MATHS 765
Mathematical Modelling

Course Notes, Weeks 1–6

Dimensional analysis, scaling, asymptotic reduction,
and the construction of models

JP-E
University of Auckland

Draft v1 — August 10, 2026

Contents

Preface

These notes cover the first half of the Mathematical Modelling (MATHS 765) course.

Structure of the course

The first half of the course is structured around four modelling skills:

1. strip a problem to dimensionless form and read off what governs it (Chapters 2–3);
2. decide what is small or large, and reduce the model to something tractable, knowing where the reduction breaks (Chapters 4–5);
3. understand where the equations came from in the first place, and read a bifurcation as a prediction (Chapter 6);
4. read, criticise, and improve a real model (Chapter 7).

Prerequisites

You are assumed to know differential equations, linear algebra, calculus, and enough numerical computation to integrate an ODE. Where a physical law is needed (a drag law, a cooling law) we quote it and get on with the mathematics.

How to use these notes

Read with a pen. Every worked example is meant to be redone by hand, and every "check" left to the reader really should be checked — that is where the understanding lives. Exercises at the end of each chapter range from routine drills to small open-ended modelling questions; the latter are marked with a dagger (†) and are good preparation for the assignments. A short list of hints for selected exercises appears at the end.

Starred sections (∗) are enrichment and may be skipped on a first reading without loss of continuity.

Notation

We write $[\,Q\,]$ for the dimension of a quantity Q, and use M, L, T, Θ for the fundamental dimensions of mass, length, time, and temperature. Dimensionless groups are written $\pi_1, \pi_2, \ldots$ A small positive parameter is ε, and $\mathcal{O}(\cdot)$, $o(\cdot)$ have their usual asymptotic meanings (recalled in Chapter 3). Dots denote time derivatives, primes denote derivatives with respect to whatever independent variable is in play at the time.

These are working notes and will contain errors and rough edges; corrections are welcome.

Chapter 1

Introduction: the modelling process

1.1 Models and modelling

By a *mathematical model* we mean an equation, or a system of equations, that describes some phenomenon drawn from science, engineering, economics, biology, or any other field. By *mathematical modelling* we mean the whole process by which such a model is formulated, analysed, and tested. It is useful to break that process into four stages:

1. identify the relevant quantities — the independent and dependent variables, and the parameters — needed to describe the phenomenon;
2. make explicit, model-specific assumptions about how those quantities are related;
3. solve the resulting equations, by analytic or numerical means;
4. compare the solution with real data and interpret the outcome.

Modelling is rarely a single pass through these stages. More often it is a *cycle* (Figure 1.1): a first model is formulated and solved, its predictions are held up against observation, the discrepancies reveal which assumptions were too crude, and a revised model is built. Progress comes from going around the loop, not from getting it right the first time.

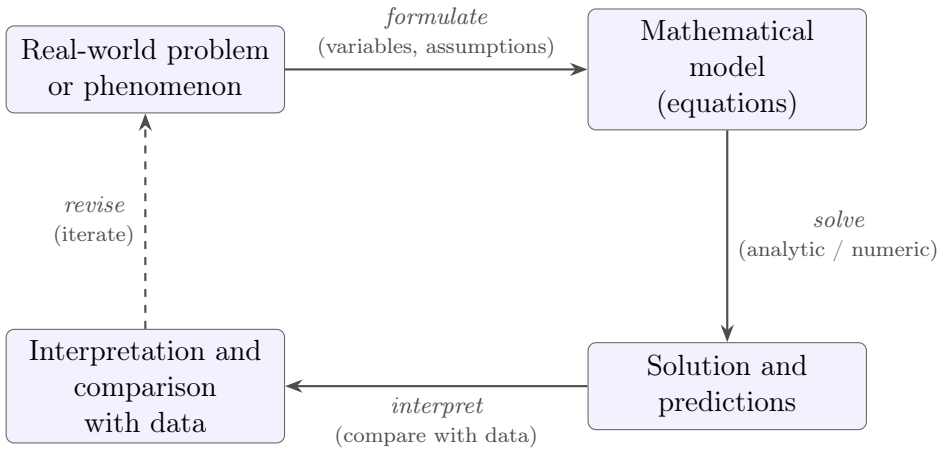


Figure 1.1: The modelling cycle. Formulation (stages 1–2) and solution (stage 3) are the focus of the first half of this course; validation against data (stage 4) closes the loop, and disagreement sends us back to revise the assumptions and go round again.

1.2 The character of modelling

Modelling draws on more than mathematics. It requires intuition about the area of application, judgement in choosing which effects to include, facility with methods of solution, and honesty in confronting the results with data. It aims to make sense of the world quantitatively as we observe it. In pursuit of that aim, scientific exactness is sometimes traded for mathematical tractability: a model that captures the essential mechanism and can actually be solved is usually more useful than an exhaustive one that cannot.

Because a model rests on assumptions, its predictions depend on them, and changing the assumptions changes the model. This is not a defect but a source of insight. When a conclusion survives a change in some assumption, we say the model is *robust* to that assumption; when it does not, we have learned that the assumption is doing real work. In this way, modelling tells us which parameters and processes matter and which do not — often, it's the most valuable output. A sound working principle is therefore to *start simple* and to build in complexity only as the data or the questions demand it. Much of this first half of the course is a toolkit for this decision: scaling, dominant balance, and asymptotic reduction are the means by which we decide, defensibly, what to keep and what to discard.

1.3 Kinds of model

Models come in many types, and it helps to know where a given model sits. Common distinctions include *deterministic* versus *stochastic* (does chance enter the description explicitly?); *continuous* versus *discrete* (do the variables vary smoothly, or in steps?); *dynamic* versus *static* (does the model describe evolution in time, or an equilibrium?); *quantitative* versus *qualitative* (do we want an analytical solution, or the behaviour of the solutions?); and *explanatory* versus *descriptive* (does the model aim at an underlying mechanism, or merely summarise the data?).

The subject is far too large for one course, so we deliberately choose a selection of examples to illustrate different models.

Chapter 2

Dimensional analysis and the Buckingham Pi theorem

Before writing a single equation, one can often extract a surprising amount of information about a model from a fact so obvious it is easy to overlook: the two sides of any physical law must have the same dimensions. This chapter turns that observation into a systematic tool. It is the ideal place to begin, because it delivers real modelling power with essentially no physical derivation.

2.1 Dimensions, units, and homogeneity

Every physical quantity has a *dimension*: the *kind* of thing it is — a length, a time, a mass. To attach a number to a quantity we must also choose a *unit*, a reference amount of that kind; a length, for instance, may be measured in metres, feet, or miles. The choice of unit is an arbitrary human convention and can be changed at will: the same length is 2 m or about 6.6 ft — the number shifts with the unit, while the length itself does not. The dimension, by contrast, is intrinsic: the quantity is a length however we choose to measure it. Dimensional analysis works at the level of dimensions, and never units, precisely because the dimension is the part of a quantity that does not depend on this arbitrary choice. Mechanics is built on a small set of *fundamental* (or *base*) dimensions — mass, length, time, and temperature — each with a corresponding base unit in the SI system (Table 2.1).

Dimension	Symbol	SI base unit	Symbol
Mass	M	kilogram	kg
Length	L	metre	m
Time	T	second	s
Temperature	Θ	kelvin	K

Table 2.1: The fundamental (base) dimensions used throughout these notes, with their SI base units. Other base dimensions exist — notably electric current (ampere, A) — and are introduced only as needed.

Every other physical quantity has a dimension that is a product of powers of these base dimensions, dictated by its defining relation. For example, speed is distance travelled per unit time, so a velocity $v(t)$ has dimension LT^{-1}; acceleration is the rate of change of velocity, giving LT^{-2}; and force, being mass times acceleration, has dimension $M \cdot LT^{-2} = MLT^{-2}$. Continuing in this way one builds up the dimensions of all the derived quantities of mechanics (Table 2.2).

We write $[\,Q\,]$ for the dimension of a quantity Q; thus $[\,\text{force}\,] = MLT^{-2}$. A quantity with $[\,Q\,] = 1$ (all exponents zero) is *dimensionless*: pure numbers, angles, and ratios of like quantities are all dimensionless.

Quantity	Defining relation	Dimension	SI unit
Velocity (v)	ℓ/t	LT^{-1}	$\mathrm{m\,s^{-1}}$
Acceleration (a)	v/t	LT^{-2}	$\mathrm{m\,s^{-2}}$
Momentum (U)	mv	MLT^{-1}	$\mathrm{kg\,m\,s^{-1}}$
Mass density (ρ)	m/V	ML^{-3}	$\mathrm{kg\,m^{-3}}$
Force (F)	ma	MLT^{-2}	newton (N)
Energy (E), work (W)	$F\,\ell$	ML^2T^{-2}	joule (J)
Power (P)	E/t	ML^2T^{-3}	watt (W)
Pressure (P), stress (σ)	F/A	$ML^{-1}T^{-2}$	pascal (Pa)
Frequency (f)	$1/t$	T^{-1}	hertz (Hz)
Viscosity (μ)	$\sigma/(v/\ell)$	$ML^{-1}T^{-1}$	pascal-second (Pa·s)

Table 2.2: Some derived quantities, with a defining relation, their dimensions in terms of the base dimensions, and their SI units. Note that energy and work share the same dimension, as do pressure and stress.

Definition 2.1 (Dimensional homogeneity)**.** An equation is *dimensionally homogeneous* if every additive term in it has the same dimension.

Principle 2.2 (Dimensional homogeneity)**.** *Any equation expressing a relationship between physical quantities, valid independently of the choice of units, must be dimensionally homogeneous.*

This innocuous-looking principle has sharp consequences. You may never add a length to a time. The argument of any transcendental function — exp, sin, log — must be dimensionless, since its power-series expansion sums terms of different powers of its argument. And a dimensionless equation can always be rearranged so that every quantity appearing is itself dimensionless; this is the engine of everything below.

Example 2.3 (Reading off the dimension of a physical constant)**.** Newton's law of gravitation $F = Gm_1m_2/r^2$ must be homogeneous. Since $[\,F\,] = MLT^{-2}$ and $[\,m_1m_2/r^2\,] = M^2L^{-2}$,

$$[\,G\,] = \frac{MLT^{-2}}{M^2L^{-2}} = M^{-1}L^3T^{-2}.$$

We have learned the dimensions of the gravitational constant without any experiment. Note that here 'constant' means it does not vary between systems, not that it is a pure number. We will distinguish *dimensionless constants* like 2π from *dimensional (physical) constants* like G or the speed of light.

Example 2.4 (Newton's second law from dimensional reasoning)**.** Suppose we know only that some physical law relates the force F on a body, its mass m, and its acceleration a — that is, $f(F, m, a) = 0$ for an unknown function f — without knowing the form of the law. From the dimensions $[\,F\,] = MLT^{-2}$, $[\,m\,] = M$, and $[\,a\,] = LT^{-2}$ we can form the dimensionless combination

$$\pi = \frac{ma}{F}, \qquad [\,\pi\,] = \frac{M \cdot LT^{-2}}{MLT^{-2}} = 1.$$

Now make a plausible assumption: because the law must hold whatever units we use, it can be expressed using dimensionless quantities alone, and so is equivalent to a relation $\mathcal{F}(\pi) = 0$. If this equation has a root, then π must equal a constant, $ma/F = k$, and therefore

$$\boxed{F = C\,m\,a}, \qquad C = 1/k.$$

Dimensional reasoning *alone* has told us that force is proportional to mass and to acceleration; the single constant C remains to be fixed by experiment — and the choice $C = 1$ is precisely what defines the unit of force, the newton.

The assumption used in this last example — that a physical law can always be recast as a relation among dimensionless quantities alone — is exactly the content of the Buckingham Π theorem, to which we now turn.

2.2 The Buckingham Pi theorem

Suppose a problem involves n dimensional quantities $q_1, \dots, q_n$, linked by some law

$$f(q_1, \dots, q_n) = 0 \tag{2.1}$$

whose form we do not yet know. Let k be the number of fundamental dimensions involved. The Buckingham Pi theorem says that (2.1) is really a relationship between far fewer *dimensionless* quantities.

To make this precise, seek dimensionless products

$$\pi = q_1^{p_1} q_2^{p_2} \cdots q_n^{p_n}.$$

Writing $[\,q_j\,] = M^{m_j} L^{\ell_j} T^{t_j} \cdots$, the product π is dimensionless iff the exponent of each fundamental dimension vanishes:

$$\sum_{j=1}^{n} m_j p_j = 0, \quad \sum_{j=1}^{n} \ell_j p_j = 0, \quad \sum_{j=1}^{n} t_j p_j = 0, \ \dots \tag{2.2}$$

This is a homogeneous linear system $A\,\mathbf{p} = \mathbf{0}$, where the *dimension matrix* A (of size $k \times n$) has the exponent vectors of the q_j as its columns.

Before stating the theorem, it is worth watching this machinery reproduce, mechanically, the group we guessed by eye in Example 2.4. There the dimensional quantities were $q_1 = F$, $q_2 = m$, $q_3 = a$ (so $n = 3$), and the unknown law $f(F, m, a) = 0$ played the role of the relation $f(q_1, \dots, q_n) = 0$ in (2.1). Rather than *spot* the dimensionless combination, let us *derive* it. Seek every dimensionless product

$$\pi = F^{p_1}\, m^{p_2}\, a^{p_3}$$

Substituting the dimensions $[\,F\,] = MLT^{-2}$, $[\,m\,] = M$, $[\,a\,] = LT^{-2}$ and collecting powers,

$$[\,\pi\,] = \big(MLT^{-2}\big)^{p_1} \big(M\big)^{p_2} \big(LT^{-2}\big)^{p_3} = M^{\,p_1+p_2}\; L^{\,p_1+p_3}\; T^{\,-2p_1-2p_3}.$$

For π to be dimensionless the exponent of each fundamental dimension must vanish — and *this is precisely where the equations in* (2.2) *come from*: one equation per dimension,

$$M:\ p_1 + p_2 = 0, \qquad L:\ p_1 + p_3 = 0, \qquad T:\ -2p_1 - 2p_3 = 0.$$

In matrix form these are $A\,\mathbf{p} = \mathbf{0}$ with

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ -2 & 0 & -2 \end{pmatrix}, \qquad \mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix},$$

whose three columns are just the dimension vectors of F, m, and a. The third row is -2 times the second, so the T-equation repeats the L-equation and carries no new information; the rank is therefore $r = 2$. Solving the two independent equations gives $p_2 = -p_1$ and $p_3 = -p_1$, so every solution is a multiple of $(1, -1, -1)$. Choosing $p_1 = 1$,

$$\pi = F^1 m^{-1} a^{-1} = \frac{F}{ma},$$

which is the reciprocal of the group ma/F we guessed earlier — the same dimensionless quantity. The solution space is one-dimensional, so there is exactly *one* independent group, matching the count $n - r = 3 - 2 = 1$.

Two lessons generalise from this. First, the abstract-looking sums in (2.2) are nothing more than "the total exponent of M must cancel, and likewise for L and for T", assembled into a linear system; finding dimensionless groups *is* finding the null space of the dimension matrix. Second, note the pattern that decides how much dimensional analysis can tell us: when a single group survives, as here, the reduced law $\mathcal{F}(\pi) = 0$ pins that group to a constant; when *several* groups survive — as for the pendulum in Example 2.8, where two remain — the reduced law is an undetermined *function* of them, and the method leaves correspondingly more to be filled in.

Theorem 2.5 (Buckingham Π)**.** *Let* $r = \operatorname{rank} A$*. Then the relation* (2.1) *is equivalent to a relation*

$$\mathcal{F}(\pi_1, \pi_2, \ldots, \pi_{n-r}) = 0$$

among $n - r$ *independent dimensionless products* $\pi_1, \ldots, \pi_{n-r}$*.*

Proof sketch. The dimensionless products correspond exactly to the null space of the dimension matrix A, by (2.2). By the rank–nullity theorem that null space has dimension $n-r$, so there are $n - r$ independent dimensionless products, and any dimensionless product is a power-product of a chosen basis $\pi_1, \ldots, \pi_{n-r}$. Because (2.1) is unit-independent (Principle 2.2), it can be written purely in terms of dimensionless quantities, hence in terms of the π_i; this yields $\mathcal{F}(\pi_1, \ldots, \pi_{n-r}) = 0$. A careful proof also uses that one may solve for the "dependent" variable, which we omit. ■

The theorem is constructive once we know how to build the π_i. The oldest practical route is:

Recipe 2.6 (Rayleigh's method of indices)**.** To express a target quantity q_0 in terms of others $q_1, \ldots, q_m$:

1. write $q_0 = C\, q_1^{p_1} \cdots q_m^{p_m}$ with C dimensionless;
2. substitute dimensions and demand homogeneity, one fundamental dimension at a time;
3. solve the resulting linear equations for the exponents p_i (some may remain free, giving dimensionless groups);
4. collect the result, leaving any undetermined dimensionless function where the analysis cannot reach.

2.3 Worked examples

Before the examples, it helps to separate two jobs that are easily confused. The Buckingham theorem tells us *how many* independent dimensionless groups a problem has — the number $n-r$ — and that the law reduces to a relation among them. It does *not* hand us the groups themselves. Constructing them is a separate task, done either by Rayleigh's method of indices (Recipe 2.6) or, when several groups are present, by the closely related *repeating-variable* procedure below.

One subtlety recurs and is worth stating once. When Rayleigh's method is applied by writing a target quantity as a single power-product times a constant C, that "constant" is constant only with respect to the *dimensional* variables; it may still depend on any leftover *dimensionless* group. So whenever a problem has more than one group, the C of Rayleigh's method is really an undetermined *function* of the remaining groups. Watch for this in the pendulum.

Recipe 2.7 (Finding all the groups: repeating variables)**.** Given n quantities whose dimension matrix has rank r, so that there are $n - r$ groups:

1. choose r "repeating" variables that between them carry all r dimensions and are dimensionally independent (form no dimensionless product on their own);

2. combine each of the remaining $n - r$ variables, one at a time, with the repeating set as a power-product, and fix the exponents by homogeneity;
3. each such combination is one dimensionless group, and together they form a basis.

Different admissible choices of repeating variables give different but equivalent bases — which is why the groups of a problem are never unique.

Example 2.8 (The period of a pendulum, before any equation)**.** What can determine the period $\mathcal{T}$ of a simple pendulum? Plausibly its length L, the gravitational acceleration g, the bob mass m, and the (dimensionless) initial amplitude θ_0; so we posit a law $\mathcal{T} = f(L, g, m, \theta_0)$, with

$$[\,\mathcal{T}\,] = T, \quad [\,L\,] = L, \quad [\,g\,] = LT^{-2}, \quad [\,m\,] = M,$$

and θ_0 already dimensionless.

Counting the groups. Form the dimension matrix, with the four dimensional quantities as columns and M, L, T as rows:

$$A = \begin{array}{c} \\ M \\ L \\ T \end{array} \begin{array}{c} \begin{array}{cccc} \mathcal{T} & L & g & m \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & -2 & 0 \end{pmatrix} \end{array}.$$

Its rank is $r = 3$, meaning the three *rows* — the fundamental dimensions M, L, T — are linearly independent as vectors: none is redundant, each genuinely does work in the problem. (This is a statement about the rows of A, not about the quantities; the *columns* may perfectly well be dependent — indeed here g's column equals L's minus twice T's, since $[\,g\,] = [\,L\,][\,T\,]^{-2}$, and it is exactly such a column relation that yields a dimensionless group.) Since $r = 3 = k$, there is $n - r = 4 - 3 = 1$ independent group among the dimensional quantities; together with the already-dimensionless θ_0 that makes *two* groups in all. Note in passing that m is the only quantity carrying mass, so it can belong to no dimensionless group and must drop out entirely — a conclusion obtained for free.

