point of $\mathcal{T}$. The integral estimate will allow us to show that $\mathcal{T}$ is a contraction on a suitable space of continuous functions and hence to construct the solution by successive approximation.

9.1.2 Function spaces and the contraction principle

To apply the fixed-point approach, we first introduce the function space in which the integral operator will act and the fixed-point theorem that will be used.

Completeness of the function space $C([a,b];\mathbb{R}^n)$. Let $[a,b] \subset \mathbb{R}$ be a closed bounded interval. We denote by

$$C([a,b];\mathbb{R}^n)$$

the vector space of continuous functions

$$\mathbf{y} : [a,b] \to \mathbb{R}^n.$$

For $\mathbf{y} \in C([a,b];\mathbb{R}^n)$, define the supremum norm by

$$\|\mathbf{y}\|_\infty = \max_{a \le t \le b} |\mathbf{y}(t)|.$$

Since $\mathbf{y}$ is continuous and $[a,b]$ is compact, this maximum exists.

The distance between two functions $\mathbf{y}$ and $\mathbf{z}$ is measured by

$$\|\mathbf{y} - \mathbf{z}\|_\infty = \max_{a \le t \le b} |\mathbf{y}(t) - \mathbf{z}(t)|.$$

Thus two functions are close in the supremum norm when they are uniformly close throughout the interval.

A sequence $\{\mathbf{y}_k\}$ converges to $\mathbf{y}$ in this norm precisely when

$$\max_{a \le t \le b} |\mathbf{y}_k(t) - \mathbf{y}(t)| \longrightarrow 0.$$

This is uniform convergence.

The function space $C([a,b];\mathbb{R}^n)$ has an important completeness property.

Proposition 9.3. *The normed vector space*

$$C([a,b];\mathbb{R}^n)$$

with the supremum norm is complete.

Proof. Let $\{\mathbf{y}_k\}$ be a Cauchy sequence in the supremum norm. For each fixed $t \in [a,b]$, the sequence $\{\mathbf{y}_k(t)\}$ is Cauchy in $\mathbb{R}^n$. Since $\mathbb{R}^n$ is complete, it has a limit, which we denote by $\mathbf{y}(t)$:

$$\mathbf{y}(t) = \lim_{k \to \infty} \mathbf{y}_k(t).$$

This defines a function

$$\mathbf{y} : [a,b] \to \mathbb{R}^n.$$

Fix $\varepsilon > 0$. Since $\{\mathbf{y}_k\}$ is Cauchy in the supremum norm, there exists N such that

$$\|\mathbf{y}_k - \mathbf{y}_\ell\|_\infty < \varepsilon$$

whenever $k, \ell \ge N$. Hence, for every fixed $t \in [a,b]$,

$$|\mathbf{y}_k(t) - \mathbf{y}_\ell(t)| < \varepsilon, \qquad k, \ell \ge N.$$

Letting $\ell \to \infty$ and using $\mathbf{y}_\ell(t) \to \mathbf{y}(t)$, we obtain

$$|\mathbf{y}_k(t) - \mathbf{y}(t)| \le \varepsilon, \qquad k \ge N.$$

Since this estimate holds for every $t \in [a,b]$,

$$\|\mathbf{y}_k - \mathbf{y}\|_\infty \le \varepsilon, \qquad k \ge N.$$

Hence $\mathbf{y}_k \to \mathbf{y}$ uniformly.

It remains to show that $\mathbf{y}$ is continuous. Fix $t_* \in [a,b]$. Choose N so that

$$\|\mathbf{y} - \mathbf{y}_N\|_\infty < \frac{\varepsilon}{3}.$$

Since $\mathbf{y}_N$ is continuous, there exists $\delta > 0$ such that

$$|t - t_*| < \delta \quad \Longrightarrow \quad |\mathbf{y}_N(t) - \mathbf{y}_N(t_*)| < \frac{\varepsilon}{3}.$$

Therefore,

$$\begin{aligned} |\mathbf{y}(t) - \mathbf{y}(t_*)| &\le |\mathbf{y}(t) - \mathbf{y}_N(t)| + |\mathbf{y}_N(t) - \mathbf{y}_N(t_*)| \\ &\quad + |\mathbf{y}_N(t_*) - \mathbf{y}(t_*)| < \varepsilon. \end{aligned}$$

Hence $\mathbf{y}$ is continuous on $[a,b]$, and the space is complete. □

A complete normed vector space is called a *Banach space*. We shall use this language only for the function space above. The essential point is that a Cauchy sequence of continuous functions in the supremum norm converges to another continuous function in the same space.

The contraction mapping theorem. We next recall the fixed-point theorem that will produce solutions of the integral equation.

Definition 9.4. Let (X,d) be a metric space. A map

$$\mathcal{T} : X \to X$$

is called a *contraction* if there exists a constant q with

$$0 \le q < 1$$

such that

$$d(\mathcal{T}u, \mathcal{T}v) \le q\, d(u,v)$$

for all $u, v \in X$.

Thus a contraction reduces distances by a uniform factor smaller than one. In particular, every contraction is Lipschitz continuous and therefore continuous.

Theorem 9.5 (Contraction mapping theorem)**.** *Let X be a complete metric space, and let*

$$\mathcal{T} : X \to X$$

be a contraction. Then $\mathcal{T}$ has a unique fixed point $u_* \in X$:

$$\mathcal{T}u_* = u_*.$$

Moreover, for every initial choice $u_0 \in X$, the iteration

$$u_{k+1} = \mathcal{T}u_k$$

converges to u_*.

Proof. Let $q < 1$ be a contraction constant. Starting from $u_0 \in X$, define

$$u_{k+1} = \mathcal{T} u_k.$$

Then

$$d(u_{k+1}, u_k) \le q\, d(u_k, u_{k-1}),$$

and therefore

$$d(u_{k+1}, u_k) \le q^k d(u_1, u_0).$$

If $m > k$, the triangle inequality gives

$$\begin{aligned} d(u_m, u_k) &\le \sum_{j=k}^{m-1} d(u_{j+1}, u_j) \\ &\le \sum_{j=k}^{m-1} q^j d(u_1, u_0) \\ &\le \frac{q^k}{1-q} d(u_1, u_0). \end{aligned}$$

Hence $\{u_k\}$ is a Cauchy sequence. Since X is complete, there exists $u_* \in X$ such that

$$u_k \to u_*.$$

Because $\mathcal{T}$ is Lipschitz continuous,

$$\mathcal{T} u_k \to \mathcal{T} u_*.$$

But $\mathcal{T} u_k = u_{k+1} \to u_*$, and therefore

$$\mathcal{T} u_* = u_*.$$

To prove uniqueness, suppose that v_* is another fixed point. Then

$$d(u_*, v_*) = d(\mathcal{T} u_*, \mathcal{T} v_*) \le q\, d(u_*, v_*).$$

Since $q < 1$, this is possible only if

$$d(u_*, v_*) = 0.$$

Hence $u_* = v_*$. □

The contraction mapping theorem contains both an existence statement and an approximation procedure. The fixed point exists and is unique in X, and repeated application of $\mathcal{T}$ converges to it. We now apply this theorem to the integral operator associated with the initial value problem.

9.1.3 The Picard operator and the Cauchy–Lipschitz theorem

We now apply the contraction mapping theorem to the integral equation

$$\mathbf{y}(t) = \mathbf{y}_0 + \int_{t_0}^{t} \mathbf{f}(s, \mathbf{y}(s))\, ds.$$

The Picard operator. Define

$$(\mathcal{T}\mathbf{y})(t) = \mathbf{y}_0 + \int_{t_0}^{t} \mathbf{f}(s, \mathbf{y}(s))\, ds.$$

Then $\mathbf{y}$ solves the initial value problem precisely when

$$\mathcal{T}\mathbf{y} = \mathbf{y}.$$

Thus the problem is reduced to finding a fixed point of $\mathcal{T}$ in a suitable complete function space.

Local Lipschitz continuity. The contraction property requires quantitative control of the dependence of $\mathbf{f}$ on the state variable.

Definition 9.6. Let $\Omega \subset \mathbb{R} \times \mathbb{R}^n$ be open. A function

$$\mathbf{f} : \Omega \to \mathbb{R}^n$$

is *locally Lipschitz continuous in* $\mathbf{y}$ if, for every $(t_*, \mathbf{y}_*) \in \Omega$, there exist a neighborhood $U \subset \Omega$ of $(t_*, \mathbf{y}_*)$ and a constant $L > 0$ such that

$$|\mathbf{f}(t, \mathbf{y}) - \mathbf{f}(t, \mathbf{z})| \le L|\mathbf{y} - \mathbf{z}|$$

whenever $(t, \mathbf{y}), (t, \mathbf{z}) \in U$.

Theorem 9.7 (Cauchy–Lipschitz theorem)**.** *Let*

$$\mathbf{f} : \Omega \subset \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$$

be continuous and locally Lipschitz continuous in $\mathbf{y}$*, where* Ω *is open. For every* $(t_0, \mathbf{y}_0) \in \Omega$*, there exists* $\delta > 0$ *such that the initial value problem*

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}), \qquad \mathbf{y}(t_0) = \mathbf{y}_0,$$

has a unique solution on

$$[t_0 - \delta, t_0 + \delta].$$

Proof. Choose $h > 0$ and $r > 0$ such that the compact cylinder

$$R = [t_0 - h, t_0 + h] \times \overline{B_r(\mathbf{y}_0)}$$

is contained in a neighborhood on which $\mathbf{f}$ is Lipschitz continuous in $\mathbf{y}$, uniformly in t, with Lipschitz constant L. Since $\mathbf{f}$ is continuous and R is compact, there is a constant $M > 0$ such that

$$|\mathbf{f}(t, \mathbf{y})| \le M$$

for all $(t, \mathbf{y}) \in R$.

We shall choose $0 < \delta \le h$ sufficiently small that

$$M\delta < r, \qquad L\delta < 1.$$

The complete function space. Let

$$I_\delta = [t_0 - \delta, t_0 + \delta]$$

and define

$$X = \{\mathbf{y} \in C(I_\delta; \mathbb{R}^n) : \|\mathbf{y} - \mathbf{y}_0\|_\infty \le r\},$$

where $\mathbf{y}_0$ also denotes the constant function

$$\mathbf{y}_0(t) \equiv \mathbf{y}_0.$$

The set X is closed in $C(I_\delta; \mathbb{R}^n)$ and is therefore complete.

Invariance of X. If $\mathbf{y} \in X$, then $(t, \mathbf{y}(t)) \in R$ for every $t \in I_\delta$. Consequently,

$$\begin{aligned} |(\mathcal{T}\mathbf{y})(t) - \mathbf{y}_0| &= \left| \int_{t_0}^{t} \mathbf{f}(s, \mathbf{y}(s))\, ds \right| \\ &\le M|t - t_0| \le M\delta. \end{aligned}$$

Hence

$$\|\mathcal{T}\mathbf{y} - \mathbf{y}_0\|_\infty \le M\delta \le r.$$

Therefore

$$\mathcal{T}(X) \subset X.$$

The contraction estimate. For $\mathbf{y}, \mathbf{z} \in X$,

$$\begin{aligned} |(\mathcal{T}\mathbf{y})(t) - (\mathcal{T}\mathbf{z})(t)| &\le \left| \int_{t_0}^{t} \big[\mathbf{f}(s, \mathbf{y}(s)) - \mathbf{f}(s, \mathbf{z}(s))\big]\, ds \right| \\ &\le L|t - t_0| \|\mathbf{y} - \mathbf{z}\|_\infty. \end{aligned}$$

Taking the supremum over $t \in I_\delta$ gives

$$\|\mathcal{T}\mathbf{y} - \mathcal{T}\mathbf{z}\|_\infty \le L\delta \|\mathbf{y} - \mathbf{z}\|_\infty.$$

Since $L\delta < 1$, the operator $\mathcal{T}$ is a contraction from the complete metric space X into itself. The contraction mapping theorem therefore gives a unique fixed point of $\mathcal{T}$ in X, and hence a unique local solution in X of the initial value problem.

It remains to verify that uniqueness is not restricted to solutions already assumed to lie in X. Let $\mathbf{z}$ be any solution on I_δ with $\mathbf{z}(t_0) = \mathbf{y}_0$. If $\mathbf{z}$ were to leave $\overline{B_r(\mathbf{y}_0)}$, let τ be the first time at which

$$|\mathbf{z}(\tau) - \mathbf{y}_0| = r.$$

Up to time τ, the graph of $\mathbf{z}$ remains in R, and therefore

$$|\mathbf{z}(\tau) - \mathbf{y}_0| = \left| \int_{t_0}^{\tau} \mathbf{f}(s, \mathbf{z}(s))\, ds \right| \le M|\tau - t_0| \le M\delta < r,$$

which is a contradiction. The same argument applies on the left of t_0. Hence every solution with the prescribed initial value belongs to X, where uniqueness follows from the contraction mapping theorem.

This proves the fundamental local existence and uniqueness theorem. □

Picard iteration. The fixed-point proof also gives a sequence of successive approximations. Starting from

$$\mathbf{y}^{(0)}(t) = \mathbf{y}_0,$$

define

$$\mathbf{y}^{(k+1)}(t) = \mathbf{y}_0 + \int_{t_0}^{t} \mathbf{f}(s, \mathbf{y}^{(k)}(s))\, ds.$$

These functions are called the *Picard iterates*. On the interval obtained in the Cauchy–Lipschitz theorem,

$$\mathbf{y}^{(k)} \longrightarrow \mathbf{y}$$

uniformly, where $\mathbf{y}$ is the unique solution.

If

$$q = L\delta < 1,$$

then the contraction estimate gives

$$\|\mathbf{y}^{(k)} - \mathbf{y}\|_\infty \le \frac{q^k}{1-q} \|\mathbf{y}^{(1)} - \mathbf{y}^{(0)}\|_\infty.$$

Thus Picard iteration provides both the existence proof and a constructive approximation of the solution.

Autonomous systems. For an autonomous system

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}), \qquad \mathbf{y}(0) = \mathbf{y}_0,$$

the theorem applies by viewing $\mathbf{f}$ as independent of t. Thus a locally Lipschitz vector field determines a unique local trajectory through every point of its domain.

9.1.4 Existence and uniqueness hypotheses

The assumptions of continuity and local Lipschitz continuity play different roles in the local theory. We now examine how these conditions can be verified and what may happen when the Lipschitz condition fails.

A convenient sufficient condition. In applications, local Lipschitz continuity is often verified through differentiability. If $\mathbf{f}$ is continuously differentiable with respect to $\mathbf{y}$, then $\mathbf{f}$ is locally Lipschitz continuous in $\mathbf{y}$.

Indeed, suppose that the line segment joining $\mathbf{y}$ and $\mathbf{z}$ remains in the domain and that

$$\|D_{\mathbf{y}}\mathbf{f}(t, \mathbf{u})\| \le L$$

along this segment. For fixed t, the mean value formula gives

$$\mathbf{f}(t, \mathbf{y}) - \mathbf{f}(t, \mathbf{z}) = \int_0^1 D_{\mathbf{y}}\mathbf{f}\big(t, \mathbf{z} + s(\mathbf{y} - \mathbf{z})\big)(\mathbf{y} - \mathbf{z})\, ds,$$

and hence

$$|\mathbf{f}(t, \mathbf{y}) - \mathbf{f}(t, \mathbf{z})| \le L|\mathbf{y} - \mathbf{z}|.$$

Thus the frequently used assumption

$$\mathbf{f} \in C^1(\Omega; \mathbb{R}^n)$$

implies the hypotheses of the Cauchy–Lipschitz theorem.

Peano's existence theorem. A classical theorem of Peano states that continuity of $\mathbf{f}$ is sufficient for local existence.

Theorem 9.8 (Peano existence theorem)**.** *Let*

$$\mathbf{f} : \Omega \subset \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$$

be continuous, where Ω is open. Then for every $(t_0, \mathbf{y}_0) \in \Omega$, the initial value problem

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}), \qquad \mathbf{y}(t_0) = \mathbf{y}_0,$$

has at least one local solution.

The proof uses compactness rather than the contraction principle and will not be developed here. Peano's theorem gives existence but not uniqueness.

Continuity without uniqueness. Continuity of $\mathbf{f}$ alone does not guarantee uniqueness. Consider

$$y' = |y|^{1/2}, \qquad y(0) = 0.$$

The function

$$f(y) = |y|^{1/2}$$

is continuous but is not locally Lipschitz at $y = 0$. Besides the solution

$$y(t) = 0,$$

for every $a \geq 0$ there is a solution

$$y_a(t) = \begin{cases} 0, & t \leq a, \\ \dfrac{(t-a)^2}{4}, & t \geq a. \end{cases}$$

Thus infinitely many solutions pass through the same initial point.

Geometrically, a solution may remain at the equilibrium for an arbitrary waiting time and then leave it. This behavior is ruled out by the local Lipschitz condition.

The two hypotheses in the Cauchy–Lipschitz theorem therefore play different roles: continuity gives local existence, while local Lipschitz continuity in $\mathbf{y}$ gives uniqueness.

Local versus global existence. The Cauchy–Lipschitz theorem is local. The vector field need only be continuous and locally Lipschitz near the initial point. These hypotheses guarantee a unique solution on some interval containing t_0, but they do not guarantee that the solution exists for all time.

For example,

$$y' = y^2, \qquad y(0) = 1,$$

has the unique solution

$$y(t) = \frac{1}{1-t},$$

which becomes unbounded as $t \to 1^-$. The vector field is smooth on all of $\mathbb{R}$, yet the solution cannot be continued beyond $t = 1$. Continuation and global existence will be studied later in this chapter.

Exercise 9.1 (Lipschitz condition)**.** Determine whether each of the following functions is locally Lipschitz continuous:

$$y^2, \qquad \sin y, \qquad |y|, \qquad \sqrt{|y|}, \qquad |y|^\alpha \quad (\alpha > 0).$$

For each case, determine the initial values y_0 for which the Cauchy–Lipschitz theorem applies to

$$\dot{y} = f(y), \qquad y(0) = y_0.$$

Exercise 9.2 (Picard iteration)**.** Consider

$$y' = y, \qquad y(0) = 1.$$

Starting with $y^{(0)}(t) = 1$, define the Picard iterates by

$$y^{(k+1)}(t) = 1 + \int_0^t y^{(k)}(s)\, ds.$$

Show by induction that

$$y^{(k)}(t) = \sum_{j=0}^{k} \frac{t^j}{j!}.$$

Deduce that the Picard iterates converge uniformly on every bounded interval to the solution $y(t) = e^t$. Give an explicit bound for

$$\|y - y^{(k)}\|_\infty$$

on $[-T, T]$.

Exercise 9.3 (Choosing the contraction interval)**.** Consider the initial value problem

$$y' = t + y^2, \qquad y(0) = 0.$$

On the rectangle

$$R = [-\delta, \delta] \times [-1, 1],$$

find constants M and L such that

$$|t + y^2| \le M$$

and

$$|(t + y^2) - (t + z^2)| \le L|y - z|$$

for all $(t, y), (t, z) \in R$.

Determine a positive value of δ for which

$$M\delta \le 1, \qquad L\delta < 1.$$

Conclude directly from the contraction argument that the initial value problem has a unique solution on $[-\delta, \delta]$.

Exercise 9.4 (Nonuniqueness and waiting times)**.** Let $0 < \alpha < 1$ and consider

$$y' = |y|^\alpha, \qquad y(0) = 0.$$

Show that $f(y) = |y|^\alpha$ is continuous on $\mathbb{R}$ but is not locally Lipschitz at $y = 0$.

For every $a \geq 0$, define

$$y_a^+(t) = \begin{cases} 0, & t \leq a, \\ \big((1-\alpha)(t-a)\big)^{1/(1-\alpha)}, & t \geq a. \end{cases}$$

Verify that y_a^+ is continuously differentiable and solves the differential equation.

Similarly, for every $b \leq 0$, define

$$y_b^-(t) = \begin{cases} -\big((1-\alpha)(b-t)\big)^{1/(1-\alpha)}, & t \leq b, \\ 0, & t \geq b. \end{cases}$$

Verify that y_b^- is also a solution. By joining these two types of solutions across an interval on which $y = 0$, describe the solutions that pass through $(0, 0)$. Explain why uniqueness fails.

9.2 Gronwall Estimates and Continuous Dependence

The contraction principle gives local existence and uniqueness. To compare two solutions, however, it is more useful to estimate how their difference evolves in time. The basic tool is Gronwall's inequality, which compares a differential inequality with a linear differential equation.

9.2.1 Gronwall's inequality

Comparison with a linear equation. Let $\eta : [t_0, T] \to \mathbb{R}$ be differentiable and suppose that

$$\eta'(t) \leq a(t)\eta(t) + b(t),$$

where a and b are continuous. Consider the linear initial value problem

$$z' = a(t)z + b(t), \qquad z(t_0) = \eta(t_0).$$

The comparison principle suggests that

$$\eta(t) \leq z(t).$$

Since the linear equation can be solved explicitly, this gives a quantitative estimate for η.

Theorem 9.9 (Gronwall's inequality)**.** *Let η be differentiable on $[t_0, T]$, and let a and b be continuous functions with*

$$\eta'(t) \le a(t)\eta(t) + b(t).$$

Then

$$\eta(t) \le \eta(t_0)e^{A(t)} + \int_{t_0}^{t} e^{A(t)-A(s)}b(s)\,ds,$$

where

$$A(t) = \int_{t_0}^{t} a(\tau)\,d\tau.$$

Proof. Multiplying the differential inequality by $e^{-A(t)}$ gives

$$e^{-A(t)}\big(\eta'(t) - a(t)\eta(t)\big) \le e^{-A(t)}b(t).$$

Since

$$\frac{d}{dt}\left(e^{-A(t)}\eta(t)\right) = e^{-A(t)}\big(\eta'(t) - a(t)\eta(t)\big),$$

integration from t_0 to t yields

$$e^{-A(t)}\eta(t) - \eta(t_0) \le \int_{t_0}^{t} e^{-A(s)}b(s)\,ds.$$

Multiplying by $e^{A(t)}$ gives the result. □

The right-hand side is precisely the solution of

$$z' = a(t)z + b(t), \qquad z(t_0) = \eta(t_0).$$

Thus Gronwall's inequality may be read as the comparison statement

$$\eta'(t) \le a(t)\eta(t) + b(t) \quad \Longrightarrow \quad \eta(t) \le z(t).$$

The homogeneous form. If $b(t) = 0$, then

$$\eta'(t) \le a(t)\eta(t)$$

implies

$$\eta(t) \le \eta(t_0)\exp\left(\int_{t_0}^{t} a(s)\,ds\right).$$

In particular, if $a(t) \le L$, then

$$\eta(t) \le \eta(t_0)e^{L(t-t_0)}.$$

A basic consequence is that if

$$\eta(t_0) = 0, \qquad \eta'(t) \le L\eta(t),$$

then

$$\eta(t) = 0$$

provided $\eta(t) \ge 0$. This observation gives another proof of uniqueness for differential equations.

The integral form. The estimate is often encountered without a derivative.

Corollary 9.10 (Integral Gronwall inequality)**.** *Let $\eta : [t_0, T] \to [0, \infty)$ be continuous and suppose that*

$$\eta(t) \le c + \int_{t_0}^{t} \big(a(s)\eta(s) + b(s)\big)\, ds,$$

where $c \ge 0$ and a, b are continuous and nonnegative. Then

$$\eta(t) \le ce^{A(t)} + \int_{t_0}^{t} e^{A(t)-A(s)} b(s)\, ds,$$

where

$$A(t) = \int_{t_0}^{t} a(\tau)\, d\tau.$$

Proof. Define

$$u(t) = c + \int_{t_0}^{t} \big(a(s)\eta(s) + b(s)\big)\, ds.$$

Then $\eta(t) \le u(t)$ and

$$u'(t) = a(t)\eta(t) + b(t) \le a(t)u(t) + b(t).$$

The differential form of Gronwall's inequality applied to u gives the result. □

In particular, if

$$\eta(t) \le c + L \int_{t_0}^{t} \eta(s)\, ds,$$

then

$$\eta(t) \le ce^{L(t-t_0)}.$$

Backward-time estimates. The same argument applies on an interval to the left of t_0. If $t \le t_0$, one may integrate from t to t_0, or reverse time by setting

$$\tau = t_0 - t.$$

Accordingly, estimates on a symmetric local interval are usually written in terms of $|t - t_0|$.

9.2.2 Uniqueness and stability estimates

Gronwall's inequality gives a direct quantitative comparison of two solutions of the same differential equation.

Proposition 9.11. *Let $\mathbf{y}$ and $\mathbf{z}$ be solutions of*

$$\dot{\mathbf{u}} = \mathbf{f}(t, \mathbf{u})$$

on $[t_0, T]$. Suppose that $\mathbf{f}$ is Lipschitz continuous in $\mathbf{y}$ with constant L on a region containing both trajectories. Then

$$|\mathbf{y}(t) - \mathbf{z}(t)| \le e^{L(t-t_0)} |\mathbf{y}(t_0) - \mathbf{z}(t_0)|$$

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

Proof. From the integral equations,

$$\mathbf{y}(t) - \mathbf{z}(t) = \mathbf{y}(t_0) - \mathbf{z}(t_0) + \int_{t_0}^{t} \big[\mathbf{f}(s, \mathbf{y}(s)) - \mathbf{f}(s, \mathbf{z}(s))\big]\, ds.$$

Hence

$$|\mathbf{y}(t) - \mathbf{z}(t)| \le |\mathbf{y}(t_0) - \mathbf{z}(t_0)| + L \int_{t_0}^{t} |\mathbf{y}(s) - \mathbf{z}(s)|\, ds.$$

The integral form of Gronwall's inequality gives

$$|\mathbf{y}(t) - \mathbf{z}(t)| \le e^{L(t-t_0)} |\mathbf{y}(t_0) - \mathbf{z}(t_0)|.$$

□

Uniqueness. If the two solutions have the same initial value, then

$$\mathbf{y}(t_0) = \mathbf{z}(t_0),$$

and the estimate gives

$$|\mathbf{y}(t) - \mathbf{z}(t)| = 0.$$

Thus

$$\mathbf{y}(t) = \mathbf{z}(t)$$

throughout their common interval of existence. This gives a second proof of uniqueness, now based on Gronwall's inequality rather than the contraction mapping theorem.

Stability of trajectories. The same estimate contains more information than uniqueness. If the initial values are close, then the corresponding solutions remain close for finite time:

$$|\mathbf{y}(t) - \mathbf{z}(t)| \le e^{L(t-t_0)} |\mathbf{y}(t_0) - \mathbf{z}(t_0)|. \tag{9.1}$$

Thus the amplification of an initial perturbation is controlled by the factor

$$e^{L(t-t_0)}.$$

The estimate does not imply that the distance between trajectories must grow exponentially. It only provides an upper bound on how rapidly the distance can grow.

An estimate with different vector fields. A similar argument compares solutions of two different equations. Suppose that

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}), \qquad \dot{\mathbf{z}} = \mathbf{g}(t, \mathbf{z}),$$

and that $\mathbf{f}$ is Lipschitz continuous in $\mathbf{y}$ with constant L. Then

$$\begin{aligned} \mathbf{f}(s, \mathbf{y}(s)) - \mathbf{g}(s, \mathbf{z}(s)) = {} & \mathbf{f}(s, \mathbf{y}(s)) - \mathbf{f}(s, \mathbf{z}(s)) \\ & + \mathbf{f}(s, \mathbf{z}(s)) - \mathbf{g}(s, \mathbf{z}(s)). \end{aligned}$$

$$
\begin{aligned}
|\mathbf{y}(t) - \mathbf{z}(t)| \le |\mathbf{y}(t_0) - \mathbf{z}(t_0)| &\\
&+ L \int_{t_0}^{t} |\mathbf{y}(s) - \mathbf{z}(s)|\, ds \\
&+ \int_{t_0}^{t} |\mathbf{f}(s, \mathbf{z}(s)) - \mathbf{g}(s, \mathbf{z}(s))|\, ds.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
|\mathbf{y}(t) - \mathbf{z}(t)| \le e^{L(t-t_0)} |\mathbf{y}(t_0) - \mathbf{z}(t_0)| &\\
&+ \int_{t_0}^{t} e^{L(t-s)} |\mathbf{f}(s, \mathbf{z}(s)) - \mathbf{g}(s, \mathbf{z}(s))|\, ds.
\end{aligned}
$$

This estimate separates the effect of perturbing the initial value from the effect of perturbing the differential equation itself.

9.2.3 Continuous dependence on initial data

The stability estimate obtained above shows that solutions depend continuously on their initial values.

Let

$$\mathbf{y}(t; t_0, \mathbf{y}_0)$$

denote the solution of

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}), \qquad \mathbf{y}(t_0) = \mathbf{y}_0.$$

Theorem 9.12 (Continuous dependence on initial data)**.** *Suppose that* $\mathbf{f}$ *is continuous and locally Lipschitz continuous in* $\mathbf{y}$*. Let* $\mathbf{y}(t; t_0, \mathbf{y}_0)$ *be a solution defined on* $[t_0, T]$*. Then, for all initial values* $\mathbf{z}_0$ *sufficiently close to* $\mathbf{y}_0$*, the solution*

$$\mathbf{y}(t; t_0, \mathbf{z}_0)$$

is also defined on $[t_0, T]$*, and*

$$\mathbf{y}(t; t_0, \mathbf{z}_0) \longrightarrow \mathbf{y}(t; t_0, \mathbf{y}_0)$$

uniformly on $[t_0, T]$ *as* $\mathbf{z}_0 \to \mathbf{y}_0$*.*

Proof. The reference trajectory

$$\Gamma = \{(t, \mathbf{y}(t; t_0, \mathbf{y}_0)) : t \in [t_0, T]\}$$

is compact. Choose compact sets

$$\Gamma \subset \operatorname{int} K_0 \subset K_0 \subset \operatorname{int} K \subset K \subset \Omega$$

such that $\mathbf{f}$ is Lipschitz continuous in $\mathbf{y}$ on K, with constant L. Let

$$\rho = \operatorname{dist}(K_0, \partial K) > 0.$$

Since the reference trajectory lies in $\Gamma \subset K_0$, its distance from ∂K is at least ρ. Choose $\mathbf{z}_0$ sufficiently close to $\mathbf{y}_0$ so that

$$|\mathbf{z}_0 - \mathbf{y}_0| < \rho e^{-L(T-t_0)}.$$

By the local existence theorem, the solution

$$\mathbf{y}(t; t_0, \mathbf{z}_0)$$

exists on some interval containing t_0.

Suppose, for contradiction, that this perturbed solution leaves K before time T. Let $\tau \le T$ be its first exit time from K. Then both the reference solution and the perturbed solution remain in K for $t_0 \le t \le \tau$, so the stability estimate (9.1) gives

$$|\mathbf{y}(t; t_0, \mathbf{z}_0) - \mathbf{y}(t; t_0, \mathbf{y}_0)| \le e^{L(t-t_0)} |\mathbf{z}_0 - \mathbf{y}_0|, \qquad t_0 \le t \le \tau.$$

In particular,

$$\begin{aligned} |\mathbf{y}(\tau; t_0, \mathbf{z}_0) - \mathbf{y}(\tau; t_0, \mathbf{y}_0)| &\le e^{L(\tau-t_0)} |\mathbf{z}_0 - \mathbf{y}_0| \\ &\le e^{L(T-t_0)} |\mathbf{z}_0 - \mathbf{y}_0| \\ &< \rho. \end{aligned}$$

But

$$\mathbf{y}(\tau; t_0, \mathbf{y}_0) \in K_0,$$

while, by definition of the first exit time,

$$\mathbf{y}(\tau; t_0, \mathbf{z}_0) \in \partial K.$$

Hence

$$|\mathbf{y}(\tau; t_0, \mathbf{z}_0) - \mathbf{y}(\tau; t_0, \mathbf{y}_0)| \ge \operatorname{dist}(K_0, \partial K) = \rho,$$

a contradiction.

Therefore the perturbed trajectory cannot leave K before time T. By repeated application of the local existence theorem, it can be continued throughout $[t_0, T]$. The stability estimate is consequently valid on the whole interval:

$$|\mathbf{y}(t; t_0, \mathbf{z}_0) - \mathbf{y}(t; t_0, \mathbf{y}_0)| \le e^{L(t-t_0)} |\mathbf{z}_0 - \mathbf{y}_0|, \qquad t_0 \le t \le T.$$

Thus

$$\max_{t_0 \le t \le T} |\mathbf{y}(t; t_0, \mathbf{z}_0) - \mathbf{y}(t; t_0, \mathbf{y}_0)| \le e^{L(T-t_0)} |\mathbf{z}_0 - \mathbf{y}_0|,$$

which tends to zero as $\mathbf{z}_0 \to \mathbf{y}_0$. This proves uniform continuous dependence on the initial data on $[t_0, T]$. □

The solution map. For fixed t and t_0, define

$$\Phi_{t,t_0}(\mathbf{y}_0) = \mathbf{y}(t; t_0, \mathbf{y}_0).$$

The theorem states that Φ_{t,t_0} is locally Lipschitz continuous wherever the solutions are defined on a common finite interval. Indeed,

$$|\Phi_{t,t_0}(\mathbf{y}_0) - \Phi_{t,t_0}(\mathbf{z}_0)| \le e^{L|t-t_0|} |\mathbf{y}_0 - \mathbf{z}_0|.$$

Hadamard well-posedness. An initial value problem is said to be *well-posed in the sense of Hadamard* if a solution exists, is unique, and depends continuously on the initial data. The Cauchy–Lipschitz theorem and the preceding continuous-dependence estimate therefore show that

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}), \qquad \mathbf{y}(t_0) = \mathbf{y}_0,$$

is locally well-posed whenever $\mathbf{f}$ is continuous and locally Lipschitz continuous in $\mathbf{y}$.

This is a finite-time statement. It does not imply that nearby trajectories remain close as $t \to \infty$.

9.2.4 Continuous dependence on parameters

Many differential equations contain parameters that represent physical constants, forcing strengths, damping coefficients, or geometric data. A well-posed model should respond continuously not only to changes in the initial value, but also to small changes in these parameters.

Consider

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}, \boldsymbol{\mu}), \qquad \mathbf{y}(t_0) = \mathbf{y}_0,$$

where

$$\boldsymbol{\mu} \in \mathbb{R}^m$$

is a parameter. Let

$$\mathbf{y}(t; \mathbf{y}_0, \boldsymbol{\mu})$$

denote the corresponding solution.

Theorem 9.13 (Continuous dependence on parameters)**.** *Suppose that* $\mathbf{f}$ *is continuous in* $(t, \mathbf{y}, \boldsymbol{\mu})$ *and locally Lipschitz continuous in* $\mathbf{y}$*, uniformly for* $\boldsymbol{\mu}$ *near* $\boldsymbol{\mu}_0$*. Assume that*

$$\mathbf{y}(t; \mathbf{y}_0, \boldsymbol{\mu}_0)$$

is defined on $[t_0, T]$*.*

Then, for $\boldsymbol{\mu}$ *sufficiently close to* $\boldsymbol{\mu}_0$*, the solution*

$$\mathbf{y}(t; \mathbf{y}_0, \boldsymbol{\mu})$$

is also defined on $[t_0, T]$*, and*

$$\mathbf{y}(t; \mathbf{y}_0, \boldsymbol{\mu}) \longrightarrow \mathbf{y}(t; \mathbf{y}_0, \boldsymbol{\mu}_0)$$

uniformly on $[t_0, T]$ *as* $\boldsymbol{\mu} \to \boldsymbol{\mu}_0$*.*

Proof. Write

$$\mathbf{z}(t) = \mathbf{y}(t; \mathbf{y}_0, \boldsymbol{\mu}_0), \qquad \mathbf{y}(t) = \mathbf{y}(t; \mathbf{y}_0, \boldsymbol{\mu}).$$

The reference trajectory

$$\Gamma = \{(t, \mathbf{z}(t)) : t \in [t_0, T]\}$$

is compact. Choose compact sets

$$\Gamma \subset \operatorname{int} K_0 \subset K_0 \subset \operatorname{int} K \subset K$$

contained in the domain of $\mathbf{f}$ for all $\boldsymbol{\mu}$ sufficiently close to $\boldsymbol{\mu}_0$. By the uniform local Lipschitz hypothesis, $\mathbf{f}$ is Lipschitz continuous in $\mathbf{y}$ on K, uniformly for such $\boldsymbol{\mu}$, with some constant L. Let

$$\rho = \operatorname{dist}(K_0, \partial K) > 0.$$

On any interval on which the perturbed solution remains in K,

$$\begin{aligned}
|\mathbf{y}(t) - \mathbf{z}(t)| &\le \int_{t_0}^{t} |\mathbf{f}(s, \mathbf{y}(s), \boldsymbol{\mu}) - \mathbf{f}(s, \mathbf{z}(s), \boldsymbol{\mu})|\, ds \\
&\quad + \int_{t_0}^{t} |\mathbf{f}(s, \mathbf{z}(s), \boldsymbol{\mu}) - \mathbf{f}(s, \mathbf{z}(s), \boldsymbol{\mu}_0)|\, ds \\
&\le L \int_{t_0}^{t} |\mathbf{y}(s) - \mathbf{z}(s)|\, ds \\
&\quad + \int_{t_0}^{t} |\mathbf{f}(s, \mathbf{z}(s), \boldsymbol{\mu}) - \mathbf{f}(s, \mathbf{z}(s), \boldsymbol{\mu}_0)|\, ds.
\end{aligned}$$

Gronwall's inequality therefore gives

$$|\mathbf{y}(t) - \mathbf{z}(t)| \le \int_{t_0}^{t} e^{L(t-s)} |\mathbf{f}(s, \mathbf{z}(s), \boldsymbol{\mu}) - \mathbf{f}(s, \mathbf{z}(s), \boldsymbol{\mu}_0)|\, ds.$$

Since $\mathbf{f}$ is continuous in $\boldsymbol{\mu}$, uniformly along the compact reference trajectory,

$$\max_{t_0 \le s \le T} |\mathbf{f}(s, \mathbf{z}(s), \boldsymbol{\mu}) - \mathbf{f}(s, \mathbf{z}(s), \boldsymbol{\mu}_0)| \longrightarrow 0 \qquad \text{as } \boldsymbol{\mu} \to \boldsymbol{\mu}_0.$$

Hence the right-hand side is smaller than ρ for all $t \in [t_0, T]$ when $\boldsymbol{\mu}$ is sufficiently close to $\boldsymbol{\mu}_0$. The perturbed trajectory therefore cannot reach ∂K before time T.

By repeated application of the local existence theorem, the perturbed solution can consequently be extended throughout $[t_0, T]$. The same estimate then yields

$$\sup_{t_0 \le t \le T} |\mathbf{y}(t) - \mathbf{z}(t)| \longrightarrow 0 \qquad \text{as } \boldsymbol{\mu} \to \boldsymbol{\mu}_0.$$

□

A quantitative estimate. Suppose, in addition, that

$$|\mathbf{f}(t, \mathbf{y}, \boldsymbol{\mu}) - \mathbf{f}(t, \mathbf{y}, \boldsymbol{\nu})| \le K |\boldsymbol{\mu} - \boldsymbol{\nu}|$$

on the region under consideration. Then

$$\begin{aligned}
&|\mathbf{y}(t; \mathbf{y}_0, \boldsymbol{\mu}) - \mathbf{y}(t; \mathbf{y}_0, \boldsymbol{\nu})| \\
&\qquad \le K |\boldsymbol{\mu} - \boldsymbol{\nu}| \int_{t_0}^{t} e^{L(t-s)}\, ds.
\end{aligned}$$

If $L > 0$, this becomes

$$|\mathbf{y}(t; \mathbf{y}_0, \boldsymbol{\mu}) - \mathbf{y}(t; \mathbf{y}_0, \boldsymbol{\nu})| \le \frac{K}{L} \left(e^{L(t-t_0)} - 1 \right) |\boldsymbol{\mu} - \boldsymbol{\nu}|.$$

For $L = 0$, the corresponding estimate is

$$|\mathbf{y}(t; \mathbf{y}_0, \boldsymbol{\mu}) - \mathbf{y}(t; \mathbf{y}_0, \boldsymbol{\nu})| \le K (t - t_0) |\boldsymbol{\mu} - \boldsymbol{\nu}|.$$

Exercise 9.5 (An inhomogeneous Gronwall estimate)**.** Let $\eta : [0,T] \to [0,\infty)$ be continuously differentiable and suppose that

$$\eta'(t) \le a\eta(t) + be^{-ct},$$

where $a, b, c > 0$.

Use Gronwall's inequality to show that

$$\eta(t) \le \eta(0)e^{at} + \frac{b}{a+c}\left(e^{at} - e^{-ct}\right).$$

Interpret the two terms as the effects of the initial value and the forcing term.

Exercise 9.6 (A contractive nonlinear equation)**.** Consider

$$y' = t - y^3.$$

Let $y(t)$ and $z(t)$ be two solutions on $[0,T]$ with

$$y(0) = y_0, \qquad z(0) = z_0.$$

Set

$$u(t) = |y(t) - z(t)|^2.$$

Show that

$$\begin{aligned} u'(t) &= -2(y(t) - z(t))(y(t)^3 - z(t)^3) \\ &= -2|y(t) - z(t)|^2\big(y(t)^2 + y(t)z(t) + z(t)^2\big) \le 0. \end{aligned}$$

Deduce that

$$|y(t) - z(t)| \le |y_0 - z_0|, \qquad 0 \le t \le T.$$

Explain why this proves uniqueness and continuous dependence on the initial value.

Exercise 9.7 (Perturbing the differential equation)**.** Let $\mathbf{y}$ and $\mathbf{z}$ satisfy

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}), \qquad \dot{\mathbf{z}} = \mathbf{f}(t, \mathbf{z}) + \mathbf{r}(t),$$

with

$$\mathbf{y}(t_0) = \mathbf{y}_0, \qquad \mathbf{z}(t_0) = \mathbf{z}_0.$$

Suppose that $\mathbf{f}$ is Lipschitz continuous in the state variable with constant L, and that

$$|\mathbf{r}(t)| \le \varepsilon$$

on $[t_0, T]$.

Show that

$$|\mathbf{y}(t) - \mathbf{z}(t)| \le e^{L(t-t_0)}|\mathbf{y}_0 - \mathbf{z}_0| + \varepsilon \int_{t_0}^{t} e^{L(t-s)}\, ds.$$

Deduce that, for $L > 0$,

$$|\mathbf{y}(t) - \mathbf{z}(t)| \le e^{L(t-t_0)}|\mathbf{y}_0 - \mathbf{z}_0| + \frac{\varepsilon}{L}\left(e^{L(t-t_0)} - 1\right).$$

State the corresponding estimate when $L = 0$.

Exercise 9.8 (Dependence on a parameter)**.** Consider

$$y' = \mu y + \sin t, \qquad y(0) = y_0.$$

Let y_μ and y_ν be the solutions corresponding to parameters μ and ν. Assume that

$$|y_\nu(t)| \le R$$

on $[0, T]$.

Show that

$$|y_\mu(t) - y_\nu(t)| \le |\mu| \int_0^t |y_\mu(s) - y_\nu(s)|\, ds + |\mu - \nu| \int_0^t |y_\nu(s)|\, ds.$$

Use Gronwall's inequality to prove

$$|y_\mu(t) - y_\nu(t)| \le R|\mu - \nu| \int_0^t e^{|\mu|(t-s)}\, ds.$$

Conclude that the solution depends continuously on the parameter μ on every common finite interval of existence.

9.3 Flows and Differentiable Dependence

For an autonomous differential equation, uniqueness organizes the solutions into a family of flow maps. When the vector field is continuously differentiable, these maps are differentiable with respect to the initial point, and their derivatives satisfy the variational equation. The resulting linearized flow describes the transport and deformation of infinitesimal vectors along trajectories.

9.3.1 Flow maps

Consider the autonomous system

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}),$$

where

$$\mathbf{f} : \Omega \subset \mathbb{R}^n \to \mathbb{R}^n$$

is locally Lipschitz continuous. For each initial value $\mathbf{y}_0 \in \Omega$, let

$$\mathbf{y}(t, \mathbf{y}_0)$$

denote the solution satisfying

$$\mathbf{y}(0, \mathbf{y}_0) = \mathbf{y}_0.$$

The local flow. For each time t for which the solution is defined, set

$$\varphi_t(\mathbf{y}_0) = \mathbf{y}(t, \mathbf{y}_0).$$

The map

$$\varphi_t : \mathbf{y}_0 \longmapsto \mathbf{y}(t, \mathbf{y}_0)$$

is called the *flow map* at time t.

For each t, define

$$\mathcal{D}_t = \{\mathbf{y}_0 \in \Omega : \text{the solution through } \mathbf{y}_0 \text{ exists up to time } t\}.$$

Then the flow map is

$$\varphi_t : \mathcal{D}_t \longrightarrow \Omega.$$

Moreover, if $\mathbf{y}_0 \in \mathcal{D}_t$, then

$$\varphi_t(\mathbf{y}_0) \in \mathcal{D}_{-t},$$

and

$$\varphi_{-t}(\varphi_t(\mathbf{y}_0)) = \mathbf{y}_0.$$

Hence

$$\varphi_t : \mathcal{D}_t \longrightarrow \mathcal{D}_{-t}$$

is a bijection with inverse φ_{-t}.

If $\Omega_0 \subset \mathcal{D}_t$ is a set of initial states, define its image at time t by

$$\Omega_t = \varphi_t(\Omega_0).$$

Thus

$$\varphi_t : \Omega_0 \longrightarrow \Omega_t$$

describes the deformation or transport of the initial set under the flow.

The flow property. Autonomy implies that restarting a trajectory at an intermediate time produces the same trajectory.

Proposition 9.14. *Whenever both sides are defined,*

$$\varphi_{t+s} = \varphi_t \circ \varphi_s.$$

Also,

$$\varphi_0 = \mathrm{id}\,.$$

Proof. Fix $\mathbf{y}_0$ and define

$$\mathbf{z}(t) = \varphi_t(\varphi_s(\mathbf{y}_0)).$$

Then $\mathbf{z}$ solves

$$\dot{\mathbf{z}} = \mathbf{f}(\mathbf{z})$$

and

$$\mathbf{z}(0) = \varphi_s(\mathbf{y}_0).$$

On the other hand,

$$\widetilde{\mathbf{z}}(t) = \varphi_{t+s}(\mathbf{y}_0)$$

solves the same equation and satisfies

$$\widetilde{\mathbf{z}}(0) = \varphi_s(\mathbf{y}_0).$$

Uniqueness therefore gives

$$\varphi_t(\varphi_s(\mathbf{y}_0)) = \varphi_{t+s}(\mathbf{y}_0).$$

The identity $\varphi_0 = \mathrm{id}$ follows from the initial condition. □

Thus the family $\{\varphi_t\}$ satisfies the composition law of time evolution.

For a nonautonomous equation, the corresponding maps depend on both the initial and final times. Writing

$$\Phi_{t,t_0}(\mathbf{y}_0) = \mathbf{y}(t; t_0, \mathbf{y}_0),$$

one has

$$\Phi_{t,s} \circ \Phi_{s,t_0} = \Phi_{t,t_0}.$$

Inverse maps. If $\Omega_0 \subset \mathcal{D}_t$ and

$$\Omega_t = \varphi_t(\Omega_0),$$

then uniqueness implies

$$\varphi_{-t} \circ \varphi_t = \mathrm{id} \qquad \text{on } \Omega_0,$$

and

$$\varphi_t \circ \varphi_{-t} = \mathrm{id} \qquad \text{on } \Omega_t.$$

Hence

$$\varphi_t : \Omega_0 \longrightarrow \Omega_t$$

is one-to-one and onto, with

$$\varphi_t^{-1} = \varphi_{-t}.$$

Nonintersection of trajectories. Suppose that two trajectories meet at the same time:

$$\varphi_t(\mathbf{y}_0) = \varphi_t(\mathbf{z}_0).$$

Viewed as solutions with initial time t, they have the same initial value. Uniqueness therefore implies that they coincide wherever both are defined. In particular, at time 0,

$$\mathbf{y}_0 = \mathbf{z}_0.$$

Thus distinct trajectories cannot intersect.

This property concerns uniqueness of trajectories through a point. It does not imply uniqueness of periodic orbits: different periodic orbits may coexist without intersecting.

Exercise 9.9 (Derivative of the inverse flow)**.** Let

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}), \qquad \mathbf{f} \in C^1(\mathbb{R}^n; \mathbb{R}^n),$$

and suppose that the flow is defined for $-T \le t \le T$.

1. Prove that
$$\varphi_{-t} = \varphi_t^{-1}.$$

2. Differentiate
$$\varphi_{-t}(\varphi_t(\mathbf{y}_0)) = \mathbf{y}_0$$
with respect to $\mathbf{y}_0$ to show that
$$D\varphi_{-t}(\varphi_t(\mathbf{y}_0)) = \big(D\varphi_t(\mathbf{y}_0)\big)^{-1}.$$

3. Explain why the variation matrix of a flow is never singular.

9.3.2 Differentiability with respect to initial data

The flow property follows from uniqueness. We now impose the stronger assumption that the vector field is continuously differentiable and study the dependence of the solution on its initial value.

Consider the nonautonomous problem
$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}), \qquad \mathbf{y}(t_0) = \mathbf{y}_0,$$
and write
$$\mathbf{y}(t, \mathbf{y}_0)$$
for the solution with initial value $\mathbf{y}_0$ at time t_0.