Constructing the group. By Rayleigh's method (Recipe 2.6) we look for a relation of the form $\mathcal{T} = C\, L^{p_1} g^{p_2}$, with C dimensionless and m omitted (it cannot appear). Taking the dimension of each side — using $[\,\mathcal{T}\,] = T$, $[\,L\,] = L$, $[\,g\,] = LT^{-2}$ — the constant drops out and we are left with a purely dimensional equation:

$$[\,\mathcal{T}\,] = [\,L\,]^{p_1}[\,g\,]^{p_2} \qquad \Longrightarrow \qquad T = L^{p_1}\left(LT^{-2}\right)^{p_2} = L^{\,p_1+p_2}\,T^{\,-2p_2}.$$

Both sides are now products of the base dimensions L and T, and the left-hand side is L^0T^1; matching the exponent of each dimension separately,

$$L: \quad p_1 + p_2 = 0, \qquad\qquad T: \quad -2p_2 = 1,$$

gives $p_2 = -\tfrac{1}{2}$ and $p_1 = \tfrac{1}{2}$. Hence, $\mathcal{T} = C\, L^{1/2} g^{-1/2} = C\sqrt{L/g}$, and rearranging shows the dimensionless group to be $C = \mathcal{T}\sqrt{g/L}$ (one checks $[\,\mathcal{T}\sqrt{g/L}\,] = 1$).

By assuming a single power-product times a constant, we quietly ignored the second group θ_0. Since the problem in fact has *two* groups, that "constant" need not be a pure number: it may depend on the leftover dimensionless group θ_0. Making the dependence explicit, $C = \Phi(\theta_0)$, so

$$\boxed{\mathcal{T} = \sqrt{L/g}\;\Phi(\theta_0)}$$

for some undetermined function Φ. This is precisely the subtlety flagged above — Rayleigh's "constant" is really a function of the remaining groups — and it is Buckingham that warned us a second group was there to worry about. Equivalently, with two groups $\pi_1 = \mathcal{T}\sqrt{g/L}$ and $\pi_2 = \theta_0$, the theorem gives $\mathcal{F}(\pi_1, \pi_2) = 0$, or $\pi_1 = \Phi(\pi_2)$.

Three lessons: (i) the mass is irrelevant — a genuine prediction; (ii) $\mathcal{T} \propto \sqrt{L/g}$ is logically dictated by dimensional homogeneity; (iii) all the amplitude dependence hides inside Φ, which dimensional analysis cannot reach. We will *compute* Φ in Chapter 4 and find the period increases with amplitude.

Example 2.9 (Terminal velocity of a sphere)**.** A sphere of radius a falls at its steady terminal velocity U through a fluid of density ρ and viscosity μ, driven by a reduced gravity $\tilde{g}$ (of dimension LT^{-2}; it packages the buoyant weight and is treated as one given quantity). We posit $U = f(a, \tilde{g}, \rho, \mu)$, with

$$[\,U\,] = LT^{-1}, \quad [\,a\,] = L, \quad [\,\tilde{g}\,] = LT^{-2}, \quad [\,\rho\,] = ML^{-3}, \quad [\,\mu\,] = ML^{-1}T^{-1}.$$

Counting the groups. There are $n = 5$ quantities, and the three rows M, L, T of the dimension matrix are linearly independent, so its rank is $r = 3 = k$. Hence there are $n - r = 5 - 3 = 2$ groups, and the law reduces to a relation between *two* of them. (Contrast the pendulum, which had one dimensional group plus θ_0; here both groups are built from dimensional variables.)

Constructing the groups. We use the repeating-variable procedure (Recipe 2.7). Take a, ρ, μ as the repeating variables — they carry L, M, T independently — and combine them in turn with the two remaining variables, U and $\tilde{g}$.

Combining with U: seek $\pi_1 = U\, a^{p_1} \rho^{p_2} \mu^{p_3}$ dimensionless. Then

$$[\,\pi_1\,] = (LT^{-1})(L)^{p_1}(ML^{-3})^{p_2}(ML^{-1}T^{-1})^{p_3} = M^{\,p_2+p_3}\, L^{\,1+p_1-3p_2-p_3}\, T^{\,-1-p_3},$$

so $p_2 + p_3 = 0$, $-1 - p_3 = 0$, $1 + p_1 - 3p_2 - p_3 = 0$, giving $p_3 = -1,\ p_2 = 1,\ p_1 = 1$ and

$$\pi_1 = \frac{\rho U a}{\mu} \equiv \mathrm{Re} \qquad \text{(the Reynolds number)}.$$

Combining with $\tilde{g}$: seek $\pi_2 = \tilde{g}\, a^{p_1} \rho^{p_2} \mu^{p_3}$; the same bookkeeping gives $p_3 = -2,\ p_2 = 2,\ p_1 = 3$ and

$$\pi_2 = \frac{\rho^2 \tilde{g}\, a^3}{\mu^2} \equiv \mathrm{Ga} \qquad \text{(the Galileo number)}.$$

These are the "two convenient" groups: *convenient* because the choice of repeating variables was ours, and a different admissible choice would give a different but equivalent pair — Re and Ga are simply the two with standard names and clear meaning. By Buckingham the law is $\mathcal{F}(\mathrm{Re}, \mathrm{Ga}) = 0$, which we solve for Re:

$$\mathrm{Re} = \Phi(\mathrm{Ga}) \qquad \Longleftrightarrow \qquad U = \frac{\mu}{\rho a}\,\Phi(\mathrm{Ga}).$$

Closing the problem. With *two* groups, dimensional analysis leaves the function Φ undetermined; only physics or experiment can pin it down. The details lie outside this part of the course, but the qualitative outcome is worth recording: a slow, viscous descent and a fast, inertial descent give *different* forms of Φ, so that $U \propto a^2$ in the one regime and $U \propto a^{1/2}$ in the other. The single relation $\mathrm{Re} = \Phi(\mathrm{Ga})$ accommodates both; which applies is a matter of physics, and the way the terminal velocity depends on radius changes character between the two.

Example 2.10 (Taylor's blast wave: estimating the energy of an explosion)**.** A very intense point explosion releases energy E into air of density ρ, producing a roughly spherical shock of radius R at time t. Posit $R = f(E, \rho, t)$, with

$$[\,E\,] = ML^2T^{-2}, \quad [\,\rho\,] = ML^{-3}, \quad [\,t\,] = T, \quad [\,R\,] = L.$$

There are $n = 4$ quantities and, the three rows M, L, T of the dimension matrix being independent, rank $r = 3 = k$, so there is $n - r = 1$ group. With a single group the reduced law forces that group to be constant, so dimensional analysis will essentially solve the problem outright.

By Rayleigh's method write $R = C\,E^{p_1}\rho^{p_2}t^{p_3}$ and demand homogeneity:

$$L = (ML^2T^{-2})^{p_1}(ML^{-3})^{p_2}(T)^{p_3} = M^{\,p_1+p_2}L^{\,2p_1-3p_2}T^{\,-2p_1+p_3},$$

so $p_1 + p_2 = 0$, $2p_1 - 3p_2 = 1$, $-2p_1 + p_3 = 0$, giving $p_1 = \frac{1}{5},\ p_2 = -\frac{1}{5},\ p_3 = \frac{2}{5}$. Hence

$$\boxed{R = C\left(\frac{Et^2}{\rho}\right)^{1/5}}$$

Here C genuinely is a pure constant — there is no second group for it to depend on; it is of order one and can be computed from the full gas dynamics, but taking $C \approx 1$ is enough for an estimate. A single group has fixed the entire time-dependence of the shock — and, remarkably, this lets one *measure* the energy E. Taking logarithms of $R = C(Et^2/\rho)^{1/5}$,

$$\log R = \tfrac{2}{5}\log t + \tfrac{1}{5}\log\frac{E}{\rho} + \log C,$$

so a plot of $\log R$ against $\log t$, assembled from a sequence of photographs of the expanding fireball, should be a straight line. Its *slope* should be $\frac{2}{5}$ — a direct check that the model is right — while its *intercept* equals $\frac{1}{5}\log(E/\rho) + \log C$. Reading that intercept off the graph, and inserting the known air density ρ together with $C \approx 1$, one solves for E. This is how G. I. Taylor famously estimated the yield of a nuclear test from published high-speed photographs, with no access to the classified physics.

Example 2.11 (Capillary rise)**.** A liquid of density ρ and surface tension σ (with $[\,\sigma\,] = MT^{-2}$) rises to height h in a narrow tube of radius r under gravity g, wetting the wall at contact angle θ (dimensionless). Posit $h = f(\sigma, \rho, g, r, \theta)$, with

$$[\,h\,] = L,\quad [\,\sigma\,] = MT^{-2},\quad [\,\rho\,] = ML^{-3},\quad [\,g\,] = LT^{-2},\quad [\,r\,] = L.$$

Counting. Five dimensional quantities; the rows M, L, T of the dimension matrix are independent, so rank $r = 3 = k$ and there are $5 - 3 = 2$ groups; with the dimensionless θ that is *three* groups in all.

Constructing. Take ρ, g, r as repeating variables (Recipe 2.7) and combine with the remaining h and σ. Requiring each product dimensionless gives

$$h\,\rho^{p_1}g^{p_2}r^{p_3} :\ \frac{h}{r}, \qquad\qquad \sigma\,\rho^{p_1}g^{p_2}r^{p_3} :\ \frac{\sigma}{\rho g r^2} = \frac{1}{\mathrm{Bo}},$$

where $\mathrm{Bo} = \rho g r^2/\sigma$ is the *Bond number*, measuring gravity against surface tension. With θ as the third group, the law reduces to

$$\frac{h}{r} = \Psi(\mathrm{Bo}, \theta)$$

for an undetermined Ψ. (Strictly, the group that emerged is $1/\mathrm{Bo}$, the reciprocal of the Bond number; since any function of $1/\mathrm{Bo}$ is equally a function of Bo, either serves.) Once again dimensional analysis supplies the *structure* but not the function.

2.4 Scaling laws and self-similarity*

When a single dimensionless group controls a problem, the solution is forced to be a *power law*, as in Taylor's $R \propto t^{2/5}$. Such power laws — also called *scaling laws*, since they say how one quantity scales with another — are ubiquitous. Consider a diffusion process with diffusivity D. Its dimension follows from the diffusion equation $\partial_t u = D\,\partial_{xx}u$: matching $[u]/T$ against $[D]\,[u]/L^2$ gives $[\,D\,] = L^2T^{-1}$. On an unbounded domain the problem has *no intrinsic length*

— nothing in its ingredients, D and the running time t, is itself a length — so the only length one can build is $\sqrt{Dt}$ (indeed $[\,Dt\,] = L^2$, so $[\,\sqrt{Dt}\,] = L$). Since there is no other length for it to be proportional to, any distance the process produces must scale with this one: one predicts a spreading front $\sim \sqrt{Dt}$ before solving any PDE, growing like the square root of time. (Give the problem an intrinsic length — a finite container, an initial blob of definite size — and a second group appears, so the answer becomes $\sqrt{Dt}$ times an undetermined function of the ratio; the clean law is special to the unbounded case.) Solutions that depend on the independent variables only through such dimensionless combinations are called *self-similar*; they reduce a PDE to an ODE in the similarity variable. This is the gateway to Barenblatt's theory of intermediate asymptotics, beyond our scope here but well worth knowing exists.

Exercises

Exercise 2.1. A mass m on a spring of stiffness κ (with $[\,\kappa\,] = MT^{-2}$) oscillates with period $\mathcal{T}$. Use dimensional analysis to determine $\mathcal{T}$ up to a dimensionless constant. Why does g not appear?

Exercise 2.2. The speed c of deep-water gravity waves depends on wavelength λ, gravity g, and water density ρ. Show $c = \sqrt{g\lambda}\,\Phi$ for a constant Φ, and explain why ρ cannot appear.

Exercise 2.3. The drag force F on a car depends on its frontal area A, its speed v, the density ρ and viscosity μ of the air, and its geometric shape.

1. Show that the relationship can be written in the dimensionless form:

$$C_D = \frac{F}{\frac{1}{2}\rho v^2 A} = \Phi(\mathrm{Re}, \mathrm{GEOM})$$

   where C_D is the drag coefficient and GEOM accounts for the dimensionless geometric ratios of the car's shape.

2. Determine the expression for the dimensionless group Re (the Reynolds number) by using $\sqrt{A}$ as the characteristic length scale.

3. Explain, in physical modelling terms, why the drag coefficient C_D must depend only on the Reynolds number and the car's geometry, rather than on the velocity v alone.

Exercise 2.4. Kepler's third law relates a planet's orbital period $\mathcal{T}$ to its orbital radius a. If you attempt $\mathcal{T} = f(a, M_\odot, G)$ with $M_\odot$ the mass of the Sun, G the gravitational (dimensional) *constant* (see Example 2.3). Show that dimensional analysis forces $\mathcal{T}^2 \propto a^3/(GM_\odot)$. What goes wrong if you forget to include G?

Exercise 2.5. The penetration depth of heat diffusing for a time t with diffusivity D (where $[\,D\,] = L^2T^{-1}$) can involve no other scale. Argue that the only possible length is $\sqrt{Dt}$, and state what this predicts for how long it takes heat to penetrate a distance ℓ.

Exercise 2.6 (†). A raindrop of radius a falls at terminal velocity. Using Example 2.9, decide whether a 1 mm drop is likely to be in the Stokes or the Newton regime (you may look up representative values of ρ, μ for air and water), and hence predict how terminal velocity scales with drop size in that regime. Comment on what your prediction assumes. *Hint: the regime is set by the Reynolds number: Stokes when* $\mathrm{Re} \ll 1$*, Newton when* $\mathrm{Re} \gg 1$.

Exercise 2.7 (†). Choose any everyday phenomenon (the pitch of a wine glass, the range of a thrown ball, the time for a kettle to boil) and carry out a dimensional analysis: list the variables you think matter, form the dimensionless groups, and state precisely what your analysis predicts and what it leaves undetermined. Then say how you would test the prediction.

Hints and answers

Ex. 1. Only m and κ can appear; $\mathcal{T} = 2\pi\sqrt{m/\kappa}$ (dimensional analysis fixes everything but the 2π). Gravity g enters no other variable, so it can form no group with $\mathcal{T}$.

Ex. 2. ρ is the sole carrier of mass and must drop out; the single remaining group gives $c \propto \sqrt{g\lambda}$.

Ex. 3. $\mathrm{Re} = \rho v\sqrt{A}/\mu$. Being itself dimensionless, C_D can depend only on other dimensionless quantities (the Reynolds number and the shape), never on v alone.

Ex. 4. Matching dimensions forces $\mathcal{T}^2 \propto a^3/(GM_\odot)$. Omit G and the three remaining quantities form no dimensionless group at all (no time can be built from a length and a mass): the method would report "no relation possible", flagging the missing variable.

Ex. 5. With $[\,D\,] = L^2T^{-1}$ the only length available is $\sqrt{Dt}$; reaching a depth ℓ therefore takes a time $t \sim \ell^2/D$.

Ex. 6. Estimate Re for a 1 mm drop: it comes out well above unity, so the drop is in the Newton regime and $U \propto a^{1/2}$.

Recommended reading

The best first companion to this chapter is Dym [2], whose Chapter 2 is devoted almost entirely to dimensional analysis and is written in much the same spirit as these notes; it is available online, and is the place to start. Similarly, Chapters 1.2 to 1.4 of Holmes [3] is a very good read that gives similar examples to the ones here. Logan [4] opens with very readable treatments of dimensional analysis. The classic short account of dimensional analysis and self-similarity is Barenblatt [1], especially its opening chapters.

References for this chapter

[1] G. I. Barenblatt. *Scaling, Self-Similarity, and Intermediate Asymptotics*. New York: Cambridge University Press, 1996.

[2] Clive L. Dym. *Principles of Mathematical Modeling*. 2nd ed. Amsterdam and Boston: Elsevier Academic Press, 2004. ISBN: 9780122265518.

[3] M. H. Holmes. *Introduction to the Foundations of Applied Mathematics*. Vol. 56. Texts in Applied Mathematics. New York: Springer, 2009.

[4] J. D. Logan. *Applied Mathematics*. 4th ed. Hoboken, NJ: Wiley, 2013.

Chapter 3

Nondimensionalisation and dominant balance

Dimensional analysis (Chapter 2) works before we have equations. Once we *do* have a model — a differential equation with parameters — the same spirit lets us clean it up: strip it to dimensionless form, expose the parameters that actually matter, and compare the sizes of terms so we can justify throwing some away. This last step, *dominant balance*, is an important habit in mathematical modelling.

3.1 Why nondimensionalise

There are three payoffs. First, *parameter reduction*: a model with many dimensional constants usually depends on far fewer dimensionless combinations, so we solve one problem instead of a family. Second, *revealing small and large*: only a dimensionless parameter can honestly be called "small"; a dimensional quantity is small only relative to something, and nondimensionalisation makes that something explicit. Third, *comparability*: once every term is dimensionless, we may legitimately compare their sizes and ask which dominates.

Recipe 3.1 (Nondimensionalisation)**.**

1. Choose a characteristic *scale* for each variable (a reference length, time, velocity, etc), built from the parameters of the problem.
2. Introduce dimensionless variables, e.g. $\hat{x} = x/L_c$, $\hat{t} = t/t_c$.
3. Substitute, and divide through by the coefficient of one chosen term.
4. Read off the dimensionless groups that remain; interpret each.

The choice of scales is not unique, and different choices highlight different regimes. A scale built from the problem's own parameters (an *intrinsic* scale, like the terminal velocity of a falling body) is usually more revealing than one imposed from outside (an initial value, say).

3.2 Nondimensionalising differential equations

Example 3.2 (The pendulum: the small-angle model as a consequence)**.** The undamped pendulum — a bob of mass m on a light rigid rod of length L — obeys Newton's second law along its circular arc. The only force with a component along the motion is gravity, whose tangential component is $-mg\sin\theta$ (restoring, hence the minus sign), while the tangential acceleration is $L\ddot{\theta}$. Thus $mL\ddot{\theta} = -mg\sin\theta$, i.e.

$$L\ddot{\theta} + g\sin\theta = 0, \qquad \theta(0) = \theta_0,\ \dot{\theta}(0) = 0.$$

(The same equation follows from a torque balance about the pivot; the general rigid-body version, for an extended pendulum, is derived in Chapter 6.)

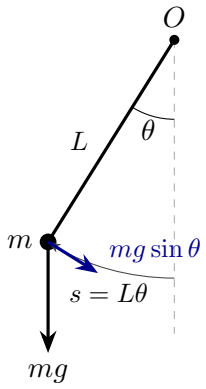


Figure 3.1: Forces on a simple pendulum. Gravity mg acts straight down; its component *along the arc* — perpendicular to the rod — is the restoring force $-mg\sin\theta$. The bob travels on a circle of radius L, so its position along the arc is $s = L\theta$ and its tangential acceleration is $\ddot{s} = L\ddot{\theta}$; balancing mass × acceleration against this force, $mL\ddot{\theta} = -mg\sin\theta$, gives the equation of motion.

The only time scale we can form from L, g and m is $t_c = \sqrt{L/g}$ (compare Example 2.8). Put $\hat{t} = t/t_c = t\sqrt{g/L}$ and use the chain rule; then $\ddot{\theta} = (g/L)\,\theta''$ (primes for $\mathrm{d}/\mathrm{d}\hat{t}$), and the equation becomes

$$\boxed{\theta'' + \sin\theta = 0}, \qquad \theta(0) = \theta_0,\ \theta'(0) = 0.$$

All the constants are gone, and the amplitude θ_0 is now the *only* parameter — and it is dimensionless, so it is meaningful to call it "small". Here comes a step that may be new. We *choose* to study small swings, that is, we *assume* the initial amplitude is small, $\theta_0 \ll 1$. This is a modelling decision, not a mathematical necessity; its reward is that the motion then stays small for *all* time. The reason is *conservation of energy*: a pendulum released from rest at angle θ_0 can never swing beyond θ_0 (to do so would require energy it does not have), so $|\theta(t)| \le \theta_0 \ll 1$ throughout the motion, not merely at the start. We make this energy argument precise in Chapter 4; here, the consequence is what matters. Because θ stays small, we may expand $\sin\theta = \theta - \frac{1}{6}\theta^3 + \cdots$ and keep only the leading term, giving the linear model.

$$\theta'' + \theta = 0.$$

So the small-angle approximation *follows* from a small-amplitude assumption together with energy conservation. Just as importantly, the expansion displays the very next term we discarded, $-\frac{1}{6}\theta^3$, whose effect on the period we will restore in Chapter 4.

Example 3.3 (The damped, driven pendulum)**.** Adding two familiar ingredients — *linear damping* (a resistive torque $-b\dot{\theta}$ that opposes the motion in proportion to the angular velocity, modelling air resistance or friction at the pivot, and draining energy) and a *periodic drive* (an external torque $F\cos\omega_d t$ that pumps energy in, as though the support were shaken back and forth at frequency ω_d) —,

$$mL\ddot{\theta} + b\dot{\theta} + mg\sin\theta = F\cos(\omega_d t),$$

and scaling time again by the intrinsic pendulum time $\sqrt{L/g}$ — the natural choice, since it is built from the *restoring term* $g\sin\theta$ (the gravitational term that pulls the pendulum back towards the bottom) and normalises its coefficient to one, leaving damping and forcing to show up in the remaining groups; a different scale (for instance the forcing period $1/\omega_d$) is equally legitimate and would foreground a different balance, as we warned above that scales are never unique — one obtains

$$\theta'' + \beta\,\theta' + \sin\theta = \gamma\cos(\Omega\hat{t}\,),$$

with three dimensionless groups: a damping number $\beta = b/(m\sqrt{gL})$, a forcing amplitude $\gamma = F/(mg)$, and a frequency ratio $\Omega = \omega_d\sqrt{L/g}$. Each has a plain meaning: β measures the strength of the damping relative to the natural swinging (how quickly the oscillations would die away), γ measures the drive against gravity's restoring pull (how hard the pendulum is being pushed), and Ω compares the frequency at which we are shaking the pendulum with the frequency at which it would naturally swing. The case $\Omega \approx 1$ — driving it at very nearly its own natural rhythm — is called *resonance*. It deserves the name: each push then arrives in step with the swing already underway, so the pushes accumulate rather than fight one another, and even a modest drive can build a very large amplitude. (It is the same effect as pushing a child on a swing: push at the right moment each cycle, and the swing grows dramatically; push at random moments, and it does not. In engineering, the effect is not always welcome — a structure driven near one of its natural frequencies can be shaken far harder than the size of the forcing would suggest.) These three dimensionless parameters — not the six original ones — are exactly the control parameters you would study in a Dynamical Systems course when it studies this system's bifurcations. (Verifying the groups with dimensional analysis is a good exercise.)

Remark 3.4 (The damped linear oscillator). Dropping the drive and taking small angles, the pendulum becomes a *linear damped oscillator* — the same equation that governs a mass on a spring with friction, a car's suspension absorbing a bump, or the current in an RLC electrical circuit — namely $\theta'' + \beta\theta' + \theta = 0$. It is customary to rewrite this in the *standard form*

$$x'' + 2\zeta x' + x = 0, \qquad \zeta = \tfrac{1}{2}\beta,$$

which is the *same* equation, but written using the conventional *damping ratio* ζ, which, by construction is non-negative. The rewrite makes ζ the natural measure of damping, because the eigenvalues come out as $-\zeta \pm \sqrt{\zeta^2 - 1}$, so the behaviour is read straight off the single threshold $\zeta = 1$. The character of the oscillator is then set entirely by ζ: *under-damped* ($\zeta < 1$, a decaying oscillation), *critically damped* ($\zeta = 1$), or *over-damped* ($\zeta > 1$, a monotone return) — the three cases classified by those eigenvalues, as in Appendix A.

Example 3.5 (A body thrown upward with linear drag: two balances in one equation)**.** A body of mass m, launched straight up with speed w_0, is slowed both by gravity and by air resistance. Taking the velocity w positive upward, Newton's second law reads

$$m\frac{\mathrm{d}w}{\mathrm{d}t} = -mg - cw, \qquad w(0) = w_0,$$

where g is the gravitational acceleration, $-mg$ the weight, and $-cw$ a linear drag with coefficient $c > 0$. Scale velocity by w_0 and time by the "gravity time" $t_g = w_0/g$. With $\hat{w} = w/w_0$, $\hat{t} = t/t_g$ one finds $\mathrm{d}w/\mathrm{d}t = g\,\hat{w}'$ and, dividing by mg,

$$\hat{w}' = -1 - \varepsilon\,\hat{w}, \qquad \hat{w}(0) = 1, \qquad \varepsilon = \frac{cw_0}{mg}.$$

The dimensionless group ε compares two *forces*: the drag at launch, cw_0, against the weight, mg. We use this differential equation to explain an important skill in applied mathematics: reduced models. We use this equation because it is analytically solvable. It is a linear differential equation with exact solution $\hat{w} = (1 + \frac{1}{\varepsilon})e^{-\varepsilon\hat{t}} - \frac{1}{\varepsilon}$. This will allow us to compare the exact solution with the approximations given by the method. In general, for other problems, no exact solution is possible, and we will need to understand the implications of reduced models.

We now *assume* $\varepsilon \ll 1$, this means that the drag force at launch is only a tiny fraction of the body's weight — physically, a not-too-fast throw through air, where air resistance is a minor effect beside gravity — and see what follows.

Now compare the *sizes* of the two terms on the right of the dimensionless equation. The first is exactly -1; the second is $-\varepsilon\hat{w}$. Which is bigger depends on how large $\hat{w}$ is, so we must ask what $\hat{w}$ actually does.

Early on. At the start we know $\hat{w}$ exactly, from the initial condition: $\hat{w}(0) = 1$. So at $\hat{t} = 0$ the velocity is of size one, since we chose to measure velocity in units of the launch speed w_0. Now substitute $\hat{w} = 1$ into the *full* equation

$$\hat{w}' = -1 - \varepsilon\hat{w} \quad \Longrightarrow \quad \hat{w}'(0) = -1 - \varepsilon \cdot 1 = -(1+\varepsilon) \approx -1,$$

since $\varepsilon \ll 1$. So the *rate of change* is also of size one. This too is a consequence of the scaling: we measured time in units of $t_g = w_0/g$, which is precisely the time gravity needs to change the velocity by an amount comparable to w_0.

Putting the two together: $\hat{w}$ starts at size one and changes at a rate of size one, so over a dimensionless time $\hat{t}$ of order one it changes by an amount of order one — and therefore *remains* of order one throughout this initial phase. (In dimensional terms, $\hat{w}' \approx -1$ says the velocity drops by roughly the launch speed w_0 in a time t_g; the ball is slowing appreciably, but has not yet done anything dramatic.) With $\hat{w} = \mathcal{O}(1)$, we can size the two terms on the right:

$$\underbrace{-1}_{\text{size 1}} \qquad \text{versus} \qquad \underbrace{-\varepsilon\hat{w}}_{\text{size } \varepsilon}.$$

The drag term is smaller than the gravity term by a factor of $\varepsilon \ll 1$ — a hundred times smaller if $\varepsilon = 0.01$. It is negligible *compared with* the -1, so we drop it:

$$\hat{w}' \approx -1 \;\Rightarrow\; \hat{w} \approx 1 - \hat{t} \qquad \text{(gravity-dominated)}.$$

This reduced model says the body decelerates under gravity *alone*, exactly as it would in a vacuum: the velocity falls at a constant rate, as though air resistance did not exist.

But look at what this reduced solution predicts: $\hat{w} \approx 1 - \hat{t}$ keeps *decreasing without bound.* (Physically: the body rises, stops, and then falls, gathering downward speed.) So $|\hat{w}|$ does not stay of order one—it grows. And once $|\hat{w}|$ has grown to be of order $1/\varepsilon$, the neglected term is no longer small:

$$\varepsilon\hat{w} \sim \varepsilon \cdot \tfrac{1}{\varepsilon} = 1,$$

which is the same size as the -1 we balanced against. The approximation has destroyed itself: *the reduced solution predicts its own breakdown*, and this is exactly where we should look to detect it. Beyond that point, a different pair of terms must balance. The only remaining possibility is that the two terms on the right cancel each other, leaving nothing for $\hat{w}'$ to be — that is, the body has stopped accelerating, $\hat{w}' \approx 0$. Setting the left-hand side to zero:

$$0 \approx -1 - \varepsilon\hat{w} \;\Rightarrow\; \hat{w} \to -\tfrac{1}{\varepsilon} \qquad \text{(drag-balanced: terminal velocity)}.$$

Both predictions are confirmed by the exact solution $\hat{w} = (1 + \frac{1}{\varepsilon})e^{-\varepsilon\hat{t}} - \frac{1}{\varepsilon}$: expanding it for small $\hat{t}$ recovers $1 - \hat{t}$, and letting $\hat{t} \to \infty$ gives $-1/\varepsilon$. Figure 3.2 shows all three curves together for $\varepsilon = 0.05$, and the division of labour between the two reduced models is plain. Early on, the straight line $1 - \hat{t}$ is indistinguishable from the exact solution: the body decelerates as if in a vacuum, crosses $\hat{w} = 0$ at the top of its flight near $\hat{t} = 1$, and begins to fall. But the straight line then keeps descending without bound, while the true solution bends away from it and flattens onto the horizontal line $\hat{w} = -1/\varepsilon = -20$. That bend is drag reasserting itself, exactly where the first model predicted its own failure — at $\hat{t} \sim 1/\varepsilon = 20$, when $|\hat{w}|$ has grown to order $1/\varepsilon$. Neither reduced model is good everywhere, and neither is wrong: each is the right description of its own regime, and together they account for the whole flight.

The lesson. One innocuous-looking equation has *two* different reduced models, each valid in its own regime and each with a clear physical meaning. Early on, gravity dominates and drag may be ignored: the body decelerates freely, as if in a vacuum. Late on, gravity and drag exactly balance: the body falls at a constant terminal velocity, $\hat{w} = -1/\varepsilon$ (dimensionally $w = -mg/c$ — weight balanced by drag, as one would expect). And the two regimes live on

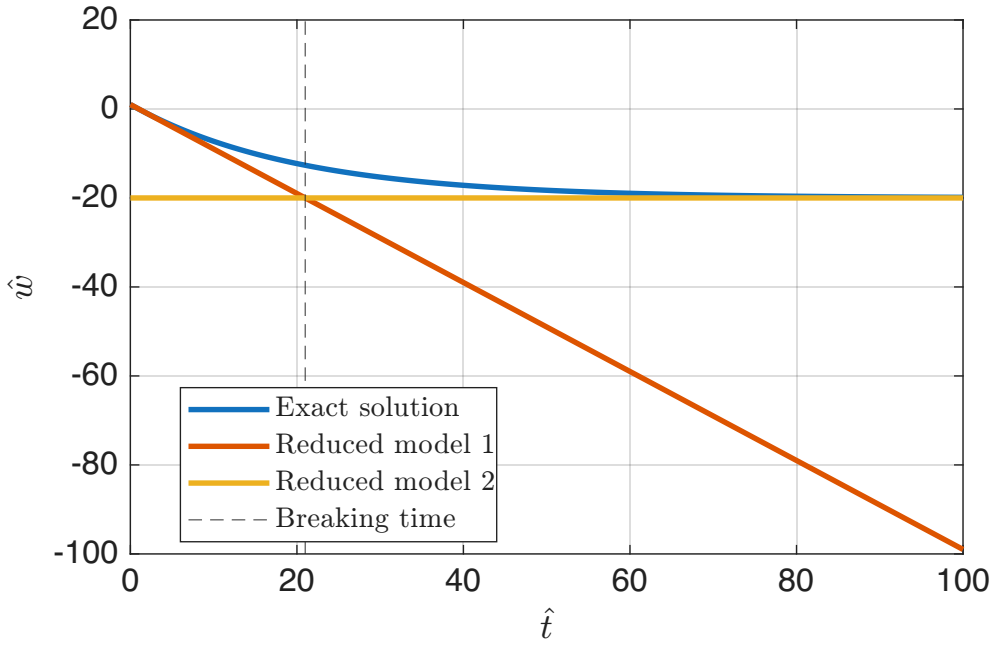


Figure 3.2: The thrown body with linear drag, for $\varepsilon = 0.05$. The exact solution $\hat{w} = (1 + 1/\varepsilon)e^{-\varepsilon\hat{t}} - 1/\varepsilon$ is shown with the two reduced models obtained by dominant balance: the gravity-dominated approximation $\hat{w} \approx 1-\hat{t}$, valid while $\hat{t} \ll 1/\varepsilon$, and the drag-balanced terminal velocity $\hat{w} \approx -1/\varepsilon$, approached only when $\hat{t} \gtrsim 1/\varepsilon$. Each captures one regime of the same equation; the crossover occurs where the neglected term $\varepsilon\hat{w}$ grows to the size of the retained -1.

different timescales: the first holds for $\hat{t} = \mathcal{O}(1)$, while the terminal state is only reached when $\hat{t} \sim 1/\varepsilon$, i.e. dimensionally $t \sim m/c$, a much longer time. So there is no single "simplest model" of a thrown ball — which simplification is legitimate depends on *when* you look. That is the lesson to carry forward.

3.3 Dominant balance and distinguished limits

Why this matters. A model worth analysing is usually a model too complicated to solve. The modeller's central move is therefore to *throw terms away* — but responsibly, keeping only those that matter and knowing when the remaining terms stop being the right ones. Dominant balance is the tool that makes this formal. It answers the three questions a modeller must be able to answer about any simplification: *which* terms may I drop, *why* is dropping them justified, and *when* does that justification not hold anymore? Example 3.5 did all three: it produced two different reduced models of a thrown ball, said exactly which forces each one keeps, and predicted the timescale on which the first gives way to the second. Every reduction in the rest of these notes — the small-angle pendulum, the quasi-steady-state approximations of Chapter 5, the closures of Chapter 6 — is an instance of this same move.

A bit of terminology. An equation is a sum of terms set to zero. A *balance* is a hypothesis that two (or more) particular terms are the largest ones present and are of comparable size, so that they must cancel one another. In contrast, all the remaining terms are smaller and may be neglected. The pair so chosen is the *dominant balance*. A given equation may admit *several* distinct balances, each valid in a different regime — as the thrown body did, having one balance early on ($\hat{w}'$ against -1) and a different one late (-1 against $-\varepsilon\hat{w}$). Those balances that pass the self-consistency test below are the *distinguished limits* (the standard name for them in asymptotics).

Methodology. Faced with an equation containing a small parameter, we *hypothesise* which

terms are the large ones, provisionally discard the rest, and solve the resulting reduced equation. We cannot know in advance that the discarded terms really are small — their size depends on the solution, which we do not yet have. That is precisely why the last step is to *check*: take the solution just found, substitute it back, and confirm that the neglected terms are indeed smaller than the ones retained. If the check succeeds, the balance is self-consistent, and the reduced model is justified. If it fails, the guess was wrong, and another pair of terms must be tried.

The simplest place to *learn* this technique is a single algebraic equation. The example below is not a model, but the rescaling 'trick' is exactly the one needed when a small parameter multiplies the highest derivative of a *differential* equation, producing the *boundary layers* of Chapter 5.

Example 3.6 (A quadratic and its lost root)**.** Consider

$$\varepsilon x^2 + x - 1 = 0, \qquad 0 < \varepsilon \ll 1.$$

Balance 1. There are three terms: εx^2, x, and -1. Hypothesise that the last two are the dominant pair, so that the first is negligible. Dropping εx^2 leaves $x - 1 \approx 0$, i.e. $x \approx 1$. Now *check* the hypothesis, which we could not do before having the solution: with $x \approx 1$ the discarded term is $\varepsilon x^2 \approx \varepsilon \cdot 1 = \varepsilon$, while the terms we kept are of size 1. Since $\varepsilon \ll 1$, the discarded term is indeed far smaller than the ones retained — the balance is self-consistent, and the reduced model is justified.

But the equation is quadratic and must have two roots, and Balance 1 has found only one. The second has escaped to large x, where dropping εx^2 was illegitimate: once x grows, the term carrying the small parameter need no longer be small. To find it, we *rescale*, magnifying the region where the missing root lives. Put $x = X/\delta$, where δ is to be chosen and $X = \mathcal{O}(1)$ — the point of the rescaling being to make the unknown of size one again, so that the sizes of the terms can be compared. Then

$$\frac{\varepsilon X^2}{\delta^2} + \frac{X}{\delta} - 1 = 0,$$

whose three terms have sizes ε/δ^2, $1/\delta$ and 1. Balancing the first two requires $\varepsilon/\delta^2 \sim 1/\delta$, that is $\delta = \varepsilon$; the pair is then of size $1/\varepsilon$ each, comfortably larger than the third, so the hypothesis survives its check. (Balancing the first against the third instead would give $\delta = \sqrt{\varepsilon}$, but the middle term would then be of size $1/\sqrt{\varepsilon}$ — *larger* than the pair supposedly dominating — so that hypothesis fails, exactly as the method requires.)

With $\delta = \varepsilon$, multiplying through by ε turns the equation into

$$X^2 + X - \varepsilon = 0.$$

This is still exact; the approximation enters, as always, by dominant balance — applied now in the new variable. The term $-\varepsilon$ is negligible beside the other two, and discarding it leaves

$$X^2 + X \approx 0, \qquad \text{that is} \qquad X(X+1) \approx 0 \quad \Longrightarrow \quad X \approx 0 \ \text{ or } \ X \approx -1.$$

Two candidates emerge, and each must face the same self-consistency test as before. Here, the test is against the premise of the rescaling: that X has size one.

The candidate $X \approx -1$ passes. It is of size one, as assumed, so $x = X/\varepsilon \approx -1/\varepsilon$ is genuinely large and the magnification was justified:

$$x \approx -\tfrac{1}{\varepsilon} \qquad \text{(the lost root)}.$$

The candidate $X \approx 0$ fails. It is *not* of size one, contradicting the assumption that the scaling was built on, so this scaling has no right to it. What has happened is that the magnification has blurred the root we already know. Since $X = \varepsilon x$, the root $x \approx 1$ of Balance 1 sits at $X \approx \varepsilon$, indistinguishable from 0 at leading order. The statement "$X \approx 0$" is therefore not false — it

says correctly that the remaining root is far smaller than $1/\varepsilon$ — but it is uninformative, and recovering the actual value requires the unscaled balance already carried out.

This is the general lesson, and it returns in Chapter 5: *each scaling resolves one root and blurs the others.* The unscaled variable x resolves the $\mathcal{O}(1)$ root and pushes the large one out of view; the magnified variable $X = \varepsilon x$ resolves the large root and collapses the $\mathcal{O}(1)$ one onto the origin. Neither view sees both, which is precisely why two balances are needed. When a rescaled equation offers a root at the origin, the reading is usually not 'a new solution' but 'a solution already found, seen from too far away'.

Both results can be confirmed by expanding the exact roots $x = \left(-1 \pm \sqrt{1+4\varepsilon}\right)/(2\varepsilon)$, which we will study in later chapters.