Suppose that $\mathbf{f}$ is continuously differentiable with respect to $\mathbf{y}$. Compare the solutions starting from $\mathbf{y}_0$ and $\mathbf{y}_0 + \mathbf{h}$. Set
$$\mathbf{y}(t) = \mathbf{y}(t, \mathbf{y}_0), \qquad \mathbf{y}_\mathbf{h}(t) = \mathbf{y}(t, \mathbf{y}_0 + \mathbf{h}),$$
and define
$$\mathbf{w}_\mathbf{h}(t) = \mathbf{y}_\mathbf{h}(t) - \mathbf{y}(t).$$
Then
$$\mathbf{w}_\mathbf{h}(t_0) = \mathbf{h}$$
and
$$\dot{\mathbf{w}}_\mathbf{h}(t) = \mathbf{f}(t, \mathbf{y}(t) + \mathbf{w}_\mathbf{h}(t)) - \mathbf{f}(t, \mathbf{y}(t)).$$
By the mean value formula,
$$\begin{aligned}\mathbf{f}(t, \mathbf{y}(t) + \mathbf{w}_\mathbf{h}(t)) &- \mathbf{f}(t, \mathbf{y}(t)) \\ &= \left[\int_0^1 D_\mathbf{y}\mathbf{f}\big(t, \mathbf{y}(t) + s\mathbf{w}_\mathbf{h}(t)\big)\, ds\right] \mathbf{w}_\mathbf{h}(t).\end{aligned}$$
Thus
$$\dot{\mathbf{w}}_\mathbf{h}(t) = \mathbf{A}_\mathbf{h}(t)\mathbf{w}_\mathbf{h}(t), \qquad \mathbf{w}_\mathbf{h}(t_0) = \mathbf{h},$$
where
$$\mathbf{A}_\mathbf{h}(t) = \int_0^1 D_\mathbf{y}\mathbf{f}\big(t, \mathbf{y}(t) + s\mathbf{w}_\mathbf{h}(t)\big)\, ds.$$

As $\mathbf{h} \to 0$, continuous dependence gives

$$\mathbf{w}_{\mathbf{h}} \to 0$$

uniformly on every common finite interval of existence. Hence

$$\mathbf{A}_{\mathbf{h}}(t) \longrightarrow D_{\mathbf{y}}\mathbf{f}(t, \mathbf{y}(t))$$

uniformly on that interval. This suggests that the first-order change in the solution should satisfy

$$\dot{\boldsymbol{\xi}} = D_{\mathbf{y}}\mathbf{f}(t, \mathbf{y}(t))\boldsymbol{\xi}.$$

Theorem 9.15 (Differentiable dependence on initial data)**.** *Suppose that* $\mathbf{f}$ *is continuous in* t *and continuously differentiable in* $\mathbf{y}$*. Let*

$$\mathbf{y}(t, \mathbf{y}_0)$$

be defined on $[t_0, T]$*. Then, for each fixed* $t \in [t_0, T]$*, the map*

$$\mathbf{y}_0 \longmapsto \mathbf{y}(t, \mathbf{y}_0)$$

is continuously differentiable near $\mathbf{y}_0$*.*

Its derivative

$$\mathbf{F}(t) = D_{\mathbf{y}_0}\mathbf{y}(t, \mathbf{y}_0)$$

satisfies

$$\dot{\mathbf{F}}(t) = D_{\mathbf{y}}\mathbf{f}(t, \mathbf{y}(t, \mathbf{y}_0))\mathbf{F}(t), \qquad \mathbf{F}(t_0) = \mathbf{I}.$$

Proof. Let $\mathbf{F}(t)$ be the solution of

$$\dot{\mathbf{F}}(t) = D_{\mathbf{y}}\mathbf{f}(t, \mathbf{y}(t, \mathbf{y}_0))\mathbf{F}(t), \qquad \mathbf{F}(t_0) = \mathbf{I}.$$

For a small increment $\mathbf{h}$, define the remainder

$$\mathbf{r}_{\mathbf{h}}(t) = \mathbf{y}(t, \mathbf{y}_0 + \mathbf{h}) - \mathbf{y}(t, \mathbf{y}_0) - \mathbf{F}(t)\mathbf{h}.$$

Using the integral equations for the three terms gives

$$\begin{aligned}\mathbf{r}_{\mathbf{h}}(t) = &\int_{t_0}^{t} D_{\mathbf{y}}\mathbf{f}(s, \mathbf{y}(s, \mathbf{y}_0))\mathbf{r}_{\mathbf{h}}(s)\, ds \\ &+ \int_{t_0}^{t} \mathbf{R}_{\mathbf{h}}(s)\, ds,\end{aligned}$$

where

$$\begin{aligned}\mathbf{R}_{\mathbf{h}}(s) = &\mathbf{f}(s, \mathbf{y}(s, \mathbf{y}_0 + \mathbf{h})) - \mathbf{f}(s, \mathbf{y}(s, \mathbf{y}_0)) \\ &- D_{\mathbf{y}}\mathbf{f}(s, \mathbf{y}(s, \mathbf{y}_0))\big(\mathbf{y}(s, \mathbf{y}_0 + \mathbf{h}) - \mathbf{y}(s, \mathbf{y}_0)\big).\end{aligned}$$

Since $D_{\mathbf{y}}\mathbf{f}$ is continuous and

$$\mathbf{y}(t, \mathbf{y}_0 + \mathbf{h}) - \mathbf{y}(t, \mathbf{y}_0) = O(|\mathbf{h}|)$$

uniformly on $[t_0, T]$, we have

$$\sup_{t_0 \le s \le T} |\mathbf{R}_{\mathbf{h}}(s)| = o(|\mathbf{h}|).$$

Gronwall's inequality then yields

$$\sup_{t_0 \le t \le T} |\mathbf{r}_\mathbf{h}(t)| = o(|\mathbf{h}|).$$

Therefore

$$\mathbf{y}(t, \mathbf{y}_0 + \mathbf{h}) = \mathbf{y}(t, \mathbf{y}_0) + \mathbf{F}(t)\mathbf{h} + o(|\mathbf{h}|),$$

which proves differentiability with respect to $\mathbf{y}_0$.

Continuity of the derivative follows from continuous dependence for the corresponding linear equation. □

The matrix

$$\mathbf{F}(t) = D_{\mathbf{y}_0}\mathbf{y}(t, \mathbf{y}_0)$$

describes the first-order response of the solution at time t to a perturbation of its initial value.

9.3.3 The variational equation and transport along flows

For an autonomous system

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}),$$

the derivative of the flow is

$$\mathbf{F}(t, \mathbf{y}_0) = D\varphi_t(\mathbf{y}_0).$$

By the preceding theorem, it satisfies

$$\dot{\mathbf{F}}(t, \mathbf{y}_0) = D\mathbf{f}(\varphi_t(\mathbf{y}_0))\mathbf{F}(t, \mathbf{y}_0), \qquad \mathbf{F}(0, \mathbf{y}_0) = \mathbf{I}.$$

This is the *variational equation* along the trajectory $\varphi_t(\mathbf{y}_0)$.

For an initial perturbation $\boldsymbol{\xi}_0$, the corresponding linearized perturbation is

$$\boldsymbol{\xi}(t) = \mathbf{F}(t, \mathbf{y}_0)\boldsymbol{\xi}_0,$$

and it satisfies

$$\dot{\boldsymbol{\xi}} = D\mathbf{f}(\varphi_t(\mathbf{y}_0))\boldsymbol{\xi}, \qquad \boldsymbol{\xi}(0) = \boldsymbol{\xi}_0.$$

Relation with linearization. The variational equation is the linearization of the differential equation along a trajectory. Writing

$$\widetilde{\mathbf{y}}(t) = \mathbf{y}(t) + \boldsymbol{\xi}(t),$$

we have

$$\begin{aligned} \dot{\widetilde{\mathbf{y}}} &= \mathbf{f}(\mathbf{y} + \boldsymbol{\xi}) \\ &= \mathbf{f}(\mathbf{y}) + D\mathbf{f}(\mathbf{y})\boldsymbol{\xi} + o(|\boldsymbol{\xi}|). \end{aligned}$$

Subtracting

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$$

and neglecting higher-order terms gives

$$\dot{\boldsymbol{\xi}} = D\mathbf{f}(\mathbf{y}(t))\boldsymbol{\xi}.$$

At an equilibrium $\mathbf{y}_*$,

$$\mathbf{f}(\mathbf{y}_*) = \mathbf{0},$$

the reference solution is constant:

$$\mathbf{y}(t) \equiv \mathbf{y}_*.$$

The variational equation becomes

$$\dot{\boldsymbol{\xi}} = D\mathbf{f}(\mathbf{y}_*)\boldsymbol{\xi},$$

which is precisely the linearization used in the study of equilibria.

Chain rule for the linearized flow. Differentiating the flow property

$$\varphi_{t+s} = \varphi_t \circ \varphi_s$$

with respect to the initial point gives

$$D\varphi_{t+s}(\mathbf{y}_0) = D\varphi_t(\varphi_s(\mathbf{y}_0))D\varphi_s(\mathbf{y}_0).$$

Thus the derivative of the flow satisfies the same composition law as successive linearized evolutions.

Since $D\varphi_t(\mathbf{y}_0)$ is invertible whenever the backward flow is defined,

$$D\varphi_t(\mathbf{y}_0)^{-1} = D\varphi_{-t}(\varphi_t(\mathbf{y}_0)).$$

Hence a C^1 local flow map is a diffeomorphism between its domain and its image.

Transport of the vector field. The vector field is tangent to its own flow. Differentiating

$$\varphi_{t+s}(\mathbf{y}_0) = \varphi_t(\varphi_s(\mathbf{y}_0))$$

with respect to s at $s = 0$ gives

$$\mathbf{f}(\varphi_t(\mathbf{y}_0)) = D\varphi_t(\mathbf{y}_0)\mathbf{f}(\mathbf{y}_0).$$

Thus the derivative of the flow transports the initial velocity $\mathbf{f}(\mathbf{y}_0)$ to the velocity along the trajectory at time t.

Transport of vorticity. An important application of the variational equation occurs in three-dimensional incompressible fluid flow. Let $\mathbf{v}$ be a sufficiently smooth solution of the incompressible Euler equations, and let

$$\frac{d}{dt}\varphi_t(\mathbf{a}) = \mathbf{v}(t, \varphi_t(\mathbf{a})), \qquad \varphi_0(\mathbf{a}) = \mathbf{a},$$

be its particle flow. Then

$$\mathbf{F}(t, \mathbf{a}) = D_{\mathbf{a}}\varphi_t(\mathbf{a})$$

satisfies

$$\frac{d}{dt}\mathbf{F}(t, \mathbf{a}) = D\mathbf{v}(t, \varphi_t(\mathbf{a}))\mathbf{F}(t, \mathbf{a}), \qquad \mathbf{F}(0, \mathbf{a}) = \mathbf{I}.$$

For the three-dimensional incompressible Euler equations, the vorticity

$$\boldsymbol{\omega} = \nabla \times \mathbf{v}$$

satisfies

$$\frac{d}{dt}\boldsymbol{\omega}(t,\varphi_t(\mathbf{a})) = D\mathbf{v}(t,\varphi_t(\mathbf{a}))\boldsymbol{\omega}(t,\varphi_t(\mathbf{a})).$$

Therefore

$$\boldsymbol{\omega}(t,\varphi_t(\mathbf{a})) = D_{\mathbf{a}}\varphi_t(\mathbf{a})\boldsymbol{\omega}_0(\mathbf{a}).$$

Thus the deformation of the fluid flow transports and stretches the vorticity vector.

Volume change and Liouville's formula. The derivative

$$\mathbf{F}(t,\mathbf{y}_0) = D\varphi_t(\mathbf{y}_0)$$

maps an infinitesimal displacement at $\mathbf{y}_0$ to the corresponding displacement at $\varphi_t(\mathbf{y}_0)$. In particular,

$$\det \mathbf{F}(t,\mathbf{y}_0)$$

is the local factor by which the flow changes volume.

Since

$$\dot{\mathbf{F}}(t,\mathbf{y}_0) = D\mathbf{f}(\varphi_t(\mathbf{y}_0))\mathbf{F}(t,\mathbf{y}_0), \qquad \mathbf{F}(0,\mathbf{y}_0) = \mathbf{I},$$

Jacobi's formula gives

$$\begin{aligned}\frac{d}{dt}\det \mathbf{F}(t,\mathbf{y}_0) &= \det \mathbf{F}(t,\mathbf{y}_0)\,\mathrm{tr}\left(\mathbf{F}(t,\mathbf{y}_0)^{-1}\dot{\mathbf{F}}(t,\mathbf{y}_0)\right)\\ &= \mathrm{div}\,\mathbf{f}(\varphi_t(\mathbf{y}_0))\det \mathbf{F}(t,\mathbf{y}_0).\end{aligned}$$

Therefore

$$\det D\varphi_t(\mathbf{y}_0) = \exp\left(\int_0^t \mathrm{div}\,\mathbf{f}(\varphi_s(\mathbf{y}_0))\,ds\right).$$

This identity is called *Liouville's formula.*

If

$$\mathrm{div}\,\mathbf{f} = 0,$$

then

$$\det D\varphi_t(\mathbf{y}_0) = 1.$$

Hence the flow preserves volume. More precisely, for every measurable set E on which the flow is defined,

$$|\varphi_t(E)| = \int_E |\det D\varphi_t(\mathbf{y}_0)|\; d\mathbf{y}_0 = |E|.$$

Thus a divergence-free vector field generates an incompressible, volume-preserving flow.

Periodic solutions and Floquet multipliers. Suppose that $\mathbf{y}_*(t)$ is a periodic solution of period T. The variational equation along this solution is

$$\dot{\boldsymbol{\xi}} = D\mathbf{f}(\mathbf{y}_*(t))\boldsymbol{\xi},$$

whose coefficient matrix is T-periodic. Let $\mathbf{X}(t)$ be its fundamental matrix with $\mathbf{X}(0) = \mathbf{I}$. The matrix

$$\mathbf{M} = \mathbf{X}(T)$$

is called the *monodromy matrix*, and its eigenvalues are the *Floquet multipliers*.

One Floquet multiplier is always equal to 1. Indeed,

$$\boldsymbol{\xi}(t) = \dot{\mathbf{y}}_*(t)$$

solves the variational equation and is T-periodic. This multiplier corresponds to a change of phase along the periodic orbit. The remaining multipliers describe growth or decay transverse to the orbit.

For a planar system, Liouville's formula gives

$$\det \mathbf{M} = \exp\left(\int_0^T \nabla \cdot \mathbf{f}(\mathbf{y}_*(t))\, dt\right).$$

Since one multiplier is 1, the other is

$$\rho = \exp\left(\int_0^T \nabla \cdot \mathbf{f}(\mathbf{y}_*(t))\, dt\right).$$

Thus the periodic orbit is transversely attracting when $\rho < 1$ and repelling when $\rho > 1$.

Exercise 9.10 (Sensitivity of an explicit nonlinear flow)**.** Consider

$$y' = -y^3, \qquad y(0) = y_0.$$

Find the solution $y(t, y_0)$ and compute

$$F(t) = \frac{\partial y(t, y_0)}{\partial y_0}$$

directly.

Next derive the variational equation

$$F'(t) = -3y(t, y_0)^2 F(t), \qquad F(0) = 1,$$

and verify that its solution agrees with the derivative obtained directly.

Explain why

$$0 < F(t) \leq 1 \qquad \text{for } t \geq 0.$$

Exercise 9.11 (Derivative of a linear flow)**.** Let

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}, \qquad \mathbf{y}(0) = \mathbf{y}_0,$$

where $\mathbf{A}$ is a constant matrix.

Show that

$$\varphi_t(\mathbf{y}_0) = e^{t\mathbf{A}}\mathbf{y}_0$$

and hence

$$D\varphi_t(\mathbf{y}_0) = e^{t\mathbf{A}}.$$

Verify directly that this matrix satisfies the variational equation

$$\dot{\mathbf{F}} = \mathbf{A}\mathbf{F}, \qquad \mathbf{F}(0) = \mathbf{I}.$$

Also verify the composition identity

$$D\varphi_{t+s} = D\varphi_t D\varphi_s.$$

Now consider the nonautonomous linear system

$$\dot{\mathbf{y}} = \mathbf{A}(t)\mathbf{y}, \qquad \mathbf{y}(t_0) = \mathbf{y}_0,$$

where $\mathbf{A}(t)$ is continuous, and let $\mathbf{F}(t, t_0)$ denote its state transition matrix.

1. Show that
$$\mathbf{y}(t) = \mathbf{F}(t, t_0)\mathbf{y}_0, \qquad D_{\mathbf{y}_0}\varphi_{t,t_0}(\mathbf{y}_0) = \mathbf{F}(t, t_0),$$
and verify that
$$\frac{\partial}{\partial t}\mathbf{F}(t, t_0) = \mathbf{A}(t)\mathbf{F}(t, t_0), \qquad \mathbf{F}(t_0, t_0) = \mathbf{I}.$$
2. Prove the identities
$$\mathbf{F}(t, s)\mathbf{F}(s, r) = \mathbf{F}(t, r), \qquad \mathbf{F}(s, t) = \mathbf{F}(t, s)^{-1}.$$
3. Explain why, in general, the evolution depends on both t and t_0, rather than only on the difference $t - t_0$.

Exercise 9.12 (Area preservation for a planar flow)**.** Consider the planar system

$$\dot{x} = \frac{\partial H}{\partial y}, \qquad \dot{y} = -\frac{\partial H}{\partial x},$$

where $H \in C^2(\mathbb{R}^2)$.

Show that the vector field is divergence free. Deduce from Liouville's formula that

$$\det D\varphi_t(x_0, y_0) = 1$$

whenever the flow is defined.

Conclude that the flow preserves planar area. Explain why this does not imply that distances or angles are preserved.

9.4 Continuation and Global Existence

The Cauchy–Lipschitz theorem gives a unique solution only on a small interval about the initial time. The next question is how far this solution can be continued. The obstruction may come from unbounded growth of the solution or from approaching the boundary of the domain of the vector field.

9.4.1 Maximal extensions and the continuation principle

Consider the initial value problem

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}), \qquad \mathbf{y}(t_0) = \mathbf{y}_0,$$

where

$$\mathbf{f} : \Omega \subset \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

is continuous and locally Lipschitz continuous in $\mathbf{y}$.

The Cauchy–Lipschitz theorem gives a unique solution on a small interval about t_0. We now determine how far this solution can be continued.

Definition 9.16. Let

$$\mathbf{y} : I \to \mathbb{R}^n$$

be a solution. A solution

$$\widetilde{\mathbf{y}} : \widetilde{I} \to \mathbb{R}^n$$

is called an *extension* of $\mathbf{y}$ if

$$I \subset \widetilde{I}$$

and

$$\widetilde{\mathbf{y}}(t) = \mathbf{y}(t), \qquad t \in I.$$

The solution $\mathbf{y}$ is called a *maximal extension* if it has no extension to a strictly larger interval. Its domain is called the *maximal interval of existence*.

Theorem 9.17 (Maximal extension)**.** *For every $(t_0, \mathbf{y}_0) \in \Omega$, the unique local solution of*

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}), \qquad \mathbf{y}(t_0) = \mathbf{y}_0,$$

has a unique maximal extension. Its maximal interval of existence has the form

$$I_{\max} = (\alpha, \beta), \qquad \alpha < t_0 < \beta,$$

where either endpoint may be finite or infinite.

Proof. Let $\mathcal{I}$ be the collection of all intervals containing t_0 on which the initial value problem has a solution, and define

$$I_{\max} = \bigcup_{I \in \mathcal{I}} I.$$

Since every interval in $\mathcal{I}$ contains t_0, their union is an interval containing t_0.

By local uniqueness, any two solutions with the same initial value agree wherever both are defined. Thus all solutions through $(t_0, \mathbf{y}_0)$ can be joined to define one solution on $I_{\max}$.

The interval $I_{\max}$ is open because the local existence theorem can be applied at every point of the solution. Hence

$$I_{\max} = (\alpha, \beta).$$

By construction, the solution cannot be extended to a larger interval. It is therefore the unique maximal extension. □

The endpoints of the maximal interval are determined by the behavior of the trajectory inside the domain Ω. A solution can fail to extend through a finite time only by escaping from every compact subset of Ω.

Theorem 9.18 (Continuation principle)**.** *Let*

$$\mathbf{y} : (\alpha, \beta) \to \mathbb{R}^n$$

be a maximal extension. If $\beta < \infty$, then

$$\big(t, \mathbf{y}(t)\big)$$

leaves every compact subset of Ω as $t \to \beta^-$.

Similarly, if $\alpha > -\infty$, then

$$\big(t, \mathbf{y}(t)\big)$$

leaves every compact subset of Ω as $t \to \alpha^+$.

Proof. Suppose that $\beta < \infty$ and that $(t, \mathbf{y}(t))$ does not leave every compact subset of Ω as $t \to \beta^-$. Then there exist a compact set $K \subset \Omega$ and a sequence

$$t_k \to \beta^-$$

such that

$$(t_k, \mathbf{y}(t_k)) \in K.$$

By compactness of K and the local existence theorem, there exists $\delta > 0$ such that every solution starting from a point of K exists for at least time δ forward. For sufficiently large k,

$$\beta - t_k < \delta.$$

Restart the solution at the point

$$(t_k, \mathbf{y}(t_k)).$$

The resulting local solution agrees with the original solution on their common interval by uniqueness, but it is defined beyond time β. This extends the original solution past β, contradicting maximality.

Therefore $(t, \mathbf{y}(t))$ leaves every compact subset of Ω as $t \to \beta^-$. The argument at the left endpoint is analogous. □

Thus a finite endpoint can occur in two basic ways: the solution may become unbounded, or the trajectory may approach the boundary of the domain of the vector field.

Finite-time blow-up. Consider

$$y' = y^2, \qquad y(0) = y_0.$$

For $y_0 \neq 0$, the solution is

$$y(t) = \frac{y_0}{1 - y_0 t}.$$

If $y_0 > 0$, then

$$I_{\max} = \left(-\infty, \frac{1}{y_0}\right),$$

and

$$y(t) \to \infty \qquad \text{as } t \to \frac{1}{y_0}^{-}.$$

Thus a smooth vector field defined on all of $\mathbb{R}$ may still produce finite-time blow-up.

If $y_0 < 0$, then

$$I_{\max} = \left(\frac{1}{y_0}, \infty \right),$$

while for $y_0 = 0$ the solution is identically zero and is defined for all time.

Approaching the boundary of the domain. Consider

$$y' = \frac{1}{1-y}, \qquad y(0) = 0,$$

with

$$\Omega = \{(t, y) : y < 1\}.$$

The solution is

$$y(t) = 1 - \sqrt{1 - 2t},$$

and its maximal interval is

$$I_{\max} = \left(-\infty, \frac{1}{2} \right).$$

As $t \to \frac{1}{2}^{-}$,

$$y(t) \to 1.$$

The solution remains bounded, but it approaches the boundary of Ω, where the vector field is not defined.

The blow-up alternative. When the vector field is defined on all of $\mathbb{R} \times \mathbb{R}^n$, there is no boundary of the domain to approach. The continuation principle therefore reduces to a blow-up criterion.

Corollary 9.19 (Blow-up alternative)**.** *Suppose that*

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

is continuous and locally Lipschitz continuous in $\mathbf{y}$*. Let*

$$\mathbf{y} : (\alpha, \beta) \to \mathbb{R}^n$$

be a maximal extension.

If $\beta < \infty$*, then*

$$|\mathbf{y}(t)| \to \infty \qquad \text{as } t \to \beta^{-}.$$

Similarly, if $\alpha > -\infty$*, then*

$$|\mathbf{y}(t)| \to \infty \qquad \text{as } t \to \alpha^{+}.$$

Proof. Suppose that $\beta < \infty$ but $|\mathbf{y}(t)|$ does not tend to infinity. Then there are $R > 0$ and times $t_k \to \beta^-$ such that

$$|\mathbf{y}(t_k)| \le R.$$

Local existence applied uniformly on the compact set

$$[\beta - 1, \beta + 1] \times \overline{B_{R+1}(\mathbf{0})}$$

gives a time $\delta > 0$ such that every solution starting from

$$[\beta - 1, \beta] \times \overline{B_R(\mathbf{0})}$$

exists for at least time δ.

For sufficiently large k,

$$\beta - t_k < \delta.$$

Restarting the solution at $(t_k, \mathbf{y}(t_k))$ then extends it beyond β, contradicting maximality. Therefore

$$|\mathbf{y}(t)| \to \infty \qquad \text{as } t \to \beta^-.$$

The proof at the left endpoint is analogous. □

Thus proving global existence is often reduced to obtaining an a priori bound for the solution on every finite time interval.

9.4.2 A priori estimates and global existence

Theorem 9.20 (Global existence from a priori bounds)**.** *Suppose that*

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

is continuous and locally Lipschitz continuous in $\mathbf{y}$*. Let*

$$\mathbf{y} : (\alpha, \beta) \to \mathbb{R}^n$$

be the maximal extension with initial condition

$$\mathbf{y}(t_0) = \mathbf{y}_0.$$

Assume that, for every $T > t_0$*, there exists a constant* $C_T < \infty$ *such that*

$$|\mathbf{y}(t)| \le C_T, \qquad t_0 \le t < \min\{T, \beta\}.$$

Then

$$\beta = \infty.$$

Hence $\mathbf{y}(t)$ *exists for all* $t \in [t_0, \infty)$*.*

Proof. If $\beta < \infty$, choose $T > \beta$. Then

$$|\mathbf{y}(t)| \le C_T, \qquad t_0 \le t < \beta,$$

which contradicts the blow-up alternative. □

We now give two common ways to obtain such bounds.

Linear growth of the vector field. A basic global existence condition is that the vector field grow at most linearly in $\mathbf{y}$.

Theorem 9.21 (Linear-growth criterion)**.** *Suppose that*

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

is continuous and locally Lipschitz continuous in $\mathbf{y}$*, and that*

$$|\mathbf{f}(t, \mathbf{y})| \leq a(t) + b(t)|\mathbf{y}|,$$

where a *and* b *are nonnegative continuous functions. Then every solution of*

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$$

is defined for all forward time.

Proof. For $t \geq t_0$, the integral equation gives

$$\begin{aligned} |\mathbf{y}(t)| &\leq |\mathbf{y}_0| + \int_{t_0}^{t} |\mathbf{f}(s, \mathbf{y}(s))|\, ds \\ &\leq |\mathbf{y}_0| + \int_{t_0}^{t} a(s)\, ds + \int_{t_0}^{t} b(s)|\mathbf{y}(s)|\, ds. \end{aligned}$$

Gronwall's inequality yields

$$|\mathbf{y}(t)| \leq \left(|\mathbf{y}_0| + \int_{t_0}^{t} a(s)\, ds \right) \exp\left(\int_{t_0}^{t} b(s)\, ds \right).$$

This is finite on every bounded time interval. Therefore the solution cannot blow up in finite time, and the continuation principle gives global existence. □

If a and b are constants, the estimate becomes

$$|\mathbf{y}(t)| \leq (|\mathbf{y}_0| + a(t - t_0))\, e^{b(t-t_0)}.$$

A slightly sharper estimate is available when $b > 0$:

$$|\mathbf{y}(t)| \leq |\mathbf{y}_0| e^{b(t-t_0)} + \frac{a}{b}\left(e^{b(t-t_0)} - 1 \right).$$

Energy and Lyapunov estimates. An a priori bound may be obtained from a nonnegative energy or Lyapunov function. The systematic study of Lyapunov functions will be given in Chapter 11.

Let

$$\Phi : \mathbb{R}^n \to [0, \infty)$$

be continuously differentiable and suppose that along solutions,

$$\frac{d}{dt}\Phi(\mathbf{y}(t)) \leq c_0 + c_1 \Phi(\mathbf{y}(t)),$$

where $c_0, c_1 \geq 0$. Gronwall's inequality gives

$$\Phi(\mathbf{y}(t)) \leq \Phi(\mathbf{y}_0)e^{c_1(t-t_0)} + \frac{c_0}{c_1}\left(e^{c_1(t-t_0)} - 1\right)$$

when $c_1 > 0$, with the usual modification

$$\Phi(\mathbf{y}(t)) \leq \Phi(\mathbf{y}_0) + c_0(t - t_0)$$

when $c_1 = 0$.

If, in addition,

$$\Phi(\mathbf{y}) \to \infty \qquad \text{as } |\mathbf{y}| \to \infty,$$

then a bound for $\Phi(\mathbf{y}(t))$ gives an a priori bound for $\mathbf{y}(t)$. This yields the following global existence criterion.

Corollary 9.22. *Suppose that*

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

is continuous and locally Lipschitz continuous in $\mathbf{y}$*. Suppose further that there is a continuously differentiable function*

$$\Phi : \mathbb{R}^n \to [0, \infty)$$

such that

$$\Phi(\mathbf{y}) \to \infty \qquad \textit{as } |\mathbf{y}| \to \infty,$$

and

$$\nabla\Phi(\mathbf{y}) \cdot \mathbf{f}(t, \mathbf{y}) \leq c_0 + c_1\Phi(\mathbf{y}).$$

Then every solution is defined for all forward time.

Example 9.23 (A priori bound for a nonlinear system)**.** Consider

$$\dot{x} = -x + y - x^3, \qquad \dot{y} = -x - y - y^3.$$

Set

$$\Phi(x, y) = \frac{1}{2}(x^2 + y^2).$$

Along a solution,

$$\begin{aligned}\frac{d}{dt}\Phi(x(t), y(t)) &= x\dot{x} + y\dot{y} \\ &= x(-x + y - x^3) + y(-x - y - y^3) \\ &= -x^2 - y^2 - x^4 - y^4 \\ &\leq 0.\end{aligned}$$

Therefore

$$\Phi(x(t), y(t)) \leq \Phi(x_0, y_0),$$

and hence

$$x(t)^2 + y(t)^2 \leq x_0^2 + y_0^2.$$

Thus the solution remains bounded for all forward time. By the continuation principle, it exists globally.

Here Φ is a Lyapunov function. Its role in this section is only to provide the a priori bound needed for global existence; Lyapunov stability will be studied systematically in Chapter 11.

Exercise 9.13 (Maximal intervals of existence)**.** Find the maximal interval of existence for each initial value problem.

1. $$y' = 1 + y^2, \qquad y(0) = y_0.$$

2. $$y' = \frac{1}{(1-y)^2}, \qquad y(0) = 0,$$

   where the vector field is defined for $y < 1$.

3. $$y' = -y^3, \qquad y(0) = y_0.$$

For each problem, determine the maximal interval of existence and whether the solution exists for all forward time. If continuation fails at a finite endpoint, determine whether the obstruction is finite-time blow-up or approach to the boundary of the domain.

Homework Problems

Homework 9.1 (Variation matrix along a planar periodic orbit)**.** Consider the planar autonomous system

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}), \qquad \mathbf{f} \in C^1(\mathbb{R}^2; \mathbb{R}^2),$$

and let

$$\mathbf{y}(t) = \varphi_t(\mathbf{y}_0)$$

be a nonconstant periodic solution of period T. Let

$$\mathbf{F}(t) = D\varphi_t(\mathbf{y}_0).$$

1. Show that

   $$\mathbf{F}(T)\mathbf{f}(\mathbf{y}_0) = \mathbf{f}(\mathbf{y}_0),$$

   and hence that 1 is an eigenvalue of $\mathbf{F}(T)$.

2. Use Liouville's formula to show that the other eigenvalue is

   $$\lambda = \exp\left(\int_0^T \operatorname{div} \mathbf{f}(\mathbf{y}(t))\, dt\right).$$

3. Explain what happens to linearized transverse perturbations when

$$\operatorname{div} \mathbf{f} < 0$$

along the periodic orbit.

Homework 9.2 (Global existence from a one-sided growth condition)**.** Let

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}), \qquad \mathbf{y}(t_0) = \mathbf{y}_0,$$

where

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

is continuous and locally Lipschitz continuous in $\mathbf{y}$. Assume that

$$\mathbf{y} \cdot \mathbf{f}(t, \mathbf{y}) \le a(t) + b(t)|\mathbf{y}|^2,$$

where a and b are nonnegative continuous functions.

1. Set

$$u(t) = |\mathbf{y}(t)|^2.$$

Show that

$$u'(t) \le 2a(t) + 2b(t)u(t).$$

2. Use Gronwall's inequality to obtain an explicit bound for $|\mathbf{y}(t)|^2$ on every finite interval $[t_0, T]$.

3. Deduce that the solution exists for all forward time.

4. Apply the result to the system

$$\dot{x} = y - x^3, \qquad \dot{y} = -x - y^3.$$

Notice that the vector field has cubic growth and therefore does not satisfy a global linear-growth bound.

Homework 9.3 (Sensitivity with respect to a parameter)**.** Consider the parameter-dependent initial-value problem

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}, \mu), \qquad \mathbf{y}(t_0, \mu) = \mathbf{g}(\mu),$$

where $\mathbf{f}$ and $\mathbf{g}$ are continuously differentiable. Assume that the solution exists on a common interval for the parameter values under consideration, and set

$$\mathbf{z}(t) = \frac{\partial \mathbf{y}}{\partial \mu}(t, \mu).$$

1. Show that $\mathbf{z}$ satisfies the inhomogeneous variational equation

$$\dot{\mathbf{z}} = D_{\mathbf{y}}\mathbf{f}(t, \mathbf{y}(t,\mu), \mu)\mathbf{z} + \frac{\partial \mathbf{f}}{\partial \mu}(t, \mathbf{y}(t,\mu), \mu), \qquad \mathbf{z}(t_0) = \mathbf{g}'(\mu).$$

2. Let $\mathbf{F}(t, s)$ be the state transition matrix for the associated homogeneous variational equation. Derive the variation-of-constants formula for $\mathbf{z}(t)$.

3. Apply the result to the logistic equation

$$y' = \mu y(1 - y), \qquad y(0) = y_0, \qquad 0 < y_0 < 1,$$

and verify the sensitivity by differentiating the explicit solution with respect to μ.

4. Interpret the sign of $\partial y/\partial \mu$ for $t > 0$.

Historical Notes and Further Reading

Historical notes. The early theory of ordinary differential equations concentrated largely on equations that could be integrated explicitly. During the nineteenth century, attention increasingly turned to a more fundamental question: under what conditions does an initial-value problem determine a solution, even when no formula for that solution is available? Cauchy gave some of the first rigorous local existence results for differential equations. Peano later showed that continuity of the vector field alone is sufficient for local existence, although uniqueness may fail. The method of successive approximations was subsequently developed by Picard and Lindelöf into the approach underlying the theorem now commonly called the Picard–Lindelöf or Cauchy–Lipschitz theorem.

In this method, the initial-value problem

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \qquad \mathbf{y}(t_0) = \mathbf{y}_0,$$

is rewritten as the integral equation

$$\mathbf{y}(t) = \mathbf{y}_0 + \int_{t_0}^{t} \mathbf{f}(s, \mathbf{y}(s))\, ds.$$

Successive substitution produces the Picard iterates. Under a Lipschitz condition, the iteration converges to a unique solution. Thus Peano's theorem separates existence from uniqueness, while the Picard–Lindelöf theorem provides both under the stronger Lipschitz hypothesis. The work of Picard (1890) and Lindelöf (1894) made successive approximation a basic tool in the theory of differential equations. Its modern formulation uses the contraction principle of Banach (1922), which places the argument in the general setting of complete metric spaces.

Existence and uniqueness alone do not fully express the reliability of a model. The solution should also vary continuously when the initial data, parameters, or vector field are perturbed. Estimates of this kind are based on an integral inequality associated with Gronwall (1919). Gronwall's inequality converts a local differential or integral estimate into quantitative control over a finite time interval. In this chapter it provides a common mechanism for proving uniqueness and continuous dependence on initial conditions and parameters.

When the vector field is differentiable, its solutions define a local flow, and the derivative of the flow with respect to the initial state satisfies the variational equation

$$\mathbf{Z}' = D_{\mathbf{y}}\mathbf{f}(t, \mathbf{y}(t))\mathbf{Z}.$$

Thus the linearized equation describes the transport of infinitesimal perturbations along a nonlinear trajectory. This connects the analytic foundation developed here with the linearization and stability methods used elsewhere in the book.

Local existence does not imply global existence. A solution may cease to exist in finite time because it becomes unbounded or approaches the boundary of the region on which the vector field is defined. Continuation theorems reduce global existence to the construction of suitable *a priori* bounds. Energy estimates, invariant regions, comparison arguments, and Lyapunov functions therefore serve not only to describe the dynamics but also to guarantee that the evolution remains defined for all relevant times.

Further reading. For classical treatments of existence, uniqueness, continuation, differential inequalities, and dependence on initial data and parameters, see Coddington and Levinson (1955) and Hartman (2002). A geometric viewpoint emphasizing flows and qualitative behavior is developed by Arnold (1992). More recent graduate-level accounts are given by Teschl (2012) and Viana and Espinar (2021), while Logemann and Ryan (2014) provides an upper-division treatment connecting rigorous initial-value theory with qualitative dynamics, applications, and control.

Chapter 10

Numerical Methods for Ordinary Differential Equations

10.1 From ODEs to Difference Equations

Many ordinary differential equations cannot be solved by explicit formulas. Even when formulas are available, numerical approximations are often needed for plotting, simulation, comparison with data, and exploration of parameter dependence. Numerical methods therefore form an essential part of the practical study of ordinary differential equations.

The purpose of this chapter is introductory. We do not develop numerical analysis as a separate subject. Instead, we explain how basic numerical schemes arise from the differential equation itself, and how one begins to judge their accuracy, stability, and qualitative behavior.

Throughout the chapter we mainly consider the scalar initial value problem

$$\dot{y} = f(t, y), \qquad y(t^0) = y^0.$$

The corresponding formulas for systems

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$$

are obtained by replacing y by $\mathbf{y}$ and f by $\mathbf{f}$. We shall not repeat this vector notation unless it is needed.

Choose time levels

$$t^0 < t^1 < \cdots < t^N,$$

and write

$$h_n = t^{n+1} - t^n$$

for the step size. A numerical method constructs approximations

$$y^n \approx y(t^n), \qquad n = 0, 1, \ldots, N.$$

Thus the continuous solution curve is replaced by a discrete sequence

$$y^0, y^1, \ldots, y^N.$$

Equivalently, the differential equation is replaced by a difference equation:

$$\dot{y} = f(t, y) \qquad \longrightarrow \qquad y^{n+1} = \Psi_{h_n}(t^n, y^n).$$

For each step size h_n, the map

$$(t^n, y^n) \longmapsto \Psi_{h_n}(t^n, y^n)$$

is the one-step numerical update.

There are two elementary ways to pass from the ODE to such a difference equation. The first is to approximate the derivative by a finite difference. For a constant step size h, the forward difference approximation is

$$\dot{y}(t^n) \approx \frac{y(t^{n+1}) - y(t^n)}{h}.$$

Substituting this into

$$\dot{y} = f(t, y)$$

gives

$$\frac{y(t^{n+1}) - y(t^n)}{h} \approx f(t^n, y(t^n)).$$

Replacing the exact values by numerical values gives

$$\frac{y^{n+1} - y^n}{h} = f(t^n, y^n),$$

or

$$y^{n+1} = y^n + hf(t^n, y^n).$$

This is Euler's method.

The second way is to use the integral formulation. Integrating the differential equation from t^n to t^{n+1} gives the exact identity

$$y(t^{n+1}) = y(t^n) + \int_{t^n}^{t^{n+1}} f(t, y(t))\, dt.$$

A numerical method is obtained by approximating the integral. If we use the left endpoint rule,

$$\int_{t^n}^{t^{n+1}} f(t, y(t))\, dt \approx hf(t^n, y(t^n)),$$

then we again obtain

$$y^{n+1} = y^n + hf(t^n, y^n).$$

Thus Euler's method may be viewed either as a finite difference method or as a quadrature method.

This second viewpoint is especially useful. More accurate methods are obtained by using better approximations of the integral. For example, one may use information at the midpoint, at a predicted endpoint, or at several intermediate points. This leads to Runge–Kutta methods.

Once an ODE has been replaced by a difference equation, new questions arise. How accurate is one step of the method? How do errors accumulate over many steps? When is the discrete evolution stable? What happens if the exact solution decays very rapidly? What happens if the exact solution oscillates for a long time? These questions lead to the main themes of this chapter:

discretization, accuracy, stability, stiffness, long-time structure.

10.2 Euler and Runge–Kutta Methods

In this section we introduce the basic explicit one-step methods. These methods compute y^{n+1} from known information at the current time step. Euler's method uses one slope. Runge–Kutta methods use several slopes inside one time step in order to approximate the accumulated change more accurately.

Throughout this section, we consider

$$\dot{y} = f(t, y), \qquad y(t^0) = y^0,$$

with constant step size

$$h = t^{n+1} - t^n.$$

10.2.1 Euler's method

Euler's method is

$$\boxed{y^{n+1} = y^n + hf(t^n, y^n).}$$

It is explicit: once y^n is known, the next value y^{n+1} is computed directly. No algebraic equation has to be solved.

Geometrically, Euler's method moves along the tangent line at the current point. At (t^n, y^n), the differential equation gives the slope

$$f(t^n, y^n).$$

Euler's method uses this slope as an approximation throughout the time interval $[t^n, t^{n+1}]$ and takes the step

$$y^n \longmapsto y^n + hf(t^n, y^n).$$

Thus the exact solution curve is approximated by a polygonal path.

Example 10.1. Consider

$$\dot{y} = y, \qquad y(0) = 1.$$

Euler's method gives

$$y^{n+1} = y^n + hy^n = (1 + h)y^n.$$

Therefore

$$y^n = (1 + h)^n.$$

At $t^n = nh$, the exact solution is

$$y(t^n) = e^{nh}.$$

Thus Euler's method replaces the exact one-step factor e^h by the numerical factor $1 + h$. The error at time t^n is therefore

$$y(t^n) - y^n = e^{nh} - (1 + h)^n.$$

For fixed $t^n = nh$, this error is of order $O(h)$ as $h \to 0$, consistent with the first-order accuracy of Euler's method.

Euler's method is simple and inexpensive. Its disadvantage is that it is only first-order accurate. More accurate methods use additional information inside each time step.

10.2.2 Second-order Runge–Kutta methods

Euler's method uses only the left endpoint slope

$$f(t^n, y^n).$$

A better approximation should use more information about the behavior of the solution inside the interval $[t^n, t^{n+1}]$. Second-order Runge–Kutta methods use two evaluations of f per step.

The midpoint method. The midpoint method is based on the midpoint approximation of the integral. Since the exact midpoint value $y(t^n + h/2)$ is unknown, we first estimate it by a half Euler step:

$$k_1 = f(t^n, y^n),$$
$$y^n + \frac{h}{2}k_1 \approx y\left(t^n + \frac{h}{2}\right).$$

Then we evaluate the slope at this approximate midpoint:

$$k_2 = f\left(t^n + \frac{h}{2}, y^n + \frac{h}{2}k_1\right).$$

The update is

$$\boxed{y^{n+1} = y^n + hk_2.}$$

Equivalently,

$$\begin{aligned} k_1 &= f(t^n, y^n), \\ k_2 &= f\left(t^n + \frac{h}{2}, y^n + \frac{h}{2}k_1\right), \\ y^{n+1} &= y^n + hk_2. \end{aligned}$$

The first slope k_1 is not used directly for the final update. It is used to estimate the midpoint value. The actual step is then made using the midpoint slope k_2.

Heun's method. Heun's method, also called the improved Euler method or the explicit trapezoidal method, is based on averaging an initial slope and an approximate endpoint slope.

First take an Euler prediction:

$$k_1 = f(t^n, y^n),$$
$$\widetilde{y}^{n+1} = y^n + hk_1.$$

Then evaluate the slope at the predicted endpoint:

$$k_2 = f(t^n + h, \widetilde{y}^{n+1}).$$

Finally, use the average of the two slopes:

$$\boxed{y^{n+1} = y^n + \frac{h}{2}(k_1 + k_2).}$$

Equivalently,

$$\begin{aligned} k_1 &= f(t^n, y^n), \\ k_2 &= f(t^n + h, y^n + hk_1), \\ y^{n+1} &= y^n + \frac{h}{2}(k_1 + k_2). \end{aligned}$$

The midpoint method and Heun's method are both second-order methods. They use two slopes, but they sample the vector field differently. The midpoint method uses a single approximate midpoint slope. Heun's method averages a beginning slope and an approximate endpoint slope.

Example 10.2. For

$$\dot{y} = ay,$$

the midpoint method gives

$$k_1 = ay^n,$$

and

$$k_2 = a\left(y^n + \frac{h}{2}k_1\right) = a\left(y^n + \frac{ah}{2}y^n\right).$$

Therefore

$$y^{n+1} = y^n + hk_2 = \left(1 + ah + \frac{(ah)^2}{2}\right)y^n.$$

Thus the numerical one-step factor agrees with

$$e^{ah} = 1 + ah + \frac{(ah)^2}{2} + O(h^3)$$

through terms of order h^2.

10.2.3 The classical fourth-order Runge–Kutta method

The most widely used elementary Runge–Kutta method is the classical fourth-order method, usually called RK4. It uses four slopes:

$$\begin{aligned} k_1 &= f(t^n, y^n), \\ k_2 &= f\left(t^n + \frac{h}{2}, y^n + \frac{h}{2}k_1\right), \\ k_3 &= f\left(t^n + \frac{h}{2}, y^n + \frac{h}{2}k_2\right), \\ k_4 &= f(t^n + h, y^n + hk_3). \end{aligned}$$

The update is

$$\boxed{y^{n+1} = y^n + \frac{h}{6}\left(k_1 + 2k_2 + 2k_3 + k_4\right).}$$

The four slopes sample the vector field at the beginning, two approximate midpoints, and the endpoint of the interval. The two midpoint slopes receive larger weights. The combination is chosen so that the method has local truncation error $O(h^5)$ and global error $O(h^4)$.

For many nonstiff problems, RK4 gives a good balance between accuracy and cost. It is explicit and easy to implement. Its limitation is that, like other explicit methods, it may require very small time steps for stiff equations.

Example 10.3. For

$$\dot{y} = ay,$$

RK4 gives

$$y^{n+1} = R(ah)y^n,$$

where

$$R(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24}.$$

Thus RK4 replaces the exact one-step factor e^z by its fourth-degree Taylor polynomial.

Exercise 10.1 (Midpoint and Heun methods for a linear equation)**.** Apply the midpoint method and Heun's method to

$$\dot{y} = ay.$$

Show that both methods give

$$y^{n+1} = \left(1 + ah + \frac{(ah)^2}{2}\right) y^n.$$

Thus, although the two methods evaluate the vector field at different points, they have the same update for this linear equation.

10.3 Accuracy and Convergence

After a numerical scheme has been constructed, the next question is accuracy. A method may look reasonable from its derivation, but we need a quantitative way to measure how well it approximates the exact solution.

There are two related but different errors. The first is the error made in one step, assuming that the method starts from the exact value. This is the *local truncation error*. The second is the accumulated error after many steps. This is the *global error*.

The distinction is important. A method may make a small error in each step, but many such errors accumulate over the whole interval. Moreover, previous errors may be amplified by the dynamics of the numerical scheme. Thus global accuracy depends not only on how well one step approximates the differential equation, but also on how errors propagate from step to step.

Consider a one-step numerical method written in the form

$$y^{n+1} = y^n + hF_h(t^n, y^n).$$

Here F_h is the numerical increment. For Euler's method,

$$F_h(t, y) = f(t, y).$$

For Runge–Kutta methods, F_h contains several evaluations of f during one time step.

10.3.1 Local truncation error

Let $y(t)$ be the exact solution of

$$\dot{y} = f(t, y).$$

If we insert the exact value $y(t^n)$ into one step of the numerical method, the method gives

$$y(t^n) + hF_h(t^n, y(t^n)).$$

The exact solution satisfies

$$\boxed{y(t^{n+1}) = y(t^n) + hF_h(t^n, y(t^n)) + \varepsilon^n.}$$

The quantity ε^n is the *local truncation error*, or one-step defect, of the method.

Method of order p. A method is said to have order p if, for sufficiently small h,

$$|\varepsilon_n| \leq Mh^{p+1}$$

uniformly on the time interval under consideration, where M is independent of h and n. Thus the local truncation error is of order $O(h^{p+1})$. For example, Euler's method is

$$y^{n+1} = y^n + hf(t^n, y^n).$$

Taylor's formula gives

$$y(t^{n+1}) = y(t^n) + h\dot{y}(t^n) + \frac{h^2}{2}\ddot{y}(t^n) + O(h^3).$$

Since $\dot{y}(t^n) = f(t^n, y(t^n))$, we obtain

$$y(t^{n+1}) = y(t^n) + hf(t^n, y(t^n)) + O(h^2).$$

Therefore Euler's method has local truncation error $O(h^2)$ and is a first-order method.

10.3.2 Global error and convergence

The global error at time t^n is

$$e^n = |y(t^n) - y^n|.$$

We now show how the global error is controlled by the local truncation errors.