The connection with the thrown body. The two examples are the same phenomenon wearing different clothes. In each, a naive first look — "the ε term is small, drop it" — captures one behaviour and *silently loses another*, and in each the lost behaviour lives out at size $1/\varepsilon$, far beyond where the first approximation is valid. There, the thrown body's velocity grew until $\hat{w} \sim -1/\varepsilon$, at which point the neglected drag term $\varepsilon\hat{w}$ became as large as the gravity term and a *new* balance took over, giving the terminal velocity. Here x likewise runs out to $x \sim -1/\varepsilon$, at which point the neglected term εx^2 becomes as large as x, and a new balance takes over, giving the lost root. In both cases, the tell-tale is identical: *the small parameter multiplies the term that would otherwise be dominant when the unknown becomes large.* So whenever ε multiplies the highest power (as here) or, for a differential equation, the highest derivative, treat it as a warning that setting $\varepsilon = 0$ has thrown away an entire solution — a root here, and, in Chapter 5, a boundary layer. The cure is the same in every case: rescale, and find the balance that the naive limit destroyed.

Remark 3.7 (The systematic search for balances). Given a sum of terms, list them, and for each candidate pair hypothesise that *those two* dominate and are of the same order; deduce the implied size of the unknown, then verify that every other term is genuinely smaller. Only self-consistent hypotheses survive. This mechanical procedure rarely fails and is worth practising until it is automatic.

In modelling terms, each surviving balance *is* a reduced model — a legitimate simplification of the original, together with a statement of the regime in which it holds. The procedure, therefore, does more than simplify algebra: it enumerates the different physical regimes a model contains, tells you which effects matter in each, and warns you when a regime is about to end. That is why a modeller runs it before doing anything else, and why the rest of these notes lean on it so heavily.

3.4 Remarks

Dominant balance answers a single question: *which terms matter?* Its output is a *reduced model* — a simpler equation — together with the regime of validity in which that equation may be trusted. Notice what kind of object that is. It is a statement about the *equation*, not about the solution; and it produces *one* description: the leading-order one.

The obvious next question — *how accurate is that model, and how do I improve it?* — is not answered here. Answering it is the topic of Chapter 4, which takes the leading-order balance as the first term of a systematic expansion and computes corrections directly from the equation, order by order.

The ordering of the two chapters is not arbitrary, and it is worth seeing why this one must come first. You cannot expand about a leading-order model until you know which model that is — and, more sharply, you cannot know whether a straightforward expansion will work at all until you know whether the problem is *regular* or *singular*. The lost-root example already showed how a naive expansion can destroy an entire solution, and how rescaling recovers it.

Exercises

Exercise 3.1. Nondimensionalise the damped oscillator $m\ddot{x} = -kx - b\dot{x}$. Show that the equation reduces to $x'' + 2\zeta x' + x = 0$, and identify the single dimensionless damping ratio ζ in terms of m, k, b.

Exercise 3.2. For the projectile of Example 3.5:

(a) Using the gravity-dominated reduced model $\hat{w} \approx 1 - \hat{t}$, find the dimensionless time $\hat{t}_{\text{top}}$ at the top of the flight. Show that the natural height scale is w_0^2/g, so that the dimensionless height climbed is $\hat{h} = \int_0^{\hat{t}_{\text{top}}} \hat{w}\,\mathrm{d}\hat{t}$, and evaluate it. Check that your answer reproduces the familiar $h = w_0^2/2g$.

(b) Over what range of $\hat{t}$ is the approximation $\hat{w} \approx 1 - \hat{t}$ valid? Is the top of the flight inside that range? (Recall that the reduced model fails once the neglected term $\varepsilon\hat{w}$ grows to the size of the retained -1.)

Exercise 3.3. Find the leading behaviour of all roots of $\varepsilon x^3 + x^2 - 1 = 0$ as $\varepsilon \to 0^+$, following the method of Example 3.6: first the balance that keeps $x = \mathcal{O}(1)$, then a rescaling $x = X/\delta$ for the root that has escaped to large x. Check that you have found as many roots as a cubic must have, and say what becomes of the $\mathcal{O}(1)$ roots when viewed in the rescaled variable.

Exercise 3.4. Consider $\varepsilon\dot{y} = -y + \cos t$ on $t \geq 0$, with $y(0) = 0$ and $0 < \varepsilon \ll 1$ (everything is already dimensionless here).

(a) *Outer regime:* here y varies on the $\mathcal{O}(1)$ timescale, $\dot{y} = \mathcal{O}(1)$ and the term $\varepsilon\dot{y}$ is negligible. Write down the resulting balance, solve it, and check that it is self-consistent.

(b) Show that this outer solution cannot meet the initial condition, and conclude that a thin layer must sit at $t = 0$.

(c) *Inner regime:* here rescale $t = \delta\tau$ and choose δ so that $\varepsilon\dot{y}$ is restored to the same size as the other terms; verify that $\delta = \varepsilon$, so the regime has width $\mathcal{O}(\varepsilon)$. Write the inner equation, simplify $\cos(\varepsilon\tau)$ for $\tau = \mathcal{O}(1)$, and solve it.

(d) Check that your inner solution satisfies $y(0) = 0$ *and* tends, as $\tau \to \infty$, to the value the outer solution takes at $t = 0$.

(e) Explain why it is the *sign* of the $-y$ term that makes part (d) succeed, and say what would go wrong were it $+y$.

Exercise 3.5 (†). A spherical raindrop of radius a falls from rest under gravity through air, slowed by linear (Stokes) drag $-kv$ with $k = 6\pi\mu a$, where μ is the viscosity of air. Being a sphere of water, the drop has mass $m = \frac{4}{3}\pi a^3\rho_w$, with ρ_w the density of water.

(a) Write down the velocity equation. From the parameters alone, identify a velocity scale v_T — the terminal velocity, at which drag balances weight — and a time scale t_c.

(b) Nondimensionalise using these scales. You should find that *every* parameter disappears, leaving $\hat{v}' = 1 - \hat{v}$ with $\hat{v}(0) = 0$. What does the absence of any remaining parameter tell you about how drops of different sizes fall?

(c) Identify the two regimes: the balance holding for $\hat{t} \ll 1$ and the balance holding for $\hat{t} \gg 1$, and interpret each physically. State, in dimensional terms, the timescale on which the drop approaches terminal velocity.

(d) Substitute for m and k to express v_T in terms of a alone (together with ρ_w, g and μ). Then, for a drop of radius 0.5 mm, evaluate v_T and compare it with the observed fall speed of raindrops, about 4 m/s. Evaluate also the Reynolds number $\mathrm{Re} = \rho_{\text{air}} v_T a/\mu$, noting that this one is built from the density of the *air* the drop moves through, not of the water it is made of. Is the Stokes-drag assumption self-consistent here, and for what

drop sizes would it be? (Compare Chapter 2. Take $\rho_w \approx 10^3\,\mathrm{kg\,m^{-3}}$, $\rho_{\mathrm{air}} \approx 1.2\,\mathrm{kg\,m^{-3}}$ and $\mu \approx 1.8\times10^{-5}\,\mathrm{Pa\,s}$.)

Exercise 3.6 (†). Take a model from another course (a logistic growth law, an RC circuit, a cooling body) and nondimensionalise it two *different* ways, using two different choices of scale. Explain what regime each choice is adapted to, and which small parameter each exposes.

Hints and answers

Ex. 1. With $t = \sqrt{m/k}\,\hat{t}$ the equation becomes $x'' + \big(b/\sqrt{mk}\big)x' + x = 0$, so $\zeta = b/(2\sqrt{mk})$. The three regimes $\zeta < 1$, $= 1$, > 1 are those of Remark 3.4.

Ex. 2. (a) The top is where $\hat{w} = 0$, i.e. $\hat{t}_{\mathrm{top}} = 1$. For the height scale: $h = \int w\,\mathrm{d}t = w_0 t_g \int \hat{w}\,\mathrm{d}\hat{t}$, and $t_g = w_0/g$, so $h = (w_0^2/g)\,\hat{h}$ as claimed. Then $\hat{h} = \int_0^1 (1-\hat{t})\,\mathrm{d}\hat{t} = \frac{1}{2}$, giving $h = w_0^2/2g$. (b) The model fails when the neglected $\varepsilon\hat{w}$ is comparable to the retained -1, i.e. when $|\hat{w}| \sim 1/\varepsilon$; since $|\hat{w}| \sim \hat{t}$ for large $\hat{t}$, this is the same as $\hat{t} \sim 1/\varepsilon$. So it is valid for $\hat{t} \ll 1/\varepsilon$ — and the top, at $\hat{t} = 1$, sits comfortably inside (for $\varepsilon = 0.01$, valid until $\hat{t} \sim 100$). This is why the leading-order answer is trustworthy for the ascent but *not* for the long descent to terminal velocity. (How much the drag actually shifts these answers is a question for Chapter 4, which computes the correction systematically.)

Ex. 3. The balance $x^2 \approx 1$ (dropping εx^3) gives $x \approx \pm 1$; the check is that $\varepsilon x^3 \approx \pm\varepsilon$, indeed small beside 1. For the escaped root put $x = X/\delta$: the terms have sizes ε/δ^3, $1/\delta^2$, 1, and balancing the first two forces $\delta = \varepsilon$. Multiplying through by ε^2 gives the exact $X^3 + X^2 - \varepsilon^2 = 0$; dropping ε^2 leaves $X^2(X+1) \approx 0$, so $X \approx -1$ — whence $x \approx -1/\varepsilon$ — or $X \approx 0$ twice. As in Example 3.6, the roots at $X \approx 0$ are not new: they are the two $\mathcal{O}(1)$ roots blurred onto the origin by the magnification. Three roots in all, as a cubic must have.

Ex. 4. (a) With $\varepsilon\dot{y}$ negligible the balance is $0 \approx -y + \cos t$, so $y \approx \cos t$; this is self-consistent, since then $\varepsilon\dot{y} = -\varepsilon\sin t = \mathcal{O}(\varepsilon)$, indeed small. (b) But $y_{\mathrm{outer}}(0) = \cos 0 = 1$, whereas $y(0) = 0$ is required: the outer solution cannot turn the corner, so a different regime is needed at $t = 0$. (c) With $t = \delta\tau$ the chain rule gives $\mathrm{d}y/\mathrm{d}t = \delta^{-1}\mathrm{d}y/\mathrm{d}\tau$, so the equation reads $(\varepsilon/\delta)\,\mathrm{d}y/\mathrm{d}\tau = -y + \cos(\delta\tau)$; the derivative term is restored to size one when $\delta = \varepsilon$. Inside this regime $\tau = \mathcal{O}(1)$, so $\cos(\varepsilon\tau) = 1 + \mathcal{O}(\varepsilon^2)$ and the inner problem is

$$\frac{\mathrm{d}y}{\mathrm{d}\tau} = -y + 1, \qquad y(0) = 0 \qquad \Longrightarrow \qquad y \approx 1 - e^{-\tau} = 1 - e^{-t/\varepsilon}.$$

(d) It vanishes at $\tau = 0$ as required, and tends to $1 = \cos 0$ as $\tau \to \infty$, which is exactly the outer solution's value at $t = 0$; the two pieces therefore join, and $y \approx \cos t - e^{-t/\varepsilon}$ is uniformly valid. In words: y climbs from 0 to ≈ 1 in a time of order ε, then simply follows $\cos t$. (e) The $-y$ is what makes the inner correction $e^{-\tau}$ *decay* onto the outer solution. Were it $+y$, the correction would grow like $e^{+\tau}$, no matching would be possible, and no layer could sit at $t = 0$ — which is the general rule for deciding where a layer may live.

Ex. 5. (a) $m\dot{v} = mg - kv$, $v(0) = 0$. Drag balances weight at $v_T = mg/k$, and the only time scale available is $t_c = m/k = v_T/g$. (b) With $v = v_T\hat{v}$, $t = t_c\hat{t}$, the equation becomes $\hat{v}' = 1 - \hat{v}$, $\hat{v}(0) = 0$, with solution $\hat{v} = 1 - e^{-\hat{t}}$: no parameter survives. Every drop therefore falls in the same way once measured in its own units — only the values of v_T and t_c differ from drop to drop. (c) For $\hat{t} \ll 1$, $\hat{v}$ is still small and the balance is $\hat{v}' \approx 1$, i.e. $\hat{v} \approx \hat{t}$ (dimensionally $v \approx gt$: free fall, drag not yet felt). For $\hat{t} \gg 1$ the acceleration has died away, $\hat{v}' \approx 0$, giving $\hat{v} \approx 1$ (terminal velocity). The crossover is at $\hat{t} \sim 1$, i.e. $t \sim m/k = v_T/g$. (d) Substituting $m = \frac{4}{3}\pi a^3\rho_w$ and $k = 6\pi\mu a$ into $v_T = mg/k$:

$$v_T = \frac{\frac{4}{3}\pi a^3\rho_w\, g}{6\pi\mu a} = \frac{4}{18}\,\frac{\rho_w g a^2}{\mu} = \frac{2}{9}\,\frac{\rho_w g a^2}{\mu}, \qquad t_c = \frac{v_T}{g} = \frac{2}{9}\,\frac{\rho_w a^2}{\mu}.$$

Note the strong dependence on size: $v_T \propto a^2$, so a drop twice as large falls four times as fast. With $a = 5\times10^{-4}\,\mathrm{m}$ this gives $v_T \approx 30\,\mathrm{m/s}$ and $t_c \approx 3\,\mathrm{s}$, while real raindrops fall at about $4\,\mathrm{m/s}$ — the model is wrong by nearly an order of magnitude. The reason shows up in the Reynolds number, $\mathrm{Re} = \rho_{\mathrm{air}} v_T a/\mu \approx 10^3$: Stokes drag assumes $\mathrm{Re} \ll 1$, so the model contradicts its own assumption. Setting $\mathrm{Re} = 1$ and solving for a (using $v_T \propto a^2$, so that $\mathrm{Re} = \frac{2}{9}\rho_{\mathrm{air}}\rho_w g a^3/\mu^2$) gives $a \approx 50\,\mu\mathrm{m}$: the model is self-consistent only for drops of a few tens of microns — fog and mist, not rain. This is the same conclusion reached by dimensional analysis in Chapter 2.

Ex. 6. For instance, the logistic law $\dot{N} = rN(1 - N/K)$ with $N(0) = N_0$. Scaling N by the carrying capacity K and t by $1/r$ gives $n' = n(1 - n)$ with $n(0) = N_0/K$: this is adapted to the *late* behaviour, the approach to saturation. Scaling N instead by the initial population N_0 gives $n' = n(1 - \varepsilon n)$ with $\varepsilon = N_0/K$ and $n(0) = 1$: if the population starts far below capacity, then $\varepsilon \ll 1$, and the leading balance $n' \approx n$ exposes the *early* regime of exponential growth: same model, two scalings, two regimes.

Recommended reading

Section 1.5 of Holmes [1] treats nondimensionalisation and scaling in a similar spirit to ours, and its Chapter 2 introduces perturbation methods — of which Section 2.1 should be readable with what we have covered here. For the same material developed in a modelling context, see Chapter 6 of Lin & Segel [2].

References for this chapter

[1] M. H. Holmes. *Introduction to the Foundations of Applied Mathematics*. Vol. 56. Texts in Applied Mathematics. New York: Springer, 2009.

[2] C. C. Lin and L. A. Segel. *Mathematics Applied to Deterministic Problems in the Natural Sciences*. Classics in Applied Mathematics. Philadelphia: Society for Industrial and Applied Mathematics, 1988.

Chapter 4

Modelling with small parameters I: regular expansions

Chapter 3 introduced dominant balance, and with it the three questions a modeller must be able to answer about any simplification: which terms may be dropped, why dropping them is justified, and when that justification expires. But its output is always a *single* reduced model — one simpler equation, one leading-order description, together with the regime in which it holds. The obvious next question was deliberately left open there: how accurate is that model, and how do we improve it?

In this chapter, we assume the solution to a mathematical model can be written as a series in the small parameter (an *expansion*), substitute the expansion into the equation, and solve order by order. The leading term reproduces the dominant balance of Chapter 3 exactly; the later terms are the corrections, and each is obtained by solving a *linear* problem, whatever the nonlinearity of the original. When this procedure works, it is wonderfully efficient. When it fails — and for oscillators it fails in a way worth understanding — the failure itself is informative, and repairing it reveals something the leading-order model could not see.

The word *regular* tells us about which failures we will meet. Chapter 3 showed one way a naive expansion can break: in the lost-root example, setting $\varepsilon = 0$ destroyed an entire solution, and only a rescaling recovered it. Problems of that kind are called *singular*, and they are the subject of Chapter 5. Here we treat the *regular* case, in which no rescaling is needed — and then discover, in the second half of the chapter, that regularity is not quite enough: a problem can be regular and still go wrong over long times. That second failure, and its cure, are the heart of what follows.

4.1 Why expand?

First, it quantifies the error. A leading-order model is a claim, and a claim without an error bar is hard to act on. If the correction is $\mathcal{O}(\varepsilon)$ and $\varepsilon = 0.01$, the reduced model is good to about a percent, and one can say so.

Second, the correction often carries the physics. This is the part that matters most for modelling, and the pendulum is the standard illustration. Dimensional analysis in Chapter 2 showed that the period must take the form $\mathcal{T} = \sqrt{L/g}\,\Phi(\theta_0)$, and left Φ undetermined. The linear model that Chapter 3 derived says Φ is a constant: the period does not depend on amplitude at all. That is a genuine prediction, and it is *wrong* — real pendulums swing more slowly when swung harder. The entire amplitude dependence, the whole of the interesting content of Φ, lives in the first correction. Compute one more term, and a qualitative feature of the system appears that the leading-order model had erased.

Third, the correction diagnoses its own breakdown. We used this idea repeatedly in Chapter 3: a reduced model fails when the term it neglected grows to the size of the terms it kept. In an

expansion, this becomes a simple, mechanical test. If

$$x = x_0 + \varepsilon x_1 + \cdots ,$$

then the expansion is trustworthy only while $\varepsilon x_1 \ll x_0$. Watching for the moment when that ordering fails is how one finds the range of validity — and, as we shall see, watching *how* it fails is how one discovers the right repair.

4.2 Making "approximately" precise: asymptotic series

We have been writing $\approx$ and $\ll$ freely since Chapter 3, trusting to context. Before building a method on top of it, we should say exactly what it means.

Definition 4.1 (Asymptotic order symbols)**.** As $\varepsilon \to 0$, we write $f = \mathcal{O}(g)$ if $|f/g|$ remains bounded, and $f = o(g)$ if $f/g \to 0$. Informally, $\mathcal{O}$ says "no bigger than, up to a constant", and o says "genuinely smaller". If $f = o(g)$ we normally write $f \ll g$.

Definition 4.2 (Asymptotic series)**.** A sequence of functions $\{\phi_n(\varepsilon)\}$ with $\phi_{n+1} = o(\phi_n)$ is an *asymptotic sequence*: each term is genuinely smaller than the one before. A formal series $\sum_n a_n \phi_n(\varepsilon)$ is an *asymptotic expansion* of $f(\varepsilon)$, written $f \sim \sum a_n \phi_n$, if for every fixed N

$$f(\varepsilon) - \sum_{n=0}^{N} a_n \phi_n(\varepsilon) = o\big(\phi_N(\varepsilon)\big) \qquad (\varepsilon \to 0).$$

Usually $\phi_n = \varepsilon^n$, and the series is an ordinary power series in ε.

Read the definition carefully and notice which quantity is held fixed. It says: *fix the number of terms, then let* $\varepsilon \to 0$, and the remaining error is smaller than the last term retained. It does *not* say anything about letting the number of terms grow at fixed ε — which is what convergence would mean. The two are different questions, and an asymptotic series answers only the first.

This is not a technicality, because the consequence is startling: *an asymptotic series need not converge*, and many of the most useful ones do not.

Example 4.3 (Asymptotic but divergent*)**.** The function

$$f(\varepsilon) = \int_0^\infty \frac{e^{-t}}{1 + \varepsilon t}\, \mathrm{d}t$$

has the formal expansion

$$f(\varepsilon) \sim \sum_{n=0}^{\infty} (-1)^n n!\, \varepsilon^n,$$

obtained by expanding $1/(1+\varepsilon t)$ and integrating term by term. Because $n!$ eventually grows faster than ε^{-n} shrinks, this series diverges for *every* $\varepsilon > 0$: adding terms indefinitely is guaranteed to destroy the answer. And yet truncating it after a few terms gives an excellent approximation to $f(\varepsilon)$ for small ε, with an error no larger than the first omitted term.

The moral is worth stating plainly, because it inverts the usual instinct. Convergence is the wrong question to ask of an approximation scheme. What a modeller wants is a good answer from a few terms, and a divergent asymptotic series routinely supplies that where a convergent Taylor series would not. There is even, for each ε, an optimal number of terms to keep — "optimal truncation" — beyond which the approximation gets worse rather than better. In practice this is rarely a live concern: one or two corrections is almost always as far as a modeller wants to go, both because the algebra grows quickly and because the model itself is seldom accurate enough to deserve more.

4.3 Regular perturbation

Definition 4.4 (Regular and singular)**.** A perturbation problem is *regular* if the straightforward ansatz $x = x_0 + \varepsilon x_1 + \varepsilon^2 x_2 + \cdots$ succeeds: each term can be computed, and the resulting expansion is a valid approximation throughout the domain of interest. It is *singular* otherwise — when the naive expansion loses a solution, or fails in part of the domain, or produces terms that grow without bound.

The tell-tale. Chapter 3 closed with an example with a warning worth repeating here, since it is exactly what distinguishes the two cases: trouble is signalled when ε multiplies the highest-order term — the highest power in an algebraic equation, the highest derivative in a differential one. Setting $\varepsilon = 0$ then lowers the order of the problem, and a solution must vanish. Compare the two quadratics:

$$\underbrace{\varepsilon x^2 + x - 1 = 0}_{\text{Chapter 3: } \textit{singular}} \qquad \text{versus} \qquad \underbrace{x^2 + \varepsilon x - 1 = 0}_{\text{here: } \textit{regular}}.$$

They look almost identical, and they behave completely differently. In the first, setting $\varepsilon = 0$ turns a quadratic into a linear equation: one root is destroyed, escaping to $x \sim -1/\varepsilon$, and only a rescaling recovers it. In the second, setting $\varepsilon = 0$ leaves $x^2 = 1$, still a quadratic, with both roots present and of size one. Nothing escapes, and no rescaling is needed. The examples below are of this second, well-behaved kind.

Recipe 4.5 (Regular perturbation)**.**

1. Substitute $x = x_0 + \varepsilon x_1 + \varepsilon^2 x_2 + \cdots$ into the equation (and into the initial or boundary conditions).
2. Expand everything in powers of ε and collect terms order by order.
3. Set the coefficient of each power of ε to zero. This gives a *sequence* of problems: one for x_0, then one for x_1, and so on.
4. Solve them in turn. Each is typically easier than the original — the $\mathcal{O}(1)$ problem is the dominant balance of Chapter 3, and every later problem is *linear* in the new unknown, driven by terms already computed.
5. Check that the ordering holds: $\varepsilon x_1 \ll x_0$ throughout the domain of interest. If it does not, the problem is not regular after all.

Step 4 contains the reason the method is worth the trouble. However nonlinear the original problem, every correction obeys a linear equation whose forcing is built from terms already known. The nonlinearity is dealt with once, at leading order; after that one is only ever solving linear problems.

Example 4.6 (A well-behaved algebraic root)**.** Take $x^2 + \varepsilon x - 1 = 0$ and write $x = x_0 + \varepsilon x_1 + \varepsilon^2 x_2 + \cdots$. Substituting and collecting powers:

$$\mathcal{O}(1): \quad x_0^2 - 1 = 0 \;\Rightarrow\; x_0 = 1$$

(taking the positive root; $x_0 = -1$ gives the other branch by the same argument. Left as an exercise). Then

$$\mathcal{O}(\varepsilon): \quad 2x_0x_1 + x_0 = 0 \;\Rightarrow\; x_1 = -\tfrac{1}{2}, \qquad \mathcal{O}(\varepsilon^2): \quad x_1^2 + 2x_0x_2 + x_1 = 0 \;\Rightarrow\; x_2 = \tfrac{1}{8}.$$

Hence $x = 1 - \frac{1}{2}\varepsilon + \frac{1}{8}\varepsilon^2 + \cdots$, which agrees term by term with the expansion of the exact root $x = \frac{1}{2}\big(-\varepsilon + \sqrt{\varepsilon^2 + 4}\big)$. Both roots are accounted for, both stay of size one, and no rescaling was required: the problem is regular, exactly as the tell-tale predicted.

Example 4.7 (A regular initial-value problem)**.** Consider $\dot{x} = -x + \varepsilon x^2$ with $x(0) = 1$ — linear decay with a weak nonlinear correction. Substituting $x = x_0 + \varepsilon x_1 + \cdots$ and collecting orders:

$$\mathcal{O}(1): \quad \dot{x}_0 = -x_0,\ x_0(0) = 1 \ \Rightarrow\ x_0 = e^{-t};$$

$$\mathcal{O}(\varepsilon): \quad \dot{x}_1 = -x_1 + x_0^2 = -x_1 + e^{-2t},\ x_1(0) = 0 \ \Rightarrow\ x_1 = e^{-t} - e^{-2t}.$$

Note the structure promised above: the $\mathcal{O}(\varepsilon)$ problem is a *linear* equation for x_1, forced by the already-known x_0. So

$$x \approx e^{-t} + \varepsilon\big(e^{-t} - e^{-2t}\big).$$

Now apply step 5 of the recipe. The correction εx_1 is bounded for all $t \geq 0$, and in fact decays, so it never overtakes x_0: the ordering holds everywhere and the expansion is *uniformly valid* on $0 \leq t < \infty$. This is what a regular problem looks like when it behaves.

4.4 Where regular perturbation fails: oscillators and secular terms

The tell-tale of the previous section is a useful warning but not a guarantee. It catches problems where ε multiplies the highest derivative, and there are none of those in what follows. Nevertheless the naive expansion is about to fail — for a different reason, on a different timescale, and with a different cure.

This second failure mode deserves attention because of where it occurs. Oscillators are among the most common models in the physical sciences, and a weak nonlinearity is the rule rather than the exception: no real spring is exactly linear, no real pendulum swings at exactly infinitesimal amplitude. If regular perturbation quietly failed for such systems we would want to know why, and what to do instead.

Example 4.8 (The Duffing oscillator: breakdown of the naive expansion)**.** Consider a mass on a slightly nonlinear spring,

$$\ddot{x} + x + \varepsilon x^3 = 0, \qquad x(0) = a,\ \dot{x}(0) = 0, \qquad 0 < \varepsilon \ll 1,$$

the *Duffing oscillator*. The cubic term is a small correction to the linear restoring force $-x$: for $\varepsilon > 0$ the spring stiffens as it stretches. Note that ε multiplies neither the highest derivative nor anything else alarming, so the tell-tale raises no objection.

Try $x = x_0 + \varepsilon x_1 + \cdots$. At $\mathcal{O}(1)$ we recover the linear oscillator, $\ddot{x}_0 + x_0 = 0$ with the given data, so

$$x_0 = a\cos t.$$

At $\mathcal{O}(\varepsilon)$,

$$\ddot{x}_1 + x_1 = -x_0^3 = -a^3\cos^3 t = -a^3\Big(\tfrac{3}{4}\cos t + \tfrac{1}{4}\cos 3t\Big),$$

using $\cos^3 t = \frac{3}{4}\cos t + \frac{1}{4}\cos 3t$. Here is the difficulty. The left-hand side is an oscillator of natural frequency 1, and the forcing on the right *contains a* $\cos t$ *term*: the system is being driven at exactly its own natural frequency. This is resonance, in the sense of Chapter 3, and a resonantly forced oscillator responds with an amplitude that grows linearly in time. The particular solution contains

$$x_1 \supset -\tfrac{3a^3}{8}\, t\sin t.$$

A term growing like $t\sin t$ is called *secular* (from the astronomers' *saeculum*, a long age — these terms were first met in planetary calculations, where they matter only over centuries).

Now apply step 5 of Recipe 4.5, and watch it fail. The "correction" is $\varepsilon x_1 \sim \varepsilon a^3 t$, while the leading term is $x_0 \sim a$. The ordering $\varepsilon x_1 \ll x_0$ therefore holds only while

$$\varepsilon a^3 t \ll a, \qquad \text{that is} \qquad t \ll \frac{1}{\varepsilon a^2}.$$

At times $t \sim 1/(\varepsilon a^2)$ the correction is as large as the thing it corrects, and the expansion has destroyed itself — in exactly the manner Chapter 3 taught us to watch for. The expansion is perfectly good over a few oscillations; it is worthless over many.

The diagnosis. The mathematics has failed, but it has failed in an informative way, and reading the failure correctly is what suggests the cure. Ask what the nonlinearity actually does to the motion. A stiffer spring pulls back harder, so the mass returns sooner: the oscillation is slightly *faster*. The nonlinearity changes the *frequency*, by a small amount proportional to ε.

Now consider what a small frequency error does over a long time. Two clocks running at frequencies 1 and $1+\delta$ stay in step at first, but the phase difference between them grows like δt, and after a time $t \sim 1/\delta$ they are in complete disagreement — one reads noon while the other reads midnight. The amplitude of neither clock has changed at all; only the timing has drifted.

That is precisely our situation, with $\delta \sim \varepsilon$. The true solution is an oscillation of very nearly constant amplitude and slightly shifted frequency. The naive expansion, having assumed the frequency is exactly 1, has no way to represent a phase drift; the only thing it can do with the mismatch is to let the amplitude of the correction grow, which is exactly what the secular term does. The growth is not physical. It is the expansion straining to express a frequency shift in the only vocabulary it has.

Written out, the point is transparent. The true solution behaves like $a\cos\big((1+\varepsilon\omega_1)t\big)$, and expanding that in powers of ε at fixed t gives

$$a\cos\big((1+\varepsilon\omega_1)t\big) = a\cos t - \varepsilon\omega_1 a\, t\sin t + \cdots ,$$

whose second term is precisely a secular term. The naive expansion was not wrong so much as *badly organised*: it was expanding a perfectly bounded function in a way that manufactures unbounded terms. Fix the organisation and the problem disappears.

4.5 Poincaré–Lindstedt: straining time

The diagnosis dictates the cure. If the trouble is that we assumed the frequency was exactly 1, then we should stop assuming it: let the frequency be unknown, expand it in ε alongside the solution, and determine it as part of the calculation. The freedom this introduces is exactly the freedom needed to suppress the resonant forcing that produced the secular terms.

Recipe 4.9 (Poincaré–Lindstedt)**.**

1. Introduce a *strained time* $\tau = \omega t$, where the frequency is unknown and is itself expanded,
$$\omega = 1 + \varepsilon\omega_1 + \varepsilon^2\omega_2 + \cdots .$$
2. Rewrite the equation in terms of τ (so $\mathrm{d}/\mathrm{d}t = \omega\,\mathrm{d}/\mathrm{d}\tau$), and expand $x = x_0 + \varepsilon x_1 + \cdots$ as usual.
3. At each order, *choose ω_j so that the resonant terms in the forcing vanish.* Suppressing resonance is what prevents secular growth.
4. Solve the resulting (now bounded) problem for x_j.

The logic of step 3 is worth pausing on, since it is what makes the method work. In the naive expansion the $\mathcal{O}(\varepsilon)$ problem was over-determined in a hidden way: it had a resonant forcing and no freedom with which to absorb it, so it responded with unbounded growth. Introducing ω_1 supplies exactly one new unknown, and demanding a bounded solution supplies exactly one new equation to fix it. The frequency is not guessed; it is *forced* on us by the requirement that the expansion stay orderly.

Example 4.10 (Duffing, properly)**.** With $\tau = \omega t$ and $x' = \mathrm{d}x/\mathrm{d}\tau$, the equation $\ddot{x} + x + \varepsilon x^3 = 0$ becomes

$$\omega^2 x'' + x + \varepsilon x^3 = 0.$$

Using $\omega^2 = 1 + 2\varepsilon\omega_1 + \mathcal{O}(\varepsilon^2)$ and $x = x_0 + \varepsilon x_1 + \cdots$:

$$\mathcal{O}(1): \quad x_0'' + x_0 = 0, \quad x_0 = a\cos\tau;$$

$$\mathcal{O}(\varepsilon): \quad x_1'' + x_1 = -2\omega_1 x_0'' - x_0^3 = \left(2\omega_1 a - \tfrac{3}{4}a^3\right)\cos\tau - \tfrac{1}{4}a^3\cos 3\tau.$$

The $\cos\tau$ term on the right is the resonant one, and now we have a lever with which to remove it. Choosing

$$2\omega_1 a - \tfrac{3}{4}a^3 = 0 \implies \omega_1 = \tfrac{3}{8}a^2$$

kills it, leaving only the harmless $\cos 3\tau$ forcing (harmless because the oscillator is not resonant at frequency 3). Therefore

$$\boxed{\omega = 1 + \tfrac{3}{8}\,\varepsilon a^2 + \mathcal{O}(\varepsilon^2)}$$

and, to leading order,

$$x \approx a\cos\Big[\big(1 + \tfrac{3}{8}\varepsilon a^2\big)t\Big],$$

which stays accurate over the long time scale $t \sim 1/\varepsilon$ on which the naive expansion collapsed. Solving for x_1 — now bounded, since the resonant term is gone — gives the small correction $\frac{a^3}{32}(\cos 3\tau - \cos\tau)$.

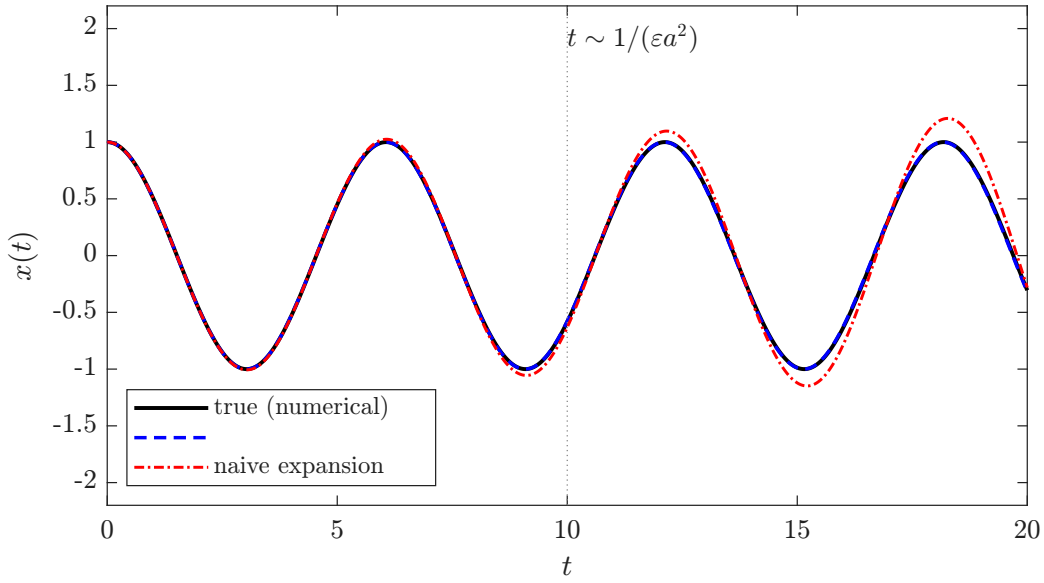


Figure 4.1: The Duffing oscillator $\ddot{x} + x + \varepsilon x^3 = 0$ with $a = 1$, $\varepsilon = 0.1$. The naive expansion (dash-dot) tracks the true solution for a few oscillations but then drifts out of phase and grows without bound, failing near $t \sim 1/(\varepsilon a^2)$ as predicted. The Poincaré–Lindstedt approximation (dashed), which corrects the frequency rather than inflating the amplitude, stays in step with the true solution (solid) throughout.

The result is worth reading as physics rather than algebra. *The frequency depends on the amplitude*: swing this oscillator harder and it oscillates faster. No linear model can produce such a statement, since for a linear oscillator the period is famously independent of amplitude. Amplitude-dependent frequency is a signature of nonlinearity, and one term of perturbation theory was enough to reveal it.

Remark 4.11 (Back to the pendulum: closing the loop). We can now finish a calculation begun two chapters ago. The nondimensional pendulum of Example 3.2 is $\theta'' + \sin\theta = 0$; expanding the sine,

$$\theta'' + \theta - \tfrac{1}{6}\theta^3 + \cdots = 0,$$

which is a Duffing oscillator with a *negative* cubic coefficient. Applying the same method — or simply using the formula $\omega = 1 + \frac{3}{8}\beta a^2$ for $\ddot{x} + x + \beta x^3 = 0$ with $\beta = -\frac{1}{6}$ — gives

$$\omega \approx 1 - \tfrac{1}{16}a^2, \qquad \mathcal{T} = \frac{2\pi}{\omega} \approx 2\pi\Big(1 + \tfrac{1}{16}a^2\Big).$$

The period *increases* with amplitude: a pendulum swung harder runs slower. And this is exactly the amplitude dependence that dimensional analysis identified but could not compute, back in Example 2.8: we have found the leading behaviour of the unknown function $\Phi(\theta_0)$, namely $\Phi(\theta_0) \approx 2\pi(1 + \theta_0^2/16)$.

It is worth appreciating the division of labour. Dimensional analysis told us the period had to be $\sqrt{L/g}$ times a function of amplitude alone — a strong structural statement, obtained before writing any equation. Nondimensionalisation reduced the model to a single parameter-free equation. Perturbation theory then computed the function that dimensional analysis had to leave blank. Three chapters, three tools, one answer.

The sign matters, and is the physical punchline. The pendulum *softens* (its restoring force grows more slowly than linearly, since $\sin\theta < \theta$), so the period lengthens; the $+\varepsilon x^3$ Duffing oscillator *stiffens*, so its period shortens. A sign error inverts the prediction, and a model that predicts the wrong direction of an effect is worse than no model at all.

4.6 The energy method: the exact period of a conservative oscillator

Poincaré–Lindstedt gives the period as a series, and buys its accuracy with two assumptions: that the nonlinearity is *weak*, and that there is a linear oscillator to perturb about. Both can fail. A strongly nonlinear oscillator has no small parameter, and — more strikingly — some oscillators have no linear behaviour to perturb about at all, because their restoring force vanishes faster than linearly at the origin. For such problems perturbation theory has nothing to say.

There is a second route to the period which has neither restriction. It applies to any *conservative* oscillator, however strongly nonlinear, and it is exact. The price is that the answer comes as an integral rather than a formula. The two methods are complementary, and it is worth having both: one gives an explicit approximation, the other an exact expression one can evaluate numerically. Since the word "conservative" is doing real work here, we build the idea up from the beginning.

Which models are conservative?

Consider models of the form

$$\ddot{x} = f(x),$$

in which there is a single position variable x (this is known as *one degree of freedom*) and the acceleration depends on the position *alone*. What matters is what is *absent*: no $\dot{x}$ on the right, so no damping or drag; and no explicit t, so no external driving. Such a system is closed — once released, nothing further is done to it.

The distinction is easy to see in the models we have already met. The undamped pendulum $\theta'' = -\sin\theta$ qualifies. The damped pendulum $\theta'' = -\sin\theta - c\theta'$ does not, because of the θ' term; nor does the thrown body of Chapter 3, $\hat{w}' = -1 - \varepsilon\hat{w}$, whose drag term depends on velocity. The difference is physically meaningful: drag drains the motion, and a system that is losing something cannot conserve it.