Assume that the one-step method is

$$y^{n+1} = y^n + hF_h(t^n, y^n),$$

and that the exact solution satisfies

$$y(t^{n+1}) = y(t^n) + hF_h(t^n, y(t^n)) + \varepsilon^n.$$

Subtracting the numerical equation from the exact one-step relation gives

$$\begin{aligned} y(t^{n+1}) - y^{n+1} = {} & y(t^n) - y^n \\ & + h\left[F_h(t^n, y(t^n)) - F_h(t^n, y^n)\right] + \varepsilon^n. \end{aligned}$$

Assume that F_h is Lipschitz continuous in y, uniformly in t and h:

$$|F_h(t,y) - F_h(t,z)| \le L|y-z|.$$

Then

$$e^{n+1} \le e^n + hLe^n + |\varepsilon^n|.$$

If the method has order p, then

$$|\varepsilon^n| \le Mh^{p+1}.$$

Hence

$$e^{n+1} \le (1+hL)e^n + Mh^{p+1}.$$

Proposition 10.4 (Discrete Gronwall inequality)**.** *Let $\{a^n\}_{n=0}^N$ and $\{b^n\}_{n=0}^{N-1}$ be nonnegative sequences satisfying*

$$a^{n+1} \le (1+hL)a^n + hb^n, \qquad n = 0, 1, \dots, N-1,$$

where $h > 0$ and $L \ge 0$. Then

$$a^n \le (1+hL)^n a^0 + h\sum_{j=0}^{n-1}(1+hL)^{n-1-j}b^j.$$

In particular, if $t^n = nh \le T$, then

$$a^n \le e^{Lt^n}\left(a^0 + h\sum_{j=0}^{n-1} b^j\right).$$

Proof. Iterating the recurrence gives

$$a^n \le (1+hL)^n a^0 + h\sum_{j=0}^{n-1}(1+hL)^{n-1-j}b^j.$$

Since

$$(1+hL)^m \le e^{Lmh},$$

and $(n-1-j)h \le t^n$, we obtain

$$a^n \le e^{Lt^n}\left(a^0 + h\sum_{j=0}^{n-1} b^j\right).$$

□

Apply the proposition with

$$a^n = e^n, \qquad b^n = Mh^p.$$

Since $t^n = nh \le T$, we obtain

$$\begin{aligned} e^n &\le e^{Lt^n}\left(e^0 + h\sum_{j=0}^{n-1} Mh^p\right) \\ &= e^{Lt^n}\left(e^0 + Mt^n h^p\right) \\ &\le e^{LT}e^0 + MTe^{LT}h^p. \end{aligned}$$

Thus, for $0 \le t^n \le T$,

$$e^n \le e^{LT} e^0 + C_T h^p,$$

where

$$C_T = MTe^{LT}.$$

If the numerical method starts from the exact initial value, then $e^0 = 0$, and therefore

$$e^n \le C_T h^p.$$

Thus a local truncation error of order $O(h^{p+1})$ leads to a global error of order $O(h^p)$ on every fixed time interval.

Theorem 10.5 (Global error estimate)**.** *Consider the one-step method*

$$y^{n+1} = y^n + hF_h(t^n, y^n).$$

Assume that F_h is Lipschitz continuous in y, uniformly in t and h:

$$|F_h(t, y) - F_h(t, z)| \le L|y - z|,$$

where L is independent of h. Suppose also that the local truncation error satisfies

$$|\varepsilon_n| \le Mh^{p+1}.$$

Then, on a fixed time interval $0 \le t^n \le T$,

$$|y(t^n) - y^n| \le e^{LT}|y(t^0) - y^0| + C_T h^p.$$

In particular, if

$$y^0 = y(t^0),$$

then

$$|y(t^n) - y^n| = O(h^p)$$

uniformly for $0 \le t^n \le T$.

Closing. The preceding argument already contains a basic stability estimate: errors made at one step remain controlled on every fixed time interval. Later, in the section on stability and stiffness, we examine stability more directly and introduce the notion of absolute stability.

10.3.3 Adaptive step-size control

So far the step size h has been fixed. In many computations, a fixed step size is inefficient. The solution may change slowly on some intervals and rapidly on others. A good solver should use larger steps where the solution is smooth and smaller steps where the solution changes rapidly.

Adaptive methods adjust the step size during the computation. The basic idea is to estimate the local error made in one step. If the estimated error is too large, the step is rejected and recomputed with a smaller step size. If the estimated error is much smaller than required, the next step size may be increased.

A common approach is to use two approximations of different orders in the same step. Suppose these give

$$y_{\text{low}}^{n+1} \qquad \text{and} \qquad y_{\text{high}}^{n+1}.$$

Then the difference

$$y_{\text{high}}^{n+1} - y_{\text{low}}^{n+1}$$

is used as an estimate of the local error. Embedded Runge–Kutta pairs are based on this idea.

A standard example is MATLAB's `ode45`. It is an adaptive explicit Runge–Kutta method based on an embedded pair of orders four and five. The two approximations share most of the same stage evaluations, so their difference provides an error estimate at relatively little additional cost. The solver uses this estimate to accept or reject the current step and to choose the next step size.

Adaptive step-size control is important in practice, and it is used in many standard ODE solvers. However, it does not remove all difficulties. In particular, stiff equations may still force explicit methods to take extremely small steps. For such problems, implicit methods are often more appropriate.

Exercise 10.2 (Global error of Euler's method on the test equation)**.** Apply forward Euler with step size $h = T/N$ to

$$\dot{y} = ay, \qquad y(0) = y_0.$$

Show that

$$y^N = (1 + ah)^N y_0$$

and prove that $y^N \to e^{aT} y_0$ as $h \to 0$. By expanding $\log(1 + ah)$, show more precisely that

$$y(T) - y^N = \frac{1}{2} a^2 T e^{aT} y_0 h + O(h^2).$$

Exercise 10.3 (Local truncation error of Heun's method)**.** Consider the initial value problem

$$\dot{y} = y^2, \qquad y(0) = 1.$$

Apply one step of Heun's method with step size h to compute y^1. Using the exact solution, determine the one-step truncation error

$$\varepsilon^0 = y(h) - y^1,$$

and verify that it is of order $O(h^3)$.

Exercise 10.4 (Local truncation error of RK4)**.** For the classical RK4 method applied to

$$\dot{y} = ay,$$

show that

$$y^{n+1} = R(ah) y^n,$$

where

$$R(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24}.$$

Starting from the exact value $y(t^n)$, compare one RK4 step with

$$y(t^{n+1}) = e^{ah} y(t^n).$$

Show that the local truncation error is

$$\varepsilon^n = \left(e^{ah} - R(ah)\right) y(t^n)$$

and verify that

$$\varepsilon^n = \frac{(ah)^5}{120} y(t^n) + O(h^6).$$

Hence confirm that RK4 has local truncation error $O(h^5)$ for this equation.

Exercise 10.5 (Step doubling and an error estimate)**.** Suppose a one-step method of order p has local error

$$Ch^{p+1} + O(h^{p+2}).$$

Compare one step of length h, producing y_h, with two steps of length $h/2$, producing $y_{h/2}$.

1. Show that the leading local error of the two half-steps is

$$\frac{C}{2^p} h^{p+1} + O(h^{p+2}).$$

2. Show that

$$\frac{y_{h/2} - y_h}{2^p - 1}$$

estimates the error in the two-half-step result, up to the sign convention used for the error.

3. Explain how this comparison can be used for adaptive step-size control.

10.4 Stability and Stiffness

Accuracy measures how well a method approximates the exact solution when the step size is small. But accuracy is not the only issue. A numerical method also has to be stable: small errors made at one step should not be amplified uncontrollably by later steps.

This question is especially important for equations with rapid decay. The exact solution may be stable and quickly approach equilibrium, but an explicit numerical method may become unstable if the step size is too large. This phenomenon leads naturally to the notion of absolute stability and to the problem of stiffness.

10.4.1 Error propagation and absolute stability

Accuracy concerns how well one step of a method approximates the exact solution. Stability concerns how errors propagate after they have been made. These errors may come from local truncation error, round-off error, or small perturbations of the initial data.

Consider a one-step method

$$y^{n+1} = y^n + hF_h(t^n, y^n).$$

Suppose two numerical solutions are computed by the same method:

$$y^{n+1} = y^n + hF_h(t^n, y^n),$$

and

$$z^{n+1} = z^n + hF_h(t^n, z^n).$$

Definition 10.6 (Stability on a fixed time interval)**.** The one-step method is said to be stable on the interval $0 \le t^n \le T$ if there exists a constant $C_T > 0$, independent of the step size h and of n, such that

$$|y^n - z^n| \le C_T |y^0 - z^0|, \qquad 0 \le t^n \le T,$$

for any two numerical solutions generated by the method.

Thus stability means that small perturbations of the initial data remain controlled throughout the computation and are not amplified without bound as the step size is refined. It is the discrete analogue of continuous dependence on initial data for differential equations.

For one-step methods, this kind of estimate follows if F_h is Lipschitz continuous in y. Indeed, if

$$|F_h(t, y) - F_h(t, z)| \le L|y - z|,$$

then

$$\begin{aligned} |y^{n+1} - z^{n+1}| &\le |y^n - z^n| + h|F_h(t^n, y^n) - F_h(t^n, z^n)| \\ &\le (1 + hL)|y^n - z^n|. \end{aligned}$$

Iterating gives

$$|y^n - z^n| \le (1 + hL)^n |y^0 - z^0|.$$

If $t^n = nh \le T$, then

$$(1 + hL)^n \le e^{LT}.$$

Therefore

$$|y^n - z^n| \le e^{LT} |y^0 - z^0|.$$

Thus the numerical solution depends continuously on the initial data over each fixed time interval.

The preceding estimate is the discrete analogue of continuous dependence on initial data. It shows how small perturbations are propagated by the numerical scheme. This is the basic idea behind stability. In the error estimate of the previous section, the perturbations came from the local truncation errors. In actual computation, they may also come from round-off error or small changes in the initial data.

There is also a different but closely related question. For long-time computation, especially for decaying linear equations, we want the numerical method to reproduce the decay of the exact solution. This leads to the test equation and the notion of absolute stability.

The test equation and absolute stability. The basic test equation for stability is

$$\dot{y} = \lambda y,$$

where $\lambda \in \mathbb{C}$. Its exact solution is

$$y(t) = e^{\lambda t} y(0).$$

If

$$\Re\lambda < 0,$$

then the exact solution decays to zero as $t \to \infty$. A numerical method applied to this equation should also produce a decaying sequence, at least when the step size is chosen appropriately.

For a one-step method applied to the test equation, the numerical solution often has the form

$$y^{n+1} = R(h\lambda) y^n,$$

where $R(z)$ is called the stability function of the method. Thus

$$y^n = R(h\lambda)^n y^0.$$

The numerical solution decays precisely when

$$|R(h\lambda)| < 1.$$

This leads to the region of absolute stability:

$$\mathcal{S} = \{z \in \mathbb{C} : |R(z)| < 1\}.$$

If $h\lambda \in \mathcal{S}$, then the numerical method damps the test equation. If

$$|R(h\lambda)| = 1,$$

the numerical solution remains bounded but does not decay to zero. If

$$|R(h\lambda)| > 1,$$

the numerical solution grows in magnitude instead. Thus, when $\Re\lambda < 0$, both of the last two cases fail to reproduce the long-time decay of the exact solution.

Euler's method. For Euler's method,

$$y^{n+1} = y^n + h\lambda y^n = (1 + h\lambda) y^n.$$

Thus

$$R(z) = 1 + z.$$

The absolute stability condition is

$$|1 + z| < 1.$$

Therefore the stability region is the disk

$$\mathcal{S}_{\text{Euler}} = \{z \in \mathbb{C} : |1 + z| < 1\}.$$

This is the disk of radius 1 centered at -1 in the complex plane.

If $\lambda < 0$ is real, then the stability condition becomes

$$|1 + h\lambda| < 1.$$

Since $\lambda < 0$, this is equivalent to

$$-1 < 1 + h\lambda < 1.$$

Hence

$$0 < h < \frac{2}{|\lambda|}.$$

Thus Euler's method is stable for the decay equation $\dot{y} = \lambda y$ only if the step size satisfies

$$h < \frac{2}{|\lambda|}.$$

This restriction is not an accuracy condition. It is a stability condition. Even if one only wants a rough approximation of the decay, the step size must be small enough to prevent numerical growth or oscillation.

Runge–Kutta methods. For the midpoint method and Heun's method, applying the method to

$$\dot{y} = \lambda y$$

gives

$$y^{n+1} = \left(1 + z + \frac{z^2}{2}\right) y^n, \qquad z = h\lambda.$$

Thus

$$R(z) = 1 + z + \frac{z^2}{2}.$$

For RK4,

$$R(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24}.$$

These stability functions agree with the Taylor expansion of e^z up to the corresponding order. However, their stability regions are still bounded. Thus explicit Runge–Kutta methods also have step-size restrictions when they are applied to rapidly decaying equations.

Higher order improves accuracy, but it does not remove the basic stability restriction for explicit methods.

Exercise 10.6 (A-stability of backward Euler)**.** Apply backward Euler to the complex test equation

$$\dot{y} = \lambda y, \qquad z = h\lambda.$$

Show that its stability function is

$$R(z) = \frac{1}{1 - z}.$$

Prove that

$$|R(z)| < 1 \qquad \text{whenever} \qquad \Re z < 0.$$

Thus backward Euler is stable throughout the open left half-plane. Contrast this result with the bounded stability region of forward Euler.

10.4.2 Stiffness and implicit methods

Stiff decay. A simple model of stiffness is the fast decay equation

$$\dot{y} = -ay, \qquad a \gg 1.$$

The exact solution is

$$y(t) = e^{-at}y(0),$$

which decays on the short time scale

$$\tau = \frac{1}{a}.$$

After a short transient, the solution is already very close to zero.

Euler's method gives

$$y^{n+1} = (1 - ah)y^n.$$

The stability condition is

$$|1 - ah| < 1,$$

or

$$0 < h < \frac{2}{a}.$$

Thus the step size must resolve the very small time scale $1/a$, even if the solution has already decayed and is no longer changing significantly.

This is the basic difficulty of stiff equations. Stability, rather than accuracy, may force the step size to be extremely small.

Example 10.7. Consider

$$\dot{y} = -100y, \qquad y(0) = 1.$$

The exact solution is

$$y(t) = e^{-100t}.$$

It decays rapidly to zero. At $t = 1$, the exact solution is already extremely small:

$$e^{-100}.$$

Euler's method gives

$$y^{n+1} = (1 - 100h)y^n.$$

For stability, we need

$$|1 - 100h| < 1,$$

hence

$$0 < h < 0.02.$$

If $h = 0.05$, then

$$1 - 100h = -4,$$

and the numerical solution grows in magnitude by a factor of 4 at each step, with alternating sign. This numerical behavior is completely different from the exact decay.

This example shows that a stable differential equation can lead to an unstable numerical computation. The instability is not caused by the equation itself; it is caused by the combination of the method and the step size.

Backward Euler and implicit stability A common way to handle stiff decay is to use an implicit method. The simplest implicit method is *backward Euler*:

$$y^{n+1} = y^n + hf(t^{n+1}, y^{n+1}).$$

Unlike forward Euler, the unknown y^{n+1} appears on both sides of the equation. Therefore one may have to solve an algebraic equation at each time step.

For the test equation

$$\dot{y} = \lambda y,$$

backward Euler gives

$$y^{n+1} = y^n + h\lambda y^{n+1}.$$

Thus

$$(1 - h\lambda)y^{n+1} = y^n,$$

and

$$y^{n+1} = \frac{1}{1 - h\lambda} y^n.$$

The stability function is

$$R(z) = \frac{1}{1 - z}.$$

The absolute stability condition is

$$\left| \frac{1}{1 - z} \right| < 1,$$

or

$$|1 - z| > 1.$$

In particular, if $\lambda < 0$ is real and $h > 0$, then

$$z = h\lambda < 0,$$

and

$$|1 - z| = 1 + |z| > 1.$$

Therefore backward Euler is stable for every step size $h > 0$ on the negative real axis.

For the stiff decay equation

$$\dot{y} = -ay, \qquad a > 0,$$

backward Euler gives

$$y^{n+1} = \frac{1}{1 + ah} y^n.$$

The factor

$$\frac{1}{1+ah}$$

always has magnitude less than 1. Thus backward Euler damps the solution for every positive time step.

This property makes backward Euler useful for stiff problems. The price is that the method is implicit. For a nonlinear equation, each step requires solving

$$y^{n+1} = y^n + hf(t^{n+1}, y^{n+1}),$$

which may require Newton's method or another nonlinear solver.

Explicit versus implicit methods. Explicit methods compute y^{n+1} directly from known quantities. They are simple and efficient when the problem is nonstiff. However, they may require very small time steps for stability.

Implicit methods require the solution of an algebraic equation at each step. They are more expensive per step, but they may allow much larger stable time steps for stiff problems.

The choice between explicit and implicit methods is therefore not determined only by accuracy. It also depends on the time scales and stability properties of the differential equation.

Closing lesson. The stability analysis of the test equation gives several important lessons.

First, a numerical method defines its own discrete dynamics. Even if the exact solution decays, the numerical solution may grow if the step size is outside the stability region.

Second, explicit methods have bounded stability regions. For rapidly decaying solutions, this may impose a severe step-size restriction.

Third, stiffness occurs when stability forces a much smaller step size than accuracy alone would require. In such cases, implicit methods are often more appropriate.

Finally, stability is a property of the pair consisting of the method and the equation. A method that works well for one problem may be inefficient or unstable for another.

Exercise 10.7 (Stability interval of a second-order Runge–Kutta method)**.** For the midpoint method or Heun's method applied to

$$\dot{y} = \lambda y, \qquad \lambda < 0,$$

the stability function is

$$R(z) = 1 + z + \frac{z^2}{2}.$$

Determine all real values $z < 0$ for which

$$|R(z)| < 1.$$

Hence determine the restriction on the step size h.

Exercise 10.8 (Backward Euler for a nonlinear equation)**.** Apply backward Euler to

$$\dot{y} = -y^2, \qquad y(0) = y_0 > 0.$$

Show that y^{n+1} satisfies a quadratic equation. Determine the root that is consistent with

$$y^{n+1} \longrightarrow y^n \qquad \text{as } h \longrightarrow 0.$$

Show that the resulting numerical sequence remains positive and decreases monotonically.

10.5 Long-Time Behavior and Structure-Preserving Methods

Accuracy and stability are often studied on a fixed time interval. In many applications, however, the main concern is the behavior of the numerical solution over much longer times. One may want to know whether the solution remains bounded, approaches an equilibrium, oscillates with the correct phase, or preserves an energy-like quantity.

A numerical method may be accurate over short time intervals and yet give a qualitatively incorrect long-time picture. It may introduce artificial damping, artificial growth, or a slowly accumulating phase error. For Hamiltonian systems, it may also fail to preserve the underlying symplectic structure. Long-time computation therefore depends not only on local accuracy and stability, but also on how well the numerical method reproduces the geometric and dynamical structure of the differential equation.

In this section, we use the harmonic oscillator as a test problem. It provides a simple setting in which artificial growth, artificial damping, phase error, and energy drift can be seen explicitly. We first examine the failures of forward and backward Euler, then introduce symplectic methods, and finally compare structure preservation with the higher-order accuracy of RK4.

Test problem: the harmonic oscillator. Consider

$$\ddot{q} + \omega^2 q = 0, \qquad \omega > 0.$$

Writing

$$p = \dot{q},$$

we obtain the first-order system

$$\dot{q} = p, \qquad \dot{p} = -\omega^2 q.$$

Its energy is

$$E(q,p) = \frac{1}{2}p^2 + \frac{1}{2}\omega^2 q^2.$$

Along exact solutions,

$$\frac{dE}{dt} = p\dot{p} + \omega^2 q\dot{q} = p(-\omega^2 q) + \omega^2 qp = 0.$$

Thus the exact motion preserves energy, and the trajectories in the (q,p)-plane are the ellipses

$$\frac{1}{2}p^2 + \frac{1}{2}\omega^2 q^2 = E.$$

The solution therefore oscillates forever with constant amplitude. A good long-time numerical method should avoid artificial damping and artificial growth, while also keeping the phase of the oscillation reasonably accurate over many periods.

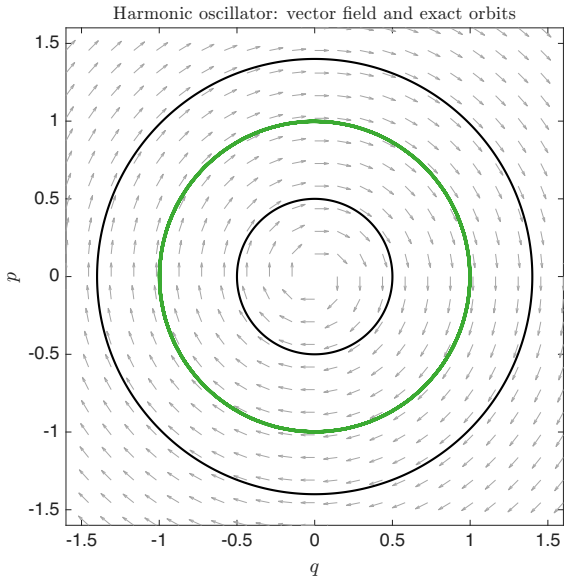


Figure 10.1: Phase plane of the harmonic oscillator $\dot{q} = p$, $\dot{p} = -\omega^2 q$. The exact trajectories are closed curves of constant energy.

The harmonic oscillator is therefore a useful test problem for long-time numerical methods. Artificial growth produces outward spiraling, artificial damping produces inward spiraling, and an incorrect numerical frequency produces accumulated phase error.

10.5.1 Artificial damping and growth

The harmonic oscillator provides a simple example in which standard numerical methods can produce qualitatively incorrect dynamics. Forward Euler creates energy, whereas backward Euler dissipates energy. Thus two closely related methods distort the same conservative system in opposite directions.

Forward Euler. Applying forward Euler to

$$\dot{q} = p, \qquad \dot{p} = -\omega^2 q,$$

gives

$$q^{n+1} = q^n + hp^n, \qquad p^{n+1} = p^n - h\omega^2 q^n.$$

For simplicity, take $\omega = 1$. Then

$$q^{n+1} = q^n + hp^n, \qquad p^{n+1} = p^n - hq^n.$$

The numerical energy is

$$E^n = \frac{1}{2}(p^n)^2 + \frac{1}{2}(q^n)^2.$$

A direct calculation gives

$$\begin{aligned} 2E^{n+1} &= (p^{n+1})^2 + (q^{n+1})^2 \\ &= (p^n - hq^n)^2 + (q^n + hp^n)^2 \\ &= (1+h^2)\left[(p^n)^2 + (q^n)^2\right]. \end{aligned}$$

Therefore

$$E^{n+1} = (1+h^2)E^n,$$

and hence

$$E^n = (1+h^2)^n E^0.$$

Thus forward Euler increases the energy at every step. The exact orbit is closed, but the numerical orbit spirals outward. The method introduces artificial amplitude growth even though the exact solution has constant amplitude.

Backward Euler. Backward Euler gives

$$q^{n+1} = q^n + hp^{n+1}, \qquad p^{n+1} = p^n - hq^{n+1}$$

when $\omega = 1$. Solving for the new variables gives

$$q^{n+1} = \frac{q^n + hp^n}{1+h^2}, \qquad p^{n+1} = \frac{p^n - hq^n}{1+h^2}.$$

Consequently,

$$\begin{aligned} 2E^{n+1} &= (p^{n+1})^2 + (q^{n+1})^2 \\ &= \frac{(p^n - hq^n)^2 + (q^n + hp^n)^2}{(1+h^2)^2} \\ &= \frac{1}{1+h^2}\left[(p^n)^2 + (q^n)^2\right]. \end{aligned}$$

Therefore

$$E^{n+1} = \frac{1}{1+h^2}E^n,$$

and hence

$$E^n = \frac{1}{(1+h^2)^n}E^0.$$

Thus backward Euler decreases the energy at every step. Its numerical orbit spirals inward toward the origin, producing artificial damping.

Forward and backward Euler therefore have opposite defects. Forward Euler creates energy, whereas backward Euler dissipates energy. Neither method reproduces the long-time conservative behavior of the harmonic oscillator.

Figure 10.2 makes the opposite defects of the two methods visible in the phase plane. Forward Euler expands the orbit from one revolution to the next, while backward Euler contracts it toward the equilibrium. These effects are not properties of the harmonic oscillator; they are artifacts of the numerical discretization.

Thus consistency and short-time accuracy alone do not guarantee correct long-time behavior. A method may approximate each individual step while systematically altering the amplitude and energy of the oscillation. This motivates numerical schemes that preserve more of the geometric structure of Hamiltonian dynamics.

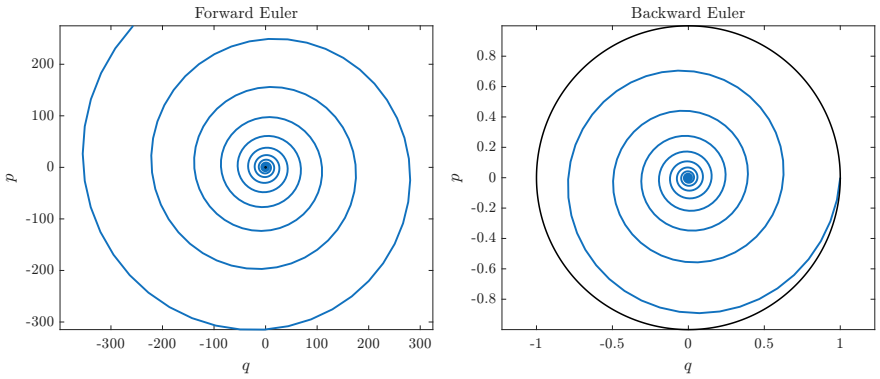


Figure 10.2: Phase-plane orbits for the harmonic oscillator computed by forward Euler and backward Euler. The exact orbit is shown in black. Forward Euler produces artificial energy growth and an outward spiral, whereas backward Euler produces artificial energy decay and an inward spiral.

10.5.2 Symplectic methods

The preceding examples show that a numerical method may systematically create or dissipate energy even when the exact system is conservative. Symplectic methods are designed to reduce this type of geometric distortion. A symplectic method need not preserve the exact energy at each step. Instead, it preserves the geometry of Hamiltonian motion, which often avoids the systematic artificial damping or growth seen in standard nonsymplectic discretizations.

Symplectic Euler. For the harmonic oscillator

$$\dot{q} = p, \qquad \dot{p} = -\omega^2 q,$$

the symplectic Euler method is

$$p^{n+1} = p^n - h\omega^2 q^n,$$
$$q^{n+1} = q^n + h p^{n+1}.$$

The momentum is first updated using the old position, and the position is then updated using the new momentum.

For $\omega = 1$, the method can be written as

$$\begin{bmatrix} q^{n+1} \\ p^{n+1} \end{bmatrix} = \begin{bmatrix} 1 - h^2 & h \\ -h & 1 \end{bmatrix} \begin{bmatrix} q^n \\ p^n \end{bmatrix}.$$

The determinant of the update matrix is

$$(1 - h^2) + h^2 = 1.$$

Thus the numerical map preserves oriented area in the (q, p)-plane and, in one degree of freedom, is symplectic. Unlike forward Euler, it does not systematically expand phase-space area, and unlike backward Euler, it does not systematically contract it.

The symplectic Euler method does not preserve the exact energy

$$E(q,p) = \frac{1}{2}p^2 + \frac{1}{2}q^2$$

at every step. For the harmonic oscillator with $\omega = 1$, the numerical orbit remains bounded when

$$0 < h < 2.$$

In this range, the eigenvalues of the update matrix lie on the unit circle, and the numerical points lie on an invariant ellipse close to the exact energy curve. The exact energy is not preserved, but its error remains bounded and oscillatory.

The symplectic Euler method is a first-order method. Its main advantage over forward and backward Euler is therefore not a higher order of accuracy, but the preservation of the symplectic structure.

Symplectic numerical schemes. In canonical coordinates, a Hamiltonian system has the form

$$\mathbf{z} = \begin{bmatrix} \mathbf{q} \\ \mathbf{p} \end{bmatrix}, \qquad \dot{\mathbf{z}} = \mathbf{J}\nabla H(\mathbf{z}),$$

where

$$\mathbf{J} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} \end{bmatrix}.$$

A one-step numerical method defines a map

$$\mathbf{z}^{n+1} = \Psi_h(\mathbf{z}^n).$$

Definition 10.8 (Symplectic numerical scheme)**.** A one-step method

$$\mathbf{z}^{n+1} = \Psi_h(\mathbf{z}^n)$$

for a Hamiltonian system is called symplectic if

$$\big(D\Psi_h(\mathbf{z})\big)^T \mathbf{J} D\Psi_h(\mathbf{z}) = \mathbf{J}$$

wherever $D\Psi_h(\mathbf{z})$ is defined.

In one degree of freedom, the symplectic condition is equivalent to

$$\det D\Psi_h(q,p) = 1.$$

Thus a symplectic numerical map preserves oriented area in the (q,p)-plane. A symplectic method need not preserve the exact energy at each step. Instead, it preserves the geometry of Hamiltonian motion, which often prevents the systematic artificial damping or growth seen in nonsymplectic methods.

The leapfrog method. For a mechanical Hamiltonian system

$$\dot{q} = p, \qquad \dot{p} = -\nabla V(q),$$

the leapfrog method is

$$p^{n+\frac{1}{2}} = p^n - \frac{h}{2}\nabla V(q^n),$$

$$q^{n+1} = q^n + hp^{n+\frac{1}{2}},$$

$$p^{n+1} = p^{n+\frac{1}{2}} - \frac{h}{2}\nabla V(q^{n+1}).$$

This method is also commonly called the Störmer–Verlet method.

For the harmonic oscillator,

$$V(q) = \frac{1}{2}\omega^2 q^2, \qquad \nabla V(q) = \omega^2 q,$$

so the method becomes

$$p^{n+\frac{1}{2}} = p^n - \frac{h}{2}\omega^2 q^n,$$

$$q^{n+1} = q^n + hp^{n+\frac{1}{2}},$$

$$p^{n+1} = p^{n+\frac{1}{2}} - \frac{h}{2}\omega^2 q^{n+1}.$$

The leapfrog method is explicit for mechanical systems of this form, second-order accurate, time-reversible, and symplectic. Thus it combines structure preservation with one order higher accuracy than symplectic Euler. These properties make it especially effective for long-time simulation of Hamiltonian systems.

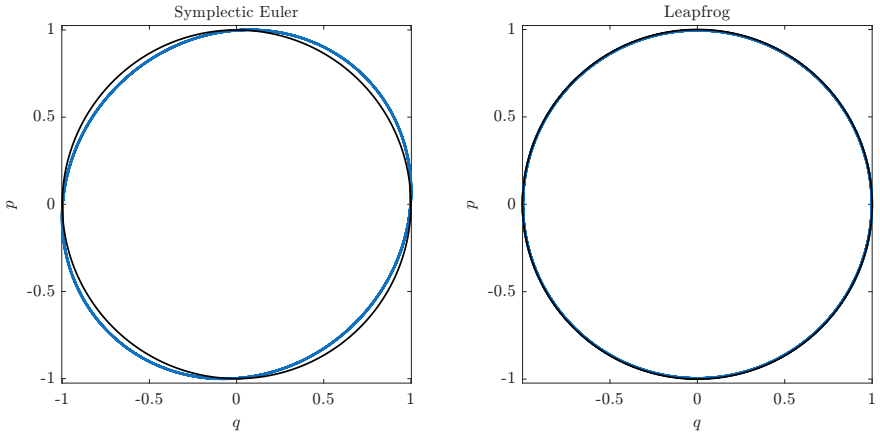


Figure 10.3: Phase-plane orbits for the harmonic oscillator computed by symplectic Euler and leapfrog. The exact orbit is shown in black. Both numerical methods avoid the systematic inward or outward spiraling of the forward and backward Euler methods. Leapfrog follows the exact orbit more closely, reflecting its higher order of accuracy.

Figure 10.3 shows the qualitative advantage of symplectic discretization. Both methods avoid the systematic inward or outward spiraling of the ordinary Euler methods. Symplectic Euler is only first-order accurate, so its orbit shows a more visible distortion from the exact ellipse. Leapfrog is second-order accurate and follows the exact orbit more closely. Thus symplecticity controls the qualitative geometry of the discrete motion, while the order of the method affects the size of the remaining numerical error.

10.5.3 Order versus structure preservation

The preceding comparison suggests that preserving the Hamiltonian structure is important for long-time computation. There is, however, another way to reduce numerical distortion: use a method of substantially higher order. To compare these two strategies, we apply the classical RK4 method to the same harmonic oscillator.

The classical RK4 method. For

$$\dot{q} = p, \qquad \dot{p} = -\omega^2 q,$$

the four RK4 stages are

$$\begin{aligned}
k_{1,q} &= p^n, & k_{1,p} &= -\omega^2 q^n, \\
k_{2,q} &= p^n + \frac{h}{2} k_{1,p}, & k_{2,p} &= -\omega^2 \left(q^n + \frac{h}{2} k_{1,q} \right), \\
k_{3,q} &= p^n + \frac{h}{2} k_{2,p}, & k_{3,p} &= -\omega^2 \left(q^n + \frac{h}{2} k_{2,q} \right), \\
k_{4,q} &= p^n + h k_{3,p}, & k_{4,p} &= -\omega^2 \left(q^n + h k_{3,q} \right).
\end{aligned}$$

The update is

$$\begin{aligned}
q^{n+1} &= q^n + \frac{h}{6} \left(k_{1,q} + 2k_{2,q} + 2k_{3,q} + k_{4,q} \right), \\
p^{n+1} &= p^n + \frac{h}{6} \left(k_{1,p} + 2k_{2,p} + 2k_{3,p} + k_{4,p} \right).
\end{aligned}$$

Since the system is linear, the update simplifies to

$$q^{n+1} = \left(1 - \frac{\omega^2 h^2}{2} + \frac{\omega^4 h^4}{24} \right) q^n + \left(h - \frac{\omega^2 h^3}{6} \right) p^n,$$

and

$$p^{n+1} = -\omega^2 \left(h - \frac{\omega^2 h^3}{6} \right) q^n + \left(1 - \frac{\omega^2 h^2}{2} + \frac{\omega^4 h^4}{24} \right) p^n.$$

Equivalently,

$$\begin{bmatrix} q^{n+1} \\ p^{n+1} \end{bmatrix} = \begin{bmatrix} \alpha_h & \beta_h \\ -\omega^2 \beta_h & \alpha_h \end{bmatrix} \begin{bmatrix} q^n \\ p^n \end{bmatrix},$$

where

$$\alpha_h = 1 - \frac{\omega^2 h^2}{2} + \frac{\omega^4 h^4}{24}, \qquad \beta_h = h - \frac{\omega^2 h^3}{6}.$$

RK4 is fourth-order accurate but is not symplectic. Its excellent behavior over the time interval considered below reflects its high order and the resulting small accumulated error, rather than exact preservation of the Hamiltonian geometry.

Amplitude, phase, and energy errors. For the harmonic oscillator with $\omega = 1$,

$$\dot{q} = p, \qquad \dot{p} = -q,$$

and initial data

$$q(0) = 1, \qquad p(0) = 0,$$

the exact solution is

$$q(t) = \cos t, \qquad p(t) = -\sin t.$$

Its amplitude and energy are constant:

$$A(t) = \sqrt{q(t)^2 + p(t)^2} = 1,$$

and

$$E(t) = \frac{1}{2}\left(q(t)^2 + p(t)^2\right) = \frac{1}{2}.$$

Its phase is $\theta(t) = t$. For a numerical solution, define

$$A^n = \sqrt{(q^n)^2 + (p^n)^2},$$

and let θ^n be the continuously unwrapped angle determined by

$$\theta^n = \operatorname{atan2}(-p^n, q^n) \pmod{2\pi},$$

with $\theta^0 = 0$. Also define

$$E^n = \frac{1}{2}\left((q^n)^2 + (p^n)^2\right).$$

We compare the three errors

$$A^n - 1, \qquad \theta^n - t^n, \qquad \frac{E^n - E^0}{E^0}.$$

The amplitude error measures artificial growth or decay of the oscillation. The phase error measures the accumulated error in the oscillation angle. Even when the amplitude remains bounded, a small error in the numerical frequency may accumulate over many periods. The relative energy error measures how well the conservative character of the system is reproduced.

Figures 10.4–10.6 compare forward Euler, backward Euler, symplectic Euler, leapfrog, and RK4. All five methods use the same initial condition, the same fixed step size

$$h = 0.15,$$

and the same final time. Thus the comparison involves methods of different orders as well as methods with different geometric properties.

Amplitude error. Forward Euler amplifies the motion, whereas backward Euler damps it. Symplectic Euler and leapfrog avoid systematic amplitude drift. RK4, although not symplectic, also preserves the amplitude extremely accurately over the displayed interval. In this experiment, its fourth-order accuracy compensates very effectively for its failure to preserve the symplectic structure exactly.

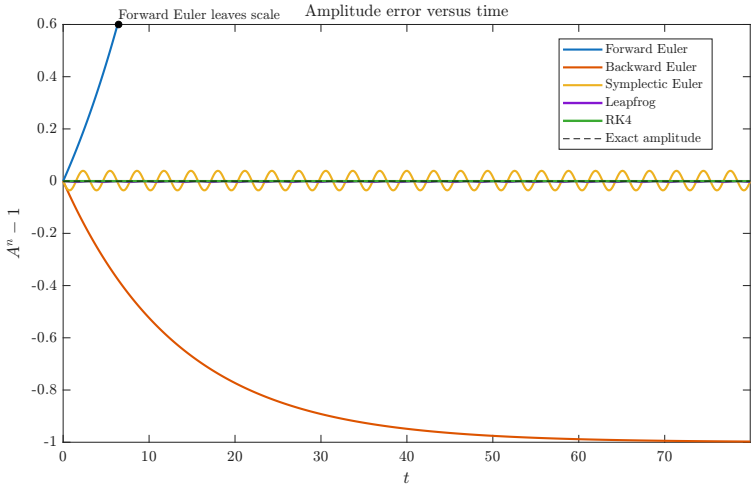


Figure 10.4: Amplitude error for the harmonic oscillator. Forward Euler produces rapid artificial growth and soon leaves the visible scale, whereas backward Euler produces artificial decay. Symplectic Euler has a bounded oscillatory amplitude error, leapfrog has a smaller bounded error, and the RK4 error is nearly indistinguishable from zero on the displayed time interval.

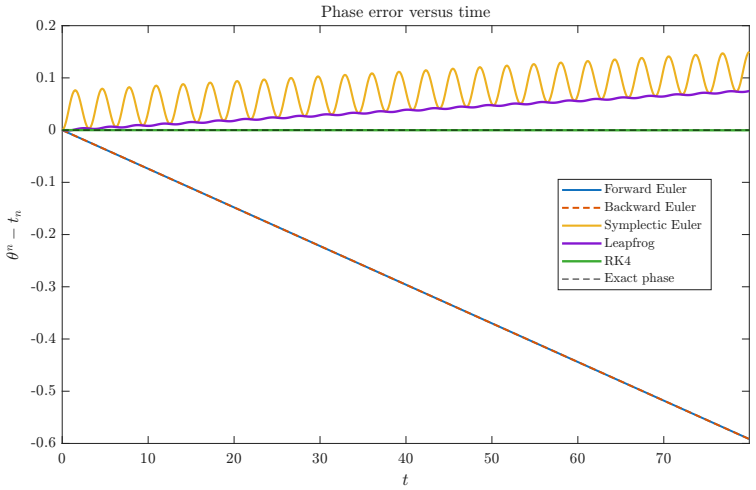


Figure 10.5: Phase error for the harmonic oscillator. Forward and backward Euler have the same steadily accumulating negative phase error. Symplectic Euler and leapfrog have smaller phase errors, while RK4 maintains an almost exact phase on the displayed time interval.

Phase error. Forward and backward Euler have the same numerical rotation angle and hence the same phase error, even though one increases the amplitude and the other decreases it. Symplectic Euler and leapfrog keep the orbit bounded, but they do not preserve the exact frequency. Their phase errors therefore accumulate with time. Leapfrog, being second-order accurate, has substantially smaller phase error than the first-order symplectic Euler method. RK4 has still smaller phase error on the displayed interval. Thus symplecticity alone does not imply the smallest phase error.

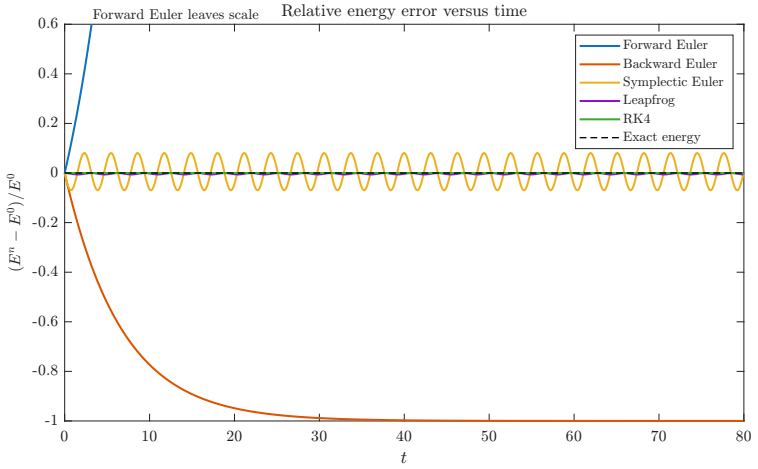


Figure 10.6: Relative energy error for the harmonic oscillator. Forward Euler creates artificial energy growth, whereas backward Euler creates artificial energy decay. The energy errors of symplectic Euler and leapfrog remain bounded and oscillatory. The RK4 energy error is extremely small on the displayed time interval.

Energy error. Forward and backward Euler show monotone energy drift. For the harmonic oscillator considered here, the symplectic methods do not preserve the exact energy at every step, but their energy errors remain bounded and oscillatory. RK4 is not symplectic and therefore has no general mechanism forcing its energy error to remain bounded for arbitrarily long times. Nevertheless, because it is fourth-order accurate, its energy drift is extremely small in the present computation.

Order versus structure preservation. These experiments show that order and structure preservation are different sources of numerical fidelity. A low-order symplectic method may preserve the Hamiltonian geometry while having a visible phase error. Conversely, a high-order nonsymplectic method may reproduce amplitude, phase, and energy with remarkable accuracy over a long but finite time interval. Thus higher order can compensate, sometimes very effectively, for failure to preserve the structure exactly.

High order and symplecticity are not interchangeable, however. Increasing the order reduces the local truncation error and can postpone visible drift for a long time. Symplecticity instead constrains the form of the discrete dynamics and often prevents systematic energy drift over much longer time scales. Which property is more important depends on the equation, the step size, the length of the simulation, and the quantity one wishes to reproduce.

Exercise 10.9 (Stability and symplecticity of the leapfrog method)**.** For the harmonic oscillator

$$\dot{q} = p, \qquad \dot{p} = -q,$$

the leapfrog method is

$$\begin{aligned} p^{n+\frac{1}{2}} &= p^n - \frac{h}{2} q^n, \\ q^{n+1} &= q^n + h p^{n+\frac{1}{2}}, \\ p^{n+1} &= p^{n+\frac{1}{2}} - \frac{h}{2} q^{n+1}. \end{aligned}$$

Eliminate the intermediate value $p^{n+\frac{1}{2}}$ and write the method as

$$\begin{bmatrix} q^{n+1} \\ p^{n+1} \end{bmatrix} = A_h \begin{bmatrix} q^n \\ p^n \end{bmatrix}.$$

Find the amplification matrix A_h and its eigenvalues. Determine the values of $h > 0$ for which the method is stable. Finally, verify that

$$A_h^T J A_h = J, \qquad J = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix},$$

and hence show that the leapfrog method is symplectic.

Exercise 10.10 (Stability of symplectic Euler)**.** For the harmonic oscillator

$$\dot{q} = p, \qquad \dot{p} = -q,$$

consider the symplectic Euler method

$$p^{n+1} = p^n - h q^n, \qquad q^{n+1} = q^n + h p^{n+1}.$$

1. Show that the update can be written as

$$\begin{bmatrix} q^{n+1} \\ p^{n+1} \end{bmatrix} = \mathbf{A}_h \begin{bmatrix} q^n \\ p^n \end{bmatrix}, \qquad \mathbf{A}_h = \begin{bmatrix} 1 - h^2 & h \\ -h & 1 \end{bmatrix}.$$

2. Compute the characteristic polynomial and show that the eigenvalues lie on the unit circle when $0 < h < 2$.

3. What happens when $h = 2$ and when $h > 2$?

10.6 Software for ODEs

In practical computations, one usually does not implement a numerical ODE solver from the beginning. Standard software provides reliable implementations of Runge–Kutta methods, implicit

methods, adaptive step-size control, event detection, and error estimation. Nevertheless, the ideas in the previous sections are important for understanding what the software is doing and for choosing a suitable solver.

Most ODE solvers require the user to provide the right-hand side of the system

$$\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y}),$$

together with an initial condition

$$\mathbf{y}(t_0) = \mathbf{y}_0$$

and a time interval

$$t_0 \le t \le T.$$

The solver then returns approximate values of $\mathbf{y}(t)$ at selected time points. In other words, a software solver usually expects the vector field $\mathbf{f}(t, \mathbf{y})$, the initial value $\mathbf{y}_0$, and the time interval of integration.

10.6.1 MATLAB

In MATLAB, the standard solver for nonstiff problems is `ode45`. It is based on an adaptive explicit Runge–Kutta method. For example, to solve

$$\dot{y} = -2y, \qquad y(0) = 1,$$

on $0 \le t \le 5$, one may write

```
f = @(t,y) -2*y;
[t,y] = ode45(f,[0 5],1);
```

The output arrays `t` and `y` contain the computed time points and solution values.

For systems, the unknown is represented by a vector. For example, the harmonic oscillator

$$\dot{q} = p, \qquad \dot{p} = -q$$

may be written as

$$\dot{\mathbf{y}} = \begin{bmatrix} y_2 \\ -y_1 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} q \\ p \end{bmatrix}.$$

In MATLAB one may write

```
f = @(t,y) [y(2); -y(1)];
[t,y] = ode45(f,[0 20],[1;0]);
```

10.6.2 Python

In Python, the function `solve_ivp` from `scipy.integrate` provides a common interface for solving initial value problems. For example,

$$\dot{y} = -2y, \qquad y(0) = 1,$$

may be solved by

```
from scipy.integrate import solve_ivp

def f(t, y):
    return -2*y

sol = solve_ivp(f, [0, 5], [1])
```

The computed time points are stored in `sol.t`, and the corresponding solution values are stored in `sol.y`.

For systems, the function returns a vector. The harmonic oscillator can be written as

```
def f(t, y):
    q, p = y
    return [p, -q]

sol = solve_ivp(f, [0, 20], [1, 0])
```

The default method is suitable for many nonstiff problems. For stiff problems, `solve_ivp` provides methods such as `Radau` and `BDF`; the option `LSODA` automatically switches between nonstiff and stiff integration methods.

10.6.3 Julia

Julia also has powerful ODE solvers, especially through the package `DifferentialEquations.jl`. A typical problem is specified by defining the right-hand side, the initial condition, and the time interval. For example, one may solve

$$\dot{y} = -2y, \qquad y(0) = 1,$$

by writing

```
using DifferentialEquations

f(u,p,t) = -2u
u0 = 1.0
tspan = (0.0,5.0)
prob = ODEProblem(f,u0,tspan)
sol = solve(prob)
```

For systems, the unknown is represented by a vector. For stiff problems, the Julia `DifferentialEquations.jl` ecosystem provides implicit solvers such as `Rodas5`, `Rosenbrock23`, and `TRBDF2`. For Hamiltonian systems, specialized packages in the same ecosystem provide symplectic methods such as `SymplecticEuler`, `VelocityVerlet`, and `VerletLeapfrog`.

The purpose here is not to teach a particular programming language, but to emphasize a common pattern. In all three environments, the mathematical problem is converted into the same basic ingredients:

right-hand side, initial condition, time interval, choice of solver.

10.6.4 Choosing a solver

The choice of solver depends on the differential equation and on the purpose of the computation. The following table gives a rough guide.

type of problem	typical choice
nonstiff problem on a finite interval	adaptive explicit Runge–Kutta method
stiff problem	implicit method or stiff solver
long-time Hamiltonian dynamics	symplectic or structure-preserving method
high accuracy on a smooth problem	higher-order adaptive method
rough exploration	standard adaptive solver

For nonstiff problems, explicit adaptive Runge–Kutta methods are often a good first choice. They are easy to use and efficient when the solution does not contain rapidly decaying components.

For stiff problems, explicit methods may take extremely small time steps because of stability restrictions. In this case, implicit solvers such as backward differentiation formula methods or implicit Runge–Kutta methods may be more effective.

For Hamiltonian or nearly conservative systems, standard adaptive solvers may produce good short-time accuracy but poor long-time behavior. If the goal is to simulate many periods or to understand qualitative dynamics, symplectic methods are often more appropriate.

A numerical solver should not be treated as a black box. The user should check whether the computed solution is reasonable, whether reducing the tolerance or step size changes the result significantly, and whether the method respects the important qualitative features of the problem.

Homework Problems

The following problems combine analysis with reproducible numerical experiments. For each computation, state the numerical method, step size, time interval, initial data, and error measure. When a reference solution is required, use an adaptive solver with relative and absolute tolerances at most 10^{-11}.