More generally — and this is where the name comes from — a system is called *conservative* when it possesses a scalar quantity $E(\mathbf{x})$, the *energy*, that stays constant along every trajectory, $\mathrm{d}E/\mathrm{d}t = 0$. The rest of this section shows that every model of the form $\ddot{x} = f(x)$ is conservative in exactly this sense, constructs the E explicitly, and then puts it to work.

The potential U

Any 'nice-enough' function f of one variable has an antiderivative, so we may always define a function U by

$$U(x) = -\int f(x)\,\mathrm{d}x, \qquad \text{equivalently} \qquad f(x) = -U'(x),$$

and rewrite the model as

$$\ddot{x} = -U'(x).$$

The minus sign is a convention, chosen so that the force pushes the system "downhill" in U — towards smaller U — which makes the pictures below read the way one expects.

For the pendulum, $\theta'' = -\sin\theta$, so $f(\theta) = -\sin\theta$ and $U(\theta) = -\int(-\sin\theta)\,\mathrm{d}\theta = -\cos\theta$. One can check directly that $U'(\theta) = \sin\theta$, as required.

Why something is conserved

To find the conserved quantity, multiply $\ddot{x} = -U'(x)$ through by $\dot{x}$:

$$\dot{x}\,\ddot{x} = -U'(x)\,\dot{x}.$$

Each side is now a derivative in disguise. On the left, $\dot{x}\ddot{x} = \dfrac{\mathrm{d}}{\mathrm{d}t}\Big(\tfrac{1}{2}\dot{x}^2\Big)$; on the right, by the chain rule, $U'(x)\dot{x} = \dfrac{\mathrm{d}}{\mathrm{d}t}U\big(x(t)\big)$. So the equation says

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(\tfrac{1}{2}\dot{x}^2\Big) = -\frac{\mathrm{d}}{\mathrm{d}t}U(x) \qquad \Longrightarrow \qquad \frac{\mathrm{d}}{\mathrm{d}t}\Big(\tfrac{1}{2}\dot{x}^2 + U(x)\Big) = 0.$$

The quantity in brackets does not change as the motion proceeds. Calling it E,

$$\boxed{E = \tfrac{1}{2}\dot{x}^2 + U(x) = \text{constant}}$$

— and this constant is fixed once and for all by the initial conditions.

The two pieces have standard names and a plain reading. The term $\frac{1}{2}\dot{x}^2$ depends only on how fast the system is moving and is called the *kinetic energy*; the term $U(x)$ depends only on where it is and is called the *potential energy*. (In our nondimensional equations the mass has been scaled away, so these are energies per unit mass; the overall constant is irrelevant to everything that follows.) Their sum E is the *total energy*, and what we have just shown is that it is *conserved* — hence the name for this class of systems. The motion does not create or destroy energy; it merely trades one form for the other, converting speed into position and back. A pendulum is fastest at the bottom of its swing, where U is least and so $\frac{1}{2}\dot{\theta}^2$ must be greatest, and momentarily still at the top, where all the energy has gone into U.

Damping breaks this. Repeating the calculation for $\ddot{x} = -U'(x) - c\dot{x}$ gives $\mathrm{d}E/\mathrm{d}t = -c\dot{x}^2 \le 0$: the energy decreases steadily, which is exactly what "the oscillation dies away" means.

Reading the motion off the potential

Conservation of E is powerful because it converts a differential equation into an algebraic relation between position and speed, and that relation can be read straight off a graph of U (Figure 4.2).

Draw the curve $U(x)$ and the horizontal line at height E. Since $\frac{1}{2}\dot{x}^2 = E - U(x)$ and the left-hand side cannot be negative, the motion is confined to those x where

$$U(x) \le E,$$

that is, to the region where the curve lies *below* the line. The vertical gap between the line and the curve is the kinetic energy: wide gap, fast motion; no gap, momentarily at rest.

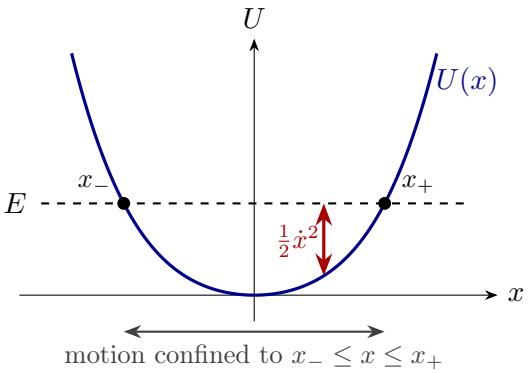


Figure 4.2: The motion read off the potential. The system is confined to the region where the curve $U(x)$ lies below the energy level E, because the kinetic energy $\frac{1}{2}\dot{x}^2 = E - U(x)$ cannot be negative. The vertical gap between line and curve is the kinetic energy, largest at the bottom of the well; it closes at the *turning points* $x_\pm$, where $U = E$, the system is momentarily at rest, and the motion reverses. Trapped between two turning points, the motion must repeat: it oscillates.

The points where the gap closes, $U(x) = E$, are the *turning points* $x_\pm$: there $\dot{x} = 0$, the system is instantaneously at rest, and — since the force $-U'$ still pushes it back down the slope — it reverses. A system released inside a well is therefore trapped between two turning points, and being trapped, it must return: the motion is periodic. This is the qualitative conclusion, obtained from a picture rather than a calculation.

The same picture also explains a statement made in Chapter 3. There, we justified the small-angle approximation by asserting that a pendulum released from rest at amplitude θ_0 can never swing beyond θ_0, so that a small initial angle stays small for *all* time and not merely at the start. That is now proved: releasing from rest at $x = a$ fixes $E = U(a)$, and the motion can never reach any point where $U(x) > U(a)$. For a symmetric well, this confines it to $|x| \le a$ for all time.

Remark 4.12 (The phase plane). The same information has a geometric form that will be useful in Chapter 5 and in Appendix A. Plotting the state of the system as a point $(x, \dot{x})$ in the *phase plane*, the relation $\frac{1}{2}\dot{x}^2 + U(x) = E$ says that a trajectory is confined to a level curve of E. Different initial conditions give different values of E and hence different curves, so the whole family of solutions can be drawn at once, without solving anything. A *closed* level curve is precisely a periodic orbit — the trajectory returns to where it began — which is the phase-plane version of the trapping argument above.

From energy to the period

So far we have the shape of the motion; now we want its period, and the energy relation delivers it. Solving $E = \frac{1}{2}\dot{x}^2 + U(x)$ for the speed gives

$$\dot{x} = \pm\sqrt{2\big(E - U(x)\big)},$$

the sign depending on the direction of travel. Since $\dot{x} = \mathrm{d}x/\mathrm{d}t$, we may invert this to give the time taken to cross each little interval of position, $\mathrm{d}t = \mathrm{d}x/\dot{x}$, and add up those times over

one full circuit of the orbit. The period is therefore a *quadrature* — an answer expressed as an integral:

$$T = \oint \frac{\mathrm{d}x}{\dot{x}} = 2\int_{x_-}^{x_+} \frac{\mathrm{d}x}{\sqrt{2\big(E - U(x)\big)}},$$

the factor 2 arising because the system traverses the interval $[x_-, x_+]$ twice per period, once in each direction, taking the same time each way. For a symmetric potential with release from rest at amplitude a (so that $E = U(a)$ and $x_\pm = \pm a$), the four quarter-swings are identical and it is tidier to write

$$\boxed{T(a) = 4\int_0^a \frac{\mathrm{d}x}{\sqrt{2\big(U(a) - U(x)\big)}}}.$$

This is exact, for any conservative oscillator, with no assumption that anything is small. The integrand does become infinite at the turning point $x = a$, where the square root vanishes, but the singularity is integrable and standard numerical routines handle it, especially after the substitution $x = a\sin\phi$ or similar.

Example 4.13 (The exact pendulum period)**.** For the pendulum $\theta'' + \sin\theta = 0$ the potential is $U(\theta) = -\cos\theta$ (indeed $U' = \sin\theta$). Released from rest at amplitude θ_0, the energy is $E = U(\theta_0) = -\cos\theta_0$, so

$$\mathcal{T}(\theta_0) = 4\int_0^{\theta_0} \frac{\mathrm{d}\theta}{\sqrt{2\big(\cos\theta - \cos\theta_0\big)}}.$$

This is a complete elliptic integral. It has no elementary closed form, and must be evaluated numerically or expanded for small θ_0 — and expanding it recovers the Poincaré–Lindstedt result $\mathcal{T} \approx 2\pi\big(1 + \frac{1}{16}\theta_0^2\big)$ of Remark 4.11, which is a satisfying check on both calculations.

So the function $\Phi(\theta_0)$ that dimensional analysis left undetermined in Example 2.8 is now known in two complementary senses: exactly, as this integral, and approximately, as an explicit formula valid at small amplitude. Neither supersedes the other. The formula is the one to think with; the integral is the one to compute with when the amplitude is not small.

Remark 4.14 (Oscillators with no linear part). The quadrature works for *any* conservative oscillator, including those that perturbation theory cannot reach. The 'cart-between-two-springs' is the standard example: its potential behaves like $U \sim x^4$ near the origin, with *no* quadratic term, so the restoring force is purely cubic and there is no small-oscillation frequency to perturb about. Poincaré–Lindstedt has no starting point. The boxed formula, by contrast, is applied unchanged and delivers the period via numerical integration.

Such oscillators also behave in a way that no linear intuition would predict. For the purely cubic oscillator $\ddot{x} = -x^3$ the period turns out to be *inversely* proportional to amplitude (Exercise 4.8): small oscillations are *slower*, not faster, and an oscillation of vanishing amplitude takes infinitely long. That is a genuine prediction, it is testable, and it is invisible to the methods in the first half of this chapter.

Exercises

Exercise 4.1. Find the first two terms in the expansion of the root of $x = 1 + \varepsilon e^x$ near $x = 1$ as $\varepsilon \to 0$. (Substitute $x = x_0 + \varepsilon x_1 + \cdots$ and expand the exponential.)

Exercise 4.2. Solve $\dot{x} = -x + \varepsilon t$ with $x(0) = 1$ by regular perturbation to $\mathcal{O}(\varepsilon)$, and compare with the exact solution.

(a) Verify that the perturbation series in fact *terminates*: it reproduces the exact solution with no error at all.

(b) Despite this, show that the expansion is *not* uniformly valid on $0 \le t < \infty$, by checking the ordering $\varepsilon x_1 \ll x_0$ of Recipe 4.5. Over what range of t does the ordering hold? (Note that x_0 decays while x_1 grows — so "exact" and "well-ordered" are not the same thing.)

Exercise 4.3. Return to the thrown body of Chapter 3, $\hat{w}' = -1 - \varepsilon\hat{w}$ with $\hat{w}(0) = 1$, whose leading-order behaviour $\hat{w} \approx 1 - \hat{t}$ was found there by dominant balance.

(a) Substitute $\hat{w} = \hat{w}_0 + \varepsilon\hat{w}_1 + \cdots$ into the *differential equation* and solve order by order. Check that the $\mathcal{O}(1)$ problem reproduces exactly the dominant balance of Chapter 3, and show that $\hat{w} \approx (1 - \hat{t}) + \varepsilon\left(\frac{1}{2}\hat{t}^2 - \hat{t}\right)$.

(b) Hence find the first drag correction to the dimensionless time and height at the top of the flight, both of which Chapter 3 obtained only to leading order. (To locate the top, solve $\hat{w}(\hat{t}_{\text{top}}) = 0$ perturbatively: write $\hat{t}_{\text{top}} = 1 + \varepsilon\hat{t}_1 + \cdots$, substitute, and collect orders — the same order-by-order idea applied to a root rather than a function.) Do they rise or fall, and is that what you would expect physically?

(c) This particular equation happens to be solvable, with exact solution $\hat{w} = (1 + \frac{1}{\varepsilon})e^{-\varepsilon\hat{t}} - \frac{1}{\varepsilon}$. Expand it in ε and confirm that you recover the same result. Which of the two routes would still have been available had the equation possessed no closed-form solution?

Exercise 4.4. The tell-tale of Section 4.3 lets one anticipate trouble before doing any work. For each problem below, predict *first* whether a straightforward expansion in powers of ε will succeed, then test the prediction.

(a) $\varepsilon x^3 + x - 2 = 0$. How many roots does the naive limit $\varepsilon = 0$ retain, how many are lost, and — by rescaling — where have the lost ones gone?

(b) $x^3 + \varepsilon x - 2 = 0$. Find the real root to $\mathcal{O}(\varepsilon)$.

(c) $\varepsilon\ddot{x} + \dot{x} + x = 0$. Setting $\varepsilon = 0$ reduces the order from two to one. Using the characteristic equation, locate both exact decay rates and identify the solution the naive limit destroys. On what timescale does it live?

(d) $\ddot{x} + \varepsilon\dot{x} + x = 0$ with $x(0) = 1$, $\dot{x}(0) = 0$ — a weakly damped oscillator. Show that the tell-tale raises no objection, yet the naive expansion produces a secular term at $\mathcal{O}(\varepsilon)$. What is the true solution doing that the expansion cannot represent?

Part (d) is the moral of the exercise: passing the tell-tale is necessary but not sufficient.

Exercise 4.5. For $\ddot{x} + x + \varepsilon x^2 = 0$ with $x(0) = a$, $\dot{x}(0) = 0$, attempt the naive expansion.

(a) Compute the $\mathcal{O}(\varepsilon)$ forcing $-x_0^2$ and write it in terms of cosines of multiples of t. Does a resonant term appear? Does a secular term therefore appear at $\mathcal{O}(\varepsilon)$?

(b) Explain, in terms of the harmonic content of x_0^2 versus x_0^3, why the answer differs from the cubic case of Example 4.8.

(c) *(Harder.)* Carry Poincaré–Lindstedt to $\mathcal{O}(\varepsilon^2)$ and show that the frequency shift, though absent at first order, appears at second.

Exercise 4.6. Apply Poincaré–Lindstedt to the Duffing oscillator $\ddot{x} + x + \varepsilon x^3 = 0$ to $\mathcal{O}(\varepsilon^2)$ and find ω_2.

Exercise 4.7 (†). A mass on a nonlinear spring obeys $\ddot{x} + x + \varepsilon x^3 = 0$ with $x(0) = a$, $\dot{x}(0) = 0$. Take $a = 1$ and $\varepsilon = 0.1$.

(a) Integrate the equation numerically, measure the period, and compare it with $2\pi/\omega$ using the Poincaré–Lindstedt frequency of Example 4.10.

(b) On one plot, overlay the true solution, the naive expansion of Example 4.8, and the Poincaré–Lindstedt solution over $0 \le t \lesssim 2/\varepsilon$. Mark where the naive curve visibly departs, and check that it happens where the theory says it should.

(c) Write one paragraph interpreting the result as a statement about modelling: what exactly does the linear (small-amplitude) model get wrong, and over what timescale is it nevertheless trustworthy?

Exercise 4.8. Use the energy method (Section 4.6) to express the period of the purely cubic oscillator $\ddot{x} = -x^3$ (potential $U = \frac{1}{4}x^4$), released from rest at amplitude a, as a quadrature. By rescaling $x = aX$, show that $T(a) \propto 1/a$: the period is inversely proportional to amplitude. Contrast this with the amplitude-independent period of a linear oscillator, and say why no method of the first half of this chapter could have produced the result.

Hints and answers

Ex. 1. At $\mathcal{O}(1)$, $x_0 = 1$. Writing $e^x = e^{x_0+\varepsilon x_1+\cdots} = e\big(1+\varepsilon x_1+\cdots\big)$ and matching at $\mathcal{O}(\varepsilon)$ gives $x_1 = e^{x_0} = e$, so $x \approx 1+\varepsilon e$. Carrying one more order gives $x_2 = e^2$, i.e. $x \approx 1+\varepsilon e+\varepsilon^2 e^2$.

Ex. 2. At $\mathcal{O}(1)$, $x_0 = e^{-t}$; at $\mathcal{O}(\varepsilon)$, $\dot{x}_1 = -x_1 + t$ with $x_1(0) = 0$, giving $x_1 = t - 1 + e^{-t}$. (a) The exact solution is $x = \varepsilon(t-1) + (1+\varepsilon)e^{-t}$, which is precisely $x_0 + \varepsilon x_1$: the series terminates, and every higher x_n vanishes. (b) Nevertheless $x_0 = e^{-t}$ *decays* while $x_1 \approx t$ for large t *grows*, so the ordering $\varepsilon x_1 \ll x_0$ reads $\varepsilon t \ll e^{-t}$, i.e. $\varepsilon t\, e^t \ll 1$. The two sides become comparable when $\varepsilon t\, e^t \sim 1$; taking logarithms, $\ln\varepsilon + \ln t + t \sim 0$, and since t dominates $\ln t$, the crossover is at $t \sim \ln(1/\varepsilon)$. So the expansion is well-ordered only on $0 \le t \lesssim \ln(1/\varepsilon)$ — an interval that grows as $\varepsilon \to 0$, so the expansion is valid on any *fixed* interval, but not *uniformly* on $[0,\infty)$: for any ε, however small, it eventually fails. The lesson is that being exact and being properly ordered are different properties: an asymptotic expansion is a statement about the *relative sizes* of its terms, and here the tiny term εx_1 eventually overtakes the exponentially small x_0.

Ex. 3. (a) $\mathcal{O}(1)$: $\hat{w}_0' = -1$ with $\hat{w}_0(0) = 1$, so $\hat{w}_0 = 1 - \hat{t}$ — precisely the gravity-dominated balance of Chapter 3. $\mathcal{O}(\varepsilon)$: $\hat{w}_1' = -\hat{w}_0 = -(1-\hat{t})$ with $\hat{w}_1(0) = 0$, so $\hat{w}_1 = \frac{1}{2}\hat{t}^2 - \hat{t}$. (b) Setting $\hat{w} = 0$ with $\hat{t}_{\text{top}} = 1 + \varepsilon\hat{t}_1$ and collecting the $\mathcal{O}(\varepsilon)$ terms gives $\hat{t}_1 = -\frac{1}{2}$, so $\hat{t}_{\text{top}} \approx 1 - \frac{1}{2}\varepsilon$, and integrating $\hat{w}$ up to that time gives $\hat{h} \approx \frac{1}{2} - \frac{1}{3}\varepsilon$. Both *fall*: drag robs the body of speed, so it neither rises as high nor takes as long getting there — as one would expect. (c) Expanding the exact solution gives the same series, as it must. But the expansion of the exact solution requires an exact solution; the order-by-order route requires only the equation, which is why it is the one that generalises. This is the whole point of the method.

Ex. 4. (a) *Singular*: ε multiplies the highest power, so $\varepsilon = 0$ leaves the linear $x = 2$ — one root retained, two lost. Rescaling $x = X/\delta$ and balancing the cubic against the linear term forces $\delta = \sqrt{\varepsilon}$; then $X^3 + X \approx 0$ gives $X \approx \pm i$, so the lost pair is $x \approx \pm i/\sqrt{\varepsilon}$ — a complex conjugate pair, far out at size $\varepsilon^{-1/2}$. (For $\varepsilon = 0.01$ the exact roots are $x \approx 1.93$ and $-0.96 \pm 10.14i$.) (b) *Regular*: $\varepsilon = 0$ leaves the full cubic $x^3 = 2$, all three roots intact. With $x = x_0 + \varepsilon x_1$, $x_0 = 2^{1/3}$ and $3x_0^2 x_1 + x_0 = 0$, so $x \approx 2^{1/3} - \varepsilon/(3\cdot 2^{1/3}) \approx 1.2573$ for $\varepsilon = 0.01$. (c) *Singular*: the characteristic equation $\varepsilon\lambda^2 + \lambda + 1 = 0$ has roots $\lambda \approx -1$ and $\lambda \approx -1/\varepsilon$. The naive limit keeps the slow decay e^{-t} and destroys the second solution, which decays like $e^{-t/\varepsilon}$ — a boundary layer of width $\mathcal{O}(\varepsilon)$ at $t = 0$, exactly as in Chapter 3, Exercise 4. (d) The tell-tale is silent, since ε multiplies only $\dot{x}$. Yet $x_0 = \cos t$ and the $\mathcal{O}(\varepsilon)$ problem is $\ddot{x}_1 + x_1 = -\dot{x}_0 = \sin t$, whose forcing is resonant, giving the secular $x_1 \supset -\frac{1}{2}t\cos t$. The true solution is $x \approx e^{-\varepsilon t/2}\cos t$: the amplitude is *slowly decaying*, and expanding $e^{-\varepsilon t/2} = 1 - \frac{1}{2}\varepsilon t + \cdots$ manufactures precisely that secular term. Note that Poincaré–Lindstedt will *not* rescue this one — it strains time to fix a frequency error, whereas here the amplitude itself is drifting. That requires the two-timescale method of Chapter 5.

Ex. 5. (a) With $x_0 = a\cos t$, the forcing is $-x_0^2 = -a^2\cos^2 t = -\frac{a^2}{2}\big(1+\cos 2t\big)$: it contains a constant and a $\cos 2t$, but *no* $\cos t$. There is no resonance, hence no secular term at $\mathcal{O}(\varepsilon)$, and the naive expansion survives this order. (b) The reason is parity. Squaring $\cos t$ produces only *even* harmonics (1 and $\cos 2t$), none of which is resonant with the natural frequency 1; cubing it produces *odd* harmonics ($\cos t$ and $\cos 3t$), one of which is exactly resonant. An even nonlinearity therefore causes no first-order frequency shift, an odd one does. (c) Solving the $\mathcal{O}(\varepsilon)$ problem gives $x_1 = -\frac{a^2}{2} + \frac{a^2}{6}\cos 2\tau + \frac{a^2}{3}\cos\tau$, and $\omega_1 = 0$. At $\mathcal{O}(\varepsilon^2)$ the term $-2x_0x_1$ now *does* generate a $\cos\tau$ component, of size $\frac{5}{6}a^3$; removing it requires $2\omega_2 a + \frac{5}{6}a^3 = 0$, so

$$\omega_2 = -\tfrac{5}{12}a^2, \qquad \omega \approx 1 - \tfrac{5}{12}\varepsilon^2 a^2.$$

The frequency is *lowered*: a quadratic nonlinearity softens the oscillator at second order.

Ex. 6. Continue Example 4.10 one order further, using $\omega^2 = 1+2\varepsilon\omega_1+\varepsilon^2(\omega_1^2+2\omega_2)$ and $x_1 = \frac{a^3}{32}(\cos 3\tau - \cos\tau)$. Collecting the $\cos\tau$ terms at $\mathcal{O}(\varepsilon^2)$ and setting their coefficient to zero gives

$$\omega_2 = -\tfrac{21}{256}a^4, \qquad \omega \approx 1 + \tfrac{3}{8}\varepsilon a^2 - \tfrac{21}{256}\varepsilon^2 a^4.$$

(A numerical check is worthwhile: for $a = 1$, $\varepsilon = 0.1$ the exact frequency is $\omega \approx 1.0367$, against 1.0375 from one term and 1.03668 from two.)

Ex. 7. (a) With $a = 1$, $\varepsilon = 0.1$, Poincaré–Lindstedt gives $\omega \approx 1.0375$ and hence $T \approx 6.056$, against a measured period of about 6.061 — agreement to better than 0.1%. (b) The naive curve tracks the true solution for the first few oscillations and then drifts visibly out of phase, its amplitude growing without bound; the departure becomes obvious around $t \sim 1/\varepsilon = 10$, exactly as Example 4.8 predicts. The Poincaré–Lindstedt curve stays in step throughout. (c) The linear model gets the *amplitude* right and the *frequency* slightly wrong; because a frequency error accumulates into a phase error, it is reliable for a few oscillations and useless after $t \sim 1/\varepsilon$. Which of those matters depends on the question being asked — another instance of the Chapter 3 lesson that a reduced model must always be quoted together with its range of validity.

Ex. 8. With $U = \frac{1}{4}x^4$ and $E = U(a) = \frac{1}{4}a^4$,

$$T(a) = 4\int_0^a \frac{\mathrm{d}x}{\sqrt{\frac{1}{2}(a^4 - x^4)}} = 4\sqrt{2}\int_0^a \frac{\mathrm{d}x}{\sqrt{a^4 - x^4}}.$$

Substituting $x = aX$ gives $\mathrm{d}x = a\,\mathrm{d}X$ and $a^4 - x^4 = a^4(1 - X^4)$, so

$$T(a) = \frac{4\sqrt{2}}{a}\int_0^1 \frac{\mathrm{d}X}{\sqrt{1 - X^4}} = \frac{C}{a}, \qquad C = 4\sqrt{2}\int_0^1 \frac{\mathrm{d}X}{\sqrt{1 - X^4}} \approx 7.416,$$

a pure number times $1/a$. So the period is inversely proportional to amplitude: *smaller* oscillations are *slower*, and the period diverges as $a \to 0$. A linear oscillator has period 2π regardless of amplitude, so this is qualitatively different behaviour, and it arises because the restoring force $-x^3$ becomes negligibly weak near the origin — the mass has almost nothing pulling it back. No method of the first half of the chapter applies, because there is no linear oscillator to perturb about: setting $\varepsilon = 0$ would leave $\ddot{x} = 0$, which does not oscillate at all.

Recommended reading

The natural companion to this chapter is Holmes [2], whose opening chapter develops asymptotic approximations, order symbols and the convergent-versus-asymptotic distinction at exactly the level we need, with many worked examples; its treatment of secular terms and of the Duffing oscillator is the standard reference. Continuing from the previous chapter, Holmes [1] covers the same perturbation ideas in a modelling setting. For the Poincaré–Lindstedt method, and for conservative systems and their phase-plane orbits — the geometric picture behind the energy method of Section 4.6 — Strogatz [4] is very readable and complements the treatment here. Lin & Segel [3] place perturbation methods in the wider modelling context.

References for this chapter

[1] M. H. Holmes. *Introduction to the Foundations of Applied Mathematics*. Vol. 56. Texts in Applied Mathematics. New York: Springer, 2009.

[2] M. H. Holmes. *Introduction to Perturbation Methods*. 2nd ed. New York: Springer, 2013.

[3] C. C. Lin and L. A. Segel. *Mathematics Applied to Deterministic Problems in the Natural Sciences*. Classics in Applied Mathematics. Philadelphia: Society for Industrial and Applied Mathematics, 1988.

[4] S. H. Strogatz. *Nonlinear Dynamics and Chaos*. 2nd ed. Boulder, CO: Westview Press, 2015.

Chapter 5

Modelling with small parameters II: multiple scales and reduction

In Chapter 4, the naive expansion failed for oscillators because a slow drift was being forced through a single fast clock, producing secular terms; Poincaré–Lindstedt fixed this by straining time — but only when the slow effect is a shift in *frequency*. A second kind of slow effect lies beyond its reach: when the *amplitude* itself drifts, as it does under damping, straining the clock cannot help, because no choice of frequency turns a constant-amplitude $a\cos(\omega t)$ into a growing or shrinking one. We begin the chapter by watching this failure happen — running both tools we already have on the simplest damped oscillator, and seeing them break.

The 'cure' we present here is the *method of multiple scales*, which introduces two independent time variables — one fast, one slow — so that a drifting amplitude or phase lives on the slow variable and never has to distort the fast oscillation. That completes the perturbation toolkit for oscillators.

The second half of the chapter is where the payoff lies. Real models routinely contain variables evolving on wildly different timescales: a fast chemical equilibrium riding on a slow reaction, a fast electrical current under a slow recovery, a fast predator response to a slow change in prey. Carrying the fast variable in full is both wasteful and numerically stiff. The remedy is *reduction*: hold the fast variable at its instantaneous equilibrium and follow only the slow one, cutting the dimension of the model. This is among the most useful things a modeller can do — and knowing *exactly when it fails* is more useful still, because the failures are where the interesting behaviour hides.

5.1 The method of multiple scales

Before constructing the 'cure', let us watch the 'disease'. We take the simplest oscillator whose amplitude drifts and try on it, in turn, both tools from Chapter 4 — the naive expansion and Poincaré–Lindstedt — and watch each break. That done, the remedy will suggest itself.

Example 5.1 (Where Poincaré–Lindstedt fails)**.** Consider the weakly damped linear oscillator

$$\ddot{x} + \varepsilon\dot{x} + x = 0, \qquad x(0) = a,\ \dot{x}(0) = 0, \qquad 0 < \varepsilon \ll 1.$$

This is simple enough that we know the exact answer, $x = a\,e^{-\varepsilon t/2}\cos\!\big(\sqrt{1-\varepsilon^2/4}\,t\big)$, so we can see precisely what any approximation must reproduce: an oscillation whose amplitude *slowly decays*, on the timescale $t \sim 1/\varepsilon$. Watch both methods fail to produce it.

The naive expansion. Put $x = x_0 + \varepsilon x_1 + \cdots$. At $\mathcal{O}(1)$, $\ddot{x}_0 + x_0 = 0$ with the given data gives $x_0 = a\cos t$. At $\mathcal{O}(\varepsilon)$,

$$\ddot{x}_1 + x_1 = -\dot{x}_0 = a\sin t.$$

The forcing is at the oscillator's own frequency — resonance — so, exactly as with Duffing in Chapter 4, a secular term is born:

$$x_1 = \tfrac{a}{2}\big(\sin t - t\cos t\big) \;\supset\; -\tfrac{a}{2}\,t\cos t.$$

The expansion is good for a few cycles and worthless once $\varepsilon t \sim 1$.

Poincaré–Lindstedt. Now strain time, $\tau = \omega t$ with $\omega = 1 + \varepsilon\omega_1 + \cdots$, hoping to absorb the resonance as we did for Duffing. The equation becomes $\omega^2 x'' + \varepsilon\omega x' + x = 0$, and at $\mathcal{O}(\varepsilon)$,

$$x_1'' + x_1 = -2\omega_1 x_0'' - x_0' = \underbrace{2\omega_1 a\cos\tau}_{\text{controllable}} + \underbrace{a\sin\tau}_{\text{not}}.$$

Here the trouble is laid bare. *Two* resonant terms sit on the right, and we have only *one* knob, ω_1. It multiplies $\cos\tau$, so it can annihilate the first term — but nothing multiplies $\sin\tau$, so the second survives whatever value we give ω_1. The best available choice, $\omega_1 = 0$, leaves $x_1'' + x_1 = a\sin\tau$: the very same secular term as before. Poincaré–Lindstedt has run, and failed.

Why it failed. The two resonant terms are physically different. The $\cos\tau$ term is *in phase with the displacement* — it is a frequency error, and straining the clock is precisely the tool for it. The $\sin\tau$ term is *in phase with the velocity*; it *is* the damping, and it asks the amplitude to change. Straining time offers no way to change an amplitude. What is missing is not a better clock but a second freedom: we must let the amplitude itself evolve.

The idea. The example tells us what the cure must provide. The true solution $x \approx a\,e^{-\varepsilon t/2}\cos t$ depends on time in *two* ways at once — a fast oscillation $\cos t$ modulated by a slow envelope $e^{-\varepsilon t/2}$ — and the two live on different timescales, order one and order $1/\varepsilon$. A single power series in ε at fixed t cannot hold both. Expanding the envelope,

$$e^{-\varepsilon t/2} = 1 - \tfrac{1}{2}\varepsilon t + \cdots,$$

manufactures exactly the secular terms we keep meeting: the function is perfectly bounded, but the *expansion* is not, because it is the wrong way to write the function.

The remedy is to stop insisting on a single time. Introduce two time variables,

$$T_0 = t \quad \text{(fast)}, \qquad T_1 = \varepsilon t \quad \text{(slow)},$$

and treat them as *independent*. Then the slow envelope is a function of T_1 alone, represented exactly rather than through its Taylor series, and the fast oscillation is a function of T_0 alone. Nothing needs to grow. The method below builds this separation into the calculation from the start, and — crucially — it does not assume in advance whether the slow drift is in amplitude, in phase, or in both. It *derives* which.

Recipe 5.2 (Two-timing)**.** Introduce a fast time $T_0 = t$ and a slow time $T_1 = \varepsilon t$, and treat them as independent variables, so that by the chain rule (Appendix B.2)

$$\frac{\mathrm{d}}{\mathrm{d}t} = \partial_{T_0} + \varepsilon\,\partial_{T_1}, \qquad \frac{\mathrm{d}^2}{\mathrm{d}t^2} = \partial_{T_0}^2 + 2\varepsilon\,\partial_{T_0}\partial_{T_1} + \varepsilon^2\partial_{T_1}^2.$$

(Despite the partial-derivative notation, no theory of partial differential equations is needed: at each order the problem is solved as an ordinary differential equation in T_0, with coefficients that depend on T_1. See Appendix B.) Expand $x = x_0(T_0, T_1) + \varepsilon x_1(T_0, T_1) + \cdots$, and collect orders as usual. At $\mathcal{O}(\varepsilon)$ a resonant (secular-producing) term appears in the forcing; *remove it* by demanding that the as-yet-unknown slow dependence of x_0 absorb the resonance. That demand is an evolution equation for the amplitude and phase on the slow time T_1 — the *amplitude equation*.

The step that does the work is the removal of the resonant term. In the naive expansion there was no freedom to absorb resonance, so it forced unbounded growth. Here the leading solution carries an unknown slow-time dependence, and requiring the $\mathcal{O}(\varepsilon)$ problem to have a bounded solution is precisely enough to pin that dependence down. The frequency shift of Poincaré–Lindstedt reappears as one part of the answer; the amplitude decay it could not reach appears as another.

Example 5.3 (Duffing by multiple scales)**.** For $\ddot{x}+x+\varepsilon x^3 = 0$, the $\mathcal{O}(1)$ equation is $\partial_{T_0}^2 x_0 + x_0 = 0$. Its general solution oscillates in T_0 with coefficients that may depend on the slow time; writing this in complex form,

$$x_0 = A(T_1)e^{iT_0} + \bar{A}(T_1)e^{-iT_0},$$

keeps the algebra compact ($\bar{A}$ is the complex conjugate, present so that x_0 is real). At $\mathcal{O}(\varepsilon)$,

$$\partial_{T_0}^2 x_1 + x_1 = -2\,\partial_{T_0}\partial_{T_1} x_0 - x_0^3.$$

Expanding x_0^3 and collecting the coefficient of e^{iT_0} — the resonant term, since the operator on the left has natural frequency 1 — gives $-2iA' - 3A^2\bar{A}$, where a prime denotes $\mathrm{d}/\mathrm{d}T_1$. A bounded x_1 requires this to vanish:

$$2iA' + 3|A|^2 A = 0.$$

Now write $A = \frac{1}{2}R\,e^{i\phi}$ with real amplitude $R(T_1)$ and phase $\phi(T_1)$, and separate real and imaginary parts:

$$R' = 0, \qquad \phi' = \tfrac{3}{8}R^2.$$

Read these as physics. The first says the amplitude does *not* drift — a result, not an assumption. The second says the phase advances at the constant slow rate $\frac{3}{8}R^2$, which is a frequency shift. With $R = a$ fixed, $x \approx a\cos\big[(1+\frac{3}{8}\varepsilon a^2)t\big]$: exactly the Poincaré–Lindstedt result of Example 4.10, now obtained by a method that was not told in advance that the amplitude stays constant.

Example 5.4 (A slowly decaying oscillator)**.** Now the case that defeated both methods at the start of the section: the weakly damped oscillator of Example 5.1,

$$\ddot{x} + \varepsilon\dot{x} + x = 0.$$

With $x_0 = A(T_1)e^{iT_0} + \text{c.c.}$, the $\mathcal{O}(\varepsilon)$ forcing is $-2\,\partial_{T_0}\partial_{T_1}x_0 - \partial_{T_0}x_0$, whose resonant coefficient is $-2iA' - iA$. Removing it gives the amplitude equation

$$2A' + A = 0 \implies A \propto e^{-T_1/2},$$

so that

$$\boxed{x \approx a\,e^{-\varepsilon t/2}\cos t}.$$

This is the answer that Example 5.1 showed both earlier methods failing to reach. The amplitude decays on the slow timescale $t \sim 1/\varepsilon$ while the oscillation proceeds on the fast one, and multiple scales has extracted the decaying *envelope* without solving the full equation — something no single-clock expansion, strained or not, could do. Here $R' \neq 0$: the method has detected a drifting amplitude, precisely the situation it was built for.

Remark 5.5 (Averaging). The amplitude and phase equations produced by two-timing are identical to those of the *method of averaging*, which replaces the fast oscillation by its average over one period and follows the resulting slow drift of the envelope. The two viewpoints are interchangeable: two-timing is the more systematic bookkeeping, averaging the more intuitive picture — "what does the oscillation do *on average* over one cycle?" Either way, the object of interest is the slow evolution of an oscillator's envelope, not the fast wiggle inside it.[1]

[1]For a great lecture on the method of averaging, see S. Strogatz's online course lecture on averaging theory: https://www.youtube.com/watch?v=ma6w0guLxnI (part of his *Nonlinear Dynamics and Chaos* series).

5.2 Boundary layers and matched asymptotics*

The other singular failure. Chapter 4's tell-tale warned that trouble comes when ε multiplies the highest-order term. For an algebraic equation that cost us a root (Example 3.6); for a differential equation it costs us more, because ε multiplying the highest *derivative* means that setting $\varepsilon = 0$ lowers the *order* of the equation. A lower-order equation has fewer constants of integration, and so cannot in general satisfy all the boundary conditions. The resolution is the exact analogue of the lost-root rescaling: the discarded term, negligible almost everywhere, comes roaring back in a thin region — a *boundary layer* — where the solution changes fast enough to make $\varepsilon \times$ (highest derivative) comparable to the rest.

Recipe 5.6 (Boundary layers: location and structure)**.** For a linear second-order problem

$$\varepsilon y'' + p(x)\,y' + q(x)\,y = 0, \qquad y(0) = \alpha,\ y(1) = \beta, \qquad 0 < \varepsilon \ll 1,$$

with $p(x) \neq 0$ on the interval, proceed as follows.

1. *Outer solution.* Set $\varepsilon = 0$: solve the reduced first-order equation $p\,y' + q\,y = 0$. It carries one constant, so it can satisfy only *one* of the two boundary conditions.
2. *Locate the layer.* A layer of width $\mathcal{O}(\varepsilon)$ sits at *one* end; the sign of p decides which:

   $$p > 0 \ \Rightarrow\ \text{layer at the } \textit{left end}\ x = 0, \qquad p < 0 \ \Rightarrow\ \text{layer at the } \textit{right end}\ x = 1.$$

   The outer solution then takes the boundary condition at the *other* end. (The rule is the decay-versus-growth test of Chapter 4, Exercise 4: the layer must sit where the fast inner correction *decays* into the interior, which is the end where p carries the sign just stated.)
3. *Inner solution.* Rescale the stretched coordinate at that end — $x = \varepsilon\xi$ for a layer at 0, or $x = 1 - \varepsilon\xi$ for a layer at 1 — so that $\varepsilon y''$ is restored to full size. At leading order the balance $\varepsilon y'' \sim p\,y'$ gives a constant-coefficient equation for the inner solution; impose the boundary condition at that end.
4. *Match and combine.* Fix the remaining inner constant by matching the inner solution (as $\xi \to \infty$) to the outer solution (as it approaches the layer). The uniformly valid *composite* is

   $$y_{\text{comp}} = y_{\text{outer}} + y_{\text{inner}} - y_{\text{overlap}},$$

   the overlap being their common limit in the matching region.