Homework 10.1 (Stiff relaxation toward a slowly varying state)**.** Consider

$$\dot{y} = -a(y - \cos t) - \sin t, \qquad a \gg 1.$$

1. Show that the general solution is

$$y(t) = \cos t + Ce^{-at}.$$

Identify the fast and slow time scales, and explain why solutions rapidly approach the slowly varying curve

$$y = \cos t.$$

2. Derive the forward Euler and backward Euler difference equations. Show that the difference between two forward Euler solutions is multiplied at each step by

$$1 - ah,$$

whereas the difference between two backward Euler solutions is multiplied by

$$\frac{1}{1+ah}.$$

Use these factors to compare the stability of the two methods.

3. Take

$$a = 1000, \qquad y(0) = 2, \qquad 0 \le t \le 2,$$

and compute the numerical solutions using

$$h = 5 \times 10^{-3}, \quad 2 \times 10^{-3}, \quad 10^{-3}, \quad 5 \times 10^{-4}.$$

Compare the results with the exact solution. For forward Euler, identify the unstable, boundary, and stable cases.

4. Repeat the computation using backward Euler. Explain why backward Euler remains stable even when the initial fast transient is not accurately resolved.

5. After the transient has decayed, measure the error in tracking $y = \cos t$. Distinguish the step-size restriction required for stability from that required for accuracy.

Homework 10.2 (Discovering fast and slow modes)**.** Consider

$$\ddot{\mathbf{q}} + \mathbf{K}\mathbf{q} = \mathbf{0}, \qquad \mathbf{K} = \frac{1}{2}\begin{bmatrix} 1+\omega^2 & 1-\omega^2 \\ 1-\omega^2 & 1+\omega^2 \end{bmatrix}, \qquad \omega \gg 1,$$

with Hamiltonian

$$H(\mathbf{q}, \mathbf{p}) = \frac{1}{2}\mathbf{p}^T\mathbf{p} + \frac{1}{2}\mathbf{q}^T\mathbf{K}\mathbf{q}, \qquad \mathbf{p} = \dot{\mathbf{q}}.$$

1. Take $\omega = 50$ and numerically solve the system with

$$\mathbf{p}(0) = \mathbf{0}$$

and, separately,

$$\mathbf{q}(0) = (1, 1)^T, \qquad \mathbf{q}(0) = (1, -1)^T.$$

From the computed motions, conjecture the two normal modes and their frequencies. Verify your conjecture by diagonalizing $\mathbf{K}$.

2. Introduce modal coordinates and express H as the sum of the two modal energies. Apply leapfrog to the coupled system and show that stability requires

$$h\omega < 2.$$

3. Choose initial data giving equal energy to the two modes. Integrate on $0 \le t \le 200$ using

$$h = 0.05, \qquad 0.03, \qquad 0.01.$$

Plot the modal energies and the total energy, and distinguish instability, phase error, and bounded energy oscillation.

4. Repeat the computations using RK4 with all three step sizes. Determine which computations are stable, and compare the phase and energy errors of RK4 and leapfrog.

Hint. Leapfrog applied to an oscillator of frequency ν is stable when $h\nu < 2$.

Project

Project 10.1 (Long-time dynamics near the pendulum separatrix)**.** Consider the simple pendulum

$$\dot{q} = p, \qquad \dot{p} = -\sin q,$$

with energy

$$E(q,p) = \frac{1}{2}p^2 + 1 - \cos q.$$

The separatrix has energy $E = 2$ and separates oscillatory motion from rotational motion.

1. Compute the energies of

$$(q_0, p_0) = (0.2, 0), \qquad (\pi - 0.05, 0), \qquad (0, 2.05),$$

and classify the corresponding motions as small-amplitude, near-separatrix, or rotational.

2. Integrate each solution to $T = 500$ using RK4 and leapfrog with

$$h = 0.2, \qquad 0.1, \qquad 0.05.$$

For $h = 0.1$, also compare forward Euler and symplectic Euler. Plot selected phase portraits and energy errors

$$E(q^n, p^n) - E(q^0, p^0)$$

that illustrate the qualitative differences among the methods.

3. Determine whether the methods produce artificial damping, systematic energy growth, or a spurious transition between oscillatory and rotational motion. Investigate whether reducing the step size removes the apparent change in the type of motion.

4. Compare representative fixed-step solutions with a high-accuracy adaptive reference solution. Discuss phase error, energy error, and any spurious crossing of the separatrix. Explain why bounded energy error does not imply small trajectory error over long times.

When comparing angles, use continuously unwrapped values of q so that accumulated phase error is retained.

Project 10.2 (Numerical predictability of the double pendulum)**.** Use the double-pendulum equations from Example 7.18, with

$$m_1 = m_2 = \ell_1 = \ell_2 = g = 1,$$

and the nearby initial conditions

$$\mathbf{y}_0 = (2, 1, 0, 0)^T, \qquad \widetilde{\mathbf{y}}_0 = (2 + 10^{-6}, 1, 0, 0)^T.$$

1. Solve both initial-value problems on $0 \le t \le 100$ using an adaptive solver with equal relative and absolute tolerances

   $$10^{-6}, \qquad 10^{-9}, \qquad 10^{-12}.$$

   Plot

   $$\|\mathbf{y}(t) - \widetilde{\mathbf{y}}(t)\|$$

   on a semilogarithmic scale, and compare the time at which this separation first reaches 10^{-1}.

2. Plot the relative energy errors. Compare each trajectory across successive tolerances, and determine whether small energy error guarantees an accurate trajectory.

3. Repeat the computation using fixed-step RK4 with

   $$h = 2^{-6}, \qquad 2^{-7}, \qquad 2^{-8}.$$

   Use refinement of the tolerance and step size to distinguish numerical error from sensitivity to the initial data.

Historical Notes and Further Reading

Historical notes. The method now called Euler's method is one of the earliest numerical procedures for initial-value problems. It appears in Euler's work on the integral calculus and approximates a solution curve by successive tangent line segments (Euler, 1768). In modern notation, the update

$$y^{n+1} = y^n + hf(t^n, y^n)$$

may be obtained either by replacing the derivative with a forward difference or by approximating the integral form of the differential equation with the left endpoint rule. These two interpretations already connect numerical integration with finite differences and quadrature.

Near the end of the nineteenth century, Runge (1895) developed methods that use several evaluations of the vector field within one time step. Kutta (1901) extended this approach and derived higher-order formulas, including the basis of the classical fourth-order Runge–Kutta method. Runge–Kutta methods achieve higher accuracy without requiring explicit derivatives of f beyond its values at selected intermediate stages.

As numerical computation became more systematic, it became necessary to distinguish consistency, convergence, and stability. The work of Dahlquist (1956) helped establish the modern stability theory of numerical ODE methods. Applying a method to the test equation

$$y' = \lambda y$$

reveals whether its discrete amplification factor reproduces the decay or growth of the exact solution. This viewpoint is especially important for stiff equations, where stability rather than accuracy alone may impose a severe restriction on the step size.

The widespread use of computers led to adaptive error control and robust software for nonstiff and stiff systems. It also made long-time simulation of mechanical and Hamiltonian systems practical. Such computations revealed that high local accuracy does not by itself guarantee correct long-time behavior: a method may introduce artificial damping, growth, phase drift, or energy drift. Geometric numerical integration therefore seeks methods that preserve important structures of the differential equation. Symplectic Euler and leapfrog, used in this chapter, are basic examples of this approach.

Further reading. For an accessible introduction to numerical methods for initial-value problems, including Runge–Kutta methods, error analysis, stability, and stiffness, see Iserles (2008). A comprehensive treatment of methods for nonstiff problems is given by Hairer et al. (1993), while stiff equations and implicit methods are developed in Hairer and Wanner (1996). For symplectic methods, backward error analysis, and the long-time numerical behavior of Hamiltonian systems, see Hairer et al. (2006).

Orientation to Part V

Part V studies the long-time behavior of autonomous differential equations. The existence and uniqueness theory of Part IV allows such an equation to generate a flow. The central question is no longer how to write an explicit solution, but where trajectories go and which geometric structures organize their motion.

The chapter begins with invariant sets and ω-limit sets, which describe the states approached along sequences of large times. Attractors and their basins then distinguish the sets that govern the asymptotic behavior of whole families of initial conditions. Absorbing sets provide a practical route to proving that such attracting dynamics is confined to a bounded region.

Dissipative systems supply an important source of this behavior. Mechanical energy, free energy, or another scalar quantity may decrease along trajectories. Lyapunov functions use such quantities to establish stability, while LaSalle's invariance principle identifies the limiting motion when the dissipation is only nonnegative rather than strictly positive. These ideas complement the conserved quantities of the variational and Hamiltonian systems studied in Part III.

The discussion then returns to the planar systems of Chapter 4. The Poincaré–Bendixson theorem severely restricts bounded recurrent motion in the plane, while Hopf bifurcation explains how a periodic orbit may be created when an equilibrium changes stability. In higher dimensions these planar restrictions disappear: contraction may coexist with stretching and folding, leading to sensitive dependence and strange attractors.

The central progression is

$$\begin{aligned}\text{flow and invariant sets} &\longrightarrow \omega\text{-limit sets} \longrightarrow \text{attractors and basins}\\ &\longrightarrow \text{Lyapunov and LaSalle methods} \longrightarrow \text{periodic orbits and Hopf bifurcation}\\ &\longrightarrow \text{higher-dimensional attracting dynamics.}\end{aligned}$$

This final part brings together the main themes of the book: modeling, geometry, stability, conservation and dissipation, bifurcation, and computation. It shows how these tools combine to describe dynamics even when individual trajectories cannot be solved explicitly.

Chapter 11

Dynamical Systems and Long-Time Behavior

Chapter 4 studied nonlinear systems through the geometry of the phase plane. Equilibria, nullclines, invariant regions, linearization, separatrices, and periodic orbits were used to construct phase portraits and to describe how trajectories are organized. That viewpoint answers the question: what does the motion look like?

The present chapter turns to a complementary question: what happens as $t \to \infty$? Although individual trajectories may be complicated, their long-time behavior is often governed by persistent structures of the flow. To describe this organization systematically, we introduce flows, invariant sets, ω-limit sets, attractors, and basins of attraction. In this way, the planar geometry of Chapter 4 becomes part of a broader theory of long-time dynamics.

Dissipative systems provide the main setting. In conservative systems, energy is preserved and trajectories often remain on energy levels. In dissipative systems, energy or another scalar quantity typically decreases along solutions, restricting the regions of phase space that can contain the future motion. Lyapunov functions turn this monotonicity into stability information, while LaSalle's invariance principle identifies the invariant set on which the dissipation vanishes.

We then return to planar systems. The Poincaré–Bendixson theorem strongly restricts the long-time behavior of bounded trajectories, while Hopf bifurcation explains a principal mechanism by which periodic orbits are created when an equilibrium loses stability. Finally, we contrast this planar theory with higher-dimensional systems, where more complicated attracting dynamics may occur.

The central progression of the chapter is therefore

flows and invariant sets $\longrightarrow$ ω-limit sets $\longrightarrow$ attractors and basins,
dissipation $\longrightarrow$ Lyapunov and LaSalle principles $\longrightarrow$ planar periodic dynamics and Hopf bifurcation $\longrightarrow$ higher-dimensional attracting dynamics.

11.1 Invariant Sets and ω-Limit Sets

Consider an autonomous system

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}), \qquad \mathbf{y} \in \mathbb{R}^n,$$

and denote by

$$\varphi_t(\mathbf{y}_0)$$

the solution at time t with initial value $\mathbf{y}_0$. Throughout this chapter, we assume that solutions exist for all $t \geq 0$, so that the system generates a forward flow, or *semiflow*, $\{\varphi_t\}_{t\geq 0}$. It satisfies

$$\varphi_0(\mathbf{y}_0) = \mathbf{y}_0, \qquad \varphi_{t+s}(\mathbf{y}_0) = \varphi_t\big(\varphi_s(\mathbf{y}_0)\big), \qquad s, t \geq 0.$$

The family $\{\varphi_t\}_{t\geq 0}$ is called the *semiflow generated by the system*. When solutions also exist for all negative times, the semiflow extends to a global flow $\{\varphi_t\}_{t\in\mathbb{R}}$.

In Chapter 4, invariant regions were used as geometric tools in the phase plane. Here the same invariance condition is applied to arbitrary subsets of the phase space. The change of terminology reflects the change of context: an invariant set need not have nonempty interior and may consist of a single equilibrium, a periodic orbit, a curve, a surface, or a more complicated subset of $\mathbb{R}^n$.

Definition 11.1. A set $A \subset \mathbb{R}^n$ is called *invariant* if

$$\varphi_t(A) = A$$

for every t for which the flow is defined. It is called *positively invariant* if

$$\varphi_t(A) \subset A, \qquad t \geq 0.$$

Equivalently, A is positively invariant if every solution starting in A remains in A for all forward time.

Full invariance means that trajectories remain in A in both forward and backward time. Positive invariance is the more natural notion in long-time dynamics, since the behavior as $t \to \infty$ depends only on the forward motion.

Typical invariant or positively invariant sets include the following.

- An equilibrium $\mathbf{y}_*$ is an invariant set consisting of one point:

  $$\varphi_t(\mathbf{y}_*) = \mathbf{y}_*.$$

- The orbit of a complete solution is invariant. In particular, a periodic orbit is an invariant closed curve.

- If H is a first integral, then each level set

  $$\{\mathbf{y} : H(\mathbf{y}) = c\}$$

  is invariant.

- If Φ is nonincreasing along trajectories, then every sublevel set

  $$\{\mathbf{y} : \Phi(\mathbf{y}) \leq c\}$$

  is positively invariant.

- The invariant regions introduced in Chapter 4 are typically positively invariant sets with nonempty interior in the phase plane.

- In population models, the nonnegative orthant

$$\mathbb{R}^n_+ = \{\mathbf{y} \in \mathbb{R}^n : y_j \geq 0,\ 1 \leq j \leq n\}$$

  is often positively invariant.

- Stable and unstable manifolds, when defined, are invariant sets.

- As we shall see, the ω-limit set of a bounded forward orbit is itself an invariant set.

Thus invariant sets provide a common language for many geometric objects encountered earlier. In Chapter 4, the emphasis was on regions that confine planar trajectories. Here the emphasis is on subsets that are preserved by the flow and that organize the long-time behavior of trajectories. An equilibrium is an invariant set consisting of a single point, while a periodic orbit is an invariant closed curve. In a conservative system, an energy level may be invariant and contain an entire family of trajectories. In a dissipative system, a positively invariant region may contain the future orbit without describing its eventual state; the trajectory may approach a smaller invariant subset, such as an equilibrium, a periodic orbit, or a more complicated attracting set. This distinction motivates the study of ω-limit sets.

11.1.1 Positive orbits and ω-limit sets

For an initial value $\mathbf{y}_0$, the set

$$\mathcal{O}^+(\mathbf{y}_0) = \{\varphi_t(\mathbf{y}_0) : t \geq 0\}$$

is called the *positive orbit* of $\mathbf{y}_0$. It records all states visited by the solution in forward time.

The positive orbit itself need not describe the long-time behavior. It contains the initial transient as well as the states visited much later. Its closure

$$\overline{\mathcal{O}^+(\mathbf{y}_0)}$$

also contains all accumulation points of the forward trajectory, but it still includes points visited only during the transient motion. To isolate the states that can be approached at arbitrarily large times, we introduce the ω-limit set.

Definition 11.2. The *ω-limit set* of $\mathbf{y}_0$ is

$$\omega(\mathbf{y}_0) = \{\mathbf{z} \in \mathbb{R}^n : \varphi_{t_k}(\mathbf{y}_0) \to \mathbf{z} \text{ for some sequence } t_k \to \infty\}.$$

Thus $\mathbf{z} \in \omega(\mathbf{y}_0)$ if the trajectory returns arbitrarily close to $\mathbf{z}$ at arbitrarily large times. The solution need not converge to $\mathbf{z}$, and it need not remain near $\mathbf{z}$ after approaching it.

There is an equivalent description of the ω-limit set that is often useful. For $T \geq 0$, define the tail of the positive orbit by

$$\mathcal{O}^+_T(\mathbf{y}_0) = \{\varphi_t(\mathbf{y}_0) : t \geq T\}.$$

Then

$$\omega(\mathbf{y}_0) = \bigcap_{T \geq 0} \overline{\mathcal{O}^+_T(\mathbf{y}_0)}.$$

Indeed, a point belongs to every closed orbit tail precisely when it can be approached by states $\varphi_t(\mathbf{y}_0)$ with arbitrarily large values of t.

The difference among these three sets is worth emphasizing:

$$\mathcal{O}^+(\mathbf{y}_0) \subset \overline{\mathcal{O}^+(\mathbf{y}_0)}, \qquad \omega(\mathbf{y}_0) \subset \overline{\mathcal{O}^+(\mathbf{y}_0)}.$$

The positive orbit records the motion, its closure adds all accumulation points, and the ω-limit set retains only those accumulation points that arise as $t \to \infty$.

Some basic cases illustrate the definition.

- If
$$\varphi_t(\mathbf{y}_0) \to \mathbf{y}_* \qquad \text{as } t \to \infty,$$
then
$$\omega(\mathbf{y}_0) = \{\mathbf{y}_*\}.$$
- If $\mathbf{y}_0$ lies on a nonconstant periodic orbit Γ, then
$$\omega(\mathbf{y}_0) = \Gamma.$$
- If $\mathbf{y}_0 = \mathbf{y}_*$ is an equilibrium, then
$$\omega(\mathbf{y}_*) = \{\mathbf{y}_*\}.$$
- If the positive orbit is unbounded, the ω-limit set may be empty.

The ω-limit set need not consist of a single point. It may be a periodic orbit, a collection of equilibria and connecting trajectories, or, in higher-dimensional systems, a more complicated invariant set. Its basic properties follow from compactness and continuity of the flow.

11.1.2 Basic properties of ω-limit sets

The usefulness of the ω-limit set comes from the fact that, under a compactness assumption, it is itself a well-structured invariant set.

Proposition 11.3. *Suppose that the positive orbit $\mathcal{O}^+(\mathbf{y}_0)$ is contained in a compact set $K \subset \mathbb{R}^n$. Then $\omega(\mathbf{y}_0)$ is nonempty, compact, connected, and invariant. Moreover,*

$$\operatorname{dist}\big(\varphi_t(\mathbf{y}_0), \omega(\mathbf{y}_0)\big) \longrightarrow 0 \qquad \text{as } t \to \infty.$$

Proof. For $T \geq 0$, let

$$K_T = \overline{\{\varphi_t(\mathbf{y}_0) : t \geq T\}}.$$

Each K_T is a nonempty compact subset of K, and the family is nested:

$$T_2 \geq T_1 \quad \Longrightarrow \quad K_{T_2} \subset K_{T_1}.$$

Since

$$\omega(\mathbf{y}_0) = \bigcap_{T \geq 0} K_T,$$

the nested-intersection property of compact sets shows that $\omega(\mathbf{y}_0)$ is nonempty and compact.

Each orbit tail

$$\{\varphi_t(\mathbf{y}_0) : t \geq T\}$$

is the continuous image of the connected interval $[T, \infty)$, and is therefore connected. Its closure K_T is also connected. Since the sets K_T are compact, connected, and nested, their intersection $\omega(\mathbf{y}_0)$ is connected.

To prove invariance, let $\mathbf{z} \in \omega(\mathbf{y}_0)$. Then there exists a sequence $t_k \to \infty$ such that

$$\varphi_{t_k}(\mathbf{y}_0) \to \mathbf{z}.$$

For any fixed $s \geq 0$, continuity of the semiflow and the semigroup property give

$$\varphi_s(\mathbf{z}) = \lim_{k\to\infty} \varphi_s\big(\varphi_{t_k}(\mathbf{y}_0)\big) = \lim_{k\to\infty} \varphi_{t_k+s}(\mathbf{y}_0).$$

Since $t_k + s \to \infty$, it follows that

$$\varphi_s(\mathbf{z}) \in \omega(\mathbf{y}_0).$$

Thus, for every $s \geq 0$,

$$\varphi_s\big(\omega(\mathbf{y}_0)\big) \subset \omega(\mathbf{y}_0).$$

To prove the reverse inclusion, let $\mathbf{z} \in \omega(\mathbf{y}_0)$. Choose a sequence $t_k \to \infty$ such that

$$\varphi_{t_k}(\mathbf{y}_0) \to \mathbf{z}.$$

For sufficiently large k, we have $t_k \geq s$. Since the positive orbit is contained in the compact set K, a subsequence of

$$\varphi_{t_k-s}(\mathbf{y}_0)$$

converges to some point $\mathbf{w} \in K$. Because $t_k - s \to \infty$, we have

$$\mathbf{w} \in \omega(\mathbf{y}_0).$$

By continuity of the semiflow,

$$\begin{aligned}\varphi_s(\mathbf{w}) &= \lim_{k\to\infty} \varphi_s\big(\varphi_{t_k-s}(\mathbf{y}_0)\big) \\ &= \lim_{k\to\infty} \varphi_{t_k}(\mathbf{y}_0) = \mathbf{z}.\end{aligned}$$

Therefore,

$$\omega(\mathbf{y}_0) \subset \varphi_s\big(\omega(\mathbf{y}_0)\big).$$

Combining the two inclusions gives

$$\varphi_s\big(\omega(\mathbf{y}_0)\big) = \omega(\mathbf{y}_0), \qquad s \geq 0.$$

Thus $\omega(\mathbf{y}_0)$ is invariant.

Finally, suppose that

$$\operatorname{dist}\big(\varphi_t(\mathbf{y}_0), \omega(\mathbf{y}_0)\big)$$

does not tend to zero. Then there exist $\varepsilon > 0$ and a sequence $t_k \to \infty$ such that

$$\operatorname{dist}\big(\varphi_{t_k}(\mathbf{y}_0), \omega(\mathbf{y}_0)\big) \geq \varepsilon.$$

Since the orbit is contained in the compact set K, a subsequence satisfies

$$\varphi_{t_{k_j}}(\mathbf{y}_0) \to \mathbf{z}$$

for some $\mathbf{z} \in K$. By the definition of the ω-limit set, $\mathbf{z} \in \omega(\mathbf{y}_0)$, contradicting the preceding inequality. Therefore

$$\operatorname{dist}\big(\varphi_t(\mathbf{y}_0), \omega(\mathbf{y}_0)\big) \to 0.$$

□

The compactness assumption is essential. Without it, a forward orbit may escape to infinity and have no accumulation point. The conclusion should also be interpreted carefully: the trajectory approaches its ω-limit set as a set, but it need not converge to a single point.

For example, if the ω-limit set is a periodic orbit Γ, then

$$\operatorname{dist}\big(\varphi_t(\mathbf{y}_0), \Gamma\big) \to 0,$$

although $\varphi_t(\mathbf{y}_0)$ continues to move around Γ and has no limit as $t \to \infty$.

Example 11.4 (An isolated periodic orbit as an ω-limit set)**.** Consider the planar system

$$\dot{x} = -y + x(1 - x^2 - y^2), \qquad \dot{y} = x + y(1 - x^2 - y^2).$$

Introduce polar coordinates

$$x = r\cos\theta, \qquad y = r\sin\theta.$$

Using

$$r\dot{r} = x\dot{x} + y\dot{y}, \qquad r^2\dot{\theta} = x\dot{y} - y\dot{x},$$

we obtain

$$\dot{r} = r(1 - r^2), \qquad \dot{\theta} = 1.$$

The circle

$$\Gamma = \{(x, y) : x^2 + y^2 = 1\}$$

is therefore invariant. On Γ, the angular variable satisfies

$$\theta(t) = t + \theta_0,$$

so Γ is a periodic orbit of period 2π.

The radial equation shows that

$$\dot{r} > 0 \quad \text{if } 0 < r < 1, \qquad \dot{r} < 0 \quad \text{if } r > 1.$$

Hence every solution with $r(0) > 0$ satisfies

$$r(t) \to 1 \qquad \text{as } t \to \infty.$$

Consequently,

$$\operatorname{dist}\big(\varphi_t(x_0, y_0), \Gamma\big) \to 0.$$

The trajectory does not converge to a single point, since $\theta(t) \to \infty$ and the solution continues to rotate around the circle. Instead,

$$\omega(x_0, y_0) = \Gamma \qquad \text{whenever } (x_0, y_0) \neq (0, 0).$$

The periodic orbit Γ is isolated: nearby circles are not periodic orbits, since their radii change according to

$$\dot{r} = r(1 - r^2).$$

Thus Γ is an attracting limit cycle rather than one member of a family of closed orbits.

Exercise 11.1 (Limit sets of a homoclinic orbit)**.** Suppose that $\mathbf{y}(t)$ is a complete homoclinic orbit to an equilibrium $\mathbf{y}_*$; that is,

$$\mathbf{y}(t) \longrightarrow \mathbf{y}_* \qquad \text{as } t \to \pm\infty.$$

The α-limit set of $\mathbf{y}(0)$ is defined by

$$\alpha(\mathbf{y}(0)) = \{\mathbf{z} : \mathbf{y}(t_k) \to \mathbf{z} \text{ for some sequence } t_k \to -\infty\}.$$

1. Show that
$$\alpha(\mathbf{y}(0)) = \omega(\mathbf{y}(0)) = \{\mathbf{y}_*\}.$$
2. Explain why the homoclinic orbit itself is neither its α-limit set nor its ω-limit set.
3. Compare this result with the periodic-orbit example, for which the entire orbit is the ω-limit set.

11.2 Attractors and Basins of Attraction

An ω-limit set describes the long-time behavior of a single trajectory, whereas an attractor describes the common long-time behavior of a family of trajectories. The corresponding basin of attraction identifies the initial conditions that share this behavior and thereby reveals how a region of phase space is organized.

11.2.1 Attraction to a Set

For a nonempty set $A \subset \mathbb{R}^n$, define the distance from $\mathbf{y} \in \mathbb{R}^n$ to A by

$$\operatorname{dist}(\mathbf{y}, A) = \inf_{\mathbf{a} \in A} |\mathbf{y} - \mathbf{a}|.$$

We say that a trajectory starting from $\mathbf{y}_0$ approaches A if

$$\operatorname{dist}\big(\varphi_t(\mathbf{y}_0), A\big) \longrightarrow 0 \qquad \text{as } t \to \infty.$$

Approach to a set does not necessarily mean convergence to a particular point. If $A = \{\mathbf{y}_*\}$ consists of one equilibrium, then

$$\operatorname{dist}\big(\varphi_t(\mathbf{y}_0), A\big) \to 0$$

is equivalent to

$$\varphi_t(\mathbf{y}_0) \to \mathbf{y}_*.$$

If A is a periodic orbit, however, a trajectory may approach A while continuing to move around it indefinitely.

Definition 11.5. A compact invariant set $A \subset \mathbb{R}^n$ is called an *attractor* if there exists a neighborhood U of A such that

$$\operatorname{dist}\big(\varphi_t(\mathbf{y}_0), A\big) \longrightarrow 0 \qquad \text{as } t \to \infty$$

for every $\mathbf{y}_0 \in U$.

Thus an attractor has two distinct properties. It is invariant, so the motion restricted to A remains in A, and it attracts all trajectories starting sufficiently near it.

Typical attractors include the following.

- An asymptotically stable equilibrium is an attractor consisting of a single point.
- An attracting limit cycle is an attractor consisting of a periodic orbit.
- An attractor may contain several equilibria together with connecting trajectories.
- An attractor may be a continuum of equilibria or other invariant states.
- In higher-dimensional systems, an attractor may have a more complicated invariant structure.

An attractor need not consist entirely of stable individual motions. It may contain unstable equilibria, connecting trajectories, or other invariant subsets. What is attracting is the set as a whole, not necessarily every trajectory contained in it.

The definition does not require an attractor to be minimal. Consequently, a larger compact invariant set may satisfy the definition together with a smaller attracting subset. When this occurs, we shall identify explicitly the smaller set that carries the essential long-time dynamics.

11.2.2 Basins of attraction

Definition 11.6. The *basin of attraction* of an attractor A is

$$\mathcal{B}(A) = \left\{\mathbf{y}_0 : \operatorname{dist}\big(\varphi_t(\mathbf{y}_0), A\big) \to 0 \text{ as } t \to \infty\right\}.$$

The basin contains a neighborhood of A, but it may extend far beyond that neighborhood. It is positively invariant: if $\mathbf{y}_0 \in \mathcal{B}(A)$, then

$$\varphi_s(\mathbf{y}_0) \in \mathcal{B}(A), \qquad s \geq 0.$$

Proposition 11.7. *The basin of an attractor is open.*

Proof. Let U be a neighborhood of A attracted to A, and let $\mathbf{y}_0 \in \mathcal{B}(A)$. Since

$$\operatorname{dist}\big(\varphi_t(\mathbf{y}_0), A\big) \to 0,$$

there is a time T such that $\varphi_T(\mathbf{y}_0) \in U$. By continuity of φ_T, all initial conditions sufficiently close to $\mathbf{y}_0$ are also mapped into U at time T. Their trajectories therefore approach A. Hence a neighborhood of $\mathbf{y}_0$ lies in $\mathcal{B}(A)$. □

When several disjoint attractors coexist, their basins are disjoint open sets. The basin boundaries contain initial conditions with different long-time behavior and are often formed by stable manifolds of saddle equilibria or unstable periodic orbits.

Example 11.8 (Two basins separated by a saddle)**.** Consider

$$V(x,y) = (x^2-1)^2 - x^2 + y^2$$

and the system

$$\dot{\mathbf{y}} = -\nabla V(\mathbf{y}), \qquad \mathbf{y} = (x,y)^T.$$

Systems of this form are called *gradient systems*. Their general properties will be studied in Section 11.3.3. Here we use this example only to illustrate attractors, basins, and basin boundaries.

Since

$$V_x = 4x^3 - 6x, \qquad V_y = 2y,$$

the system is

$$\dot{x} = 2x(3-2x^2), \qquad \dot{y} = -2y.$$

The equilibria are

$$P_\pm = \left(\pm\sqrt{\frac{3}{2}}, 0\right), \qquad P_0 = (0,0).$$

The equilibria $P_\pm$ are asymptotically stable, whereas P_0 is a saddle.

Since the sign of $x(t)$ is preserved and the scalar equation for x shows that

$$x(t) \longrightarrow \begin{cases} \sqrt{3/2}, & x(0) > 0, \\ -\sqrt{3/2}, & x(0) < 0, \end{cases}$$

while $y(t) = y(0)e^{-2t}$, we obtain

$$\mathcal{B}(P_+) = \{(x,y) : x > 0\}, \qquad \mathcal{B}(P_-) = \{(x,y) : x < 0\}.$$

The common basin boundary is

$$\{(x,y) : x = 0\} = W^s(P_0).$$

Every initial condition on this boundary approaches the saddle rather than either attracting equilibrium. Thus a basin boundary may itself be an invariant set with a distinct long-time behavior.

Example 11.9 (A circle of attracting equilibria)**.** Consider the gradient system

$$\dot{\mathbf{y}} = -\nabla V(\mathbf{y}), \qquad V(x,y) = (x^2+y^2-1)^2 - \alpha(x^2+y^2), \qquad \alpha > -2.$$

Writing

$$r_* = \sqrt{1 + \frac{\alpha}{2}},$$

we obtain

$$\dot{r} = 4r(r_*^2 - r^2), \qquad \dot{\theta} = 0.$$

The origin is an unstable equilibrium, and every nonzero solution remains on a fixed ray and satisfies

$$r(t) \longrightarrow r_*.$$

Hence

$$\mathcal{C} = \{(x, y) : x^2 + y^2 = r_*^2\}$$

is a circle of equilibria and an attractor with basin

$$\mathcal{B}(\mathcal{C}) = \mathbb{R}^2 \setminus \{(0, 0)\}.$$

No individual point of $\mathcal{C}$ is an attractor. Only initial conditions on the same ray approach that point, and a ray does not contain a neighborhood of it. Thus an invariant set may be attracting even though none of its individual equilibria is attracting.

Example 11.10 (Two attractors separated by an unstable periodic orbit)**.** Consider

$$\dot{x} = -y + x(x^2 + y^2 - 1)(4 - x^2 - y^2),$$
$$\dot{y} = x + y(x^2 + y^2 - 1)(4 - x^2 - y^2).$$

In polar coordinates,

$$\dot{r} = r(r^2 - 1)(4 - r^2), \qquad \dot{\theta} = 1.$$

The origin and the circle

$$\Gamma_2 = \{r = 2\}$$

are attractors. Their basins are

$$\mathcal{B}(\{0\}) = \{r < 1\}, \qquad \mathcal{B}(\Gamma_2) = \{r > 1\}.$$

The circle

$$\Gamma_1 = \{r = 1\}$$

is an unstable periodic orbit and forms their common basin boundary.

In contrast with the first example, the boundary is not the stable manifold of a saddle equilibrium. It is an invariant set on which the motion remains periodic.

These examples exhibit three different attracting structures: isolated equilibria, a continuum of equilibria, and a periodic orbit. They also show that a basin boundary may contain an equilibrium, a separatrix, or an entire periodic orbit.

The definitions identify attractors once the long-time behavior is known. They do not yet provide a general method for proving attraction. We next develop such a method by studying scalar quantities that decrease along solutions.

Exercise 11.2 (A basin boundary in a coupled system)**.** Consider

$$\dot{x} = x(1 - x^2), \qquad \dot{y} = -y + x.$$

1. Find and classify the equilibria.
2. Show that the sign of $x(t)$ is preserved.
3. Determine the basins of the two asymptotically stable equilibria.
4. Show that the stable manifold of the saddle is their common basin boundary.
5. Determine the ω-limit set of every initial condition.

Exercise 11.3 (A parameter-dependent basin boundary)**.** Let $0 < a < b$ and consider

$$\dot{r} = r(r^2 - a^2)(b^2 - r^2), \qquad \dot{\theta} = 1.$$

1. Determine all equilibria and periodic orbits.
2. Identify the attractors and find their basins.
3. Identify the common basin boundary.
4. Describe what happens to the attracting periodic orbit and its basin as $b \downarrow a$.

Exercise 11.4 (Basins in the competition model)**.** Revisit the symmetric strong-competition model from Section 4.2.2,

$$\dot{x} = x(1 - x - sy), \qquad \dot{y} = y(1 - y - sx), \qquad s > 1.$$

Give a rigorous determination of the basins of the two stable boundary equilibria and their common boundary.

Hint. Introduce

$$u = x + y, \qquad w = \frac{x - y}{x + y},$$

and study the sign and limiting behavior of w.

11.3 Monotone Quantities and Dissipative Structure

The preceding sections described ω-limit sets, attractors, and basins of attraction. We now turn to an important mechanism governing long-time behavior: the existence of a scalar quantity that is conserved or decreases along trajectories.

In a conservative system, such a quantity remains constant and confines the motion to a level set. In a dissipative system, it decreases and directs the motion toward lower level sets. This decrease may produce convergence to an equilibrium, but it may also lead to a periodic orbit,

a continuum of equilibria, or a more complicated invariant set. Thus dissipation should not be identified simply with convergence to rest.

The term *dissipation* is used in several related but distinct senses. It may refer to the decay of energy, contraction of phase-space volume, or the eventual confinement of all bounded sets to an absorbing region. These properties are not equivalent. The present section concentrates on monotone scalar quantities; volume contraction and absorbing sets will be considered later.

11.3.1 Conservation and Monotonicity

Consider the autonomous system

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}).$$

For a continuously differentiable function

$$E : \mathbb{R}^n \to \mathbb{R},$$

its derivative along the flow is

$$\dot{E}(\mathbf{y}) = \nabla E(\mathbf{y}) \cdot \mathbf{f}(\mathbf{y}).$$

If

$$\dot{E} = 0,$$

then E is a *first integral*, and each level set

$$\{\mathbf{y} : E(\mathbf{y}) = c\}$$

is invariant. If instead

$$\dot{E} \leq 0,$$

then E is nonincreasing along trajectories. The sublevel sets

$$\{\mathbf{y} : E(\mathbf{y}) \leq c\}$$

are positively invariant, although the individual level sets need not be invariant.

The distinction may be summarized as

$$\dot{E} = 0 \quad \Longrightarrow \quad \text{motion on a fixed level set},$$

whereas

$$\dot{E} \leq 0 \quad \Longrightarrow \quad \text{motion toward lower sublevel sets}.$$

A monotone quantity also restricts the possible ω-limit set.

Proposition 11.11 (Monotonicity principle)**.** *Suppose that the positive orbit of* $\mathbf{y}_0$ *has compact closure and that*

$$\dot{E} \leq 0$$

along the orbit. Then $E(\varphi_t(\mathbf{y}_0))$ *has a limit as* $t \to \infty$*, and* E *is constant on* $\omega(\mathbf{y}_0)$*. Consequently,*

$$\omega(\mathbf{y}_0) \subset \{\mathbf{y} : \dot{E}(\mathbf{y}) = 0\}.$$

Proof. Since the orbit has compact closure, E is bounded below on that closure. The function

$$t \longmapsto E(\varphi_t(\mathbf{y}_0))$$

is nonincreasing and therefore converges to some value E_∞.

If $\mathbf{z} \in \omega(\mathbf{y}_0)$, then for some sequence $t_j \to \infty$,

$$\varphi_{t_j}(\mathbf{y}_0) \longrightarrow \mathbf{z}.$$

By continuity,

$$E(\mathbf{z}) = \lim_{j\to\infty} E(\varphi_{t_j}(\mathbf{y}_0)) = E_\infty.$$

Thus E is constant on $\omega(\mathbf{y}_0)$. Since the ω-limit set is invariant, the trajectory through every point of $\omega(\mathbf{y}_0)$ remains in this set, and E remains equal to E_∞ along that trajectory. Hence

$$\dot{E} = 0$$

on $\omega(\mathbf{y}_0)$. □

The inclusion

$$\omega(\mathbf{y}_0) \subset \{\dot{E} = 0\}$$

is useful but generally not sufficient to determine the limiting behavior. The zero-dissipation set may contain points that do not form complete trajectories within it. The dynamically relevant object is the largest invariant subset of $\{\dot{E} = 0\}$. This observation will be developed systematically in LaSalle's invariance principle.

11.3.2 Mechanical Dissipation and Gyroscopic Forces

Consider a mechanical system with several degrees of freedom,

$$\mathbf{M}\ddot{\mathbf{q}} + \mathbf{D}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} + \nabla V(\mathbf{q}) = \mathbf{0}.$$

Here $\mathbf{q} \in \mathbb{R}^m$ is the configuration vector, $\mathbf{M}$ is a symmetric positive definite mass matrix, $\mathbf{D}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}}$ is a velocity-dependent force, and $V(\mathbf{q})$ is the potential energy. The mechanical energy is

$$E(\mathbf{q}, \dot{\mathbf{q}}) = \frac{1}{2}\dot{\mathbf{q}}^T\mathbf{M}\dot{\mathbf{q}} + V(\mathbf{q}).$$

Along a solution,

$$\begin{aligned}\frac{dE}{dt} &= \dot{\mathbf{q}}^T\mathbf{M}\ddot{\mathbf{q}} + \nabla V(\mathbf{q}) \cdot \dot{\mathbf{q}} \\ &= -\dot{\mathbf{q}}^T\mathbf{D}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}}.\end{aligned}$$

Decompose

$$\mathbf{D} = \mathbf{S} + \mathbf{A}, \qquad \mathbf{S} = \frac{1}{2}(\mathbf{D} + \mathbf{D}^T), \qquad \mathbf{A} = \frac{1}{2}(\mathbf{D} - \mathbf{D}^T).$$

Since $\mathbf{A}$ is antisymmetric,

$$\dot{\mathbf{q}}^T\mathbf{A}\dot{\mathbf{q}} = 0,$$

and therefore

$$\frac{dE}{dt} = -\dot{\mathbf{q}}^T \mathbf{S} \dot{\mathbf{q}}.$$

Thus only the symmetric part of $\mathbf{D}$ contributes to energy dissipation. If $\mathbf{S}$ is positive semidefinite, then

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

The energy identity separates three different mechanisms:

$$\begin{aligned} \nabla V(\mathbf{q}) &: \text{conservative force,} \\ \mathbf{S}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} &: \text{dissipative force,} \\ \mathbf{A}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} &: \text{gyroscopic force.} \end{aligned}$$

The gyroscopic part may change the trajectory without changing its energy.

If

$$V(\mathbf{q}) \longrightarrow \infty \qquad \text{as } |\mathbf{q}| \to \infty,$$

then the energy sublevel sets are bounded. Since they are also positively invariant, the energy identity confines every trajectory to a compact region of phase space. It does not, however, determine which invariant subset of

$$\left\{ (\mathbf{q}, \dot{\mathbf{q}}) : \dot{\mathbf{q}}^T \mathbf{S} \dot{\mathbf{q}} = 0 \right\}$$

is approached.

Oriented graphs and incidence matrices. A network of interacting components can be represented by a graph whose vertices are the components and whose edges are their pairwise connections. Choose an arbitrary orientation for each edge. If the graph has N vertices and m edges, its oriented incidence matrix

$$\mathbf{B} \in \mathbb{R}^{N \times m}$$

is defined by

$$B_{ie} = \begin{cases} -1, & \text{if vertex } i \text{ is the tail of edge } e, \\ 1, & \text{if vertex } i \text{ is the head of edge } e, \\ 0, & \text{otherwise.} \end{cases}$$

For a vector of vertex values $\mathbf{q} \in \mathbb{R}^N$, the vector

$$\mathbf{B}^T \mathbf{q}$$

contains the oriented differences across the edges. Reversing the orientation of an edge changes the sign of the corresponding column of $\mathbf{B}$, but quantities such as

$$\mathbf{B}\mathbf{R}\mathbf{B}^T$$

are unchanged when $\mathbf{R}$ is diagonal. Thus the chosen orientations do not affect the underlying mechanical system.

Example 11.12 (Damping through a network)**.** Consider N particles connected by springs and dampers along the edges of a graph with oriented incidence matrix $\mathbf{B}$. Then

$$\mathbf{B}^T\mathbf{q} \qquad \text{and} \qquad \mathbf{B}^T\dot{\mathbf{q}}$$

are the relative displacements and velocities across the edges. Suppose that the equations of motion are

$$\mathbf{M}\ddot{\mathbf{q}} + \mathbf{B}\mathbf{R}\mathbf{B}^T\dot{\mathbf{q}} + \nabla V(\mathbf{q}) = \mathbf{0},$$

where $\mathbf{M}$ is positive definite and $\mathbf{R}$ is a diagonal matrix with nonnegative entries. The energy

$$E(\mathbf{q}, \dot{\mathbf{q}}) = \frac{1}{2}\dot{\mathbf{q}}^T\mathbf{M}\dot{\mathbf{q}} + V(\mathbf{q})$$

satisfies

$$\dot{E} = -(\mathbf{B}^T\dot{\mathbf{q}})^T\mathbf{R}(\mathbf{B}^T\dot{\mathbf{q}}) \le 0.$$

Energy is dissipated only through relative motion along damped edges. Consequently,

$$\dot{E} = 0$$

does not necessarily imply $\dot{\mathbf{q}} = \mathbf{0}$. Undamped collective motions may remain in the zero-dissipation set. Determining which of these motions are invariant is a central issue in the later application of LaSalle's principle.

Remark 11.13 (Loss of time-reversal symmetry). Dissipation gives the motion a preferred direction in time. For example, under $t \mapsto -t$, the equation

$$\ddot{\mathbf{q}} + \gamma\dot{\mathbf{q}} + \nabla V(\mathbf{q}) = \mathbf{0}$$

becomes

$$\ddot{\mathbf{q}} - \gamma\dot{\mathbf{q}} + \nabla V(\mathbf{q}) = \mathbf{0}.$$

Damping is replaced by antidamping, so the reversed trajectory does not satisfy the original equation. This contrasts with the reversible Hamiltonian systems discussed in Section 8.1.3.

Remark 11.14 (Gyroscopic forces). In three dimensions, every antisymmetric matrix $\mathbf{A}$ can be represented by a vector $\boldsymbol{\Omega}$ such that

$$\mathbf{A}\mathbf{v} = \boldsymbol{\Omega} \times \mathbf{v}.$$

The resulting force is perpendicular to the velocity and therefore does no work. The magnetic Lorentz force from Example 8.5 and the Coriolis force are of this form. They may bend trajectories and alter their stability without directly dissipating energy.

11.3.3 Gradient Systems

A basic class of systems with a monotone quantity is formed by gradient systems,

$$\dot{\mathbf{y}} = -\nabla V(\mathbf{y}),$$

where $V : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable. Along every solution,

$$\frac{d}{dt}V(\mathbf{y}(t)) = \nabla V(\mathbf{y}(t)) \cdot \dot{\mathbf{y}}(t) = -|\nabla V(\mathbf{y}(t))|^2 \le 0.$$

Thus trajectories cross the level sets of V in the direction of decreasing potential.

The equilibria are precisely the critical points of V:

$$\nabla V(\mathbf{y}_*) = \mathbf{0}.$$

Strict local minima are natural candidates for stable equilibria, whereas local maxima and saddle points are generally unstable. A complete justification of these statements will follow from Lyapunov's theorem.

A gradient system cannot possess a nonconstant periodic orbit. Indeed, if $\mathbf{y}(t)$ has period $T > 0$, then

$$\begin{aligned} 0 &= V(\mathbf{y}(T)) - V(\mathbf{y}(0)) \\ &= -\int_0^T |\nabla V(\mathbf{y}(t))|^2 \, dt. \end{aligned}$$

Hence $\nabla V(\mathbf{y}(t)) = \mathbf{0}$ for all t, and the solution is constant.

More generally, a nonconstant trajectory cannot return to a previously visited state. If a trajectory connects two equilibria $\mathbf{y}_-$ and $\mathbf{y}_+$, then

$$V(\mathbf{y}_+) < V(\mathbf{y}_-).$$

Gradient dynamics therefore impose an ordering on connecting trajectories: they can move only from higher to lower values of the potential.

Example 11.15 (The Kuramoto model on a network)**.** Consider the Kuramoto model for identical phase oscillators coupled along the edges of an undirected weighted graph:

$$\dot{\theta}_i = K \sum_{j=1}^{N} a_{ij} \sin(\theta_j - \theta_i), \qquad i = 1, \dots, N,$$

where $K > 0$ and

$$a_{ij} = a_{ji} \ge 0.$$

Define

$$V(\boldsymbol{\theta}) = K \sum_{\{i,j\}} a_{ij}\big(1 - \cos(\theta_i - \theta_j)\big),$$

where each undirected edge is counted once. Then

$$\dot{\boldsymbol{\theta}} = -\nabla V(\boldsymbol{\theta}),$$

and hence

$$\dot{V} = -|\nabla V(\boldsymbol{\theta})|^2 = -\sum_{i=1}^{N} \dot{\theta}_i^2 \le 0.$$

If the graph is connected, the global minima of V are the synchronized states

$$\theta_1 = \cdots = \theta_N \pmod{2\pi}.$$

These states form a circle rather than an isolated equilibrium because the system is unchanged when the same constant is added to every phase.

The energy identity rules out nonconstant periodic motion in the reduced phase-difference variables. It does not by itself prove that every trajectory converges to synchronization, since V may possess other critical points. The limiting behavior depends on the invariant structure of the set

$$\{\boldsymbol{\theta} : \dot{V} = 0\} = \{\boldsymbol{\theta} : \nabla V = 0\}.$$

Gradient flows with a state-dependent metric. A useful generalization is

$$\dot{\mathbf{y}} = -\mathbf{G}(\mathbf{y})\nabla V(\mathbf{y}),$$

where $\mathbf{G}(\mathbf{y})$ is symmetric positive semidefinite. In this case,

$$\dot{V} = -\nabla V(\mathbf{y})^T \mathbf{G}(\mathbf{y})\nabla V(\mathbf{y}) \leq 0.$$

The matrix $\mathbf{G}$ determines the rate and preferred directions of descent. If $\mathbf{G}$ is singular, the zero-dissipation set may contain nonstationary motions or continua of states.

If $\mathbf{G}(\mathbf{y})$ is positive definite, it defines a state-dependent inner product

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{y}} = \mathbf{u}^T \mathbf{G}(\mathbf{y})^{-1} \mathbf{v}.$$

The vector $\mathbf{G}(\mathbf{y})\nabla V(\mathbf{y})$ is then the gradient of V with respect to this metric. Thus changing $\mathbf{G}$ changes the geometry of steepest descent, not merely its speed. When $\mathbf{G}$ is singular, the same interpretation may hold on a lower-dimensional constraint manifold.

Example 11.16 (The replicator equation as a gradient system)**.** The replicator equation models the evolution of competing types or strategies in a large population. Its purpose is to determine which types persist, which become extinct, and whether the population approaches a stable mixture of strategies.

Suppose that the population contains N types, and let $n_i(t)$ denote the number of individuals of type i. The population fractions are

$$x_i = \frac{n_i}{\sum_{j=1}^{N} n_j},$$

so the state belongs to the simplex

$$\Delta = \left\{ \mathbf{x} \in \mathbb{R}^N : x_i \geq 0, \quad \sum_{i=1}^{N} x_i = 1 \right\}.$$

Let a_{ij} be the payoff obtained by type i when interacting with type j. In a population with distribution $\mathbf{x}$, the expected payoff, or fitness, of type i is

$$f_i(\mathbf{x}) = (\mathbf{A}\mathbf{x})_i,$$

and the mean fitness of the population is

$$\overline{f}(\mathbf{x}) = \sum_{i=1}^{N} x_i f_i(\mathbf{x}) = \mathbf{x}^T \mathbf{A}\mathbf{x}.$$

The modeling assumption is that a type grows relative to the population when its fitness exceeds the mean fitness. This gives the replicator equation

$$\dot{x}_i = x_i\big(f_i(\mathbf{x}) - \overline{f}(\mathbf{x})\big) = x_i\left((\mathbf{A}\mathbf{x})_i - \mathbf{x}^T\mathbf{A}\mathbf{x}\right).$$

Indeed, if the population numbers satisfy

$$\dot{n}_i = f_i(\mathbf{x}) n_i,$$

then differentiating

$$x_i = \frac{n_i}{\sum_j n_j}$$

gives precisely the equation above.

The equation preserves the simplex:

$$\sum_{i=1}^{N} \dot{x}_i = 0.$$

Moreover, if $x_i = 0$, then $\dot{x}_i = 0$. Thus a type that is absent cannot appear spontaneously, and every face of the simplex is invariant.

We now ask whether this evolutionary dynamics has a monotone quantity. When the payoff matrix is symmetric,

$$\mathbf{A} = \mathbf{A}^T,$$

the answer is supplied by a gradient structure. Define

$$\mathbf{G}(\mathbf{x}) = \operatorname{diag}(\mathbf{x}) - \mathbf{x}\mathbf{x}^T, \qquad V(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}.$$

Then

$$\dot{\mathbf{x}} = -\mathbf{G}(\mathbf{x})\nabla V(\mathbf{x}).$$

Moreover,

$$\begin{aligned}\dot{V} &= -(\mathbf{A}\mathbf{x})^T\mathbf{G}(\mathbf{x})(\mathbf{A}\mathbf{x}) \\ &= -\sum_{i=1}^{N} x_i\left((\mathbf{A}\mathbf{x})_i - \mathbf{x}^T\mathbf{A}\mathbf{x}\right)^2 \le 0.\end{aligned}$$

Thus the mean payoff is nondecreasing along solutions.

On the interior of the simplex, $\mathbf{G}$ induces the metric

$$\langle \mathbf{u}, \mathbf{v}\rangle_{\mathbf{x}} = \sum_{i=1}^{N} \frac{u_i v_i}{x_i}, \qquad \sum_{i=1}^{N} u_i = \sum_{i=1}^{N} v_i = 0.$$

This is the *Shahshahani metric*. Thus, for a symmetric game, the replicator equation is a gradient flow on the simplex: evolution proceeds in the direction of increasing mean fitness, but steepest ascent is measured by a non-Euclidean, state-dependent geometry rather than by the Euclidean geometry of $\mathbb{R}^N$. The factors x_i slow the evolution of rare types and leave every boundary face invariant, features that are naturally encoded by the Shahshahani metric.

The replicator equation and its gradient geometry are classical topics in evolutionary game theory and population genetics; see Taylor and Jonker (1978); Shahshahani (1979).

11.3.4 Three Notions of Dissipation

The term *dissipation* is used for several related but distinct properties. A monotone energy or entropy describes the evolution of individual trajectories, divergence describes the deformation of phase-space volumes, and an absorbing set describes the eventual behavior of bounded families of trajectories.

Energy decay or entropy production. A system may possess a scalar function that is monotone along solutions. Depending on the physical interpretation, this may take the form

$$\dot{E} \le 0$$

for a decreasing energy or free energy, or

$$\dot{S} \ge 0$$

for an increasing entropy. Mathematically the two cases are equivalent after replacing S by $-S$.

Such a monotone quantity orders the motion along each trajectory. If a decreasing function E is coercive, its sublevel sets are bounded and provide a priori bounds. Monotonicity does not, however, imply that every small phase-space volume contracts.

For example,

$$\dot{x} = x - x^3$$

has the decreasing potential

$$V(x) = \frac{1}{4}(x^2 - 1)^2, \qquad \dot{V} = -|V'(x)|^2 \le 0.$$

Nevertheless,

$$f'(x) = 1 - 3x^2 > 0$$

near $x = 0$, so nearby intervals initially expand there.

Phase-volume contraction. For

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}),$$

the Jacobian of the flow satisfies

$$\det D\varphi_t(\mathbf{y}_0) = \exp\left(\int_0^t \nabla \cdot \mathbf{f}\big(\varphi_s(\mathbf{y}_0)\big)\, ds\right).$$

Thus the divergence measures the instantaneous rate of change of infinitesimal phase-space volume.

For a Hamiltonian system,

$$\nabla \cdot \mathbf{f} = 0,$$

and Liouville's theorem states that phase-space volume is preserved:

$$\det D\varphi_t = 1.$$

By contrast, if

$$\nabla \cdot \mathbf{f} \leq -c < 0,$$

then infinitesimal volumes contract at least at the rate e^{-ct}.

Volume contraction does not imply that trajectories remain bounded. For example,

$$\dot{x} = x, \qquad \dot{y} = -2y$$

has

$$\nabla \cdot \mathbf{f} = -1,$$

but every trajectory with $x(0) \neq 0$ is unbounded. Negative divergence therefore does not by itself imply the existence of an attractor.

Absorbing sets. A bounded set B is called an *absorbing set* if, for every bounded set D, there exists $T_D \geq 0$ such that

$$\varphi_t(D) \subset B, \qquad t \geq T_D.$$

This property describes eventual uniform confinement of bounded families of initial conditions. A system possessing a bounded absorbing set is often called *dissipative in the dynamical sense*.

A monotone quantity need not produce an absorbing set. For example,

$$\dot{x} = -x, \qquad \dot{y} = 0$$

has

$$E(x, y) = \frac{1}{2}x^2, \qquad \dot{E} = -x^2,$$

but it has no bounded set that absorbs every bounded set of initial conditions, because y is conserved.

Conversely, an absorbing set may exist even if the divergence is positive in some parts of phase space. The equation

$$\dot{x} = x - x^3$$

eventually confines every bounded set to a fixed bounded interval, although

$$f'(x) > 0$$

near the origin.

These three notions therefore describe different levels of the dynamics:

Property	What it controls
$\dot{E} \leq 0$ or $\dot{S} \geq 0$	evolution along individual trajectories
$\nabla \cdot \mathbf{f} < 0$	evolution of infinitesimal phase-space volumes
bounded absorbing set	eventual confinement of bounded families

None of these properties is equivalent to the others without additional hypotheses. In applications, however, they often occur together: an energy or entropy estimate may produce an absorbing set, while volume contraction helps explain how a bounded region can evolve toward a lower-dimensional attracting structure. Absorbing sets and their relation to global attractors will be studied after LaSalle's invariance principle.

Exercise 11.5 (Gyroscopic forces in three dimensions)**.** Consider

$$\ddot{\mathbf{q}} + \mathbf{D}(\dot{\mathbf{q}})\dot{\mathbf{q}} + \nabla V(\mathbf{q}) = \mathbf{0}, \qquad \mathbf{q} \in \mathbb{R}^3.$$

Let

$$\mathbf{A}(\mathbf{v}) = \frac{1}{2}\big(\mathbf{D}(\mathbf{v}) - \mathbf{D}(\mathbf{v})^T\big).$$

Show that there is a vector $\boldsymbol{\Omega}(\mathbf{v})$ such that

$$\mathbf{A}(\mathbf{v})\mathbf{w} = \boldsymbol{\Omega}(\mathbf{v}) \times \mathbf{w}$$

for every $\mathbf{w} \in \mathbb{R}^3$, and deduce that

$$\mathbf{v}^T \mathbf{A}(\mathbf{v})\mathbf{v} = 0.$$

Relate this representation to the magnetic Lorentz force and the Coriolis force.

Exercise 11.6 (Entropy production for diffusion on a graph)**.** Revisit the diffusion equation on a connected weighted graph from Section 5.1.4,

$$\dot{x}_i = \sum_{j=1}^{N} a_{ij}(x_j - x_i), \qquad a_{ij} = a_{ji} \geq 0,$$

where $x_i(t) > 0$ represents the amount or concentration at vertex i.

Show that the total mass

$$M = \sum_{i=1}^{N} x_i$$

is conserved. For the entropy

$$S(\mathbf{x}) = -\sum_{i=1}^{N} x_i \log x_i,$$

prove that

$$\dot{S} = \frac{1}{2}\sum_{i,j=1}^{N} a_{ij}(x_i - x_j)(\log x_i - \log x_j) \geq 0.$$

Determine when $\dot{S} = 0$, and show that the entropy is maximized, subject to the conserved total mass, at the uniform state

$$x_1 = \cdots = x_N = \frac{M}{N}.$$

Exercise 11.7 (Volume contraction in Chua's circuit)**.** Chua's circuit is a nonlinear electrical circuit containing two capacitors, an inductor, a linear resistor, and a nonlinear resistor known as Chua's diode. After nondimensionalization, its equations can be written as

$$\dot{x} = \alpha\big(y - x - h(x)\big), \qquad \dot{y} = x - y + z, \qquad \dot{z} = -\beta y.$$

Here x and y represent scaled capacitor voltages, z represents a scaled inductor current, and $\alpha, \beta > 0$ are dimensionless circuit parameters. The current–voltage relation of the nonlinear resistor is modeled by

$$h(x) = m_1 x + \frac{1}{2}(m_0 - m_1)\big(|x+1| - |x-1|\big),$$

where m_0 and m_1 are its inner and outer slopes.

Away from the switching planes $x = \pm 1$, show that

$$\nabla \cdot \mathbf{f} = -\alpha\big(1 + h'(x)\big) - 1.$$

Find conditions on m_0 and m_1 under which the divergence is negative in every linearity region.

The vector field is continuous and piecewise differentiable, and the switching planes have zero volume. Explain why Liouville's formula may be applied along trajectory segments in the individual linearity regions and combined across their switching times. Why does the resulting phase-volume contraction not, by itself, prove the existence of a chaotic attractor?

11.4 Lyapunov Functions and Stability

The preceding section studied scalar quantities that are monotone along trajectories. Such a quantity may confine the motion or restrict its ω-limit set, but it need not measure distance from a particular equilibrium. Lyapunov's direct method adds precisely this local structure: a Lyapunov function is positive relative to an equilibrium and nonincreasing along the flow.

This method can establish stability and attraction without solving the differential equation. It can also provide positively invariant regions, estimates of basins of attraction, and quantitative decay rates.

Consider

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}),$$

and let $\mathbf{y}_*$ be an equilibrium:

$$\mathbf{f}(\mathbf{y}_*) = \mathbf{0}.$$

11.4.1 Stability of an Equilibrium

Definition 11.17. The equilibrium $\mathbf{y}_*$ is *stable* if, for every $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|\mathbf{y}_0 - \mathbf{y}_*| < \delta$$

implies

$$|\varphi_t(\mathbf{y}_0) - \mathbf{y}_*| < \varepsilon \qquad \text{for all } t \geq 0.$$

Stability means that trajectories starting near the equilibrium remain near it for all future time. It does not require them to converge to the equilibrium.

Definition 11.18. The equilibrium $\mathbf{y}_*$ is *asymptotically stable* if it is stable and there exists $\delta_0 > 0$ such that

$$|\mathbf{y}_0 - \mathbf{y}_*| < \delta_0$$

implies

$$\varphi_t(\mathbf{y}_0) \longrightarrow \mathbf{y}_* \qquad \text{as } t \to \infty.$$

Thus asymptotic stability combines two different properties:

$$\text{stability} \quad + \quad \text{local attraction.}$$

An asymptotically stable equilibrium is an attractor consisting of one point.

The equilibrium $\mathbf{y}_*$ is called *globally asymptotically stable* if it is stable and every initial condition in the phase space satisfies

$$\varphi_t(\mathbf{y}_0) \longrightarrow \mathbf{y}_* \qquad \text{as } t \to \infty.$$

11.4.2 Lyapunov Functions

A Lyapunov function acts as a generalized local energy centered at the equilibrium.

Definition 11.19. Let $\mathbf{y}_*$ be an equilibrium. A continuously differentiable function Φ defined in a neighborhood U of $\mathbf{y}_*$ is called a *Lyapunov function* for $\mathbf{y}_*$ if

$$\Phi(\mathbf{y}_*) = 0,$$

$$\Phi(\mathbf{y}) > 0 \qquad \text{for } \mathbf{y} \in U \setminus \{\mathbf{y}_*\},$$

and

$$\dot{\Phi}(\mathbf{y}) := \nabla\Phi(\mathbf{y}) \cdot \mathbf{f}(\mathbf{y}) \le 0 \qquad \text{for } \mathbf{y} \in U.$$

The first two conditions say that Φ is *positive definite* with respect to $\mathbf{y}_*$. If

$$\dot{\Phi}(\mathbf{y}) < 0 \qquad \text{for } \mathbf{y} \neq \mathbf{y}_*,$$

then $\dot{\Phi}$ is *negative definite*.

The sublevel sets

$$\Omega_c = \{\mathbf{y} \in U : \Phi(\mathbf{y}) \le c\}$$

are positively invariant whenever they are contained in U. Indeed, Φ cannot increase along a trajectory, so a solution cannot cross outward through the boundary

$$\Phi(\mathbf{y}) = c.$$

This gives the geometric idea behind Lyapunov's method. Positive definiteness makes sufficiently small sublevel sets surround the equilibrium, while

$$\dot{\Phi} \le 0$$

prevents trajectories from leaving them. If the decay is strict away from the equilibrium, it also drives trajectories toward progressively smaller sublevel sets.

A Lyapunov function is therefore more specialized than the monotone quantities of Section 11.3. A mechanical energy, entropy, or potential may be monotone without being positive definite relative to a specified equilibrium. Such a quantity becomes a Lyapunov function only when its local geometry singles out that equilibrium.

11.4.3 Lyapunov's Stability Theorem

The geometry of invariant sublevel sets leads to the basic stability theorem.

Theorem 11.20 (Lyapunov's stability theorem)**.** *Let* $\mathbf{y}_*$ *be an equilibrium of*

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}),$$

and suppose that Φ *is continuously differentiable in a neighborhood of* $\mathbf{y}_*$*. Assume that*

$$\Phi(\mathbf{y}_*) = 0, \qquad \Phi(\mathbf{y}) > 0 \quad \textit{for } \mathbf{y} \neq \mathbf{y}_*,$$

and

$$\dot{\Phi}(\mathbf{y}) \leq 0.$$

Then $\mathbf{y}_*$ *is stable.*

If, in addition,

$$\dot{\Phi}(\mathbf{y}) < 0 \qquad \textit{for } \mathbf{y} \neq \mathbf{y}_*,$$

then $\mathbf{y}_*$ *is asymptotically stable.*

Proof. Choose $\varepsilon > 0$ sufficiently small that the closed ball

$$\overline{B_\varepsilon(\mathbf{y}_*)}$$

lies in the domain of Φ. Since Φ is continuous and positive away from $\mathbf{y}_*$,

$$m_\varepsilon = \min_{|\mathbf{y}-\mathbf{y}_*|=\varepsilon} \Phi(\mathbf{y}) > 0.$$

By continuity at $\mathbf{y}_*$, there exists $\delta > 0$ such that

$$|\mathbf{y}_0 - \mathbf{y}_*| < \delta \quad \Longrightarrow \quad \Phi(\mathbf{y}_0) < m_\varepsilon.$$

Since Φ is nonincreasing along solutions,

$$\Phi(\varphi_t(\mathbf{y}_0)) \leq \Phi(\mathbf{y}_0) < m_\varepsilon.$$

The trajectory cannot reach the sphere

$$|\mathbf{y} - \mathbf{y}_*| = \varepsilon,$$

because $\Phi \geq m_\varepsilon$ there. Hence

$$|\varphi_t(\mathbf{y}_0) - \mathbf{y}_*| < \varepsilon \qquad \text{for all } t \geq 0,$$

which proves stability.

Now suppose that $\dot{\Phi} < 0$ away from $\mathbf{y}_*$. Choose

$$0 < c < m_\varepsilon$$

so that the component of

$$\{\mathbf{y} : \Phi(\mathbf{y}) \le c\}$$

containing $\mathbf{y}_*$ lies inside $B_\varepsilon(\mathbf{y}_*)$. This component is compact and positively invariant. Therefore every trajectory starting sufficiently near $\mathbf{y}_*$ has compact positive-orbit closure.

By the monotonicity principle,

$$\omega(\mathbf{y}_0) \subset \{\mathbf{y} : \dot{\Phi}(\mathbf{y}) = 0\}.$$

Strict negativity gives

$$\{\mathbf{y} : \dot{\Phi}(\mathbf{y}) = 0\} = \{\mathbf{y}_*\}$$

within this neighborhood. Hence

$$\omega(\mathbf{y}_0) = \{\mathbf{y}_*\},$$

and therefore

$$\varphi_t(\mathbf{y}_0) \longrightarrow \mathbf{y}_* \qquad \text{as } t \to \infty.$$

Thus $\mathbf{y}_*$ is asymptotically stable. □

The two parts of the theorem use the Lyapunov function differently. Positivity and nonincrease trap trajectories in small sublevel sets and give stability. Strict decrease identifies the equilibrium as the only possible ω-limit set and gives attraction.

Corollary 11.21 (Global asymptotic stability)**.** *Suppose that* $\Phi : \mathbb{R}^n \to \mathbb{R}$ *satisfies*

$$\Phi(\mathbf{y}_*) = 0, \qquad \Phi(\mathbf{y}) > 0 \quad \textit{for } \mathbf{y} \neq \mathbf{y}_*,$$

$$\dot{\Phi}(\mathbf{y}) < 0 \quad \textit{for } \mathbf{y} \neq \mathbf{y}_*,$$

and

$$\Phi(\mathbf{y}) \longrightarrow \infty \qquad \textit{as } |\mathbf{y}| \to \infty.$$

Then $\mathbf{y}_*$ *is globally asymptotically stable.*

Proof. For any initial condition $\mathbf{y}_0$, the trajectory remains in the sublevel set

$$\{\mathbf{y} : \Phi(\mathbf{y}) \le \Phi(\mathbf{y}_0)\}.$$

The coercivity of Φ makes this set compact. The trajectory therefore exists for all $t \ge 0$ and has a nonempty ω-limit set. By the monotonicity principle and strict decay,

$$\omega(\mathbf{y}_0) = \{\mathbf{y}_*\}.$$

Hence every trajectory converges to $\mathbf{y}_*$. Stability follows from the local part of Lyapunov's theorem. □

The condition

$$\Phi(\mathbf{y}) \longrightarrow \infty \qquad \text{as } |\mathbf{y}| \to \infty$$

is called *coercivity* or *radial unboundedness*. It ensures that every sublevel set of Φ is bounded.

11.4.4 Applications of Lyapunov's Theorem

A Lyapunov function is often suggested by the structure of the system: physical energy for a mechanical system, a weighted quadratic form for a linear system, or a quantity specifically designed to control nonlinear coupling. The following examples illustrate these possibilities.

Example 11.22 (Stability without attraction in a conservative system)**.** Consider

$$\mathbf{M}\ddot{\mathbf{q}} + \nabla V(\mathbf{q}) = \mathbf{0},$$

where $\mathbf{M}$ is symmetric positive definite, and suppose that $\mathbf{q}_*$ is a strict local minimum of V. The state

$$(\mathbf{q}, \dot{\mathbf{q}}) = (\mathbf{q}_*, \mathbf{0})$$

is an equilibrium.

Define the shifted mechanical energy

$$\Phi(\mathbf{q}, \dot{\mathbf{q}}) = \frac{1}{2}\dot{\mathbf{q}}^T\mathbf{M}\dot{\mathbf{q}} + V(\mathbf{q}) - V(\mathbf{q}_*).$$

It is positive definite near $(\mathbf{q}_*, \mathbf{0})$, and

$$\begin{aligned}\dot{\Phi} &= \dot{\mathbf{q}}^T\big(\mathbf{M}\ddot{\mathbf{q}} + \nabla V(\mathbf{q})\big)\\ &= 0.\end{aligned}$$

Lyapunov's theorem therefore proves that the equilibrium is stable.

The conclusion is stability rather than asymptotic stability. The energy is constant, so nearby trajectories remain on fixed energy levels instead of approaching the equilibrium. This illustrates why

$$\dot{\Phi} \le 0$$

and

$$\dot{\Phi} < 0$$

have fundamentally different dynamical consequences.

Example 11.23 (The Lyapunov equation for a stable linear system)**.** Consider the constant-coefficient linear system

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}, \qquad \mathbf{A} \in \mathbb{R}^{n\times n},$$

and suppose that $\mathbf{A}$ is Hurwitz; that is, every eigenvalue of $\mathbf{A}$ has negative real part.

We seek a quadratic Lyapunov function of the form

$$\Phi(\mathbf{y}) = \mathbf{y}^T\mathbf{P}\mathbf{y},$$

where $\mathbf{P}$ is symmetric positive definite. Along a solution,

$$\begin{aligned}\dot{\Phi} &= \dot{\mathbf{y}}^T\mathbf{P}\mathbf{y} + \mathbf{y}^T\mathbf{P}\dot{\mathbf{y}} \\ &= \mathbf{y}^T\big(\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A}\big)\mathbf{y}.\end{aligned}$$

Thus we want to choose $\mathbf{P}$ so that

$$\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A}$$

is negative definite. Prescribing a symmetric positive definite matrix $\mathbf{Q}$ and requiring

$$\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{Q}$$

leads to the *Lyapunov equation*.

Because $\mathbf{A}$ is Hurwitz, this equation has the symmetric positive definite solution

$$\mathbf{P} = \int_0^\infty e^{t\mathbf{A}^T}\mathbf{Q}e^{t\mathbf{A}}\,dt.$$

The integral converges because $e^{t\mathbf{A}}$ decays exponentially. Moreover, for $\mathbf{y} \neq \mathbf{0}$,

$$\mathbf{y}^T\mathbf{P}\mathbf{y} = \int_0^\infty (e^{t\mathbf{A}}\mathbf{y})^T\mathbf{Q}(e^{t\mathbf{A}}\mathbf{y})\,dt > 0,$$

so $\mathbf{P}$ is positive definite.

Finally,

$$\begin{aligned}\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} &= \int_0^\infty \frac{d}{dt}\left(e^{t\mathbf{A}^T}\mathbf{Q}e^{t\mathbf{A}}\right)\,dt \\ &= -\mathbf{Q}.\end{aligned}$$

Consequently,

$$\dot{\Phi} = -\mathbf{y}^T\mathbf{Q}\mathbf{y} < 0 \qquad \text{for } \mathbf{y} \neq \mathbf{0}.$$

Thus every Hurwitz constant matrix admits a positive definite quadratic Lyapunov function.

This construction is especially useful when the Euclidean norm does not decrease. For example, let

$$\mathbf{A} = \begin{bmatrix} -1 & K \\ 0 & -1 \end{bmatrix}, \qquad K > 2.$$

Although both eigenvalues are -1,

$$\frac{d}{dt}|\mathbf{y}|^2 = 2(-y_1^2 + Ky_1y_2 - y_2^2)$$

is positive in some directions. Solutions may therefore exhibit transient growth in the Euclidean norm.

Taking $\mathbf{Q} = \mathbf{I}$, the Lyapunov equation gives

$$\mathbf{P} = \begin{bmatrix} \frac{1}{2} & \frac{K}{4} \\ \frac{K}{4} & \frac{K^2+2}{4} \end{bmatrix}.$$

This matrix is positive definite, and

$$\frac{d}{dt}(\mathbf{y}^T\mathbf{P}\mathbf{y}) = -|\mathbf{y}|^2.$$

The ellipsoids

$$\mathbf{y}^T\mathbf{P}\mathbf{y} = c$$

are adapted to the dynamics even though Euclidean circles are not. The Lyapunov function identifies a geometry in which decay is monotone despite possible transient growth in the Euclidean norm.

Example 11.24 (A basin estimate for a nonlinear system)**.** Consider

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y} + \mathbf{g}(\mathbf{y}),$$

where $\mathbf{A}$ is a constant Hurwitz matrix and

$$\mathbf{g}(\mathbf{0}) = \mathbf{0}, \qquad D\mathbf{g}(\mathbf{0}) = \mathbf{0}.$$

Thus $\mathbf{g}(\mathbf{y})$ contains terms of at least second order. Suppose that

$$|\mathbf{g}(\mathbf{y})| \le C|\mathbf{y}|^2 \qquad \text{for } |\mathbf{y}| \le R.$$

Let $\mathbf{P}$ be the positive definite solution of

$$\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{I},$$

and define

$$\Phi(\mathbf{y}) = \mathbf{y}^T\mathbf{P}\mathbf{y}.$$

Along a solution,

$$\begin{aligned} \dot{\Phi} &= \mathbf{y}^T(\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A})\mathbf{y} + 2\mathbf{y}^T\mathbf{P}\mathbf{g}(\mathbf{y}) \\ &= -|\mathbf{y}|^2 + 2\mathbf{y}^T\mathbf{P}\mathbf{g}(\mathbf{y}). \end{aligned}$$

Hence

$$\begin{aligned} \dot{\Phi} &\le -|\mathbf{y}|^2 + 2\|\mathbf{P}\|\,|\mathbf{y}|\,|\mathbf{g}(\mathbf{y})| \\ &\le -|\mathbf{y}|^2\left(1 - 2C\|\mathbf{P}\|\,|\mathbf{y}|\right). \end{aligned}$$

Therefore $\dot{\Phi} < 0$ whenever

$$0 < |\mathbf{y}| < \rho, \qquad \rho < \min\left\{R, \frac{1}{2C\|\mathbf{P}\|}\right\}.$$

Since

$$\Phi(\mathbf{y}) \geq \lambda_{\min}(\mathbf{P})|\mathbf{y}|^2,$$

the ellipsoid

$$\Omega_\rho = \left\{\mathbf{y} : \Phi(\mathbf{y}) \leq \lambda_{\min}(\mathbf{P})\rho^2\right\}$$

is contained in the ball $|\mathbf{y}| \leq \rho$. It is positively invariant, and every trajectory starting in its interior converges to the origin. Thus

$$\Omega_\rho \subset \mathcal{B}(\mathbf{0}).$$

This argument does more than reproduce the linearization theorem. It constructs an explicit region guaranteed to lie inside the basin of attraction. The estimate need not give the entire basin, and different choices of the matrix $\mathbf{Q}$ in

$$\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{Q}$$

may produce different basin estimates.

Example 11.25 (Flocking in the Cucker–Smale model)**.** Consider N agents moving in $\mathbb{R}^d$, with positions $\mathbf{x}_i$ and velocities $\mathbf{v}_i$. The Cucker–Smale model is

$$\dot{\mathbf{x}}_i = \mathbf{v}_i, \qquad \dot{\mathbf{v}}_i = \frac{1}{N}\sum_{j=1}^{N} \psi(|\mathbf{x}_i - \mathbf{x}_j|)(\mathbf{v}_j - \mathbf{v}_i),$$

where $\psi : [0, \infty) \to (0, \infty)$ is a nonincreasing *interaction function.* The value

$$\psi(|\mathbf{x}_i - \mathbf{x}_j|)$$

measures the strength with which agents i and j adjust their velocities toward one another. Thus nearby agents interact at least as strongly as more distant agents. The mean velocity

$$\overline{\mathbf{v}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{v}_i$$

is conserved. Flocking means that the relative positions remain bounded and the velocities align:

$$\sup_{t \geq 0} |\mathbf{x}_i(t) - \mathbf{x}_j(t)| < \infty,$$

and

$$|\mathbf{v}_i(t) - \mathbf{v}_j(t)| \longrightarrow 0.$$

Define the position and velocity fluctuations

$$X(t) = \left(\frac{1}{N} \sum_{i=1}^{N} |\mathbf{x}_i - \overline{\mathbf{x}}|^2 \right)^{1/2}, \qquad \overline{\mathbf{x}} = \frac{1}{N} x_i,$$

and

$$V(t) = \left(\frac{1}{N} \sum_{i=1}^{N} |\mathbf{v}_i - \overline{\mathbf{v}}|^2 \right)^{1/2}.$$

Let $\mathbf{y}_i = \mathbf{x}_i - \overline{\mathbf{x}}$ and $\mathbf{u}_i = \mathbf{v}_i - \overline{\mathbf{v}}$. When $X > 0$,

$$\dot{X} = \frac{1}{NX} \sum_i \mathbf{y}_i \cdot \mathbf{u}_i \le \frac{1}{NX} \left(\sum_i |\mathbf{y}_i|^2 \right)^{1/2} \left(\sum_i |\mathbf{u}_i|^2 \right)^{1/2} = V.$$

At $X = 0$, the same conclusion holds in the Dini sense:

$$D^+ X \le V.$$

Here D^+ denotes the upper right Dini derivative along a trajectory:

$$D^+ g(t) = \limsup_{h \downarrow 0} \frac{g(t+h) - g(t)}{h}.$$

For V, conservation of mean velocity gives $\dot{\mathbf{v}} = 0$. Therefore

$$V\dot{V} = -\frac{1}{N^2} \sum_{i,j} \psi(|\mathbf{x}_i - \mathbf{x}_j|) |\mathbf{u}_i - \mathbf{u}_j|^2.$$

Using

$$|\mathbf{x}_i - \mathbf{x}_j| \le 2\sqrt{N} X.$$

the monotonicity of ψ, and

$$\sum_{i,j} |\mathbf{u}_i - \mathbf{u}_j|^2 = 2N \sum_i |\mathbf{u}_i|^2 = 2N^2 V^2,$$

we obtain the differential inequalities

$$D^+ V \le -\psi(2\sqrt{N}\, X) V.$$

The velocity fluctuation alone decreases, but this does not immediately imply alignment because the interaction strength may decay as the agents separate. To control both effects, define

$$\mathcal{L}(t) = V(t) + \int_0^{X(t)} \psi(2\sqrt{N}\, s)\, ds.$$

Here s is a scalar integration variable representing spatial spread, and the upper limit is the current value $X(t)$. Along a solution,

$$\begin{aligned} D^+\mathcal{L}(t) &\le D^+V(t) + \psi(2\sqrt{N}\,X(t))D^+X(t) \\ &\le -\psi(2\sqrt{N}\,X(t))V(t) + \psi(2\sqrt{N}\,X(t))V(t) \\ &= 0. \end{aligned}$$

Thus $\mathcal{L}$ is a Lyapunov function coupling spatial spread and velocity disagreement.

Suppose that

$$V(0) < \int_{X(0)}^{\infty} \psi(2\sqrt{N}\,s)\,ds.$$

Then there exists $X_* > X(0)$ such that

$$V(0) = \int_{X(0)}^{X_*} \psi(2\sqrt{N}\,s)\,ds.$$

Since $\mathcal{L}(t) \le \mathcal{L}(0)$, the position fluctuation cannot exceed X_*. Consequently,

$$\psi(2\sqrt{N}\,X(t)) \ge \psi(2\sqrt{N}\,X_*) > 0,$$

and hence

$$V(t) \le V(0)e^{-\psi(2\sqrt{N}\,X_*)t}.$$

The velocities therefore align exponentially, while the relative positions remain bounded.

If

$$\int^{\infty} \psi(s)\,ds = \infty,$$

the flocking condition holds for every initial configuration. This is called *unconditional flocking*. If the integral is finite, flocking is guaranteed only when the initial velocity disagreement is sufficiently small relative to the initial spatial spread.

This example illustrates the constructive role of Lyapunov's method. The decreasing velocity fluctuation is not sufficient by itself; the decisive quantity combines velocity alignment with the weakening of communication as the group disperses.

Nonsmooth Lyapunov functions. The preceding flocking argument uses quantities that may fail to be differentiable at isolated states. More generally, a locally Lipschitz function Φ may serve as a Lyapunov function if its upper Dini derivative along the flow satisfies

$$D^+\Phi(\mathbf{y}) = \limsup_{h\downarrow 0} \frac{\Phi(\varphi_h(\mathbf{y})) - \Phi(\mathbf{y})}{h} \le 0.$$

The invariant-sublevel-set argument remains valid because this inequality still prevents Φ from increasing along trajectories. Norms, maxima of several functions, and diameter-type quantities frequently lead to nonsmooth Lyapunov functions.

Remark 11.26 (Inertial spin models of flocking). The Cucker–Smale model directly relaxes velocity differences and is essentially first order in the velocity variables. A more recent class of models introduces an internal angular momentum, or *spin*, to describe the inertia of turning.

For planar motion at fixed speed, write

$$\mathbf{v}_i = v_0(\cos\theta_i, \sin\theta_i)$$

and introduce a spin variable s_i. On a fixed interaction network, a simplified inertial spin model has the form

$$\dot\theta_i = \frac{s_i}{\chi}, \qquad \dot s_i = -\frac{\partial U}{\partial \theta_i} - \frac{\eta}{\chi} s_i,$$

where $\chi > 0$ is the spin inertia, $\eta > 0$ is rotational damping, and

$$U(\boldsymbol{\theta}) = \frac{J}{2N} \sum_{i,j=1}^{N} a_{ij}\big(1 - \cos(\theta_i - \theta_j)\big)$$

is the alignment potential.

The energy

$$H(\boldsymbol{\theta}, \mathbf{s}) = \sum_{i=1}^{N} \frac{s_i^2}{2\chi} + U(\boldsymbol{\theta})$$

satisfies

$$\dot H = -\frac{\eta}{\chi^2} \sum_{i=1}^{N} s_i^2 \le 0.$$

Thus the model combines a Hamiltonian turning mechanism with dissipation. Unlike purely relaxational alignment models, it permits turning information to propagate through the flock as a wave before being damped.

For moving agents, the interaction network changes with their positions, and the analysis becomes substantially more difficult. Clustering, collective turning, propagation of directional information, and the long-time behavior of inertial spin systems remain active research topics; see Cavagna et al. (2015).

Exercise 11.8 (Characterization of stable linear systems)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be constant. Prove that $\mathbf{A}$ is Hurwitz if and only if, for every symmetric positive definite matrix $\mathbf{Q}$, the Lyapunov equation

$$\mathbf{A}^T\mathbf{P} + \mathbf{P}\mathbf{A} = -\mathbf{Q}$$

has a unique symmetric positive definite solution $\mathbf{P}$.

Hint. For uniqueness, study

$$e^{t\mathbf{A}^T}\mathbf{P}e^{t\mathbf{A}}.$$

For the converse, apply the Lyapunov equation to a possibly complex eigenvector of $\mathbf{A}$.

Exercise 11.9 (A nonlinear basin estimate)**.** Consider

$$\dot x = -x + y, \qquad \dot y = -x - y + x^2.$$

Using

$$\Phi(x, y) = x^2 + y^2,$$

find a positive number R such that

$$\dot{\Phi} < 0 \qquad \text{whenever } 0 < x^2 + y^2 < R^2.$$

Deduce that the disk

$$x^2 + y^2 < R^2$$

lies in the basin of attraction of the origin.

Hint. Maximize $x^2 y$ subject to $x^2 + y^2 = r^2$.

Exercise 11.10 (Communication rate and unconditional flocking)**.** For the Cucker–Smale model, let

$$\psi(r) = \frac{1}{(1 + r^2)^{\beta}}, \qquad \beta > 0.$$

Determine the values of β for which

$$\int^{\infty} \psi(r)\, dr = \infty$$

and hence flocking occurs for every initial configuration. When the integral is finite, express a sufficient flocking condition in terms of $X(0)$ and $V(0)$.

11.5 LaSalle's Invariance Principle

Lyapunov's theorem gives asymptotic stability when

$$\dot{\Phi}(\mathbf{y}) < 0 \qquad \text{for } \mathbf{y} \neq \mathbf{y}_*.$$

In many important systems, however, one can prove only

$$\dot{\Phi} \leq 0.$$

The derivative may vanish on a large set even though most trajectories in that set immediately leave it. Mechanical systems with partial damping, network models, and systems with symmetry commonly have this structure.

LaSalle's invariance principle replaces strict decay by a dynamical question: which complete trajectories can remain inside the set where the dissipation vanishes?

11.5.1 The Zero-Dissipation Set

Let Ω be a compact positively invariant set, and suppose that $\Phi \in C^1(\Omega)$ satisfies

$$\dot{\Phi}(\mathbf{y}) = \nabla \Phi(\mathbf{y}) \cdot \mathbf{f}(\mathbf{y}) \leq 0 \qquad \text{on } \Omega.$$

Define the zero-dissipation set

$$E = \{\mathbf{y} \in \Omega : \dot{\Phi}(\mathbf{y}) = 0\}.$$

The set E need not be invariant. A trajectory may pass through a point of E and immediately enter a region where $\dot{\Phi} < 0$. What matters is the largest invariant subset of E.

Definition 11.27. The *largest invariant subset* of E is the union of all complete trajectories that remain in E for all time. It will be denoted by

$$\mathcal{M}(E).$$

To determine $\mathcal{M}(E)$, one imposes the condition defining E and then checks whether the differential equation preserves it. This may require differentiating the constraint or substituting it into the equations of motion.

For example, suppose that

$$\dot{\Phi}(\mathbf{y}) = -g(\mathbf{y})^2.$$

Then

$$E = \{\mathbf{y} : g(\mathbf{y}) = 0\}.$$

A trajectory can remain in E only if

$$\frac{d}{dt} g(\mathbf{y}(t)) = \nabla g(\mathbf{y}) \cdot \mathbf{f}(\mathbf{y}) = 0$$

along that trajectory. This additional condition often reduces the largest invariant subset far below the full zero-dissipation set.

11.5.2 LaSalle's Invariance Principle

Theorem 11.28 (LaSalle's invariance principle)**.** *Let $\Omega \subset \mathbb{R}^n$ be compact and positively invariant, and let $\Phi \in C^1(\Omega)$ satisfy*

$$\dot{\Phi} \le 0 \qquad \text{on } \Omega.$$

Let

$$E = \{\mathbf{y} \in \Omega : \dot{\Phi}(\mathbf{y}) = 0\},$$

and let $\mathcal{M}(E)$ be the largest invariant subset of E. Then every trajectory starting in Ω approaches $\mathcal{M}(E)$:

$$\operatorname{dist}\big(\varphi_t(\mathbf{y}_0), \mathcal{M}(E)\big) \longrightarrow 0 \qquad \text{as } t \to \infty.$$

Proof. Let $\mathbf{y}_0 \in \Omega$. Since Ω is compact and positively invariant, the positive orbit of $\mathbf{y}_0$ has compact closure, and its ω-limit set is nonempty, compact, and invariant.

By the monotonicity principle, Φ is constant on $\omega(\mathbf{y}_0)$ and

$$\omega(\mathbf{y}_0) \subset E.$$

Because $\omega(\mathbf{y}_0)$ is invariant, it must lie in the largest invariant subset of E:

$$\omega(\mathbf{y}_0) \subset \mathcal{M}(E).$$

A trajectory with compact positive-orbit closure approaches its ω-limit set. Therefore

$$\operatorname{dist}\big(\varphi_t(\mathbf{y}_0), \mathcal{M}(E)\big) \longrightarrow 0.$$

□

LaSalle's principle does not necessarily imply convergence to a single equilibrium. Its conclusion is convergence to an invariant set. Further information about that set is needed to determine whether the trajectory approaches an equilibrium, a continuum of equilibria, a periodic orbit, or some other invariant motion.

In the important special case

$$\mathcal{M}(E) = \{\mathbf{y}_*\},$$

LaSalle's principle gives

$$\varphi_t(\mathbf{y}_0) \longrightarrow \mathbf{y}_*.$$

Thus the central step in applying the theorem is not merely computing $\dot{\Phi}$, but identifying the largest invariant subset of $\{\dot{\Phi} = 0\}$.

11.5.3 Applications of LaSalle's Principle

The main step in applying LaSalle's principle is to distinguish the zero-dissipation set from its largest invariant subset. The following example illustrates this distinction with a nonlinear friction law.

Example 11.29 (A pendulum with nonlinear friction)**.** Consider

$$\dot{q} = p, \qquad \dot{p} = -\sin q - \gamma p^3, \qquad \gamma > 0.$$

The mechanical energy

$$\Phi(q,p) = \frac{1}{2}p^2 + 1 - \cos q$$

satisfies

$$\begin{aligned}\dot{\Phi} &= p\dot{p} + \sin q\, \dot{q} \\ &= -\gamma p^4 \leq 0.\end{aligned}$$

Hence the zero-dissipation set is

$$E = \{(q,p) : p = 0\}.$$

The set E is not invariant. At a point with $p = 0$,

$$\dot{p} = -\sin q,$$

so a trajectory can remain in E only if

$$\sin q = 0.$$

Therefore the largest invariant subset of E consists of the equilibria

$$(q,p) = (k\pi, 0), \qquad k \in \mathbb{Z}.$$

For $0 < c < 2$, let Ω_c be the component containing $(0,0)$ of

$$\{(q,p) : \Phi(q,p) \leq c\}.$$

The set Ω_c is compact and positively invariant. Moreover, the only invariant subset of $E \cap \Omega_c$ is the downward equilibrium $(0,0)$. LaSalle's invariance principle therefore gives

$$(q(t), p(t)) \longrightarrow (0,0) \qquad \text{as } t \to \infty$$

for every initial condition in Ω_c.

Example 11.30 (Indirect damping of coupled oscillators)**.** Consider two masses connected to fixed supports by springs and coupled to each other by an additional spring. Suppose that damping acts only on the first mass:

$$\begin{aligned} m_1\ddot{q}_1 + \gamma\dot{q}_1 + k_1 q_1 + \kappa(q_1 - q_2) &= 0, \\ m_2\ddot{q}_2 + k_2 q_2 + \kappa(q_2 - q_1) &= 0, \end{aligned}$$

where

$$m_1, m_2, k_1, k_2, \kappa, \gamma > 0.$$

Here k_1 and k_2 are the grounding stiffnesses, κ is the coupling stiffness, and γ is the damping coefficient.

The total mechanical energy is

$$\begin{aligned} E = \frac{1}{2}m_1\dot{q}_1^2 + \frac{1}{2}m_2\dot{q}_2^2 \\ + \frac{1}{2}k_1 q_1^2 + \frac{1}{2}k_2 q_2^2 + \frac{1}{2}\kappa(q_1 - q_2)^2. \end{aligned}$$

Along solutions,

$$\dot{E} = -\gamma\dot{q}_1^2 \leq 0.$$

Thus the zero-dissipation set is

$$\{\dot{E} = 0\} = \{\dot{q}_1 = 0\}.$$

Suppose that a trajectory remains in this set. Then

$$\ddot{q}_1 = 0,$$

and the first equation gives

$$k_1 q_1 + \kappa(q_1 - q_2) = 0.$$

Differentiating this relation and using $\dot{q}_1 = 0$ gives

$$-\kappa\dot{q}_2 = 0,$$

so

$$\dot{q}_2 = 0.$$

The second equation then yields

$$k_2 q_2 + \kappa(q_2 - q_1) = 0.$$

Consequently, (q_1, q_2) satisfies

$$\begin{bmatrix} k_1+\kappa & -\kappa \\ -\kappa & k_2+\kappa \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The stiffness matrix is positive definite because

$$k_1 + \kappa > 0$$

and

$$(k_1+\kappa)(k_2+\kappa) - \kappa^2 = k_1 k_2 + \kappa(k_1+k_2) > 0.$$

Hence

$$q_1 = q_2 = 0.$$

The largest invariant subset of $\{\dot{E} = 0\}$ is therefore the origin. Since E is positive definite and coercive, its sublevel sets are compact and positively invariant. LaSalle's invariance principle gives

$$(q_1, \dot{q}_1, q_2, \dot{q}_2) \longrightarrow (0,0,0,0).$$

Thus damping applied to one oscillator stabilizes the entire system: coupling transfers energy from the undamped oscillator to the damped one.

Example 11.31 (Entropy and a reversible chemical reaction)**.** Consider the reversible reaction

$$A + B \underset{k_r}{\overset{k_f}{\rightleftharpoons}} C,$$

where $a(t)$, $b(t)$, and $c(t)$ are the concentrations. Under mass-action kinetics,

$$\begin{aligned} \dot{a} &= -k_f ab + k_r c, \\ \dot{b} &= -k_f ab + k_r c, \\ \dot{c} &= k_f ab - k_r c. \end{aligned}$$

The quantities

$$a + c = A_0, \qquad b + c = B_0$$

are conserved. Hence the corresponding nonnegative stoichiometric compatibility class

$$\big\{(a,b,c) \in \mathbb{R}^3_+ : a + c = A_0, \quad b + c = B_0\big\}$$

is compact and positively invariant. If the initial concentrations are strictly positive, they remain positive. Moreover, the vector field points strictly inward at every boundary point compatible with $A_0, B_0 > 0$, so the positive orbit has compact closure in the interior of its compatibility class.

The equilibrium in this class is determined by

$$a_* + c_* = A_0, \qquad b_* + c_* = B_0,$$

and the detailed-balance condition

$$k_f a_* b_* = k_r c_*.$$

These equations determine a unique positive equilibrium (a_*, b_*, c_*).

Define the relative entropy

$$\begin{aligned}\Phi(a,b,c) = a\log\frac{a}{a_*} - a + a_* \\ + b\log\frac{b}{b_*} - b + b_* \\ + c\log\frac{c}{c_*} - c + c_*.\end{aligned}$$

This function is nonnegative and vanishes only at the equilibrium. Its derivative is

$$\begin{aligned}\dot\Phi &= -(k_f ab - k_r c)\log\frac{k_f ab}{k_r c} \\ &\le 0,\end{aligned}$$

because

$$(u-v)\log\frac{u}{v} \ge 0 \qquad \text{for } u, v > 0.$$

Moreover,

$$\dot\Phi = 0 \quad\Longleftrightarrow\quad k_f ab = k_r c.$$

Within a fixed compatibility class, this condition identifies only the detailed-balance equilibrium. LaSalle's invariance principle therefore gives

$$(a(t), b(t), c(t)) \longrightarrow (a_*, b_*, c_*).$$

This example shows how conservation laws first restrict the dynamics to a compact compatibility class, while entropy production selects the equilibrium approached within that class.

Example 11.32 (A charged particle attracted to a curve of equilibria)**.** Consider an overdamped charged particle moving in the plane under a potential $U(x, y)$ and a constant magnetic field perpendicular to the plane. Balancing drag, the Lorentz force, and the potential force gives

$$(\gamma\mathbf{I} + \beta\mathbf{J})\dot{\mathbf{q}} = -\nabla U(\mathbf{q}), \qquad \mathbf{J} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix},$$

where $\gamma > 0$ is the drag coefficient and β measures the magnetic field strength.

Since

$$(\gamma\mathbf{I} + \beta\mathbf{J})^{-1} = \frac{\gamma\mathbf{I} - \beta\mathbf{J}}{\gamma^2 + \beta^2},$$

the equation can be written as

$$\dot{\mathbf{q}} = -a\nabla U(\mathbf{q}) + b\mathbf{J}\nabla U(\mathbf{q}),$$

where

$$a = \frac{\gamma}{\gamma^2 + \beta^2} > 0, \qquad b = \frac{\beta}{\gamma^2 + \beta^2}.$$

The first term moves down the potential, while the second is perpendicular to ∇U and produces a transverse magnetic drift.

Let

$$U(x, y) = \frac{1}{2}\big(x^2 + 2y^2 - 1\big)^2.$$

The minimum set of U is the ellipse

$$\mathcal{E} = \{(x, y) : x^2 + 2y^2 = 1\},$$

and every point of $\mathcal{E}$ is an equilibrium.

Because

$$\nabla U \cdot \mathbf{J}\nabla U = 0,$$

we have

$$\begin{aligned} \dot{U} &= \nabla U \cdot \dot{\mathbf{q}} \\ &= -a|\nabla U|^2 \leq 0. \end{aligned}$$

Thus drag dissipates potential energy, whereas the magnetic term changes the direction of motion without changing U.

Choose $0 < c_1 < 1 < c_2$ and let

$$\Omega = \{(x, y) : c_1 \leq x^2 + 2y^2 \leq c_2\}.$$

This set is compact and positively invariant. Within Ω,

$$\dot{U} = 0 \quad \Longleftrightarrow \quad (x, y) \in \mathcal{E}.$$

Since $\mathcal{E}$ is invariant, LaSalle's principle gives

$$\operatorname{dist}\big((x(t), y(t)), \mathcal{E}\big) \longrightarrow 0.$$

The trajectory generally approaches the ellipse while drifting tangentially under the magnetic force. This example illustrates how dissipative and gyroscopic effects can coexist: one determines approach toward the equilibrium manifold, while the other changes the path by which it is approached.

Example 11.33 (Convergence to an invariant periodic orbit)**.** Consider the system on

$$\mathbb{R}^2 \times \mathbb{S}^1$$

given by

$$\dot{q} = p, \qquad \dot{p} = -V'(q) - \gamma p, \qquad \dot{\theta} = \omega(q),$$

where $\gamma > 0$. Suppose that V is coercive, has a unique critical point q_*, and that this point is

a strict minimum. Assume also that

$$\omega(q_*) \neq 0.$$

Define

$$\Phi(q,p,\theta) = \frac{1}{2}p^2 + V(q) - V(q_*).$$

Then

$$\dot{\Phi} = -\gamma p^2 \leq 0.$$

The zero-dissipation set is

$$E = \{p = 0\}.$$

A trajectory can remain in E only if

$$\dot{p} = -V'(q) = 0,$$

and hence $q = q_*$. Therefore the largest invariant subset of E is

$$\Gamma = \{(q_*, 0, \theta) : \theta \in \mathbb{S}^1\}.$$

On Γ, the remaining dynamics are

$$\dot{\theta} = \omega(q_*),$$

so Γ is a periodic orbit. LaSalle's principle gives

$$\operatorname{dist}\big((q(t),p(t),\theta(t)),\Gamma\big) \longrightarrow 0.$$

In particular,

$$q(t) \to q_*, \qquad p(t) \to 0,$$

while the phase $\theta(t)$ continues to rotate. Thus dissipation removes the transverse motion but leaves persistent motion on the limiting invariant set.

Exercise 11.11 (An undamped mode in a partially damped network)**.** Consider three identical oscillators,

$$\begin{aligned} \ddot{q}_1 + q_1 + \kappa(q_1 - q_2) &= 0, \\ \ddot{q}_2 + \gamma\dot{q}_2 + q_2 + \kappa(2q_2 - q_1 - q_3) &= 0, \\ \ddot{q}_3 + q_3 + \kappa(q_3 - q_2) &= 0, \end{aligned}$$

where $\gamma, \kappa > 0$. Show that the mechanical energy is nonincreasing, but that damping the middle oscillator does not force every solution to the origin. Find a nontrivial invariant motion in the zero-dissipation set and explain its mechanical meaning.

Exercise 11.12 (Cluster consensus)**.** Revisit the diffusion equation on an undirected weighted graph from Section 5.1.4,

$$\dot{\mathbf{x}} = -\mathbf{L}\mathbf{x},$$

but now suppose that the graph is not necessarily connected. Use LaSalle's invariance principle to determine the largest invariant subset of

$$\{\mathbf{x} : \mathbf{x}^T \mathbf{L} \mathbf{x} = 0\}.$$

Show that the solution approaches a constant value on each connected component, and determine each limiting value from the initial data.

Exercise 11.13 (A reversible reaction chain)**.** Consider

$$A \underset{k_1^-}{\overset{k_1^+}{\rightleftharpoons}} B \underset{k_2^-}{\overset{k_2^+}{\rightleftharpoons}} C$$

with mass-action kinetics and positive initial concentrations. Write the differential equations and identify the conserved total concentration. Let (a_*, b_*, c_*) be the detailed-balance equilibrium in the corresponding compatibility class. Show that

$$\Phi = \sum_{u \in \{a,b,c\}} \left(u \log \frac{u}{u_*} - u + u_* \right)$$

is nonincreasing, and use LaSalle's principle to prove convergence to (a_*, b_*, c_*).

Exercise 11.14 (Attraction to a quasiperiodic torus)**.** Consider

$$\dot{r} = -r(r-1), \qquad \dot{\theta}_1 = \omega_1, \qquad \dot{\theta}_2 = \omega_2$$

on

$$(0, \infty) \times \mathbb{T}^2.$$

Use

$$\Phi(r) = (r-1)^2$$

and LaSalle's principle to show that every trajectory approaches the invariant torus

$$\mathcal{T} = \{r = 1\}.$$

Describe the motion on $\mathcal{T}$ when ω_1/ω_2 is rational and when it is irrational. Does a trajectory necessarily converge to a single orbit or point of $\mathcal{T}$?

11.6 Absorbing Sets and Global Attractors

The preceding sections used monotone quantities to confine individual trajectories and identify their limiting invariant sets. We now ask for a stronger form of confinement: can one bounded region eventually contain every bounded family of initial conditions?

This question leads to absorbing sets and global attractors. An absorbing set describes eventual confinement, whereas the global attractor contains the invariant dynamics that remain after

transient behavior has disappeared.

Throughout this section, suppose that

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$$

generates a continuous forward flow φ_t on $\mathbb{R}^n$.

11.6.1 Absorbing Sets

Definition 11.34. A bounded set $B \subset \mathbb{R}^n$ is called an *absorbing set* if, for every bounded set $D \subset \mathbb{R}^n$, there exists a time $T_D \geq 0$ such that

$$\varphi_t(D) \subset B, \qquad t \geq T_D.$$

The time T_D may depend on the initial set D. Thus an absorbing set need not contain every trajectory initially; it must contain every bounded family of trajectories after a sufficiently long time.

An absorbing set is stronger than a bounded positively invariant set. A positively invariant set retains trajectories that begin inside it, whereas an absorbing set eventually captures trajectories that begin outside it. For example, every disk centered at the origin is positively invariant for the harmonic oscillator, but no such disk is absorbing because trajectories outside it remain on larger energy circles.

Likewise, boundedness of each individual trajectory does not imply the existence of a bounded absorbing set. The harmonic oscillator again provides an example: every trajectory is bounded, but the radii of its periodic orbits are not bounded uniformly over all initial data. By contrast, for

$$\dot{x} = x - x^3,$$

every interval $[-R, R]$ with $R > 1$ is absorbing, since the vector field points toward the interval outside its endpoints.

A common way to construct an absorbing set is through a differential inequality. Suppose that a nonnegative coercive function Φ satisfies

$$\dot{\Phi} \leq -c\Phi + C, \qquad c > 0, \quad C \geq 0.$$

Multiplying by e^{ct} and integrating gives

$$\Phi(\varphi_t(\mathbf{y}_0)) \leq e^{-ct}\Phi(\mathbf{y}_0) + \frac{C}{c}(1 - e^{-ct}).$$

Consequently, for every $R > C/c$, the sublevel set

$$B_R = \{\mathbf{y} : \Phi(\mathbf{y}) \leq R\}$$

absorbs every bounded set. Coercivity makes B_R bounded.

11.6.2 Uniform Attraction of Sets

To describe attraction of a whole set, define the one-sided distance

$$\operatorname{dist}(D, A) = \sup_{\mathbf{y}\in D} \inf_{\mathbf{a}\in A} |\mathbf{y} - \mathbf{a}|.$$

Thus

$$\operatorname{dist}\big(\varphi_t(D), A\big) \to 0$$

means that every point of $\varphi_t(D)$ eventually lies uniformly close to A. This does not require every point of A to lie close to $\varphi_t(D)$.

Definition 11.35. A compact invariant set $\mathcal{A}$ is called a *global attractor* if it attracts every bounded set:

$$\operatorname{dist}\big(\varphi_t(D), \mathcal{A}\big) \longrightarrow 0 \qquad \text{as } t \to \infty$$

for every bounded set $D \subset \mathbb{R}^n$.

This definition is stronger than pointwise attraction of every initial condition. The attraction must be uniform over each bounded family of initial conditions.

A global attractor may contain several local attractors, unstable equilibria, periodic orbits, and trajectories connecting them. It represents the complete bounded invariant dynamics of the system, rather than one particular asymptotic state.

11.6.3 Construction of the Global Attractor

An absorbing set describes eventual confinement but need not itself be the final attracting object. Once trajectories enter it, they may continue to approach a smaller invariant set.

For a bounded set B, define

$$\omega(B) = \bigcap_{T\ge 0} \overline{\bigcup_{t\ge T} \varphi_t(B)}.$$

Equivalently, $\mathbf{z} \in \omega(B)$ if there exist sequences

$$\mathbf{y}_j \in B, \qquad t_j \to \infty,$$

such that

$$\varphi_{t_j}(\mathbf{y}_j) \to \mathbf{z}.$$

If B is compact and positively invariant, then

$$\varphi_{t+s}(B) = \varphi_t\big(\varphi_s(B)\big) \subset \varphi_t(B),$$

so the compact sets $\varphi_t(B)$ decrease with t. In this case,

$$\omega(B) = \bigcap_{t\ge 0} \varphi_t(B).$$

Theorem 11.36 (Existence of a global attractor)**.** *Suppose that the flow possesses a compact, positively invariant absorbing set B. Then*

$$\mathcal{A} = \omega(B)$$

is a global attractor. Moreover, $\mathcal{A}$ is the largest compact invariant set contained in B and is independent of the particular absorbing set used in its construction.

Proof. Because the sets $\varphi_t(B)$ are nonempty, compact, and decreasing, their intersection

$$\mathcal{A} = \bigcap_{t \geq 0} \varphi_t(B)$$

is nonempty and compact.

We first show that $\mathcal{A}$ attracts B. Suppose otherwise. Then there exist $\varepsilon > 0$, times $t_j \to \infty$, and points $\mathbf{y}_j \in B$ such that

$$\operatorname{dist}\big(\varphi_{t_j}(\mathbf{y}_j), \mathcal{A}\big) \geq \varepsilon.$$

Since

$$\varphi_{t_j}(\mathbf{y}_j) \in B$$

and B is compact, a subsequence converges to some $\mathbf{z} \in B$.

For every fixed $T > 0$, we have $t_j \geq T$ for sufficiently large j, and hence

$$\varphi_{t_j}(\mathbf{y}_j) \in \varphi_T(B).$$

Since $\varphi_T(B)$ is compact,

$$\mathbf{z} \in \varphi_T(B).$$

This holds for every $T \geq 0$, so

$$\mathbf{z} \in \mathcal{A},$$

contradicting the choice of the sequence. Therefore

$$\operatorname{dist}\big(\varphi_t(B), \mathcal{A}\big) \longrightarrow 0.$$

Now let D be bounded. Since B is absorbing, there exists T_D such that

$$\varphi_t(D) \subset B, \qquad t \geq T_D.$$

For $t \geq T_D$,

$$\varphi_t(D) = \varphi_{t-T_D}\big(\varphi_{T_D}(D)\big) \subset \varphi_{t-T_D}(B).$$

It follows that

$$\operatorname{dist}\big(\varphi_t(D), \mathcal{A}\big) \longrightarrow 0.$$

The flow property immediately gives

$$\varphi_s(\mathcal{A}) \subset \mathcal{A}.$$

For the reverse inclusion, let $\mathbf{z} \in \mathcal{A}$ and fix $s > 0$. Choose

$$t_j \to \infty, \qquad \mathbf{y}_j \in B,$$

such that

$$\varphi_{t_j}(\mathbf{y}_j) \to \mathbf{z}.$$

For $t_j > s$, set

$$\mathbf{z}_j = \varphi_{t_j - s}(\mathbf{y}_j).$$

By compactness, a subsequence converges to some $\mathbf{z}_* \in \mathcal{A}$. Continuity of the flow gives

$$\varphi_s(\mathbf{z}_*) = \mathbf{z}.$$

Hence

$$\varphi_s(\mathcal{A}) = \mathcal{A},$$

so $\mathcal{A}$ is invariant.

Finally, let $K \subset B$ be compact and invariant. Then

$$K = \varphi_t(K) \subset \varphi_t(B)$$

for every $t \geq 0$. Therefore

$$K \subset \bigcap_{t \geq 0} \varphi_t(B) = \mathcal{A}.$$

Thus $\mathcal{A}$ is the largest compact invariant subset of B.

Since every compact invariant set is bounded and is eventually absorbed by B, the same argument shows that every compact invariant set lies in $\mathcal{A}$. The global attractor is therefore unique and does not depend on the choice of the absorbing set. □

The theorem gives the general route

$$\text{a priori estimate} \longrightarrow \text{compact absorbing set} \longrightarrow \text{global attractor.}$$

In finite dimensions, compactness follows from boundedness and closedness. For evolution equations in infinite-dimensional spaces, bounded sets need not be compact, and an additional property called asymptotic compactness is usually required.

11.6.4 Examples of Global Attractors

Example 11.37 (A bistable global attractor)**.** Consider

$$\dot{x} = x - x^3.$$

For every $R > 1$, the interval

$$B_R = [-R, R]$$

is compact, positively invariant, and absorbing. The global attractor is therefore

$$\mathcal{A} = \omega(B_R).$$

The equilibria are

$$-1, \qquad 0, \qquad 1.$$

The equilibria ± 1 are asymptotically stable, while 0 is unstable. Every point of $(0, 1)$ lies on a trajectory connecting 0 to 1, and every point of $(-1, 0)$ lies on a trajectory connecting 0 to -1. Consequently,

$$\mathcal{A} = [-1, 1].$$

Thus the global attractor is larger than the union of the two stable equilibria. It contains the unstable equilibrium and the connecting trajectories required to make the set invariant.

Although each individual trajectory approaches one equilibrium, uniform attraction of bounded sets requires the entire interval.

Example 11.38 (An absorbing set from a nonlinear estimate)**.** Consider

$$\dot{x} = -x^3 + y, \qquad \dot{y} = -y - x + F,$$

where F is a constant forcing parameter. Define

$$\Phi(x, y) = \frac{1}{2}(x^2 + y^2).$$

Then

$$\begin{aligned}\dot{\Phi} &= x(-x^3 + y) + y(-y - x + F) \\ &= -x^4 - y^2 + Fy.\end{aligned}$$

Using

$$Fy \le \frac{1}{2}y^2 + \frac{1}{2}F^2$$

and

$$x^4 \ge x^2 - \frac{1}{4},$$

we obtain

$$\begin{aligned}\dot{\Phi} &\le -x^2 - \frac{1}{2}y^2 + \frac{1}{4} + \frac{1}{2}F^2 \\ &\le -\Phi + \frac{1}{4} + \frac{1}{2}F^2.\end{aligned}$$

Therefore

$$\Phi(t) \le e^{-t}\Phi(0) + \left(\frac{1}{4} + \frac{1}{2}F^2\right)(1 - e^{-t}).$$

It follows that, for every

$$R > \frac{1}{4} + \frac{1}{2}F^2,$$

the disk

$$B_R = \{(x, y) : \Phi(x, y) \le R\}$$

is absorbing. Enlarging R slightly if necessary gives a compact, positively invariant absorbing set. Hence the system possesses a global attractor.

This estimate proves the existence of a global attractor without first determining its detailed structure. It also shows how the size of an absorbing region depends on the forcing amplitude F.

The structure of the global attractor can also be determined. The system has a unique equilibrium (x_*, y_*), where

$$x_*^3 + x_* = F, \qquad y_* = x_*^3.$$

For

$$W(x,y) = \frac{1}{2}(x-x_*)^2 + \frac{1}{2}(y-y_*)^2,$$

we have

$$\dot{W} = -(x-x_*)(x^3-x_*^3) - (y-y_*)^2 < 0$$

away from the equilibrium. Hence (x_*, y_*) is globally asymptotically stable, and the global attractor is

$$\mathcal{A} = \{(x_*, y_*)\}.$$

Example 11.39 (The global attractor of a damped double-well oscillator)**.** Consider the unforced damped Duffing equation

$$\ddot{q} + \gamma\dot{q} + q^3 - q = 0, \qquad \gamma > 0.$$

Writing $p = \dot{q}$, the system becomes

$$\dot{q} = p, \qquad \dot{p} = -\gamma p - q^3 + q.$$

Its mechanical energy is

$$E(q,p) = \frac{1}{2}p^2 + \frac{1}{4}(q^2-1)^2,$$

and

$$\dot{E} = -\gamma p^2 \le 0.$$

The equilibria are

$$(-1,0), \qquad (0,0), \qquad (1,0).$$

The equilibria $(\pm 1, 0)$ are asymptotically stable, while $(0,0)$ is a saddle. Their energy values are

$$E(\pm 1, 0) = 0, \qquad E(0,0) = \frac{1}{4}.$$

The coercivity of E gives uniform bounds on bounded families of trajectories. Using LaSalle's principle and compactness, one finds that, for every $R > 1/4$, the sublevel set

$$B_R = \{(q,p) : E(q,p) \le R\}$$

absorbs every bounded set. Hence the system possesses a global attractor.

The two branches of the unstable manifold of $(0,0)$ leave the saddle in opposite directions. Since the energy immediately decreases below $1/4$, neither branch can cross the barrier $q = 0$ again. LaSalle's principle then gives

$$W^u_+(0,0) \longrightarrow (1,0), \qquad W^u_-(0,0) \longrightarrow (-1,0)$$

as $t \to \infty$. These branches are heteroclinic orbits.

The global attractor is

$$\mathcal{A} = \{(-1,0),(0,0),(1,0)\} \cup W^u(0,0).$$

Thus $\mathcal{A}$ contains the two stable equilibria, the unstable saddle, and the heteroclinic trajectories connecting the saddle to the stable equilibria.

This example illustrates why a global attractor is not merely the set of possible limits as $t \to \infty$. The heteroclinic trajectories must also be included so that the attractor is invariant and contains all bounded complete dynamics.

Remark 11.40 (Equilibria, invariant sets, and attractors). These notions describe different properties of the dynamics.

An *equilibrium* is a point $\mathbf{y}_*$ satisfying

$$\mathbf{f}(\mathbf{y}_*) = \mathbf{0}.$$

The singleton $\{\mathbf{y}_*\}$ is invariant, but it need not attract nearby trajectories. It may be stable, unstable, or asymptotically stable.

An *invariant set* A satisfies

$$\varphi_t(A) = A.$$

Invariant sets may consist of equilibria, periodic orbits, invariant manifolds, connecting trajectories, or more complicated motions. Invariance alone says nothing about attraction.

An *attractor* is a compact invariant set that attracts a neighborhood. It need not be an equilibrium or even a set of equilibria: an attracting periodic orbit is an attractor with persistent motion. Moreover, an attractor may contain unstable equilibria and heteroclinic orbits, provided that the set as a whole is attracting.

Thus

$$\text{equilibrium} \quad \Longrightarrow \quad \text{invariant singleton},$$

and

$$\text{attractor} \quad \Longrightarrow \quad \text{compact invariant set},$$

but the converses do not hold. An equilibrium is an attractor precisely when it is locally attracting under the convention adopted in this chapter, while a general invariant set need not attract anything.

A *global attractor* is stronger: it attracts every bounded set uniformly. It may contain several local attractors together with unstable equilibria and the connecting trajectories that organize their basins. For the damped Duffing equation, the two stable equilibria are local attractors, whereas the global attractor also contains the saddle and its heteroclinic orbits.

Exercise 11.15 (Pointwise attraction need not be uniform)**.** Consider

$$\dot{x} = x^2(1-x).$$

Show that every solution is bounded and that

$$x(t) \longrightarrow \begin{cases} 0, & x(0) \le 0, \\ 1, & x(0) > 0. \end{cases}$$

Thus the equilibrium set

$$E = \{0, 1\}$$

attracts every individual trajectory.

Show, however, that E does not attract the interval $[0, 1]$ uniformly. In particular, for every sufficiently large t, find an initial condition $x_0(t) \in (0, 1)$ whose solution satisfies

$$x(t; x_0(t)) = \frac{1}{2}.$$

Determine the global attractor and explain the role of the trajectories in $(0, 1)$.

Exercise 11.16 (A planar global attractor)**.** Consider

$$\dot{x} = y - x^3, \qquad \dot{y} = x - y.$$

Find and classify the equilibria, construct a compact absorbing set, and determine the global attractor. Identify its heteroclinic orbits and the common boundary of the two basins of attraction.

Hint. The system is a gradient system with potential

$$V(x, y) = \frac{1}{4}(x^2 - 1)^2 + \frac{1}{2}(y - x)^2.$$

Use $\dot{V} = -|\nabla V|^2$ and the invariance of the positive and negative quadrants.

Exercise 11.17 (A nonlinear absorbing set)**.** Consider

$$\dot{x} = -x^3 + y + F, \qquad \dot{y} = -y^3 - x + G,$$

where F and G are constants.

Using

$$\Phi(x, y) = \frac{1}{2}(x^2 + y^2),$$

derive a dissipative estimate and construct a compact absorbing set. Show that the system has a unique equilibrium and use a shifted quadratic Lyapunov function to determine the global attractor.

Exercise 11.18 (Global attractors of coercive gradient systems)**.** Consider

$$\dot{\mathbf{y}} = -\nabla V(\mathbf{y}),$$

where $V \in C^2(\mathbb{R}^n)$ is bounded below and coercive. Suppose that

$$|\nabla V(\mathbf{y})|^2 \geq cV(\mathbf{y}) - C$$

for some $c > 0$ and $C \geq 0$.

Show that a sufficiently large sublevel set of V is absorbing and hence that the system possesses a global attractor. If V has only finitely many critical points, explain why every complete trajectory in the global attractor is either an equilibrium or a trajectory connecting critical points with decreasing values of V.

11.7 Periodic Attractors and Planar Long-Time Dynamics

Example 11.4 gave an explicit planar system whose ω-limit set is an attracting periodic orbit rather than an equilibrium. Periodic attractors also appeared geometrically in Chapter 4 in the Van der Pol oscillator, the Brusselator, and the FitzHugh–Nagumo system. In those models, phase portraits together with energy or nullcline arguments suggested sustained oscillations and attracting limit cycles, but did not yet provide a general theorem explaining why periodic motion must occur.

We now develop the planar long-time theory that fills this gap. The central result is the Poincaré–Bendixson theorem, which places strong restrictions on the ω-limit sets of bounded planar trajectories and provides a basic method for proving the existence of periodic orbits.

11.7.1 Periodic attractors

Let Γ be a periodic orbit of

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}).$$

The orbit Γ is called an *attracting periodic orbit* if there exists a neighborhood U of Γ such that

$$\operatorname{dist}\big(\varphi_t(\mathbf{y}_0), \Gamma\big) \to 0 \qquad \text{as } t \to \infty$$

for every $\mathbf{y}_0 \in U$.

If Γ is isolated among periodic orbits, it is called a *limit cycle*. An attracting limit cycle is therefore a periodic attractor.

Convergence to a periodic orbit is convergence to a set, not to a point. A trajectory approaching Γ continues to move around the cycle, and its phase along the cycle generally depends on the initial value. Thus

$$\operatorname{dist}\big(\varphi_t(\mathbf{y}_0), \Gamma\big) \to 0$$

does not imply that $\varphi_t(\mathbf{y}_0)$ has a limit as $t \to \infty$.

For the Van der Pol oscillator, the Brusselator, and the FitzHugh–Nagumo system studied in Chapter 4, the periodic orbits are not available in explicit form. Their phase portraits suggest attracting limit cycles, raising a more general question: under what conditions can one prove that the long-time behavior of a bounded planar trajectory is periodic?

The answer is provided by the Poincaré–Bendixson theorem.

11.7.2 The Poincaré–Bendixson theorem

The examples above suggest that bounded planar trajectories have rather restricted long-time behavior. The Poincaré–Bendixson theorem makes this precise: if the limiting motion contains no equilibrium, then the only remaining possibility is periodic motion.

Theorem 11.41 (Poincaré–Bendixson theorem)**.** *Consider a C^1 planar system*

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}), \qquad \mathbf{y} \in \mathbb{R}^2.$$

Suppose that the positive orbit of $\mathbf{y}_0$ is contained in a compact set and that its ω-limit set contains no equilibrium. Then

$$\omega(\mathbf{y}_0)$$

is a periodic orbit.

Thus, for a bounded planar trajectory, the absence of equilibria in its ω-limit set forces periodic behavior. This is one of the strongest special features of two-dimensional dynamics.

Proof sketch. Let

$$\Omega = \omega(\mathbf{y}_0).$$

Since the positive orbit of $\mathbf{y}_0$ is contained in a compact set, Ω is nonempty, compact, connected, and invariant.

Choose a point

$$\mathbf{z} \in \Omega.$$

By assumption,

$$\mathbf{f}(\mathbf{z}) \neq \mathbf{0}.$$

Take a short line segment Σ through $\mathbf{z}$ perpendicular to $\mathbf{f}(\mathbf{z})$. Since $\mathbf{f}$ is continuous, after shortening Σ and restricting to a sufficiently small neighborhood of $\mathbf{z}$, the vector field has a nonzero component across Σ and points across it in the same direction. Thus every trajectory passing sufficiently close to $\mathbf{z}$ crosses Σ transversely.

Because $\mathbf{z} \in \Omega$, there exists a sequence $t_n \to \infty$ such that $\varphi_{t_n}(\mathbf{y}_0) \to \mathbf{z}$. Hence the trajectory through $\mathbf{y}_0$ passes through this neighborhood infinitely many times and therefore intersects Σ infinitely many times. Let

$$\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \cdots$$

be the successive intersections.

The crucial planar observation is that these intersection points cannot occur in an arbitrary order along the one-dimensional segment Σ. After orienting Σ appropriately, the successive intersections are ordered monotonically,

$$\mathbf{p}_1 < \mathbf{p}_2 < \mathbf{p}_3 < \cdots,$$

or monotonically in the opposite direction. Since they remain in a compact subsegment of Σ, they converge to a point

$$\mathbf{p}_* \in \Sigma.$$

For points of Σ sufficiently near $\mathbf{p}_*$, follow the trajectory until its next return to Σ. This defines a local Poincaré return map

$$P : \Sigma \longrightarrow \Sigma.$$

For the successive intersections,

$$P(\mathbf{p}_n) = \mathbf{p}_{n+1}.$$

Continuous dependence of solutions on initial values implies that P is continuous. Therefore

$$P(\mathbf{p}_*) = \lim_{n\to\infty} P(\mathbf{p}_n) = \lim_{n\to\infty} \mathbf{p}_{n+1} = \mathbf{p}_*.$$

Thus $\mathbf{p}_*$ is a fixed point of the return map.

A fixed point of P corresponds to a trajectory that returns to the same point after some positive time T:

$$\varphi_T(\mathbf{p}_*) = \mathbf{p}_*.$$

Hence the trajectory through $\mathbf{p}_*$ is periodic. Denote this periodic orbit by Γ. Since $\mathbf{p}_* \in \Omega$ and Ω is invariant,

$$\Gamma \subset \Omega.$$

Finally, the same planar ordering argument shows that no additional orbit can remain in Ω outside Γ. Otherwise its intersections with a transversal to Γ would violate the ordering forced by uniqueness of planar trajectories. Consequently,

$$\Omega = \Gamma.$$

Thus $\omega(\mathbf{y}_0)$ is a periodic orbit. □

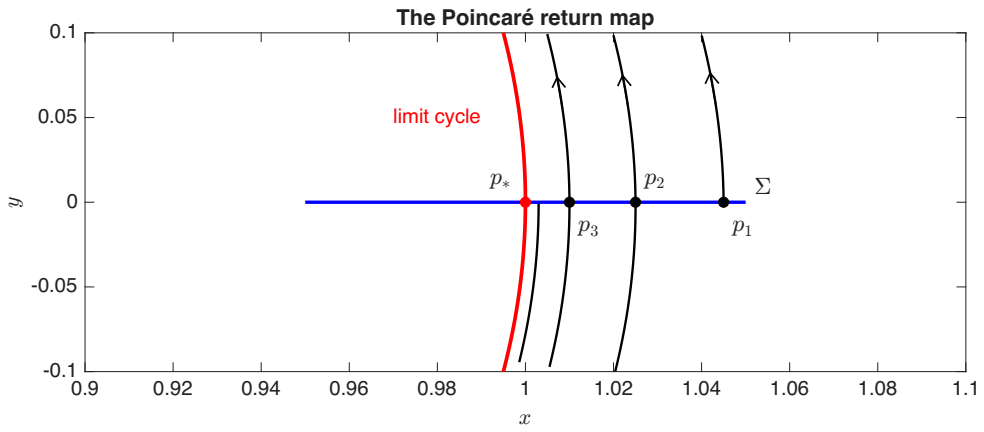


Figure 11.1: The Poincaré return map near an attracting periodic orbit. The red curve represents the limit cycle, and Σ is a local transversal. Successive intersections $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \ldots$ of a nearby trajectory with Σ approach the fixed point $\mathbf{p}_*$ of the return map. The fixed point corresponds to the periodic orbit itself.

Remark 11.42 (Where planarity enters). The key step in the proof is the ordering of successive intersections with the transversal Σ. This uses two specifically planar facts. First, uniqueness of solutions prevents two distinct trajectories from crossing. Second, an orbit segment together with the corresponding segment of Σ forms a simple closed curve, and the Jordan curve theorem separates the plane into an inside and an outside. These facts force later returns to occur in a consistent order along Σ.

This argument has no direct analogue in dimension three or higher. There a transversal has dimension at least two, so intersection points have no natural linear ordering and much more complicated recurrent behavior is possible.

Remark 11.43. The theorem does not say that every bounded planar trajectory approaches an equilibrium or a periodic orbit. If equilibria belong to the ω-limit set, the limiting set may also contain trajectories connecting equilibria. The homoclinic and heteroclinic structures introduced in Chapter 4 are examples of this possibility.

Corollary 11.44 (Trapping-region criterion)**.** *Consider a C^1 planar system*

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}).$$

Suppose that there exists a compact positively invariant region $\Omega \subset \mathbb{R}^2$ containing no equilibrium. Then Ω contains at least one periodic orbit.

Proof. Choose any $\mathbf{y}_0 \in \Omega$. Since Ω is compact and positively invariant, the positive orbit of $\mathbf{y}_0$ remains in a compact set, and hence $\omega(\mathbf{y}_0)$ is nonempty. Because Ω contains no equilibrium, neither does $\omega(\mathbf{y}_0)$. The Poincaré–Bendixson theorem therefore implies that $\omega(\mathbf{y}_0)$ is a periodic orbit contained in Ω. □

This gives a practical method for proving the existence of periodic motion: construct a compact positively invariant region and exclude equilibria from it. The Van der Pol oscillator provides an important example of this strategy.

11.7.3 The Van der Pol oscillator revisited

Chapter 4 introduced the Van der Pol oscillator

$$\ddot{x} - \mu(1 - x^2)\dot{x} + x = 0, \qquad \mu > 0,$$

and showed that the nonlinear damping injects energy when $|x| < 1$ and removes energy when $|x| > 1$. This strongly suggests the existence of a periodic orbit of preferred amplitude, but does not by itself prove that such an orbit exists. We now complete the argument using the Poincaré–Bendixson theorem.

Proposition 11.45. *For every $\mu > 0$, the Van der Pol equation*

$$\ddot{x} - \mu(1 - x^2)\dot{x} + x = 0$$

possesses at least one periodic orbit.

Proof by a trapping region and the Poincaré–Bendixson theorem. Writing

$$y = \dot{x},$$

the equation becomes

$$\dot{x} = y, \qquad \dot{y} = \mu(1 - x^2)y - x.$$

The origin is the only equilibrium. Its linearization is

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

with

$$\operatorname{tr} A = \mu > 0, \qquad \det A = 1.$$

Hence the origin is a hyperbolic source for every $\mu > 0$. More precisely, it is an unstable spiral for $0 < \mu < 2$, an improper unstable node for $\mu = 2$, and an unstable node for $\mu > 2$. Therefore there is a sufficiently small compact neighborhood D_{in} of the origin whose boundary is crossed outward by the flow.

The main task is to construct an outer boundary that prevents trajectories from escaping to infinity. The usual oscillator energy

$$E(x, y) = \frac{1}{2}(x^2 + y^2)$$

satisfies

$$\dot{E} = \mu(1 - x^2)y^2.$$

Thus energy is injected when $|x| < 1$ and dissipated when $|x| > 1$. In particular, a large circle does not provide an immediate trapping boundary, since $\dot{E} > 0$ near the y-axis.

To see the geometry more clearly, introduce the Liénard coordinate

$$v = y + \mu\left(\frac{x^3}{3} - x\right).$$

Then

$$\dot{x} = v - F(x), \qquad \dot{v} = -x,$$

where

$$F(x) = \mu\left(\frac{x^3}{3} - x\right).$$

The nullclines are

$$\dot{x} = 0 \quad \Longleftrightarrow \quad v = F(x), \qquad \dot{v} = 0 \quad \Longleftrightarrow \quad x = 0.$$

Since

$$\dot{v} = -x,$$

the flow points downward in the right half-plane and upward in the left half-plane. Likewise,

$$\dot{x} = v - F(x),$$

so the flow points to the right above the cubic $v = F(x)$ and to the left below it. These sign rules suggest the trapping geometry shown in Figure 11.3, which we now make explicit.

Choose $R > 0$ sufficiently large that

$$|F(x) - 2x| \le F(R) - 2R, \qquad |x| \le R,$$

and then choose $b > 0$ so that

$$F(R) - 2R < b < F(R) - R.$$

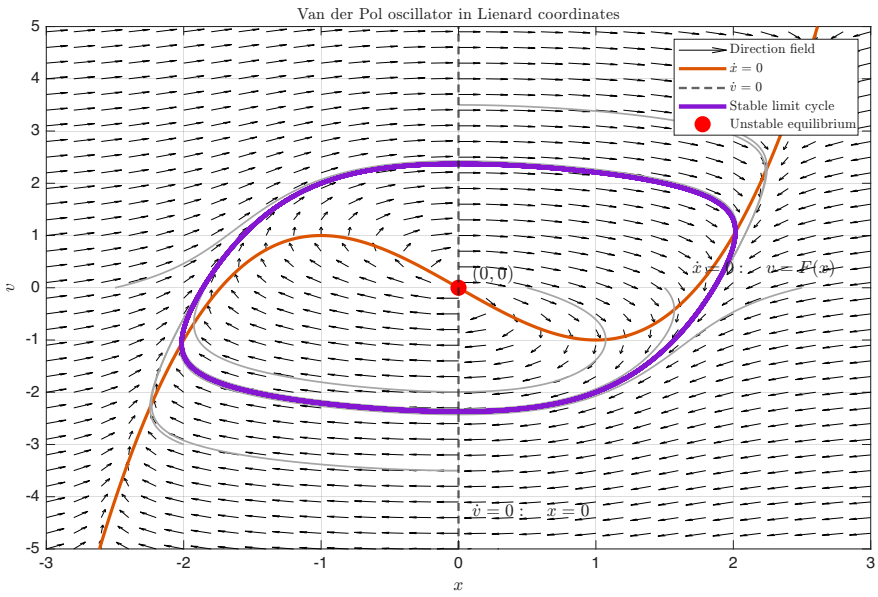


Figure 11.2: Phase portrait of the Van der Pol oscillator in Liénard coordinates for $\mu = 1.5$. The nullclines are $\dot{x} = 0$, given by the cubic $v = F(x) = \mu(x^3/3 - x)$, and $\dot{v} = 0$, given by $x = 0$. The purple curve is the numerically observed stable limit cycle whose existence is proved below.

Since $F(x) - 2x$ is an odd cubic with positive leading coefficient and has only bounded local extrema, such an R exists.

Consider the parallelogram

$$D_{\text{out}} = \{(x, v) : |x| \le R, \; |v - x| \le b\}.$$

Because F is odd, both the system and D_{out} are invariant under the central symmetry

$$(x, v) \longmapsto (-x, -v).$$

It is therefore enough to check the right vertical side and the upper oblique side.

On $x = R$, we have $v \le R + b$, and hence

$$\dot{x} = v - F(R) \le R + b - F(R) < 0.$$

Thus the vector field points into D_{out}.

On the upper boundary $v = x + b$, let

$$H(x, v) = v - x - b.$$

Since D_{out} lies in $H \le 0$, it is enough to show that $\dot{H} < 0$. Along this boundary,

$$\dot{H} = \dot{v} - \dot{x} = F(x) - 2x - b.$$

By the choice of R and b,

$$\dot{H} \le |F(x) - 2x| - b \le F(R) - 2R - b < 0.$$

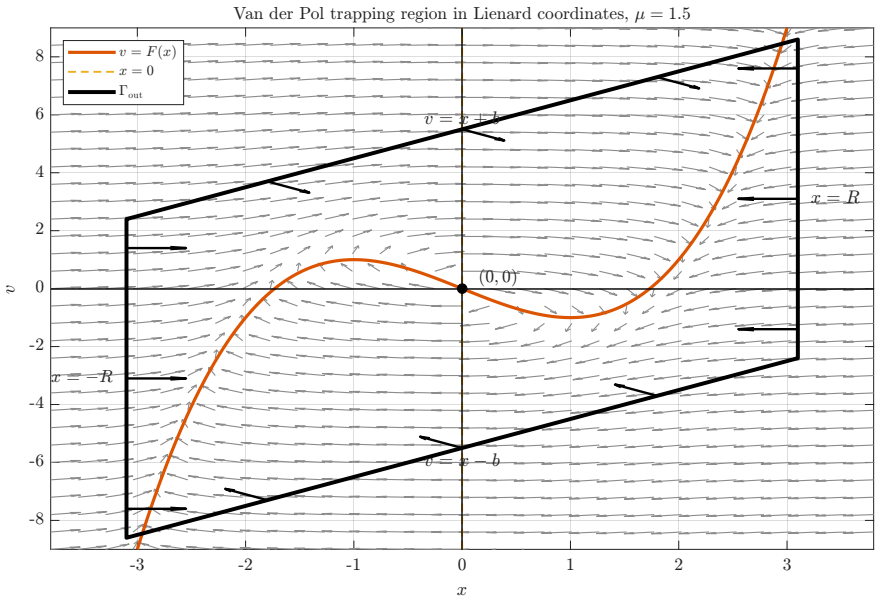


Figure 11.3: An outer trapping region for the Van der Pol oscillator in Liénard coordinates.

Thus the flow points inward along the upper side as well. By central symmetry, the same is true on the left vertical side and the lower oblique side. Hence D_{out} is positively invariant.

Since D_{in} was chosen sufficiently small around the origin, we may assume

$$D_{\text{in}} \subset D_{\text{out}}.$$

Define

$$\Omega = D_{\text{out}} \setminus \operatorname{int} D_{\text{in}}.$$

Along the outer boundary the flow points into Ω, while along the inner boundary it points away from the origin and hence again into Ω. Therefore Ω is compact and positively invariant. Moreover, the origin is the only equilibrium and has been removed from Ω, so Ω contains no equilibrium.

The trapping-region criterion, and hence the Poincaré–Bendixson theorem, now implies that Ω contains a periodic orbit. Therefore the Van der Pol equation possesses at least one periodic orbit. □

The Poincaré–Bendixson theorem proves the existence of at least one periodic orbit, but does not imply uniqueness or asymptotic stability. For the Van der Pol equation, the classical Liénard theorem (Liénard, 1928) gives the stronger conclusion that the periodic orbit is unique and asymptotically stable. The purple curve in Figure 11.2 illustrates this attracting limit cycle.

Remark 11.46 (LaSalle versus Poincaré–Bendixson). For the Liénard equation

$$\ddot{x} + f(x)\dot{x} + g(x) = 0,$$

define

$$G(x) = \int_0^x g(s)\,ds, \qquad E(x, \dot{x}) = \frac{1}{2}\dot{x}^2 + G(x).$$

Then

$$\dot{E} = -f(x)\dot{x}^2.$$

If $f(x) \geq 0$, the energy is nonincreasing, and LaSalle's invariance principle can often be used to prove convergence toward equilibria.

If f changes sign, however, E is no longer a Lyapunov function. Energy may be supplied in one region of the phase plane and dissipated in another. This is the mechanism behind self-sustained oscillations such as the Van der Pol limit cycle. In this situation, one instead seeks a compact trapping region and applies the Poincaré–Bendixson theorem to prove the existence of periodic motion.

Thus the two methods reflect two different asymptotic mechanisms: monotone dissipation toward equilibrium, and a balance of energy input and dissipation that can sustain oscillation.

11.7.4 Why the planar theory is special

The Poincaré–Bendixson theorem places strong restrictions on the long-time behavior of bounded planar trajectories. If the ω-limit set of such a trajectory contains no equilibrium, then it must be a periodic orbit. Thus, in the plane, bounded trajectories cannot have arbitrarily complicated long-time behavior away from equilibria.

The reason is fundamentally geometric. A local transversal to a planar flow is one-dimensional, so successive crossings can be ordered. Together with uniqueness of solutions and separation properties of simple closed curves in the plane, this ordering severely restricts the possible recurrence of trajectories. Here, *recurrent* behavior refers informally to motion that returns repeatedly and arbitrarily close to previously visited states, rather than settling to a single equilibrium. These are precisely the geometric features used in the proof of the Poincaré–Bendixson theorem.

This restriction disappears in dimension three and higher. A transversal then has dimension at least two, so there is no corresponding ordering of successive intersections. Bounded trajectories may approach invariant sets that are neither equilibria nor periodic orbits, and the motion on such sets may remain complicated for all future time.

It is important, however, not to confuse this dimensional distinction with the existence of oscillation itself. Planar systems can have periodic orbits, attracting limit cycles, homoclinic and heteroclinic connections, and multiple basins of attraction. What the Poincaré–Bendixson theorem rules out is bounded nonequilibrium long-time behavior more complicated than a periodic orbit, provided the ω-limit set contains no equilibrium. Such more complicated behavior can occur in higher-dimensional autonomous systems.

The next section studies a different question: how a periodic orbit can be created when an equilibrium changes stability. Later in the chapter, the Lorenz system will illustrate how much richer bounded long-time dynamics can become once the planar restriction is removed.

Exercise 11.19 (Stability of a planar periodic orbit)**.** Let Γ be a periodic orbit of period T for the planar system $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y})$. Using the variational equation and Liouville's formula from Section 9.3.3, show that its transverse Floquet multiplier is

$$\rho = \exp\left(\int_0^T \nabla \cdot \mathbf{f}(\mathbf{y}(t))\, dt\right).$$

Deduce a criterion for Γ to be attracting or repelling.

Exercise 11.20 (Why a large ellipse does not give a trapping region)**.** Consider the Van der Pol system

$$\dot{x} = y, \qquad \dot{y} = \mu(1 - x^2)y - x, \qquad \mu > 0.$$

A natural attempt to construct an outer trapping region is to use the level sets of the tilted quadratic function

$$E_\gamma(x, y) = \frac{1}{2}(x^2 + y^2) + \gamma xy, \qquad |\gamma| < 1.$$

Show that no sufficiently large level set of E_γ is an outer trapping boundary. Explain why the Liénard coordinates used above are better adapted to the nonlinear flow.

Hint. Compute $\dot{E}_\gamma$. If $\gamma \neq 0$, examine rays $y = kx$ with $k(k + \gamma) < 0$; if $\gamma = 0$, examine the portions of large circles where $|x| < 1$.

11.8 Hopf Bifurcation and the Birth of Periodic Orbits

The bifurcations studied in Chapter 1 describe changes in the equilibria of scalar equations. In planar systems, a different mechanism can occur: an equilibrium may lose stability and give rise to a periodic orbit. This transition is called a *Hopf bifurcation*.

This phenomenon already appeared in the Brusselator in Chapter 4. There the equilibrium changes stability when a complex-conjugate pair of eigenvalues passes through the imaginary axis. At the critical parameter value, the eigenvalues are purely imaginary, and oscillatory behavior emerges as the parameter crosses the threshold.

Hopf bifurcation provides one of the principal mechanisms by which sustained oscillations arise in mechanical, chemical, and biological systems. Examples include self-excited mechanical vibrations, oscillating chemical reactions, and repetitive electrical activity in neurons.

11.8.1 From an equilibrium to periodic motion

Consider a parameter-dependent planar system

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}; \mu), \qquad \mathbf{y} \in \mathbb{R}^2,$$

and suppose that it has an equilibrium $\mathbf{y}_*(\mu)$. The local behavior near this equilibrium is governed initially by the Jacobian matrix

$$\mathbf{A}(\mu) = D_{\mathbf{y}}\mathbf{f}(\mathbf{y}_*(\mu); \mu).$$

Suppose that its eigenvalues form a complex-conjugate pair

$$\lambda_\pm(\mu) = \alpha(\mu) \pm i\omega(\mu), \qquad \omega(\mu) > 0.$$

When $\alpha(\mu) < 0$, the equilibrium is locally asymptotically stable, and nearby solutions spiral toward it. When $\alpha(\mu) > 0$, the equilibrium is unstable, and nearby solutions spiral away from it. Thus a change of sign in $\alpha(\mu)$ signals a change in stability.

The critical case occurs at a parameter value $\mu = \mu_*$ for which

$$\alpha(\mu_*) = 0, \qquad \omega(\mu_*) \neq 0.$$

At this value, the linearized system has eigenvalues

$$\lambda_{\pm}(\mu_*) = \pm i\omega(\mu_*).$$

Linearization alone is then inconclusive: it detects the loss of stability, but it does not determine whether a periodic orbit appears near the equilibrium. That question depends on the nonlinear terms.

Under suitable nondegeneracy conditions, a family of small periodic orbits is created or destroyed as μ passes through μ_*. This transition is called a *Hopf bifurcation*. The periodic orbit represents a self-sustained oscillation whose amplitude is determined by the nonlinear dynamics rather than by the initial data.

11.8.2 The supercritical Hopf normal form

The simplest system exhibiting a Hopf bifurcation is

$$\begin{aligned} \dot{x} &= \mu x - \omega y - (x^2 + y^2)x, \\ \dot{y} &= \omega x + \mu y - (x^2 + y^2)y, \end{aligned} \qquad \omega > 0,$$

where μ is the bifurcation parameter.

The origin is an equilibrium for every μ. Its Jacobian matrix is

$$\mathbf{A}(\mu) = \begin{bmatrix} \mu & -\omega \\ \omega & \mu \end{bmatrix},$$

with eigenvalues

$$\lambda_{\pm} = \mu \pm i\omega.$$

Thus the origin changes from asymptotically stable to unstable as μ passes through zero.

To determine the nonlinear behavior, introduce polar coordinates

$$x = r\cos\theta, \qquad y = r\sin\theta.$$

Using

$$r\dot{r} = x\dot{x} + y\dot{y}, \qquad r^2\dot{\theta} = x\dot{y} - y\dot{x},$$

we obtain

$$\boxed{\begin{aligned} \dot{r} &= \mu r - r^3, \\ \dot{\theta} &= \omega. \end{aligned}}$$

Hence the angular motion is uniform, while the amplitude satisfies

$$\dot{r} = r(\mu - r^2).$$

For $\mu < 0$, every solution with $r > 0$ approaches the origin. At $\mu = 0$,

$$\dot{r} = -r^3,$$

and therefore

$$r(t) = \frac{r(0)}{\sqrt{1 + 2r(0)^2 t}}.$$

The origin is still asymptotically stable, although the decay is algebraic rather than exponential. This behavior is not detected by the linearized system, which is a center at $\mu = 0$.

For $\mu > 0$, the radial equilibria are

$$r = 0, \qquad r = \sqrt{\mu}.$$

The origin is unstable, while $r = \sqrt{\mu}$ is asymptotically stable. Consequently,

$$x^2 + y^2 = \mu$$

is a stable periodic orbit of period

$$P = \frac{2\pi}{\omega}.$$

Thus a stable periodic orbit emerges continuously as the equilibrium loses stability. This is a *supercritical Hopf bifurcation*. Its amplitude satisfies

$$r = \sqrt{\mu},$$

which gives the characteristic square-root scaling.

The radial equation

$$\dot{r} = \mu r - r^3$$

has the form of the supercritical pitchfork normal form. Here, however, its nonzero equilibrium represents a periodic orbit of the planar system rather than an equilibrium.

11.8.3 A self-excited mechanical oscillator

The Hopf normal form describes the local geometry of the bifurcation, but its variables do not have an immediate mechanical interpretation. We now consider a nonlinear oscillator in which the bifurcation arises from a change in the damping mechanism:

$$\ddot{x} - \mu\dot{x} + x + x^2\dot{x} = 0.$$

Equivalently,

$$\ddot{x} + x + (x^2 - \mu)\dot{x} = 0.$$

The displacement is x, and the coefficient $x^2 - \mu$ acts as a state-dependent damping coefficient.

Introducing the velocity

$$y = \dot{x},$$

gives the planar system

$$\dot{x} = y,$$
$$\dot{y} = -x + \mu y - x^2 y.$$

The origin is an equilibrium for every value of μ. The Jacobian matrix at the origin is

$$\mathbf{A}(\mu) = \begin{bmatrix} 0 & 1 \\ -1 & \mu \end{bmatrix}.$$

Its trace and determinant are

$$T = \mu, \qquad D = 1,$$

and its characteristic equation is

$$\lambda^2 - \mu\lambda + 1 = 0.$$

At $\mu = 0$, the eigenvalues are

$$\lambda_\pm = \pm i.$$

For small $\mu < 0$, the origin is a stable spiral, while for small $\mu > 0$, it is an unstable spiral. Thus a complex-conjugate pair of eigenvalues crosses the imaginary axis as μ passes through zero.

The mechanism can be understood through the mechanical energy

$$E(x, y) = \frac{1}{2}x^2 + \frac{1}{2}y^2.$$

Along a solution,

$$\begin{aligned}\dot{E} &= x\dot{x} + y\dot{y} \\ &= xy + y(-x + \mu y - x^2 y) \\ &= (\mu - x^2)y^2.\end{aligned}$$

When $\mu < 0$, one has

$$\dot{E} \le 0,$$

and the largest invariant subset of $\{\dot{E} = 0\}$ is the origin. LaSalle's principle therefore gives asymptotic stability.

When $\mu > 0$, the energy change is no longer always negative. In the region

$$|x| < \sqrt{\mu},$$

one has

$$\dot{E} > 0$$

whenever $y \neq 0$, so small oscillations gain energy. When

$$|x| > \sqrt{\mu},$$

one has

$$\dot{E} < 0,$$

so sufficiently large oscillations lose energy. The system therefore supplies energy at small amplitudes and dissipates energy at large amplitudes. This balance suggests a stable periodic oscillation; averaging quantifies its amplitude near the bifurcation.

To estimate its amplitude, suppose that $\mu > 0$ is small and that the motion is approximately harmonic:

$$x(t) \approx r\cos(t + \theta_0), \qquad y(t) \approx -r\sin(t + \theta_0),$$

where the amplitude r varies slowly compared with the oscillation itself. Since

$$E \approx \frac{1}{2}r^2,$$

we have

$$\dot{E} \approx r\dot{r}.$$

Averaging the energy equation over one oscillation gives

$$r\dot{r} \approx \mu \langle y^2 \rangle - \langle x^2 y^2 \rangle ,$$

where $\langle \cdot \rangle$ denotes the average over one period.

Using

$$\langle \sin^2 t \rangle = \frac{1}{2}$$

and

$$\langle \cos^2 t \sin^2 t \rangle = \frac{1}{8},$$

we obtain

$$\langle y^2 \rangle = \frac{1}{2} r^2, \qquad \langle x^2 y^2 \rangle = \frac{1}{8} r^4.$$

Therefore

$$r\dot{r} \approx \frac{\mu}{2} r^2 - \frac{1}{8} r^4,$$

and hence

$$\dot{r} \approx \frac{\mu}{2} r - \frac{1}{8} r^3.$$

For $\mu > 0$, the averaged amplitude equation has equilibria

$$r = 0, \qquad r = 2\sqrt{\mu}.$$

The equilibrium $r = 0$ is unstable, while $r = 2\sqrt{\mu}$ is asymptotically stable. Thus a stable periodic orbit appears as μ increases through zero, with approximate amplitude

$$r \approx 2\sqrt{\mu}.$$

Its period is close to

$$2\pi.$$

The oscillator therefore undergoes a supercritical Hopf bifurcation at $\mu = 0$. As in the normal form, the amplitude of the periodic orbit is proportional to $\sqrt{\mu}$. Here the bifurcation has a direct mechanical interpretation: negative damping amplifies small motions, while nonlinear damping limits their growth and selects a stable oscillation.

11.8.4 A chemical oscillator: the Brusselator

Reconsider the Brusselator from Section 4.2.4,

$$\begin{aligned} \dot{x} &= A - (B+1)x + x^2 y, \\ \dot{y} &= Bx - x^2 y, \end{aligned} \qquad A > 0, \quad B > 0,$$

where x and y represent concentrations and A and B are fixed input parameters.

As found there, the unique positive equilibrium is

$$(x_*, y_*) = \left(A, \frac{B}{A} \right).$$

At the equilibrium,

$$\mathbf{J}_* = \begin{bmatrix} B-1 & A^2 \\ -B & -A^2 \end{bmatrix}.$$

Its eigenvalues are

$$\lambda_\pm = \frac{T \pm \sqrt{T^2 - 4D}}{2}$$

where T and D are the trace and determinant:

$$T = B - 1 - A^2, \qquad D = A^2.$$

Because $D > 0$, the stability of the equilibrium changes when the trace passes through zero. The critical value is

$$B_c = 1 + A^2.$$

At $B = B_c$, the Jacobian has trace zero and determinant A^2, so its eigenvalues are

$$\lambda_\pm = \pm iA.$$

Moreover,

$$\frac{dT}{dB} = 1 \neq 0,$$

so the eigenvalues cross the imaginary axis as B passes through B_c.

A nonlinear normal-form calculation shows that the bifurcation is supercritical: as B increases through $B = 1 + A^2$, a stable periodic orbit emerges from the equilibrium. The linear frequency at onset is A, so its limiting period is $2\pi/A$. Thus the phase portraits of Section 4.2.4 fit the general Hopf mechanism: the parameter B drives the transition from a stable steady concentration to a sustained chemical oscillation.

11.8.5 The subcritical Hopf normal form

A Hopf bifurcation need not create a stable periodic orbit. Consider

$$\begin{aligned} \dot{x} &= \mu x - \omega y + (x^2 + y^2)x, \\ \dot{y} &= \omega x + \mu y + (x^2 + y^2)y, \end{aligned} \qquad \omega > 0.$$

The linearization at the origin has eigenvalues

$$\lambda_\pm = \mu \pm i\omega,$$

while in polar coordinates the system becomes

$$\begin{aligned} \dot{r} &= \mu r + r^3, \\ \dot{\theta} &= \omega. \end{aligned}$$

For $\mu < 0$, the radial equilibria are

$$r = 0, \qquad r = \sqrt{-\mu}.$$

The origin is asymptotically stable, whereas

$$r = \sqrt{-\mu}$$

is unstable. Thus the circle

$$x^2 + y^2 = -\mu$$

is an unstable periodic orbit.

As μ increases toward zero, this periodic orbit shrinks into the origin. At $\mu = 0$,

$$\dot{r} = r^3,$$

so the origin is unstable. For $\mu > 0$, all solutions with $r > 0$ move outward.

This is a *subcritical Hopf bifurcation*: an unstable periodic orbit exists on the stable side of the bifurcation and disappears when it reaches the equilibrium. In contrast, a supercritical Hopf bifurcation creates a stable periodic orbit on the unstable side.

Subcritical Hopf bifurcations may produce abrupt transitions. Once the equilibrium loses stability, nearby solutions need not approach a small periodic orbit and may instead move toward a distant attractor.

11.8.6 The Hopf bifurcation theorem

Consider a smooth one-parameter family of planar systems

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}; \mu),$$

with an equilibrium $\mathbf{y} = \mathbf{y}_*(\mu)$. Suppose that the eigenvalues of the Jacobian matrix at the equilibrium are

$$\lambda_\pm(\mu) = \alpha(\mu) \pm i\omega(\mu).$$

A Hopf bifurcation may occur at $\mu = \mu_c$ when

$$\alpha(\mu_c) = 0, \qquad \omega(\mu_c) \neq 0,$$

and the eigenvalues cross the imaginary axis:

$$\alpha'(\mu_c) \neq 0.$$

Theorem 11.47 (Hopf bifurcation theorem)**.** *Suppose that* $\mathbf{f}$ *is sufficiently smooth and that* $\mathbf{y}_*(\mu_c)$ *has eigenvalues*

$$\lambda_\pm(\mu_c) = \pm i\omega_0, \qquad \omega_0 > 0.$$

Assume that

$$\frac{d}{d\mu} \Re\lambda_+(\mu)\bigg|_{\mu=\mu_c} \neq 0,$$

and that the first nonlinear coefficient in the Hopf normal form is nonzero. Then a family of small periodic orbits exists near $\mathbf{y}_*(\mu_c)$ *for parameter values on one side of* μ_c*. Their amplitudes are proportional to* $\sqrt{|\mu - \mu_c|}$*, and their periods approach* $\frac{2\pi}{\omega_0}$ *as* $\mu \to \mu_c$*.*

Proof sketch. Near the critical equilibrium, first translate the equilibrium to the origin. Since the linearization at $\mu = \mu_c$ has eigenvalues $\pm i\omega_0$, an invertible real linear change of coordinates puts the linear part into the form

$$\dot{x} = -\omega_0 y, \qquad \dot{y} = \omega_0 x.$$

Introducing the complex coordinate

$$z = x + iy,$$

this becomes

$$\dot{z} = i\omega_0 z.$$

The nonlinear terms can be expanded in monomials $z^p \bar{z}^q$. By a sequence of smooth near-identity changes of variables, all nonresonant terms can be removed to the desired order. For the Hopf eigenvalues $\pm i\omega_0$, a monomial $z^p \bar{z}^q$ in the equation for z is resonant only when

$$(p-q)i\omega_0 = i\omega_0,$$

or equivalently,

$$p - q = 1.$$

There are no quadratic resonant terms, while at cubic order the first resonant monomial is

$$z^2 \bar{z} = z|z|^2.$$

Thus, after reduction to normal form, the equation has the structure

$$\dot{z} = \big(\alpha(\mu) + i\omega(\mu)\big)z + c\, z|z|^2 + \text{higher-order terms},$$

where

$$\alpha(\mu_c) = 0, \qquad \alpha'(\mu_c) \neq 0.$$

Writing

$$z = re^{i\theta}$$

gives

$$\dot{r} = \alpha(\mu)r + \ell r^3 + O\big(r^5 + |\mu - \mu_c| r^3\big),$$
$$\dot{\theta} = \omega_0 + O\big(r^2 + |\mu - \mu_c|\big).$$

The real coefficient $\ell = \operatorname{Re} c$ is, up to the chosen normalization, the first Lyapunov coefficient. The nondegeneracy assumption in the theorem is precisely

$$\ell \neq 0.$$

The periodic orbits correspond to nonzero equilibria of the radial equation. To leading order they satisfy

$$\alpha(\mu) + \ell r^2 = 0,$$

and hence

$$r^2 \sim -\frac{\alpha(\mu)}{\ell}.$$

Since $\alpha'(\mu_c) \neq 0$, the amplitude is therefore of order

$$r \sim C|\mu - \mu_c|^{1/2}.$$

If $\ell < 0$, the bifurcation is supercritical and the emerging periodic orbit is stable. If $\ell > 0$, it is subcritical and the periodic orbit is unstable. Moreover, since

$$\dot{\theta} \to \omega_0$$

as $(r, \mu) \to (0, \mu_c)$, the period tends to

$$\frac{2\pi}{\omega_0}.$$

□

The theorem is local: it describes only the periodic orbits near the equilibrium and for parameter values near μ_c.

11.8.7 Recognizing a Hopf bifurcation in a planar system

Let

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}; \mu)$$

have an equilibrium $\mathbf{y} = \mathbf{y}_*(\mu)$, and let

$$\mathbf{A}(\mu) = D_{\mathbf{y}}\mathbf{f}(\mathbf{y}_*(\mu); \mu).$$

Write

$$T(\mu) = \operatorname{tr} \mathbf{A}(\mu), \qquad D(\mu) = \det \mathbf{A}(\mu).$$

The eigenvalues satisfy

$$\lambda^2 - T(\mu)\lambda + D(\mu) = 0.$$

At a possible Hopf bifurcation point $\mu = \mu_c$, the eigenvalues must be purely imaginary. For a planar system, this is equivalent to

$$T(\mu_c) = 0, \qquad D(\mu_c) > 0.$$

The eigenvalues are then

$$\lambda_{\pm}(\mu_c) = \pm i\sqrt{D(\mu_c)},$$

so the linear frequency is

$$\omega_0 = \sqrt{D(\mu_c)}$$

and the limiting period is

$$\frac{2\pi}{\sqrt{D(\mu_c)}}.$$

Since

$$\Re\lambda_{\pm}(\mu) = \frac{T(\mu)}{2}$$

when the eigenvalues are complex, the crossing condition is

$$T'(\mu_c) \neq 0.$$

Thus the linear test is

$$\boxed{T(\mu_c) = 0, \qquad D(\mu_c) > 0, \qquad T'(\mu_c) \neq 0.}$$

These conditions identify a possible Hopf bifurcation. The nonlinear terms must still be examined to verify the creation of periodic orbits and to determine whether the bifurcation is supercritical or subcritical.

Practical procedure. To analyze a possible Hopf bifurcation:

1. Find the equilibrium $\mathbf{y}_*(\mu)$.
2. Compute the Jacobian matrix $\mathbf{A}(\mu)$ at the equilibrium.
3. Find the parameter value μ_c for which
$$T(\mu_c) = 0.$$
4. Check that
$$D(\mu_c) > 0 \qquad \text{and} \qquad T'(\mu_c) \neq 0.$$
5. Use the nonlinear terms, an amplitude equation, or numerical evidence to determine the existence and stability of the periodic orbits.

Exercise 11.21 (Amplitude and nonlinear frequency shift)**.** Consider
$$\begin{aligned} \dot{x} &= \mu x - \omega y - a(x^2+y^2)x + b(x^2+y^2)y, \\ \dot{y} &= \omega x + \mu y - b(x^2+y^2)x - a(x^2+y^2)y, \end{aligned} \qquad a > 0, \quad \omega > 0.$$
Derive the polar equations. For sufficiently small $\mu > 0$, find the amplitude and frequency of the attracting periodic orbit. What effect does b have on the bifurcation diagram and on the period?

Exercise 11.22 (A saturated subcritical Hopf bifurcation)**.** Consider the polar system
$$\dot{r} = r(\mu + r^2 - r^4), \qquad \dot{\theta} = 1.$$
Find all periodic orbits and determine their stability as μ varies. Locate the saddle-node bifurcation of periodic orbits and identify the parameter interval in which a stable equilibrium and a stable periodic orbit coexist. Describe the resulting hysteresis.

Exercise 11.23 (A degenerate Hopf point)**.** Consider
$$\begin{aligned} \dot{x} &= \mu x - y - (x^2+y^2)^2 x, \\ \dot{y} &= x + \mu y - (x^2+y^2)^2 y. \end{aligned}$$
Show that the linear crossing conditions hold at $\mu = 0$, but the cubic coefficient in the Hopf normal form vanishes. Determine the periodic orbit for $\mu > 0$ and its amplitude scaling. Explain why this does not contradict the Hopf theorem stated above.

Exercise 11.24 (The Rayleigh oscillator near onset)**.** For $0 < \varepsilon \ll 1$, consider

$$\ddot{x} + x = \varepsilon(\mu - \dot{x}^2)\dot{x}.$$

Using $E = (x^2 + \dot{x}^2)/2$ and the harmonic approximation $x = r\cos t$, derive the averaged amplitude equation. Determine the leading amplitude and stability of the periodic oscillation for $\mu > 0$.

11.9 Higher-Dimensional Attracting Dynamics

The Poincaré–Bendixson theorem gives a strong classification of bounded long-time behavior in the plane. In dimensions three and higher, no corresponding classification holds. A bounded trajectory may approach an invariant set that is neither an equilibrium nor a periodic orbit, and the motion within that set may remain complicated for all future time.

The Lorenz system provides a classical illustration:

$$\dot{x} = \sigma(y - x),$$

$$\dot{y} = rx - y - xz,$$

$$\dot{z} = xy - bz,$$

where

$$\sigma > 0, \qquad r > 0, \qquad b > 0.$$

The system arose as a simplified model of thermal convection. It is only three-dimensional and involves quadratic nonlinearities, yet it exhibits long-time behavior that is impossible for an autonomous planar system.

Our purpose is not to develop a complete theory of chaotic dynamics. Instead, we use the Lorenz system to illustrate four features of higher-dimensional dissipative systems:

1. equilibria may lose stability as a parameter varies,
2. phase-space volumes may contract,
3. all trajectories may eventually enter a bounded region,
4. the resulting attractor need not be an equilibrium or a periodic orbit.

11.9.1 The Lorenz system and its equilibria

Symmetry and equilibria. The Lorenz system is invariant under the transformation

$$(x, y, z) \longmapsto (-x, -y, z).$$

Thus any equilibrium away from the z-axis occurs together with its symmetric counterpart.

At an equilibrium,

$$\sigma(y - x) = 0,$$

so that

$$y = x.$$

The remaining equations become

$$x(r-1-z)=0, \qquad x^2-bz=0.$$

The origin

$$\mathbf{0}=(0,0,0)$$

is therefore an equilibrium for every value of r. When $r>1$, there are also two nonzero equilibria

$$\mathbf{y}_{\pm}=\left(\pm\sqrt{b(r-1)},\pm\sqrt{b(r-1)},r-1\right).$$

These equilibria are exchanged by the symmetry of the system.

The Jacobian matrix of the Lorenz system is

$$D\mathbf{f}(x,y,z)=\begin{bmatrix}-\sigma & \sigma & 0\\ r-z & -1 & -x\\ y & x & -b\end{bmatrix}.$$

Stability of the origin. At the origin,

$$D\mathbf{f}(\mathbf{0})=\begin{bmatrix}-\sigma & \sigma & 0\\ r & -1 & 0\\ 0 & 0 & -b\end{bmatrix}.$$

One eigenvalue is $-b$, while the other two satisfy

$$\lambda^2+(\sigma+1)\lambda+\sigma(1-r)=0.$$

Hence the origin is asymptotically stable when

$$0<r<1.$$

At $r=1$, one eigenvalue passes through zero, and for $r>1$ the origin has one positive eigenvalue. It is then unstable.

Stability of the nonzero equilibria. The two nonzero equilibria appear precisely when the origin loses stability. This is consistent with the symmetry

$$(x,y,z)\longmapsto(-x,-y,z):$$

the equilibria $\mathbf{y}_+$ and $\mathbf{y}_-$ are created as a symmetric pair.

To study their stability, set

$$a=\sqrt{b(r-1)}.$$

At $\mathbf{y}_+$, the Jacobian is

$$D\mathbf{f}(\mathbf{y}_+)=\begin{bmatrix}-\sigma & \sigma & 0\\ 1 & -1 & -a\\ a & a & -b\end{bmatrix}.$$

The Jacobian at $\mathbf{y}_-$ has the same characteristic polynomial. A direct calculation gives

$$\lambda^3+(\sigma+b+1)\lambda^2+b(\sigma+r)\lambda+2\sigma b(r-1)=0.$$

The coefficients are positive for $r>1$.

Lemma 11.48 (Routh–Hurwitz criterion for a cubic)**.** *For*

$$p(\lambda) = \lambda^3 + a_1\lambda^2 + a_2\lambda + a_3,$$

all roots satisfy $\Re\lambda < 0$ *if and only if*

$$a_1 > 0, \qquad a_2 > 0, \qquad a_3 > 0, \qquad a_1 a_2 > a_3.$$

By the Routh–Hurwitz criterion, the equilibria $\mathbf{y}_\pm$ are asymptotically stable provided that

$$(\sigma + b + 1)b(\sigma + r) > 2\sigma b(r - 1).$$

After simplification, this becomes

$$r < r_H := \frac{\sigma(\sigma + b + 3)}{\sigma - b - 1},$$

assuming

$$\sigma > b + 1.$$

Thus the nonzero equilibria are stable for

$$1 < r < r_H$$

and lose stability when r passes through r_H.

The classical parameters. For the classical parameter values

$$\sigma = 10, \qquad b = \frac{8}{3},$$

the critical value is

$$r_H = \frac{\sigma(\sigma + b + 3)}{\sigma - b - 1} \approx 24.74.$$

The commonly studied value

$$r = 28$$

lies beyond this threshold, so all three equilibria are unstable.

This calculation does not by itself determine the long-time behavior. Instability of the equilibria only shows that nearby trajectories do not remain close to them. To understand what happens instead, we next establish two global properties of the Lorenz flow: phase-space volumes contract, and all trajectories eventually enter a bounded region.

11.9.2 Volume contraction and boundedness

Volume contraction. The instability of all equilibria at the classical parameter values does not explain where trajectories go. Before discussing the attractor, we establish two global properties: phase-space volumes contract, and every trajectory is eventually confined to a bounded region.

The divergence of the Lorenz vector field is

$$\begin{aligned}\nabla \cdot \mathbf{f} &= \frac{\partial}{\partial x}\big(\sigma(y - x)\big) + \frac{\partial}{\partial y}(rx - y - xz) + \frac{\partial}{\partial z}(xy - bz) \\ &= -\sigma - 1 - b < 0.\end{aligned}$$

Let Ω_0 be a small region of initial conditions, and let $\Omega_t = \varphi_t(\Omega_0)$ be its image under the flow. Liouville's formula gives

$$\frac{d}{dt}\operatorname{Vol}(\Omega_t) = \int_{\Omega_t} \nabla \cdot \mathbf{f}\, d\mathbf{y}.$$

Since the divergence is constant,

$$\operatorname{Vol}(\Omega_t) = e^{-(\sigma+1+b)t}\operatorname{Vol}(\Omega_0).$$

Thus every three-dimensional volume contracts exponentially under the Lorenz flow.

This contraction is a form of dissipation. It does not, however, imply that the distance between every pair of trajectories decreases. The flow may contract strongly in some directions while stretching in another. Nor does negative divergence alone prove that individual trajectories remain bounded.

Boundedness of solutions. To obtain boundedness, define

$$\Phi(x,y,z) = \frac{r}{\sigma}x^2 + y^2 + (z-2r)^2.$$

Along a solution,

$$\begin{aligned}\dot\Phi &= 2\frac{r}{\sigma}x\dot x + 2y\dot y + 2(z-2r)\dot z \\ &= 2rx(y-x) + 2y(rx - y - xz) + 2(z-2r)(xy - bz).\end{aligned}$$

The mixed terms cancel, leaving

$$\dot\Phi = -2rx^2 - 2y^2 - 2bz(z-2r).$$

Since

$$z(z-2r) = (z-r)^2 - r^2,$$

we obtain

$$\dot\Phi = -2rx^2 - 2y^2 - 2b(z-r)^2 + 2br^2.$$

Consequently, $\dot\Phi < 0$ whenever

$$rx^2 + y^2 + b(z-r)^2 > br^2.$$

In particular, $\dot\Phi$ is negative outside a sufficiently large bounded region.

Equivalently, there exist constants $c > 0$ and $C > 0$ such that

$$\dot\Phi \le -c\Phi + C.$$

Multiplying by e^{ct} and integrating gives

$$\Phi(t) \le e^{-ct}\Phi(0) + \frac{C}{c}\bigl(1 - e^{-ct}\bigr).$$

Therefore

$$\limsup_{t\to\infty} \Phi(t) \le \frac{C}{c}.$$

Every trajectory eventually enters a sufficiently large ellipsoid

$$\{\Phi \le R\}$$

and remains there.

A bounded set that eventually captures every bounded collection of initial conditions is called an *absorbing region*. The Lorenz system therefore has a bounded absorbing region.

The two conclusions obtained above play different roles:

$$\begin{aligned} &\text{negative divergence} &&\Longrightarrow \text{ contraction of phase-space volume,}\\ &\text{absorbing region} &&\Longrightarrow \text{ boundedness of long-time motion.} \end{aligned}$$

Together they show that the Lorenz system is dissipative. All its long-time dynamics are confined to a bounded region whose volumes are continually contracted by the flow.

11.9.3 The Lorenz attractor

For the classical values

$$\sigma = 10, \qquad b = \frac{8}{3}, \qquad r = 28,$$

the eigenvalues at the origin are

$$\lambda_1 = \frac{-11 + \sqrt{1201}}{2} \approx 11.828,$$

$$\lambda_2 = \frac{-11 - \sqrt{1201}}{2} \approx -22.828, \qquad \lambda_3 = -\frac{8}{3} \approx -2.667.$$

Thus the origin is a saddle with a one-dimensional unstable direction and a two-dimensional stable manifold.

At each of the symmetric equilibria $\mathbf{y}_\pm$, the eigenvalues are approximately

$$\lambda_1 = -13.855, \qquad \lambda_{2,3} = 0.094 \pm 10.195i.$$

Thus each nonzero equilibrium is a saddle-focus with a one-dimensional stable direction and a two-dimensional unstable spiral.

A numerical trajectory is shown in Figure 11.4. After an initial transient, the trajectory remains in a bounded region and moves repeatedly around two lobes. It spends some time rotating near one of the nonzero equilibria, then crosses to the other lobe, and continues this switching in an irregular manner.

Rotation and switching between the two lobes. The two lobes of the attractor are organized around the nonzero equilibria $\mathbf{y}_+$ and $\mathbf{y}_-$. For the classical parameter values, each of these equilibria is a saddle-focus. Its eigenvalues consist of one negative real eigenvalue and a complex conjugate pair with positive real part. Thus a trajectory may approach the equilibrium in one direction, but it spirals away in a two-dimensional unstable direction.

Consequently, a trajectory typically makes several turns around one equilibrium while gradually moving away from it. It is then carried through the central region of phase space. From there, it

may return to the same lobe or cross to the opposite lobe and begin spiraling around the other equilibrium. The symmetry

$$(x, y, z) \longmapsto (-x, -y, z)$$

accounts for the two corresponding wings of the attractor.

The number of turns made around one lobe before switching is not fixed. A trajectory may circle one equilibrium several times, cross to the other lobe, remain there for a different number of turns, and then switch again. This produces the irregular sequence of left and right excursions visible in numerical solutions.

Sensitivity near the central region. The switching is particularly sensitive when a trajectory passes near the central part of the attractor. The stable manifold of the origin separates points that are carried toward different sides of the flow. Two nearby trajectories may initially follow almost the same path, but a very small difference in their positions relative to this separating surface may send one trajectory toward the right lobe and the other toward the left lobe.

Once the two trajectories enter different lobes, their separation becomes macroscopic. Even when they enter the same lobe, the unstable spiral near the corresponding saddle-focus enlarges small differences in their phase. Repeated passages through this mechanism lead to sensitive dependence on initial data.

Thus a small perturbation does not merely change the trajectory by a small amount for all later times. It may eventually change the entire sequence of lobe visits:

$$+, +, -, +, -, -, \dots$$

for one trajectory and

$$+, +, +, -, +, -, \dots$$

for another initially nearby trajectory.

Contraction, stretching, and folding. The negative divergence

$$\nabla \cdot \mathbf{f} = -\sigma - 1 - b < 0$$

implies that every small three-dimensional volume contracts exponentially. The flow therefore compresses clouds of initial conditions strongly in some directions. At the same time, the unstable directions near the saddle-foci stretch nearby trajectories.

The global flow then bends and folds the stretched set back into the bounded absorbing region. Repetition of

$$\text{stretching} \longrightarrow \text{contraction} \longrightarrow \text{folding}$$

produces the two-winged geometry of the Lorenz attractor.

The attractor is therefore thin in phase space and has zero three-dimensional volume. It should not, however, be regarded as a smooth two-dimensional manifold. Its folded layers have a fractal structure, and the familiar butterfly-shaped picture is a projection of this complicated invariant set.

The picture should not be interpreted as a single periodic orbit. A periodic trajectory would eventually trace the same closed curve repeatedly. The Lorenz trajectory instead continues to

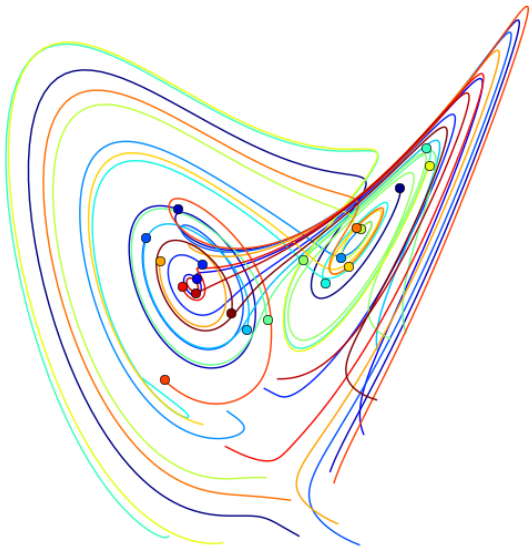


Figure 11.4: A trajectory of the Lorenz system for $\sigma = 10$, $r = 28$, and $b = 8/3$. After a transient, the trajectory remains near a bounded two-lobed attracting set.

produce new turns and switches between the two lobes without settling onto an evident closed orbit.

The bounded invariant set suggested by this computation is called the *Lorenz attractor*. It is an example of a *strange attractor*: an attracting set with complicated internal dynamics and a geometry that is not that of a smooth curve or surface.

There are two different statements here:

$$\begin{aligned} &\text{attraction:} && \operatorname{dist}\big(\varphi_t(\mathbf{y}_0), A\big) \to 0, \\ &\text{convergence to a trajectory:} && \varphi_t(\mathbf{y}_0) \to \varphi_t(\mathbf{z}_0). \end{aligned}$$

The first may hold while the second fails. Nearby trajectories can both remain close to the same attractor while separating rapidly from each other within that attractor. Thus attraction to a set does not mean that the motion on the set is stable.

The attractor may also contain unstable invariant objects. In particular, the equilibria may lie in the larger invariant structure even though nearby generic trajectories do not converge to them. Likewise, unstable periodic orbits may be embedded in the attractor without attracting a neighborhood. The attractor describes the organization of the entire limiting set, not the stability of each orbit contained in it.

Remark 11.49. A numerical plot strongly suggests the geometry and dynamics of the Lorenz attractor, but it is not by itself a proof of chaos. A rigorous analysis requires substantially deeper tools. Here the Lorenz system serves to show what becomes possible once the planar restrictions of the Poincaré–Bendixson theorem no longer apply.

11.9.4 Beyond the planar theory

The Lorenz system shows that bounded dissipative dynamics in three dimensions can be fundamentally different from bounded planar dynamics. The Poincaré–Bendixson theorem no longer restricts an ω-limit set to an equilibrium, a periodic orbit, or a collection of connecting orbits.

For the classical Lorenz parameters, trajectories remain bounded and phase-space volumes contract, yet typical trajectories do not settle onto an equilibrium or a periodic orbit. Instead, they approach a more complicated invariant set and continue to move irregularly within it.

The contrast may be summarized as follows:

Planar systems	Systems in dimension three or higher
Poincaré–Bendixson restricts bounded recurrent motion	No analogous general classification of bounded recurrent motion
Periodic orbits are the principal nonequilibrium recurrent sets	Attractors may have complicated internal dynamics
Bounded recurrent dynamics is strongly restricted	Complicated recurrent dynamics may occur

The term *chaos* is used for dynamics that combine deterministic evolution with persistent unpredictability. Although there are several precise mathematical definitions, three features commonly appear:

1. sensitive dependence on initial data,
2. complicated recurrent motion,
3. the presence of many unstable periodic orbits.

The Lorenz attractor is the classical model exhibiting these features.

It is important to distinguish unpredictability from randomness. The Lorenz equations are autonomous and deterministic: a given initial value determines a unique trajectory. The practical loss of predictability comes from the amplification of small errors in the initial data, not from a random forcing term.

The Lorenz example also clarifies the meaning of dissipation. Dissipation does not necessarily force convergence to a simple state. It may instead confine the motion to a bounded attracting set while allowing complicated dynamics to persist on that set. Thus

$$\text{dissipation} \not\Longrightarrow \text{simple asymptotic behavior.}$$

Chaos in mechanical and electrical systems. Chaotic behavior is not restricted to the Lorenz equations. A classical mechanical example is the Hénon–Heiles system

$$\ddot{q}_1 = -q_1 - 2q_1q_2, \qquad \ddot{q}_2 = -q_2 - q_1^2 + q_2^2,$$

with conserved energy

$$H = \frac{1}{2}\big(\dot{q}_1^2 + \dot{q}_2^2\big) + \frac{1}{2}\big(q_1^2 + q_2^2\big) + q_1^2q_2 - \frac{1}{3}q_2^3.$$

This is a two-degree-of-freedom Hamiltonian system, or equivalently a four-dimensional autonomous system. Its trajectories may display chaotic motion on an energy surface, although the system cannot possess a dissipative attractor with an open basin, because phase-space volume is preserved.

An important electrical example is Chua's circuit, a nonlinear circuit containing capacitors, an inductor, resistive elements, and an active nonlinear element. Its three-dimensional autonomous model can possess the well-known double-scroll attractor. Unlike the Hénon–Heiles system, Chua's circuit is dissipative in suitable parameter regimes: phase-space volumes contract, while the active element continually supplies energy. These examples distinguish chaotic trajectories in conservative systems from strange attractors in dissipative systems.

A rigorous theory of chaotic attractors requires tools beyond the scope of this book.

Exercise 11.25 (Invariant volume and the Lorenz flow)**.** Let A be a bounded measurable invariant set for the Lorenz flow. Use Liouville's formula to prove that

$$\mathrm{Vol}(A) = 0.$$

Explain why this conclusion neither implies that A is a smooth surface nor that the distance between every pair of trajectories decreases with time.

Chapter Summary

This chapter developed a framework for long-time behavior in autonomous differential equations. A flow organizes solutions, invariant sets describe subsets preserved by the dynamics, and ω-limit sets record the states approached by a trajectory as $t \to \infty$.

An attractor describes the limiting behavior of a family of trajectories, and its basin consists of all initial values whose solutions approach it. Dissipative systems often possess monotone quantities whose sublevel sets confine the motion.

Lyapunov functions convert such monotonicity into stability information. When $\dot{\Phi} < 0$ away from an equilibrium, asymptotic stability follows. When only $\dot{\Phi} \leq 0$ is available, LaSalle's invariance principle shows that trajectories approach the largest invariant subset of

$$\{\mathbf{y} : \dot{\Phi}(\mathbf{y}) = 0\}.$$

For planar systems, the Poincaré–Bendixson theorem strongly restricts bounded recurrent motion and identifies periodic orbits as the principal nonequilibrium limiting sets. Hopf bifurcation explains how such periodic orbits may be created when an equilibrium loses stability as a parameter passes through a critical value. In dimension three and higher, the planar restriction fails. The Lorenz system illustrates how a dissipative system may have bounded trajectories and contracting phase-space volume while supporting a complicated attractor.

The central progression is

$$\begin{aligned}\text{flow} &\longrightarrow \text{invariant sets} \longrightarrow \omega\text{-limit sets} \longrightarrow \text{attractors} \longrightarrow \text{Lyapunov and LaSalle methods} \\ &\longrightarrow \text{periodic attractors and Hopf bifurcation} \longrightarrow \text{higher-dimensional attracting dynamics.}\end{aligned}$$

Homework Problems

Homework 11.1 (Global convergence in a logistic predator–prey model)**.** Consider

$$\dot{x} = x\left[r\left(1 - \frac{x}{K}\right) - ay\right], \qquad \dot{y} = y(-d + bx),$$

where all parameters are positive and $d < bK$. Let (x_*, y_*) be the positive equilibrium and define

$$\Phi(z; z_*) = z - z_* - z_* \log \frac{z}{z_*}.$$

Find positive constants A and B such that

$$V(x, y) = A\Phi(x; x_*) + B\Phi(y; y_*)$$

satisfies

$$\dot{V} = -C(x - x_*)^2$$

for some $C > 0$. Show that positive solutions remain in compact subsets of the positive quadrant, and use LaSalle's invariance principle to prove that every positive solution converges to (x_*, y_*). Deduce that the positive quadrant contains no nonconstant periodic orbit.

Homework 11.2 (A two-well gradient landscape)**.** Let $0 < \alpha < 1$ and consider the gradient system

$$\dot{\mathbf{y}} = -\nabla V(\mathbf{y}), \qquad V(x, y) = \frac{1}{4}(x^2 + y^2 - 1)^2 + \frac{\alpha}{2}y^2.$$

Find and classify all equilibria. Prove that every trajectory is bounded and converges to an equilibrium. Determine the basins of the two minima and show that their common boundary is the stable set of the remaining equilibria. Describe the equilibria and connecting orbits contained in the global attractor.

Homework 11.3 (Selective damping in an oscillator chain)**.** Let

$$\ddot{\mathbf{q}} + \gamma \mathbf{e}_j \mathbf{e}_j^T \dot{\mathbf{q}} + \mathbf{K}\mathbf{q} = \mathbf{0}, \qquad \mathbf{K} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}, \qquad \gamma > 0.$$

Use the mechanical energy and LaSalle's principle to compare damping the middle mass ($j = 2$) with damping an end mass ($j = 1$). Identify every undamped normal mode and determine the long-time behavior in each case. Explain how the placement of a single damper controls whether all solutions converge to equilibrium.

Homework 11.4 (Amplitude scaling in the Brusselator)**.** Take $A = 1$ and $B = 2 + \delta$ in the Brusselator. For

$$\delta = 0.01,\ 0.02,\ 0.04,\ 0.08,$$

integrate on $0 \le t \le 2000$ from

$$(x(0), y(0)) = (1.1, B).$$

Use an adaptive solver with relative and absolute tolerances 10^{-9} and maximum step 0.05. From the interval $1000 \le t \le 2000$, estimate the half peak-to-peak amplitude of x and the period. Plot the amplitude against $\sqrt{\delta}$, fit its leading coefficient, and determine whether the computed scaling agrees with the Hopf theorem. Report the periods and one representative phase portrait.

Projects

The computational projects require reproducible numerical experiments. State the method, time interval, step size or error tolerances, initial data, and error measure, and provide the requested plots and tables.

Project 11.1 (Regular and chaotic motion in the Hénon–Heiles system)**.** For the Hénon–Heiles Hamiltonian

$$H = \frac{1}{2}(p_1^2 + p_2^2) + \frac{1}{2}(q_1^2 + q_2^2) + q_1^2 q_2 - \frac{1}{3} q_2^3,$$

construct Poincaré sections on $q_2 = 0$ with $p_2 > 0$ at the energy levels $H = 1/12$ and $H = 1/8$. Use initial data

$$q_2(0) = p_1(0) = 0, \qquad q_1(0) \in \{0.05, 0.15, 0.25\},$$

with $p_2(0) > 0$ determined by the energy. Integrate on $0 \le t \le 5000$ by leapfrog with

$$h = 2^{-6},\ 2^{-7},\ 2^{-8}.$$

Plot the sections in the (q_1, p_1)-plane, tabulate the maximum energy error, and distinguish regular invariant curves from apparently chaotic regions. Explain why energy conservation alone does not determine whether a computed trajectory is reliable.

Project 11.2 (Sensitivity and reproducibility in the Lorenz system)**.** For

$$\sigma = 10, \qquad b = \frac{8}{3}, \qquad r = 28,$$

compare the solutions starting from

$$\mathbf{y}_0 = (1, 1, 1)^T, \qquad \widetilde{\mathbf{y}}_0 = (1 + 10^{-8}, 1, 1)^T$$

on $0 \le t \le 100$. Use the same adaptive Runge–Kutta solver with relative and absolute tolerances 10^{-7}, 10^{-9}, and 10^{-11}. Plot the separation of the two solutions on a semilogarithmic scale and estimate its initial exponential growth rate. Then compare the time averages of x^2, z, and xyz over $50 \le t \le 100$. Discuss why individual trajectories lose reproducibility while statistical quantities may be more robust, and separate sensitive dependence from numerical error.

Historical Notes and Further Reading

Historical notes. The qualitative theory of differential equations initiated by Poincaré (1881) and developed further by Bendixson (1901) led to the planar result now called the Poincaré–Bendixson theorem. It shows that bounded recurrent motion in the plane is much more restricted than in higher dimensions.

Lyapunov (1892) developed the direct method of stability in his study of the stability of motion. Instead of solving the equations explicitly, one seeks a scalar function that is positive near an equilibrium and nonincreasing along trajectories. An important extension was given by LaSalle (1960). When the derivative is only nonpositive, the long-time behavior is determined, under suitable compactness and invariance assumptions, by the largest invariant subset of the zero-dissipation set. This extension is especially natural for mechanical systems with partial damping and for circuit and control models.

The local bifurcation of periodic motion from an equilibrium was placed in a general form by Hopf (1942). Under suitable nondegeneracy conditions, a conjugate pair of eigenvalues crossing the imaginary axis gives rise to a branch of periodic solutions. This mechanism underlies the onset of self-sustained oscillation in the mechanical and chemical examples of this chapter.

Several models discussed in this chapter arose from attempts to describe collective behavior. The phase-oscillator model of Kuramoto (1975) became a fundamental model of synchronization. The flocking model of Cucker and Smale (2007) provided a mathematical framework for the emergence of velocity alignment. A simple Lyapunov approach to flocking and the mean-field limit was developed by Ha and Liu (2009), while Ha and Tadmor (2008) connected the particle model with kinetic and hydrodynamic descriptions of flocking. The inertial spin model of Cavagna et al. (2015) incorporated turning dynamics and the propagation of orientation changes through a flock. In evolutionary dynamics, the replicator equation was introduced by Taylor and Jonker (1978), and its gradient structure on the probability simplex was developed geometrically by Shahshahani (1979) and Akin (1979).

A decisive example of higher-dimensional dynamics was given by Lorenz (1963). Starting from a simplified model of atmospheric convection, Lorenz obtained a three-dimensional system whose bounded deterministic trajectories may remain nonperiodic and exhibit strong sensitivity to initial data. The Lorenz equations show that dissipation and boundedness do not necessarily imply convergence to an equilibrium or a periodic orbit.

Further reading. For an accessible introduction to attractors, limit cycles, Hopf bifurcation, and the Lorenz equations, see Strogatz (2024). A systematic mathematical treatment of flows, planar dynamics, periodic orbits, and bifurcations is given by Hale and Koçak (1991). Lyapunov functions and LaSalle's invariance principle, with emphasis on engineering and control applications, are developed in Khalil (2002). For the physical derivation of the Lorenz equations and their role in climate dynamics, see Kaper and Engler (2013).

For synchronization and coupled phase oscillators, see Kuramoto (1984) and the review by Strogatz (2000). The original mathematical flocking model and its unconditional and conditional alignment regimes are developed by Cucker and Smale (2007). The inertial spin model, which supplements velocity alignment with an internal turning variable, is introduced and analyzed physically by Cavagna et al. (2015). For replicator dynamics, evolutionary stability, and the geometry of the probability simplex, see Taylor and Jonker (1978), Shahshahani (1979), Akin (1979), and Hofbauer and Sigmund (1998).

Appendix A

Computing Phase Portraits with Python

Phase portraits are geometric objects, but numerical computation is often the most convenient way to visualize them. This appendix illustrates a basic workflow in Python using the Brusselator, a nonlinear chemical-reaction model studied in Chapters 4 and 11.

The purpose here is not to develop numerical analysis, which is treated in Chapter 10, but to show how a vector field, nullclines, equilibria, and solution trajectories can be displayed in the phase plane with a relatively small amount of code.

We consider the Brusselator

$$\dot{x} = A - (B+1)x + x^2 y, \qquad \dot{y} = Bx - x^2 y,$$

where $A > 0$ and $B > 0$ are parameters.

The unique positive equilibrium is

$$(x_*, y_*) = \left(A, \frac{B}{A}\right).$$

As discussed in Chapters 4 and 11, its stability changes when

$$B = 1 + A^2.$$

For

$$A = 1,$$

the critical value is therefore

$$B = 2.$$

We shall use

$$A = 1, \qquad B = 3,$$

so that the equilibrium is unstable and a stable periodic orbit is observed.

A.1 The complete Python program

The following program uses three standard Python packages:

- `NumPy` for numerical arrays,
- `Matplotlib` for plotting,
- `SciPy` for numerical integration of the differential equation.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.integrate import solve_ivp

# ------------------------------------------------------------
# Parameters
# ------------------------------------------------------------

A = 1.0
B = 3.0

# ------------------------------------------------------------
# The Brusselator
# ------------------------------------------------------------

def brusselator(t, z):
    x, y = z
    dx = A - (B + 1)*x + x*x*y
    dy = B*x - x*x*y
    return [dx, dy]

# ------------------------------------------------------------
# Computational region
# ------------------------------------------------------------

x_min, x_max = 0.0, 4.0
y_min, y_max = 0.0, 5.0

x = np.linspace(x_min, x_max, 25)
y = np.linspace(y_min, y_max, 25)

X, Y = np.meshgrid(x, y)

U = A - (B + 1)*X + X**2 * Y
V = B*X - X**2 * Y

# ------------------------------------------------------------
```

```
# Normalize the arrows
# ------------------------------------------------------------

speed = np.sqrt(U**2 + V**2)

# Avoid division by zero at points where the vector field vanishes.
speed[speed == 0] = 1.0

Un = U / speed
Vn = V / speed

# ------------------------------------------------------------
# Create the figure
# ------------------------------------------------------------

fig, ax = plt.subplots(figsize=(8, 6))

# ------------------------------------------------------------
# Direction field
# ------------------------------------------------------------

ax.quiver(
    X, Y, Un, Vn,
    angles='xy',
    scale_units='xy',
    scale=8,
    width=0.003,
    color='0.30'
)

# ------------------------------------------------------------
# Nullclines
# ------------------------------------------------------------

xx = np.linspace(0.15, x_max, 500)

# x-nullcline: dx/dt = 0
# A - (B+1)x + x^2 y = 0

x_nullcline = ((B + 1)*xx - A) / xx**2

ax.plot(
    xx,
    x_nullcline,
    color='black',
```

```
    linewidth=2,
    label=r'$dx/dt=0$'
)

# y-nullcline: dy/dt = 0
# Bx - x^2 y = 0
# For x > 0, y = B/x

y_nullcline = B / xx

ax.plot(
    xx,
    y_nullcline,
    color='black',
    linewidth=2,
    linestyle='--',
    label=r'$dy/dt=0$'
)

# ------------------------------------------------------------
# Equilibrium
# ------------------------------------------------------------

x_eq = A
y_eq = B / A

ax.plot(x_eq, y_eq, 'ko', markersize=7)

ax.text(
    x_eq + 0.08,
    y_eq + 0.08,
    r'$(A,B/A)$'
)

# ------------------------------------------------------------
# Numerical solution trajectories
# ------------------------------------------------------------

initial_conditions = [
    [0.4, 0.5],
    [0.7, 2.0],
    [1.5, 1.0],
    [2.5, 4.0],
    [3.0, 1.0]
]
```

```
t_span = (0.0, 40.0)

t_eval = np.linspace(
    t_span[0],
    t_span[1],
    4000
)

for z0 in initial_conditions:

    sol = solve_ivp(
        brusselator,
        t_span,
        z0,
        t_eval=t_eval,
        rtol=1e-8,
        atol=1e-10
    )

    ax.plot(
        sol.y[0],
        sol.y[1],
        linewidth=1.5
    )

    # Mark the initial point.
    ax.plot(
        z0[0],
        z0[1],
        'o',
        markersize=4
    )

# ------------------------------------------------------------
# Formatting
# ------------------------------------------------------------

ax.set_xlim(x_min, x_max)
ax.set_ylim(y_min, y_max)

ax.set_xlabel(r'$x$')
ax.set_ylabel(r'$y$')

ax.set_title(
```

```
    rf'Brusselator phase portrait: $A={A}$, $B={B}$'
)

ax.legend()
ax.grid(alpha=0.25)

plt.tight_layout()
plt.show()
```

The program produces the phase portrait shown in Figure A.1. It contains the direction field, the two nullclines, the equilibrium, and several numerically computed solution trajectories.

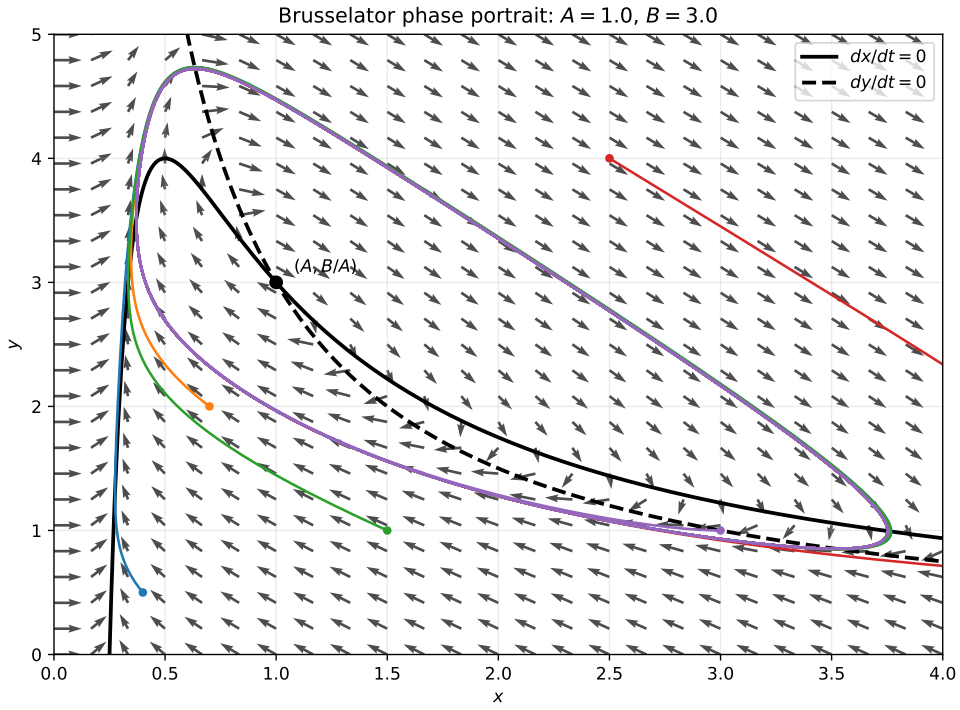


Figure A.1: Phase portrait of the Brusselator with $A = 1$ and $B = 3$, generated by the Python program above. The computation uses the rectangle $0 \le x \le 4$, $0 \le y \le 5$ and a 25×25 grid for the normalized direction field. The black solid and black dashed curves are the nullclines $\dot{x} = 0$ and $\dot{y} = 0$, respectively, and intersect at the unstable equilibrium $(x_*, y_*) = (1, 3)$. The solution trajectories are computed on $0 \le t \le 40$ from the initial conditions $(0.4, 0.5)$, $(0.7, 2)$, $(1.5, 1)$, $(2.5, 4)$, and $(3, 1)$. They approach the stable periodic orbit surrounding the unstable equilibrium.

A.2 Building the phase portrait

The vector field. The function

```
def brusselator(t, z):
    x, y = z
```

```
    dx = A - (B + 1)*x + x*x*y
    dy = B*x - x*x*y
    return [dx, dy]
```

defines the right-hand side of the differential equation. The argument `z` represents the state vector

$$\mathbf{z} = \begin{pmatrix} x \\ y \end{pmatrix},$$

and the returned vector is

$$\mathbf{f}(x, y) = \begin{pmatrix} A - (B+1)x + x^2 y \\ Bx - x^2 y \end{pmatrix}.$$

The variable `t` is included because the numerical solver expects a function of the form $\mathbf{f}(t, \mathbf{z})$; for an autonomous system, the right-hand side does not depend explicitly on t.

The computational grid. To draw a direction field, the vector field is evaluated at a rectangular grid of points:

```
x = np.linspace(x_min, x_max, 25)
y = np.linspace(y_min, y_max, 25)
X, Y = np.meshgrid(x, y)
```

The commands

```
U = A - (B + 1)*X + X**2 * Y
V = B*X - X**2 * Y
```

then evaluate the two components of the vector field at all grid points.

Normalization. Since the magnitude of the vector field may vary considerably across the phase plane, we normalize the arrows:

```
speed = np.sqrt(U**2 + V**2)
speed[speed == 0] = 1.0
Un = U / speed
Vn = V / speed
```

This changes the displayed lengths of the arrows but not their directions, and therefore gives a clearer picture of the phase-plane geometry. The direction field is drawn with

```
ax.quiver(X, Y, Un, Vn, ...)
```

Nullclines. For the Brusselator,

$$\dot{x} = A - (B+1)x + x^2 y, \qquad \dot{y} = Bx - x^2 y.$$

In the positive quadrant, the nullclines are

$$\dot{x} = 0 \quad \Longleftrightarrow \quad y = \frac{(B+1)x - A}{x^2},$$

and

$$\dot{y} = 0 \quad \Longleftrightarrow \quad y = \frac{B}{x}.$$

Their intersection gives

$$(x_*, y_*) = \left(A, \frac{B}{A}\right).$$

They are plotted by

```
x_nullcline = ((B + 1)*xx - A) / xx**2
y_nullcline = B / xx
```

Numerical trajectories. The SciPy routine

```
solve_ivp
```

computes numerical solutions of

$$\dot{\mathbf{z}} = \mathbf{f}(t, \mathbf{z}), \qquad \mathbf{z}(0) = \mathbf{z}_0.$$

For each prescribed initial condition, the commands

```
sol = solve_ivp(
    brusselator,
    t_span,
    z0,
    t_eval=t_eval,
    rtol=1e-8,
    atol=1e-10
)

ax.plot(sol.y[0], sol.y[1])
```

plot the trajectory

$$t \longmapsto (x(t), y(t))$$

in the phase plane.

A.3 Interpreting the phase portrait

For the parameter values

$$A = 1, \qquad B = 3,$$

the equilibrium is

$$(x_*, y_*) = (1, 3).$$

The Jacobian matrix of the Brusselator is

$$D\mathbf{f}(x, y) = \begin{pmatrix} -(B+1) + 2xy & x^2 \\ B - 2xy & -x^2 \end{pmatrix}.$$

At the equilibrium,

$$D\mathbf{f}\left(A, \frac{B}{A}\right) = \begin{pmatrix} B-1 & A^2 \\ -B & -A^2 \end{pmatrix}.$$

Therefore

$$\operatorname{tr} D\mathbf{f} = B - 1 - A^2,$$

and

$$\det D\mathbf{f} = A^2.$$

For

$$A = 1,$$

the trace becomes

$$B - 2.$$

Thus the critical parameter value is

$$B = 2.$$

When

$$B < 2,$$

the equilibrium is locally asymptotically stable. When

$$B > 2,$$

it becomes unstable. The Hopf bifurcation discussed in Chapter 11 explains the appearance of periodic motion near the critical value.

For $B = 3$, trajectories starting sufficiently near the unstable equilibrium spiral outward, while trajectories starting farther away may approach the attracting closed curve from the outside. The computed trajectories therefore suggest the presence of a stable periodic orbit surrounding the equilibrium.

The numerical picture does not by itself constitute a proof of the existence or stability of the periodic orbit. Its purpose is to visualize the dynamics predicted by the analytical theory.

A.4 Changing the parameter

The same program can be used to observe the transition through the Hopf bifurcation.

Keep

$$A = 1$$

and run the code for several values of B.

Case $B = 1.5$. Since

$$B < 2,$$

the equilibrium is stable. Nearby trajectories should approach

$$(1, 1.5).$$

Case $B = 2$. This is the critical value

$$B = 1 + A^2.$$

The linearization has purely imaginary eigenvalues. Linear stability alone is therefore inconclusive, and nonlinear terms determine the bifurcation.

Case $B = 3$. Now

$$B > 2,$$

and the equilibrium is unstable. The computed trajectories approach a stable periodic orbit.

Thus a sequence of phase portraits provides a visual representation of the transition

$$\text{stable equilibrium} \longrightarrow \text{loss of stability} \longrightarrow \text{stable periodic oscillation}.$$

A.5 A shorter version using `streamplot`

Matplotlib also provides the command

```
streamplot
```

for drawing approximate flow lines directly from a vector field.

For example, one may replace the `quiver` command by

```
ax.streamplot(
    X, Y, U, V,
    density=1.2,
    linewidth=0.8
)
```

This gives a visually smooth representation of the direction field.

However, a streamline plot should not be confused with the numerical integration of specified initial value problems. The curves produced by `streamplot` are primarily a visualization of the vector field, whereas `solve_ivp` computes approximations to actual solution trajectories with prescribed initial values.

For studying a dynamical system, it is often useful to use both.

A.6 Saving the figure

To include the resulting phase portrait in a LaTeX document, the figure may be saved in PDF format. Insert the `savefig` command between `tight_layout` and `show`:

```
plt.tight_layout()
plt.savefig(
    '../Fig/brusselator_python.pdf',
    bbox_inches='tight'
)
plt.show()
```

The resulting file can then be included using

```
\begin{figure}[htbp]
    \centering
    \includegraphics[width=0.72\textwidth]
        {../Fig/brusselator_python.pdf}
    \caption{Phase portrait of the Brusselator generated with Python.}
    \label{Fig:BrusselatorPython}
\end{figure}
```

For publication-quality figures, PDF output is preferable because the geometric elements remain in vector form.

A.7 Further experiments

The program can be modified easily to explore other systems in this book. The essential steps remain the same:

$$\text{define the vector field} \longrightarrow \text{choose a computational region} \longrightarrow \text{plot the direction field}$$
$$\longrightarrow \text{plot nullclines and equilibria} \longrightarrow \text{integrate selected trajectories.}$$

For example, one may replace the function `brusselator` by the Van der Pol system

$$\dot{x} = y, \qquad \dot{y} = \mu(1 - x^2)y - x,$$

or by one of the population, competition, epidemic, or FitzHugh–Nagumo systems studied in Chapter 4.

The computational procedure is almost unchanged. What changes is the mathematical interpretation of the resulting phase portrait.

Exercise A.1. Keep $A = 1$ and generate phase portraits of the Brusselator for

$$B = 1, \qquad B = 1.8, \qquad B = 2, \qquad B = 2.2, \qquad B = 3.$$

For each case:

1. locate the equilibrium;
2. compute the eigenvalues of the linearization and determine what linearization implies about stability;
3. compare the analytical prediction with the numerical phase portrait;
4. describe how the observed dynamics changes as B crosses the critical value $B = 2$.

Exercise A.2. Modify the code to plot the Van der Pol oscillator

$$\dot{x} = y, \qquad \dot{y} = \mu(1 - x^2)y - x$$

for

$$\mu = 0.5, \qquad \mu = 1.5, \qquad \mu = 5.$$

Compare the shape of the attracting periodic orbit as μ increases. Relate the numerical pictures to the discussion of relaxation oscillation in Chapters 4 and 11.

Exercise A.3. Modify the program so that the time series $x(t)$ and $y(t)$ are plotted in addition to the phase portrait. Explain what information is easier to see in the time plots and what information is easier to see in the phase portrait.

References

Ethan Akin. *The Geometry of Population Genetics*, volume 31 of *Lecture Notes in Biomathematics*. Springer, Berlin, 1979.

Warder Clyde Allee. Animal aggregations. *The Quarterly Review of Biology*, 2(3):367–398, 1927.

Warder Clyde Allee. *Animal Aggregations: A Study in General Sociology*. University of Chicago Press, Chicago, 1931.

Vladimir I. Arnold. *Mathematical Methods of Classical Mechanics*, volume 60 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2nd edition, 1989. doi: 10.1007/978-1-4757-2063-1.

Vladimir I. Arnold. *Ordinary Differential Equations*. Universitext. Springer-Verlag, Berlin, 1992. Translated by Roger Cooke.

Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3:133–181, 1922.

Ivar Bendixson. Sur les courbes définies par des équations différentielles. *Acta Mathematica*, 24: 1–88, 1901. doi: 10.1007/BF02403068.

Jacob Bernoulli. Explicationes, annotationes et additiones ad ea quae in actis superiorum annorum de curva elastica, isochrona paracentrica, et velaria hinc inde memorata et partim controversa leguntur; ubi de linea mediarum directionum, aliisque novis. *Acta Eruditorum*, pages 537–553, 1695.

Stephen R. Carpenter, Donald Ludwig, and William A. Brock. Management of eutrophication for lakes subject to potentially irreversible change. *Ecological Applications*, 9(3):751–771, 1999. doi: 10.1890/1051-0761(1999)009[0751:MOEFLS]2.0.CO;2.

Andrea Cavagna, Lorenzo Del Castello, Irene Giardina, Tomás Grigera, Asja Jelić, Stefania Melillo, Thierry Mora, Leonardo Parisi, Edmondo Silvestri, Massimiliano Viale, and Aleksandra M. Walczak. Flocking and turning: A new model for self-organized collective motion. *Journal of Statistical Physics*, 158:601–627, 2015.

Shui-Nee Chow and Jack K. Hale. *Methods of Bifurcation Theory*, volume 251 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, New York, 1982.

Fan R. K. Chung. *Spectral Graph Theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, Providence, RI, 1997.

Alexis-Claude Clairaut. Solution de plusieurs problèmes où il s'agit de trouver des courbes dont la propriété consiste dans une certaine relation entre leurs branches, exprimée par une équation donnée. *Histoire de l'Académie Royale des Sciences*, pages 196–215, 1734.

Colin W. Clark. *Mathematical Bioeconomics: The Mathematics of Conservation*. John Wiley & Sons, Hoboken, NJ, 3rd edition, 2010. ISBN 978-0-470-37299-9.

Earl A. Coddington and Norman Levinson. *Theory of Ordinary Differential Equations*. McGraw–Hill, New York, 1955.

Franck Courchamp, Ludek Berec, and Joanna Gascoigne. *Allee Effects in Ecology and Conservation*. Oxford University Press, Oxford, 2008. ISBN 978-0-19-857030-1. doi: 10.1093/acprof:oso/9780198570301.001.0001.

Felipe Cucker and Steve Smale. Emergent behavior in flocks. *IEEE Transactions on Automatic Control*, 52(5):852–862, 2007. doi: 10.1109/TAC.2007.895842.

Germund Dahlquist. Convergence and stability in the numerical integration of ordinary differential equations. *Mathematica Scandinavica*, 4:33–53, 1956. doi: 10.7146/math.scand.a-10454.

Jacob P. Den Hartog. *Mechanical Vibrations*. Dover Publications, New York, 4th edition, 1985. Reprint of the 1956 edition.

Paul A. M. Dirac. *The Principles of Quantum Mechanics*. International Series of Monographs on Physics. Clarendon Press, Oxford, 1930.

Gustav Doetsch. *Introduction to the Theory and Application of the Laplace Transformation*. Springer-Verlag, Berlin, 1974. doi: 10.1007/978-3-642-65690-3.

Richard C. Dorf and James A. Svoboda. *Introduction to Electric Circuits*. Wiley, Hoboken, NJ, 9th edition, 2014.

Jean-Marie-Constant Duhamel. Mémoire sur la méthode générale relative au mouvement de la chaleur dans les corps solides plongés dans des milieux dont la température varie avec le temps. *Journal de l'École Polytechnique*, 14(22):20–77, 1833.

Leah Edelstein-Keshet. *Mathematical Models in Biology*, volume 46 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, 2005. ISBN 978-0-89871-554-5. doi: 10.1137/1.9780898719147.

Klaus-Jochen Engel and Rainer Nagel. *One-Parameter Semigroups for Linear Evolution Equations*, volume 194 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2000. doi: 10.1007/b97696.

Leonhard Euler. *Methodus Inveniendi Lineas Curvas Maximi Minimive Proprietate Gaudentes, Sive Solutio Problematis Isoperimetrici Latissimo Sensu Accepti*. Marcum-Michaelem Bousquet, Lausanne and Geneva, 1744.

Leonhard Euler. *Institutionum Calculi Integralis*, volume 1. Academia Imperialis Scientiarum, Petropoli, 1768.

I. M. Gelfand and S. V. Fomin. *Calculus of Variations*. Dover Publications, New York, 2000.

Herbert Goldstein, Charles P. Poole, and John L. Safko. *Classical Mechanics*. Addison-Wesley, San Francisco, 3rd edition, 2002. ISBN 978-0-201-65702-9.

George Green. *An Essay on the Application of Mathematical Analysis to the Theories of Electricity and Magnetism*. T. Wheelhouse, Nottingham, 1828.

Thomas H. Gronwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 20(4):292–296, 1919. doi: 10.2307/1967124.

Seung-Yeal Ha and Jian-Guo Liu. A simple proof of the Cucker–Smale flocking dynamics and mean-field limit. *Communications in Mathematical Sciences*, 7(2):297–325, 2009. doi: 10.4310/CMS.2009.v7.n2.a2.

Seung-Yeal Ha and Eitan Tadmor. From particle to kinetic and hydrodynamic descriptions of flocking. *Kinetic and Related Models*, 1(3):415–435, 2008. doi: 10.3934/krm.2008.1.415.

Ernst Hairer and Gerhard Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2nd edition, 1996. doi: 10.1007/978-3-642-05221-7.

Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*, volume 8 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2nd edition, 1993. doi: 10.1007/978-3-540-78862-1.

Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, volume 31 of *Springer Series in Computational Mathematics*. Springer, Berlin, 2nd edition, 2006. doi: 10.1007/3-540-30666-8.

Jack K. Hale and Hüseyin Koçak. *Dynamics and Bifurcations*, volume 3 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 1991. ISBN 978-0-387-97141-4. doi: 10.1007/978-1-4612-4426-4.

William Rowan Hamilton. On a general method in dynamics. *Philosophical Transactions of the Royal Society of London*, 124:247–308, 1834.

William Rowan Hamilton. Second essay on a general method in dynamics. *Philosophical Transactions of the Royal Society of London*, 125:95–144, 1835.

Philip Hartman. *Ordinary Differential Equations*, volume 38 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, 2nd edition, 2002. doi: 10.1137/1.9780898719222.

Oliver Heaviside. *Electromagnetic Theory*, volume 2. The Electrician Printing and Publishing Company, London, 1899.

Hermann Helmholtz. Über integrale der hydrodynamischen gleichungen, welche den wirbelbewegungen entsprechen. *Journal für die reine und angewandte Mathematik*, 55:25–55, 1858.

Morris W. Hirsch, Stephen Smale, and Robert L. Devaney. *Differential Equations, Dynamical Systems, and an Introduction to Chaos*. Academic Press, Waltham, MA, 3rd edition, 2013.

Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge, 1998.

Robert Hooke. *Lectures de Potentia Restitutiva, or of Spring: Explaining the Power of Springing Bodies*. John Martyn, London, 1678.

Eberhard Hopf. Abzweigung einer periodischen lösung von einer stationären lösung eines differentialsystems. *Berichte der Mathematisch-Physikalischen Klasse der Sächsischen Akademie der Wissenschaften zu Leipzig*, 94:3–22, 1942.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 2nd edition, 2013.

Christiaan Huygens. *Horologium Oscillatorium, Sive de Motu Pendulorum ad Horologia Aptato Demonstrationes Geometricae*. F. Muguet, Paris, 1673.

Edward L. Ince. *Ordinary Differential Equations*. Dover Publications, New York, 1956. Reprint of the 1926 edition.

Arieh Iserles. *A First Course in the Numerical Analysis of Differential Equations*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2nd edition, 2008.

Camille Jordan. *Traité des substitutions et des équations algébriques*. Gauthier-Villars, Paris, 1870.

Hans G. Kaper and Hans Engler. *Mathematics and Climate*. Society for Industrial and Applied Mathematics, Philadelphia, 2013. ISBN 978-1-61197-260-3. doi: 10.1137/1.9781611972610.

William O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772):700–721, 1927. doi: 10.1098/rspa.1927.0118.

Hassan K. Khalil. *Nonlinear Systems*. Prentice Hall, Upper Saddle River, NJ, 3rd edition, 2002. ISBN 978-0-13-067389-3.

Gustav Kirchhoff. Über den durchgang eines elektrischen stromes durch eine ebene, insbesondere durch eine kreisförmige. *Annalen der Physik und Chemie*, 64(4):487–514, 1845.

Gustav Kirchhoff. Über die auflösung der gleichungen, auf welche man bei der untersuchung der linearen verteilung galvanischer ströme geführt wird. *Annalen der Physik und Chemie*, 72:497–508, 1847.

Gustav Robert Kirchhoff. *Vorlesungen über mathematische Physik, Band 1: Mechanik*. B. G. Teubner, Leipzig, 1876.

Yvette Kosmann-Schwarzbach. *The Noether Theorems: Invariance and Conservation Laws in the Twentieth Century*. Sources and Studies in the History of Mathematics and Physical Sciences. Springer, New York, 2011. doi: 10.1007/978-0-387-87868-3.

Yoshiki Kuramoto. Self-entrainment of a population of coupled non-linear oscillators. In *International Symposium on Mathematical Problems in Theoretical Physics*, volume 39 of *Lecture Notes in Physics*, pages 420–422. Springer, Berlin, 1975. doi: 10.1007/BFb0013365.

Yoshiki Kuramoto. *Chemical Oscillations, Waves, and Turbulence*, volume 19 of *Springer Series in Synergetics*. Springer-Verlag, Berlin, 1984.

Wilhelm Kutta. Beitrag zur näherungsweisen integration totaler differentialgleichungen. *Zeitschrift für Mathematik und Physik*, 46:435–453, 1901.

Yuri A. Kuznetsov. *Elements of Applied Bifurcation Theory*, volume 112 of *Applied Mathematical Sciences*. Springer, Cham, 4th edition, 2023. ISBN 978-3-031-22006-7. doi: 10.1007/978-3-031-22007-4.

Joseph-Louis Lagrange. *Méchanique analitique*. Desaint, Paris, 1788.

Cornelius Lanczos. *The Variational Principles of Mechanics*. Dover Publications, New York, 4th edition, 1986. ISBN 978-0-486-65067-8.

Pierre-Simon Laplace. *Théorie analytique des probabilités*. Veuve Courcier, Paris, 1812.

Joseph P. LaSalle. Some extensions of Liapunov's second method. *IRE Transactions on Circuit Theory*, 7(4):520–527, 1960. doi: 10.1109/TCT.1960.1086720.

Alfred Liénard. Étude des oscillations entretenues. *Revue Générale de l'Électricité*, 23:901–912, 946–954, 1928.

M. J. Lighthill. *An Introduction to Fourier Analysis and Generalised Functions*. Cambridge Monographs on Mechanics and Applied Mathematics. Cambridge University Press, Cambridge, 1958. doi: 10.1017/CBO9781139171427.

Ernst Lindelöf. Sur l'application des méthodes d'approximations successives à l'étude des intégrales réelles des équations différentielles ordinaires. *Journal de Mathématiques Pures et Appliquées*, 10:117–128, 1894.

Joseph Liouville. Note sur la théorie de la variation des constantes arbitraires. *Journal de Mathématiques Pures et Appliquées*, 3:342–349, 1838.

Jaume Llibre and Lidia Jiménez-Lara. Periodic orbits and non-integrability of hénon–heiles systems. *Journal of Physics A: Mathematical and Theoretical*, 44(20):205103, 2011. doi: 10.1088/1751-8113/44/20/205103.

Hartmut Logemann and Eugene P. Ryan. *Ordinary Differential Equations: Analysis, Qualitative Theory and Control*. Springer Undergraduate Mathematics Series. Springer, London, 2014. doi: 10.1007/978-1-4471-6398-5.

Edward N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2): 130–141, 1963. doi: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Alfred J. Lotka. *Elements of Physical Biology*. Williams & Wilkins, Baltimore, 1925.

Aleksandr M. Lyapunov. *The General Problem of the Stability of Motion*. Kharkov Mathematical Society, Kharkov, 1892. Original Russian edition.

Jerrold E. Marsden and Tudor S. Ratiu. *Introduction to Mechanics and Symmetry: A Basic Exposition of Classical Mechanical Systems*, volume 17 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 2nd edition, 1999. doi: 10.1007/978-0-387-21792-5.

Leonard Meirovitch. *Fundamentals of Vibrations*. McGraw-Hill, New York, 2001.

James D. Murray. *Mathematical Biology: I. An Introduction*, volume 17 of *Interdisciplinary Applied Mathematics*. Springer, New York, 3rd edition, 2002. ISBN 978-0-387-95223-9. doi: 10.1007/b98868.

Paul K. Newton. *The N-Vortex Problem: Analytical Techniques*, volume 145 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2001. doi: 10.1007/978-1-4684-9290-3.

Emmy Noether. Invariante variationsprobleme. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, pages 235–257, 1918.

Johann Friedrich Pfaff. *Methodus Generalis, Aequationes Differentiarum Partialium, Nec Non Aequationes Differentiales Vulgares, Utrasque Primi Ordinis, Inter Quotcunque Variabiles, Complete Integrandi*. Königliche Akademie der Wissenschaften, Berlin, 1815.

Émile Picard. Mémoire sur la théorie des équations aux dérivées partielles et la méthode des approximations successives. *Journal de Mathématiques Pures et Appliquées*, 6:145–210, 1890.

Henri Poincaré. Mémoire sur les courbes définies par une équation différentielle. *Journal de Mathématiques Pures et Appliquées*, 7:375–422, 1881.

Henri Poincaré. Sur l'équilibre d'une masse fluide animée d'un mouvement de rotation. *Acta Mathematica*, 7:259–380, 1885. doi: 10.1007/BF02402204.

Siméon-Denis Poisson. Mémoire sur la variation des constantes arbitraires dans les questions de mécanique. *Journal de l'École Polytechnique*, 8(15):266–344, 1809.

Jacopo Francesco Riccati. Animadversiones in aequationes differentiales secundi gradus. *Actorum Eruditorum Quae Lipsiae Publicantur, Supplementa*, 8:66–73, 1724.

Carl Runge. Über die numerische auflösung von differentialgleichungen. *Mathematische Annalen*, 46:167–178, 1895.

Milner B. Schaefer. Some aspects of the dynamics of populations important to the management of the commercial marine fisheries. *Bulletin of the Inter-American Tropical Tuna Commission*, 1 (2):27–56, 1954.

Marten Scheffer, Steve Carpenter, Jonathan A. Foley, Carl Folke, and Brian Walker. Catastrophic shifts in ecosystems. *Nature*, 413:591–596, 2001. doi: 10.1038/35098000.

Laurent Schwartz. *Théorie des distributions*. Actualités scientifiques et industrielles. Hermann, Paris, 1950–1951.

Siavash Shahshahani. *A New Mathematical Framework for the Study of Linkage and Selection*. Number 211 in Memoirs of the American Mathematical Society. American Mathematical Society, 1979.

Ivar Stakgold and Michael Holst. *Green's Functions and Boundary Value Problems*. Wiley, Hoboken, NJ, 3rd edition, 2011.

Steven H. Strogatz. From Kuramoto to Crawford: Exploring the onset of synchronization in populations of coupled oscillators. *Physica D: Nonlinear Phenomena*, 143(1–4):1–20, 2000. doi: 10.1016/S0167-2789(00)00094-4.

Steven H. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Chapman and Hall/CRC, New York, 3rd edition, 2024. ISBN 978-0-367-02650-9. doi: 10.1201/9780429398490.

Peter D. Taylor and Leo B. Jonker. Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences*, 40:145–156, 1978.

Gerald Teschl. *Ordinary Differential Equations and Dynamical Systems*, volume 140 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

J. M. T. Thompson and G. W. Hunt. *A General Theory of Elastic Stability*. John Wiley & Sons, London, 1973.

Ka-Kit Tung. *Topics in Mathematical Modeling*. Princeton University Press, Princeton, NJ, 2007. ISBN 978-0-691-11642-6.

Balthasar van der Pol. On "relaxation-oscillations". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):978–992, 1926. doi: 10.1080/14786442608564127.

Pierre-François Verhulst. Notice sur la loi que la population suit dans son accroissement. *Correspondance Mathématique et Physique*, 10:113–121, 1838.

Pierre-François Verhulst. Recherches mathématiques sur la loi d'accroissement de la population. *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, 18:1–40, 1845. doi: 10.3406/marb.1845.3438.

Marcelo Viana and José M. Espinar. *Differential Equations: A Dynamical Systems Approach to Theory and Practice*, volume 212 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2021.

Vito Volterra. Fluctuations in the abundance of a species considered mathematically. *Nature*, 118 (2972):558–560, 1926. doi: 10.1038/118558a0.

David V. Widder. *The Laplace Transform*, volume 6 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, 1946.

Index

a priori estimate, **403**
  bounds, 137
  energy method, 404
  nonlinear system, 405
absorbing set, **489**
  Lorenz system, 519
  nonlinear estimate, 494
action functional, **313**
adaptive step-size control, **419**, 441
  error estimator, 420
  local error control, 419
  step-size selection, 420
adjacency matrix, 198
Allee effect, **19**
  strong, **19**
  threshold, **19**
angular frequency, 98
angular impulse
  point vortices, 361
attractor, **455, 456**
  continuum of equilibria, 457
  distinguished from invariant set, 496
  invariant, 456
  periodic, **498**
  strange, 520
autonomous ODE, **3**
autonomous system
  planar, **127**

backward Euler method, 441
bacteria growth model
  with injection, 49
Banach space, **374**
barrier solution, **82**
basin of attraction, **455**
  boundary, 147, 457
  competition model, 459
  Lyapunov estimate, 476
  unstable periodic-orbit boundary, 458
beating, **109**, 110
Beltrami identity, **304**
Bernoulli equation, **52**
bifurcation, **21**
bistability, 20, **28**
  competition model, 147
  genetic toggle switch, 182
  lake model, 41
blow-up
  finite-time, **80**, 401
blow-up alternative, **402**
brachistochrone, 306
Brusselator, **153, 154**
  chemical oscillator, 510
  limit cycle, 157, 510

calculus of variations, **295**
canonical matrix, 336
canonical transformation, **340**
  change of configuration coordinates, 343
  nonlinear coordinate, 343
  position–momentum interchange, 344
capillary interface, 302
carrying capacity, **16**
catenary, 301, 304
Cauchy–Lipschitz theorem, **376**
causal integration, **264**
causal inverse operator, 280
Cayley–Hamilton theorem, **224**
center, **118**
central force, 320
central-force problem, 322
chaos

electrical systems, 523
geometric mechanism, 521
higher-dimensional systems, 523
mechanical systems, 523
characteristic equation, **94**
characteristic polynomial, **93**
characteristic root, 94
classification, 97
complex conjugate roots, 98
distinct real roots, 97
exponential mode, 95
repeated real root, 97
charged particle, 337
overdamped, 486
chemical reaction
reversible, 485
reversible chain, 489
chemical reaction chain, 196
general, 198
chemical reaction system, **153**
autocatalytic, **155**
chemostat model, 182
Chua circuit
volume contraction, 470
circuit duality, 195
circuit graph, 192
Clairaut equation, **54**
envelope, 54
family of lines, 70
singular solution, 75
classical solution methods, **43**
closed orbit, *see* periodic orbit
closed trajectory, *see* periodic orbit
coercive function, 473
collective turning, 479
collision orbit, 327
commutator, 266
commuting matrices
exponentials, 250
competition model, **144**
bistability, 459
strong competition, 147
weak competition, 146
competition, interspecific, *see* competition model
competitive exclusion, 144
species x excludes species y, 147
species y excludes species x, 147
complete metric space, **373**
closed subset, 377
conjugate momentum, **317**
connecting orbit, **139**
consensus dynamics
cluster consensus, 488
conservative system, **159**
constant-coefficient differential equation, 267
constrained variation, **299**
continuation by boundedness, **81**
continuation principle, **401**
boundary obstruction, 402
continuous dependence
initial data, **385**
parameters, **387**
stability estimate, 384
vector-field perturbation, 384
contraction map, **374**
contraction mapping theorem, **374**
convergence
global, **417**
convolution, **273**
causal, **275**
convolution theorem, **275**
Coriolis force, 463
coupled oscillator
Hamiltonian formulation, 338
coupled system, 189
critical damping
mode coalescence, 215
Cucker–Smale model, 477
cyclic coordinate, **302**, 317
spherical pendulum, 318
cycloid, 78
brachistochrone, 306

damped oscillation, **101**
damped pendulum, **163**
linearization, 163
damping
critical damping, **101**
indirect, 484
overdamped motion, **102**

partial, 488
underdamped motion, **101**
damping coefficient, 90
decoupled subsystem, 234
decoupling, 221
by invariant subspaces, 221
degree matrix, 198
delayed step input, 278
delta distribution, **277**
shifted, 278
detailed balance, 485
determinant, 120
diagonal system
resolvent, 282
diagonalizable matrix, 221
resolvent, 287
differentiable dependence
initial data, **394**
differential operator, 93
diffusion
on a graph, 198
Dirac delta, 108, *see also* delta distribution
direct-sum decomposition, 223
direction field, **55**, *see also* vector field
normalization, 535
discontinuous forcing, **277**
discrete Gronwall inequality, **418**
discrete Laplacian
fixed–free boundary conditions, 192
fixed-end chain, 191
LC ladder, 195
reaction chain, 198
discriminant, 97
disease threshold, *see* epidemic threshold
dissipation
energy, 459
dissipative system, **159**, **459**
damped pendulum, 163
phase-space contraction, 349
divergence
and volume change, 397
Lorenz system, 518
double pendulum, 315
Hamiltonian, 334
numerical predictability, 444
double-well potential, 166, 313, 495
doublet flow, 60
streamlines, 68
Duffing oscillator, **166**, 495
linearization, 166
Duhamel's formula, **105**
Duhamel's principle, **49**, **103**, **209**, **275**, 276

edge current, 192
effective potential, 318
eigenline, 117
eigenvalue, 117, 211
complex, 118, 212
distinct real, 118
Hamiltonian symmetry, 346
repeated, 120
eigenvector, 117, 211
electrical oscillator, 91
endemic equilibrium, 153
energy
mechanical, 461
numerical error, 437
simple pendulum, 159
energy conservation
Lagrangian system, 317
energy dissipation
damped pendulum, 163
energy surface, **345**
entropy
chemical reaction, 485
entropy production, 469
envelope
wave fronts, 86
envelope condition, **74**
envelope of a curve family, **73**
Clairaut family, 75
lines, 74
translated parabolas, 77
epidemic model, **150**
epidemic threshold, 153
equilibrium, **10**
asymptotically stable, **11**, 237, 471
globally asymptotically stable, **471**
Hamiltonian system, 345
invariant set, 496
local classification, 129

planar system, **129**
semistable, **11**
stable, **470**
equilibrium manifold
curved, 486
equilibrium solution, 44
Euler method
backward, **426**
forward, **413**
stability region, 423
symplectic, **431**
Euler–Lagrange equation, **297**
eutrophication, 41
evolute, 78
evolution operator, 245, 246, 256, 260
evolutionary game, 465
exact differential form, **62**
local versus global, 64
exact equation, *see* exact differential form
exactness test, **63**
excitable system, 172
excitation pulse, 179
existence
Cauchy–Lipschitz theorem, 376
global, 81, 137, **399**
global from a priori bounds, **403**
linear-growth criterion, 404
local, **371**
without uniqueness, 379
existence and uniqueness, **79**
linear system, 204
exponential Euler method, 252
exponential mode, **94**, 211
exponential order, **261**

family of integral curves, 59
family of plane curves, **67**
fast and slow modes, 442
fast–slow system, 172
final-size relation, **153**
first integral, **59**, **138**, **302**
Hamiltonian system, 352
in involution, 356
Lotka–Volterra model, **141**
first Lyapunov coefficient, 512
first variation, **296**
first-order linear system
Laplace transform, 281
FitzHugh–Nagumo system, **172**
fast–slow dynamics, 175
time evolution, 179
fixed point, **374**
flocking, 477
spin, 479
Floquet multiplier, **397**
stability of a periodic orbit, 505
flow
generated by an autonomous system, 450
flow map, **391**
derivative, 394
inverse, 392
flow property, **391**
fold bifurcation, *see* saddle-node bifurcation
folding of solution curves, *see also* envelope of a curve family
forced damped oscillator, 91
forced motion, 103
forced oscillation
damped, 112
undamped, 109
forced response, 268, 277
forcing term, 47
free response, 268, 277
frequency response, **113**
functional, **296**
functional response
saturating, 181
fundamental lemma of the calculus of variations, **297**
fundamental matrix, 206
nonautonomous system, 256
solution representation, 208
fundamental set of solutions, **96**
fundamental solution, 206
fundamental solution matrix, *see* fundamental matrix

gauge transformation, 338
canonical transformation, 366
general solution
Jordan-chain form, 236
generalized coordinate, **314**

generalized coordinates, 313
generalized eigenspace, 223
generalized eigenvector, 215
generalized exponential mode, 97
generalized mode, 218, 260, 262, *see also* Jordan chain
  complex, 254
genetic toggle switch, 182
geometric numerical integration
  structure preservation, 431
  structure preservation: versus order, 434
  structure-preserving methods, **428**
geometric viewpoint of ODEs, 43, **55**
global asymptotic stability, **473**
global attractor, **489**
  bistable system, 493
  existence, 491
  heteroclinic orbits, 495
  planar gradient system, 497
global error estimate, **419**
glycolytic oscillation, 183
gradient flow, **463**
gradient system, **463**
  replicator equation, 465
graph
  diffusion, 469
  oriented, 462
graph Fourier mode, 239
graph Laplacian, 198
  disconnected graph, 488
Green's function, **105**, **273**
  causal, 108, **274**
  fourth-order operator, 274
  impulse equation, 274
  inverse operator, 273
Gronwall inequality, **382**
  backward-time estimate, 383
  homogeneous form, 382
  integral form, 383
ground acceleration
  response to, 276
gyroscopic force, 461, 463, 469

Hadamard well-posedness, **387**
Hall drift, 486
Hamilton's equations, **335**
  canonical form, 336
Hamilton's principle, **313, 314**
Hamiltonian, **333**
  level set, 345
  positive-definite quadratic, 347
  time-dependent, 336
Hamiltonian flow
  canonical structure, **340**
  symplecticity, 341
Hamiltonian formulation, **332**
Hamiltonian generator, **354**
Hamiltonian matrix
  spectral symmetry, 346
Hamiltonian system, **331**, **336**
  absence of asymptotically stable equilibria, 349
  linear, 345
  natural mechanical system, 333
  noncanonical, 358
hardening spring, 166
harmonic oscillator, 308
  impulse response, 279
  Laplace transform, 271
harvesting model
  constant, **22**
Heaviside function, **277**
heteroclinic orbit, **310**
  explicit connection, 181
  global attractor, 493, 497
  phase-field transition, 312
homoclinic orbit, **310**
  double-well potential, 313
  Duffing oscillator, 168
homogeneous equation, 47, **93**
homogeneous first-order equation, **53**
homogeneous linear system, 203
homogeneous motion, 94
Hooke's law, **90**
Hopf bifurcation, **506**
  Brusselator, 510
  nonlinear oscillator, 508
  recognition in planar systems, 514
  subcritical, 511
  supercritical, 507
  transversality, 512

Hopf bifurcation theorem, **512**
Hopf normal form
  subcritical, 511
  supercritical, 507
hyperbolic equilibrium, **129, 130**
hysteresis, **28**
  lake model, 41

implicit solution, 44
impulse, **273**
  unit, 278
impulse response, **105**
  RLC circuit, 289
impulsive forcing, **277**
incidence matrix, 462
inertial spin model, 479
inhomogeneous equation, **93**
  general solution, 103
initial data
  Laplace transform, 264
initial value problem, **4**
  integral formulation, **371**
  linear system, 204
  well-posedness, 387
initial-data polynomial, 268
integral curve, **56**
integrating factor, **50**
  differential form, 62
  Pfaffian equation, **65**
integration operator, 264
invariant curve, **131**, **137**
invariant decomposition, 221
invariant direction, **117**
invariant interval, 83
invariant region, **136**
invariant set, **450**
  attractor, 496
  positively invariant, 136, **450**
    boundary test, 137
  under the Picard operator, 377
invariant subspace, 221
  real two-dimensional, 212
invariant torus, **350**
inverse Laplace transform, **263**
involution
  first integrals, **356**
isoperimetric constraint, 302

Jacobian matrix, **129**
Jordan basis, 227
Jordan block, 227, 260
  pole order, 285
  resolvent, 283
  size, 225
Jordan canonical form, **232**, 260
Jordan chain, 218
Jordan coordinates, 233, 234
Jordan normal form, *see* Jordan canonical form
jump condition
  impulse, 280

Kepler problem, **322**
  center-of-mass reduction, 322
  radial motion, 327
  Runge–Lenz vector, 365
Kepler's laws, 325
kernel filtration, 225
kernel increment, 225
Kirchhoff's current law, **91**, 192
Kirchhoff's voltage law, **91**, 192
Kuramoto model, 464

Lagrange multiplier, **300**
Lagrangian, **296**
Lagrangian mechanics, 295, **313**
  Hamiltonian formulation, 332
Laplace inversion, *see* inverse Laplace transform
Laplace transform, **260**
  definition, **261**
  differentiation rule, **264**
  exponential function, 261
  exponential shift rule, **265**
  linearity, 263
  multiplication by time, 265
  operational rules, 263
  power function, 262
  sine and cosine, 262
  uniqueness, **263**
LaSalle invariance principle, **481**
  partial damping, 484
  periodic orbit, 487

LC circuit
coupled through inductors, 195
leapfrog method, **433**
stability, 442
Legendre transform, **333**
libration, 162
Liénard equation, **170**
Liénard system, 501
limit cycle, 157, 159
stable, 454
linear dynamical system
modeling, 189, 253
linear equation
first order, *see* linear first-order equation
linear first-order equation, **47**
linear impulse
point vortices, 361
linear independence, 95
linear oscillator, **89**, 99
linear stability, **237**
linear system, 189
abstract theory, 203
forced response, 282
Laplace transform, 280
linearization, **12**
inconclusive, **14**, 134
near an equilibrium, **129**
linearized flow, **395**
chain rule, 396
Liouville formula, **206**, **397**
Liouville integrability, **356**
Liouville's theorem, **347**
Lipschitz condition
failure of, 84
Lipschitz continuity
bounded-derivative criterion, 378
local, **376**
local existence and uniqueness theorem, **80**
local flow, **391**
logistic model, **17**
logistic predator–prey model, 143
Lorentz force, 337, 463, 486
Lorenz attractor, **520**
lobe switching, 520
Lorenz system, **516**
boundedness, 519
classical parameters, 518
equilibria, 516
nonzero equilibria, 517
stability of the origin, 517
symmetry, 516
Lotka–Volterra model, **140**
periodic orbits, 142
time-dependent solution, 176
Lyapunov equation, 474, 480
Lyapunov function, **470**
flocking, 477
linear system, 474
mechanical energy, 165, 474
nonlinear perturbation, 476
Lyapunov stability
linear system, 237
Lyapunov stability theorem, **472**

Mach cone, 86
Malthusian growth model, **15**
mass matrix, **314**, 333
double pendulum, 315
mass-action law, **153**, 196
mathematical model, **4**
mathematical modeling, **4**
matrix exponential, **246**
as inverse transform of resolvent, 282
from Jordan form, 249
Jordan block, 247
properties, 250
matrix structure, 221
maximal extension, **400**
maximal interval of existence, **80**, **400**
maximal solution, **400**
mechanical oscillator, 90
Hopf bifurcation, 508
mechanical potential, **159**
mechanical system
dissipative, 461
method of integrating factors, *see* integrating factor
minimal polynomial
resolvent, 288
modal analysis, 189
coupled systems, 239

modal decomposition, 285
modal representation, 211
  complete, **236**
modal solution, *see* mode
modal structure
  resolvent representation, 286
mode, 189
  complex, **212**
  generalized, 211
  ordinary, **211**
  real modal pair, **212**
mode coalescence, 215
momentum
  angular, 318, 320
  canonical, 337
  conserved, 302
  linear, 320
  mechanical, 337
momentum–velocity map, 332
monodromy matrix, **397**
monotone quantity, **459**
monotonicity principle, 460
moving source, 86
multiscale oscillation, 442
mutualism model, 183

natural frequency, **99**
network model, 189
Neumann series, 284
Newton's law of heating and cooling, **6**
nilpotent matrix
  exponential, 247
node
  stable, **118**
  unstable, **118**
Noether's theorem, **319**
  Hamiltonian form, **354**
non-crossing of solutions, **82**, 392
nonautonomous linear system, 252
noncommuting matrices, 256
nonhomogeneous linear system, 203, 209
nonhyperbolic equilibrium, **134**
nonlinear damping
  Van der Pol oscillator, 170
nonlinear oscillator, **159**
  self-excited, 508
nonlinear restoring force, 166
nonlinear system
  planar, **127**
nonunique initial value problem, **83**
nonuniqueness
  failure of Lipschitz condition, 84, 379
  relation to envelopes, **83**
normal mode, 239
  canonical coordinates, 338
normal modes, 442
nullcline, **136**
  computation, 535
  distinction from orbit, 136
  x-nullcline, **136**
  y-nullcline, **136**
numerical error
  accumulation, **417**
  amplitude, 435
  analysis, **416**
  dispersion, 437
numerical method
  for ordinary differential equations, **411**
  implicit, 425, 441
  long-time behavior, **428**
  time stepping, 411
numerical predictability, 444
numerical stability
  absolute, **423**
  test equation, 423
numerical trajectory
  computation, 536

observable, **352**
ODE software, **438**
  DifferentialEquations.jl, 440
  Julia, 440
  MATLAB, 439
  Matplotlib, 530
    streamplot, 538
  NumPy, 530
  Python, 439, 529
  SciPy, 439, 530
Ohm's law, **92**
ω-limit set, **451**
  compactness, 452
  connectedness, 452

invariance, 452
one-parameter family of curves, **67**
one-parameter group, 246
open reaction system, 154
operational calculus, 259
operator algebra, 263
operator dictionary, 266
orbit, **129**
orbit, periodic, *see* periodic orbit
orbital stability, 498
order of accuracy, **417**
versus structure preservation, **434**
orthogonal trajectories, **71**
circles and radial lines, 72
confocal ellipses and hyperbolas, 72
oscillator
coupled, 484
oscillator network
damping, 463

parameter dependence
quantitative estimate, 388
parameter elimination, 67
parameter study
phase portrait, 537
parameter-dependent ODE, **21**
partial fractions, 44
partial-fraction decomposition, **269**
particular solution, 103
passive tracer, 358
Peano existence theorem, **379**
pendulum
cycloidal, 309
nonlinear friction, 483
separatrix, 311
simple, 308
spherical, 318
period
closed orbit, 307
double-well potential, 313
harmonic oscillator, 308
simple pendulum, 308
periodic orbit, **159**
as an ω-limit set, 454
as invariant limit set, 487
Floquet multiplier, 397
stability, 505
Pfaffian equation, **57**
phase curve, *see* orbit
phase error, 443
phase line, **10**
direction of motion, 10
phase oscillator
gradient structure, 464
phase plane, **127**
phase portrait, **128**
basic types, 118
computation with Python, **529**
one-dimensional, 11
planar linear system, **114**
phase space, 331, **336**
phase-plane analysis
scope and limitations, 180
phase-plane method, **127**
phase-volume contraction, 470
Picard iteration, **378**
Picard operator, **376**
pitchfork bifurcation, **33**
subcritical, **36**
planar linear system, 114
reduction to second order, 116
planar system
bounded orbit, 499
long-time behavior, 505
special role, 505
Poincaré recurrence theorem, **350**
Poincaré return map, 499
Poincaré–Bendixson theorem, **499**
Poincaré–Bendixson theorem
failure in higher dimensions, 523
point vortex, 56, 359
regular polygon, 366
point-vortex dynamics, **358**
conserved quantities, 361
four vortices, 364
integrability, 364
interaction Hamiltonian, 359
three vortices, 364
two vortices, 362
Poisson bracket, **352**
antisymmetry, 353

closure of first integrals, 356
Jacobi identity, 353
Leibniz rule, 353
pole
complex-conjugate, **270**
higher-order, 260
order, 262
repeated, **269**
shift, 265
simple, 260
simple real, **269**
pole–mode correspondence, 269, **270**
polynomial differential operator, **267**
Laplace representation, 267
polynomial–exponential mode, 218
population dynamics, **15**
population threshold, **19**
position-dependent mass, 343
positive orbit, **451**
potential energy, **159**
potential function, 62
gradient system, 463
predator saturation, 181
predator–prey equations, *see* Lotka–Volterra model
predator–prey model, **140**
global convergence, 524
prey self-limitation, 143
primary decomposition, 223
product approximation, 252, 256
propagator, **47**
purely imaginary eigenvalues
nonlinear behavior, 134

quadrature, **302**
quasiperiodic motion, **351**

radial lines, 68
radially unbounded function, 473
reducible nonlinear equation, **52**
region of convergence, **261**
regular Lagrangian, **332**
relative equilibrium
point vortices, 366
relaxation, **4**
linear process, 4
rate, **4**
time, **5**
relaxation oscillation, 170, 172
time evolution, 179
replicator equation, 465
resolvent, 260, **280**
complex-conjugate poles, 287
definition, **281**
derivative, 284
Jordan-block example, 286
large-s expansion, **284**
pole order, 285
poles and Jordan structure, **285**
properties, 283
resonance, **109**
repeated pole, 272
repeated poles, 289
transform-domain interpretation, 271
undamped oscillator, 111
resonance frequency
displacement, 114
effect of damping, 114
resonant limit, 110
Riccati equation, **52**
RLC circuit, **91**
coupled, 192
switched and impulsive input, 289
rotation
pendulum, 162
rotational symmetry, **320**
angular momentum, 355
Runge–Kutta method
classical fourth-order, **415**, 434
embedded, 420
Heun, 414
midpoint, 414
second-order, **414**
stability function, 424
Runge–Lenz vector, 365

saddle, **118**
stable and unstable curves, 131
saddle-node bifurcation, **22**
saturation
mutualism model, 183
scalar first-order ODE, **43**

seasonal growth model, 48
second-order equation
  first-order system formulation, 115
second-order linear equation, **89**
  Laplace transform, 271
  physical origins, 90
second-order system
  first-order formulation, 201
self-sustained oscillation, 170
Selkov model, 183
semiflow
  generated by an autonomous system, 450
semigroup property, 250
sensitive dependence on initial conditions
  Lorenz system, 521
sensitive dependence on initial data, 444
separable equation, **43**
separatrix, **139**, **310**
  competition model, 147
  Duffing oscillator, 168
  numerical crossing, 443
  simple pendulum, 162, 311
serial connection, 190
Shahshahani metric, 465
Shockley–Read–Hall recombination model, 184
similarity transformation, 249
simple pendulum, **159**
  linearization, 159
  numerical dynamics, 443
singular solution, *see also* Clairaut equation; envelope of a curve family
sinusoidal forcing, 109
SIR model, **150**
  time evolution, 177
  with demography, 153
slow manifold
  numerical tracking, 441
softening spring, 166
Solow growth model, **7**
  exact solution, 15
solution map, **386**
solution representation, 208
solution space, 95
  linear system, 205
source term, 47
spectral criterion for stability, 237
spiral
  stable, **118**
  unstable, **118**
spring–mass system
  coupled, 190
  fixed-end chain, 191
  free-end chain, 192
spring–mass–dashpot system, **90**
stability
  derivative test, **12**
  error propagation, 422
  Hamiltonian equilibrium, 347
  local, **12**
  Lyapunov, **470**
  numerical, **421**
  on a fixed time interval, **422**
  without attraction, 474
stability function, **423**
stability region, **423**
stable curve, 131
stable manifold, *see* stable curve
  basin boundary, 457
standard oscillator equation, **93**
state transition matrix, *see* state-transition matrix
state variable, **4**
state-space formulation, 115, 201
state-transition matrix, **208**, 245
step function, *see* Heaviside function
step response, 107
stiff decay, **425**
stiff equation, 441
stiffness, **425**
stiffness matrix, 191
Strang splitting, 256
stream function, 358
stretching and folding, 521
structure-preserving method, 443
superposition principle, 205
supremum norm, **373**
symmetry, 319
  Lorenz system, 516
symplectic flow

elementary examples, 341
symplectic form
preservation, 341, 432
symplectic map, **432**
symplectic matrix, **342**
symplectic method, **431**, 441
symplectic structure, **340**
synchronization, **109**
phase oscillators, 464
synchronized solution, **112**

time parametrization of orbit, **176**
time reversal, 335
loss under dissipation, 463
time-ordered product, 252
time-shift theorem, 278
total circulation
nonzero, 363
zero, 363
trace, 120
trace–determinant classification, **120**
trace–determinant plane, 125
transcritical bifurcation, **26**
translated parabolas, 69
translation symmetry, **320**
conserved momentum, 355
trap saturation, 184
trapping region
Lorenz system, 519
Van der Pol oscillator, 501
tridiagonal matrix, 196
turning point, 307

undamped oscillation, **100**
uniform convergence, **373**
uniqueness
Cauchy–Lipschitz theorem, 376
Gronwall estimate, 384
local, **371**
nonintersection of trajectories, 392
unit step function, *see* Heaviside function
unstable curve, 131
explicit example, 132
unstable manifold, *see* unstable curve

Van der Pol oscillator, **170**
limit cycle, 170
linearization, 170
periodic orbit, 501
variation, **296**
variation of constants, **48**, **104**, 209, 256
variation of parameters, *see* variation of constants
variation-of-parameters formula, *see* variation of constants
variational equation, **395**
along a periodic orbit, 397
relation to linearization, 395
variational principle, 295
variational problem, 295
isoperimetric, 301
phase-field model, 312
shortest curve, 298
vector field, **128**
planar, 56
transport by the flow, 396
voltage drop, 192
volume contraction, 349, **518**
volume-preserving flow, 347, 397
vortex dipole, 363
vorticity, 396

washout equilibrium, 182
Wronskian, **95**, **206**

zero eigenvalue
nonlinear behavior, 134
zero-dissipation set, **481**
pendulum, 483