The rest of this section carries out the recipe in full on one example. We take a problem with $p(x) = 2$ (so, by step 2 of the recipe, a layer at the left end):

$$\varepsilon y'' + 2y' + 2y = 0, \qquad y(0) = 0,\ y(1) = 1, \qquad 0 < \varepsilon \ll 1. \tag{5.1}$$

What setting $\varepsilon = 0$ does. Formally putting $\varepsilon = 0$ drops the equation from second order to first, $2y_0' + 2y_0 = 0$, i.e. $y_0' + y_0 = 0$ with solution $y_0 = Ce^{-x}$. This loss of order is the signature of a singular perturbation. There is now only *one* constant C but *two* boundary conditions, and no single choice can meet both: $y(0) = 0$ forces $C = 0$ and the trivial solution $y_0 \equiv 0$ (which fails $y(1) = 1$), while $y(1) = 1$ forces $C = e$ and $y_0 = e^{1-x}$ (which fails $y(0) = 0$). The reduced solution is therefore good on only part of the interval.

Remark 5.7. The same conclusion follows from attempting a regular expansion $y \sim y_0 + \varepsilon y_1 + \cdots$: substituting into (5.1) and collecting the $\mathcal{O}(1)$ terms gives exactly $y_0' + y_0 = 0$, one constant for two conditions. The regular expansion of Chapter 4 simply cannot be valid across the whole interval; something more is needed near one end.

Seeing the layer. This problem is simple enough to solve exactly, which lets us watch the layer form as ε shrinks. The characteristic equation $\varepsilon r^2 + 2r + 2 = 0$ has roots

$$r_{1,2} = \frac{-1 \pm \sqrt{1 - 2\varepsilon}}{\varepsilon},$$

real and distinct for small ε, so $y = Ae^{r_1 x} + Be^{r_2 x}$; imposing the two boundary conditions gives

$$y(x) = \frac{e^{r_1 x} - e^{r_2 x}}{e^{r_1} - e^{r_2}}.$$

Figure 5.1 plots this for $\varepsilon = 0.4,\ 0.1,\ 0.01$. For small ε the solution hugs e^{1-x} across most of the interval but swings rapidly down to meet $y(0) = 0$ in a thin region near $x = 0$ — the boundary layer, of width about ε.

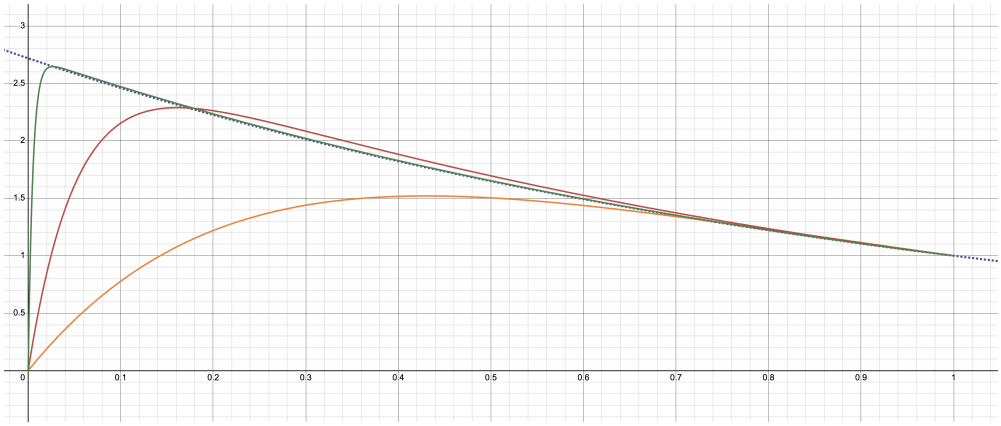


Figure 5.1: Exact solution of (5.1) for $\varepsilon = 0.4,\ 0.1$, and 0.01, with e^{1-x} dashed for comparison. As $\varepsilon \to 0$ the solution approaches e^{1-x} over most of the interval but changes ever more rapidly across a layer of width $\sim \varepsilon$ at $x = 0$.

We now build the approximation the recipe prescribes: an outer solution valid away from the layer, an inner solution valid within it, matched in their overlap and combined into one uniformly valid composite.

Step 1: outer solution

Away from $x = 0$, y and its derivatives are $\mathcal{O}(1)$, so $\varepsilon y''$ is negligible and the outer problem is $y_0' + y_0 = 0$, giving $y_{\text{outer}} = Ce^{-x}$. Since the layer sits at $x = 0$, the outer solution must carry the condition at the *other* end, $y(1) = 1$, so $C = e$ and

$$y_{\text{outer}}(x) = e^{1-x}.$$

Step 2: inner solution

To resolve the layer we stretch. Introduce $\xi = x/\delta$ with the layer width δ as yet unknown; then $\mathrm{d}/\mathrm{d}x = \delta^{-1}\mathrm{d}/\mathrm{d}\xi$ and $\mathrm{d}^2/\mathrm{d}x^2 = \delta^{-2}\mathrm{d}^2/\mathrm{d}\xi^2$, so (5.1) becomes

$$\frac{\varepsilon}{\delta^2} y'' + \frac{2}{\delta} y' + 2y = 0.$$

The width δ is fixed by dominant balance. Balancing the first term against the third would leave the middle term of size $\varepsilon^{-1/2}$, which *grows* as $\varepsilon \to 0$ — inadmissible. The only consistent choice balances the first two terms,

$$\frac{\varepsilon}{\delta^2} \sim \frac{1}{\delta} \quad \Rightarrow \quad \delta \sim \varepsilon,$$

so $\xi = x/\varepsilon$ (the $\mathcal{O}(\varepsilon)$ width promised by the recipe). Multiplying through by ε turns the equation into $y'' + 2y' + 2\varepsilon y = 0$; at leading order,

$$y''(\xi) + 2y'(\xi) = 0, \qquad y(0) = 0,$$

with solution $y_{\text{inner}} = C_1 + C_2 e^{-2\xi}$. The condition $y(0) = 0$ gives $C_1 + C_2 = 0$, so

$$y_{\text{inner}}(\xi) = C_2\big(e^{-2\xi} - 1\big).$$

Step 3: matching

The outer and inner solutions describe the same function on adjoining parts of the interval, so they must agree in the *overlap region* between them — where $x \to 0$ (leaving the outer solution's domain) and $\xi \to \infty$ (leaving the inner solution's domain), sketched in Figure 5.2. There the outer solution tends to $y_{\text{outer}}(0) = e$, while the inner solution tends to $-C_2$. Equating,

$$\lim_{x\to 0} y_{\text{outer}} = \lim_{\xi\to\infty} y_{\text{inner}} \quad \Rightarrow \quad e = -C_2,$$

so $C_2 = -e$ and $y_{\text{inner}}(\xi) = e\big(1 - e^{-2\xi}\big)$.

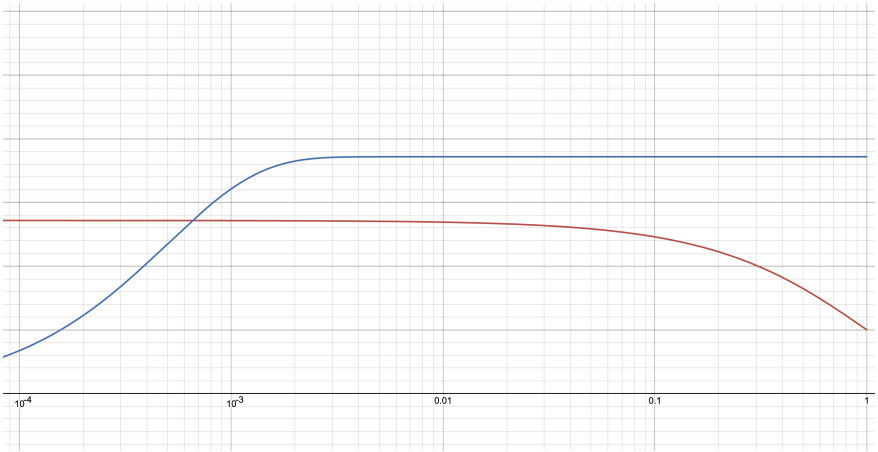


Figure 5.2: The matching. The outer solution is valid away from $x = 0$, the inner solution near it; in the overlap region ($x \to 0$, $\xi \to \infty$) both approach the common value e, which is what fixes the last inner constant.

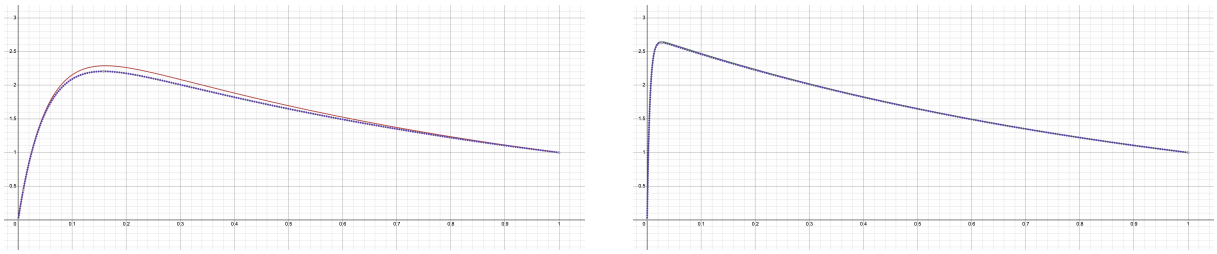
Step 4: composite solution

Neither piece is valid on the whole interval, so we combine them as in the recipe, adding the two and subtracting their common overlap value $L = e$ (counted twice otherwise):

$$y_{\text{comp}}(x) = y_{\text{outer}}(x) + y_{\text{inner}}(x/\varepsilon) - L = e^{1-x} + e\big(1 - e^{-2x/\varepsilon}\big) - e = e^{1-x} - e^{1-2x/\varepsilon}.$$

Figure 5.3 compares this composite with the exact solution: the agreement is already good at $\varepsilon = 0.1$ and excellent at $\varepsilon = 0.01$, with the error $\mathcal{O}(\varepsilon)$ as expected for a leading-order composite. Note that y_{comp} satisfies $y(0) = 0$ *exactly* but $y(1) = 1$ only asymptotically — a common and harmless feature of composite solutions.

Why the layer is at the left. The recipe placed it there because $p = 2 > 0$; the worked solution shows why. A layer can sit only where its inner correction *decays* into the interior, so that it can match the outer solution. Here the inner correction is $e^{-2\xi}$, which indeed decays as $\xi \to \infty$; had we instead stretched about $x = 1$, the corresponding correction would have *grown*, no matching would be possible, and no layer could live there. This is the decay-versus-growth test of Chapter 4, Exercise 4, and tracing its sign through the rescaling is exactly what gives the "$p > 0 \Rightarrow$ left" rule.



(a) $\varepsilon = 0.1$. (b) $\varepsilon = 0.01$.

Figure 5.3: Exact solution of (5.1) against the composite $e^{1-x} - e^{1-2x/\varepsilon}$, for decreasing ε. The composite is uniformly valid, and its error shrinks like $\mathcal{O}(\varepsilon)$.

5.3 Slow–fast systems and quasi-steady-state reduction

We now put a separation of timescales to work as a modelling tool in its own right. Suppose a model contains one variable that relaxes far faster than another:

$$\varepsilon\,\dot{x} = f(x,y), \qquad \dot{y} = g(x,y), \qquad 0 < \varepsilon \ll 1. \tag{5.2}$$

The factor ε on the left makes $\dot{x}$ of order $1/\varepsilon$: given any displacement, x moves fast. So x is the *fast* variable and y the *slow* one.

The reduction. Watch the fast variable from the slow variable's point of view. Because x relaxes almost instantly, at any moment it has essentially already reached whatever value makes its own rate vanish. Setting $\varepsilon = 0$ in (5.2) makes this precise: the fast equation collapses to the algebraic constraint

$$f(x,y) = 0,$$

the *critical manifold*. Solving it for a branch $x = h(y)$ and substituting into the slow equation[2] reduces the whole model to a single equation,

$$\dot{y} = g\big(h(y), y\big). \tag{5.3}$$

This is the *quasi-steady-state* (QSS) approximation: the fast variable is taken to sit at its instantaneous equilibrium, and we follow only the slow variable along the manifold. A two-dimensional problem has become one-dimensional — the dimension reduction promised at the outset.

When is it legitimate? Only if the equilibrium the fast variable is supposed to sit at is actually *stable* in the fast dynamics — otherwise x runs away from $h(y)$ rather than settling onto it. Freeze y and view the fast subsystem on its own fast clock $\sigma = t/\varepsilon$: it is $\mathrm{d}x/\mathrm{d}\sigma = f(x,y)$, a one-dimensional flow whose equilibrium $x = h(y)$ is attracting where $\partial f/\partial x < 0$ and repelling where $\partial f/\partial x > 0$ (the one-dimensional stability test of Appendix A). QSS is trustworthy only on the attracting branches.

Theorem 5.8 (Tikhonov, informal)**.** *If the branch $x = h(y)$ of the critical manifold is attracting ($\partial_x f < 0$ there) and the initial fast transient brings x onto it, then as $\varepsilon \to 0$ the solution of (5.2) converges — away from a thin initial layer — to the solution of the reduced equation (5.3) on that branch.*

The theorem also tells us exactly where QSS *fails*, and these are worth naming because a modeller must recognise them:

[2] This step is exactly the implicit function theorem: if f is smooth and $\partial f/\partial x \neq 0$ at a point of $f(x,y) = 0$, then near that point the constraint can be solved for a unique smooth branch $x = h(y)$. The condition $\partial_x f \neq 0$ is the same non-degeneracy that appears just below in the stability test — and it is precisely where it *fails*, $\partial_x f = 0$, that the branch cannot be continued: those are the *folds* of the manifold, the crux of the Van der Pol example to come.

(i) *The initial layer.* If the system does not start on the manifold, there is a brief fast transient — of duration $\mathcal{O}(\varepsilon)$ — during which x races to the manifold. QSS misses this; the reduced model is valid only afterwards.

(ii) *Repelling branches.* Where $\partial_x f > 0$ the fast dynamics flees the manifold, and QSS is simply wrong: the true solution is nowhere near $x = h(y)$.

(iii) *Folds.* Where an attracting branch bends round and loses stability, the reduced dynamics can no longer continue along the manifold, and the state *jumps* on the fast timescale to a distant branch. The reduction cannot describe the jump — but, as the next example shows, the jumps are the whole story.

Example 5.9 (The Van der Pol relaxation oscillator)**.** The Van der Pol equation

$$\ddot{v} - \mu(1 - v^2)\dot{v} + v = 0$$

becomes, at large μ, a classic slow–fast system. Writing it in Liénard form with $\varepsilon = 1/\mu^2$ and rescaling time, one obtains

$$\varepsilon\,\dot{x} = y - \left(\tfrac{x^3}{3} - x\right), \qquad \dot{y} = -x,$$

with $x = v$. The critical manifold $f = 0$ is the cubic

$$y = \tfrac{x^3}{3} - x,$$

an S-shaped curve. Its fast stability is set by $\partial_x f = -(x^2 - 1) = 1 - x^2$: the two outer branches $|x| > 1$ are *attracting*, the middle branch $|x| < 1$ is *repelling*, and the two meet at *folds* where $x = \pm 1$, i.e. at $(x, y) = (1, -\frac{2}{3})$ and $(-1, \frac{2}{3})$.

Now read the motion off the manifold. Land on the right attracting branch ($x > 1$). There the slow equation $\dot{y} = -x < 0$ drives y steadily *down* the branch until it reaches the lower fold at $(1, -\frac{2}{3})$. The branch ends; the fast dynamics takes over and flings the state horizontally across to the left attracting branch. On that branch $\dot{y} = -x > 0$ drives y *up* to the upper fold at $(-1, \frac{2}{3})$, whence another fast jump back. The result is a *relaxation oscillation*: slow crawls along the two outer branches, punctuated by near-instant jumps at the folds. QSS describes the crawls exactly and fails precisely at the folds — which is not a defect but the point, since the fast jumps it cannot capture are the defining feature of the motion.

Remark 5.10. The same mechanism drives excitable and oscillatory models across science: neurons firing, chemical clocks, and the electrical oscillator of Chapter 7. What transfers is not the algebra but the *picture*. An S-shaped critical manifold with two folds forces relaxation oscillations; a manifold that is monotone forces simple relaxation to equilibrium; a fold reached under slow drive signals an imminent fast jump. A modeller who can sketch $f = 0$, mark its stable and unstable branches, and follow the slow flow along it can often read off the qualitative dynamics of a slow–fast model *before* integrating anything.

Exercises

Exercise 5.1. Use multiple scales on $\ddot{x} + \varepsilon\dot{x}^3 + x = 0$ to derive the amplitude equation, and show that the amplitude decays *algebraically*, like $t^{-1/2}$. Contrast this with the *exponential* decay of the linear-damping case of Example 5.4, and explain physically why cubic damping is less effective at small amplitude.

Exercise 5.2. For the Van der Pol oscillator at *small* ε, $\ddot{x} + \varepsilon(x^2 - 1)\dot{x} + x = 0$, use multiple scales to show that the amplitude obeys $R' = \frac{1}{2}R\big(1 - \frac{1}{4}R^2\big)$, so that $R \to 2$ as $t \to \infty$ whatever the (nonzero) starting amplitude. Interpret this as the birth of a limit cycle: a single attracting closed orbit that the system approaches from inside *and* outside.

Exercise 5.3. Solve the boundary-layer problem $\varepsilon y'' + y' + y = 0$, $y(0) = 0$, $y(1) = 1$ by matched asymptotics — the analogue, with $p = 1$, of the worked example of Section 5.2. Show that the composite is $y \approx e^{1-x} - e^{1-x/\varepsilon}$. Verify it numerically at $\varepsilon = 0.05$, and confirm that the layer width scales with ε by comparing $\varepsilon = 0.05$ and $\varepsilon = 0.01$.

Exercise 5.4. For the linear slow–fast system $\varepsilon\dot{x} = y - x$, $\dot{y} = -x$, find the critical manifold, apply QSS, and solve the reduced equation. Then find the exact eigenvalues of the full linear system and show that the reduction captures the slow eigenvalue but misses the fast one. Why is QSS valid here — what is the sign of $\partial_x f$ on the manifold?

Exercise 5.5 (†). *Michaelis–Menten kinetics.* A nondimensional model of an enzyme-catalysed reaction is

$$\dot{s} = -s + (s + \kappa - \lambda)c, \qquad \varepsilon\,\dot{c} = s - (s + \kappa)c,$$

where s is substrate, c is the enzyme–substrate complex, and $\varepsilon \ll 1$ because the enzyme is scarce. Identify the fast variable, find the critical manifold $c = h(s)$, and show that QSS reduces the pair to the single *Michaelis–Menten rate law*

$$\dot{s} = -\frac{\lambda s}{s + \kappa}.$$

Check that the manifold is attracting (so the reduction is valid), and identify the brief initial layer the reduction omits. This reduction is one of the most-used approximations in mathematical biology.

Exercise 5.6 (†). For the slow–fast Van der Pol system of Example 5.9, integrate the full system numerically for $\varepsilon = 0.05$ and $\varepsilon = 0.01$. Overlay each trajectory on the cubic critical manifold in the (x, y) plane, and plot $x(t)$ to display the slow–fast alternation. Then demonstrate one failure of QSS numerically — for instance by starting off the manifold and showing the fast initial transient, or by zooming in near a fold to watch the jump. Referring to Theorem 5.8, explain exactly why QSS is invalid there.

Hints and answers

Ex. 1. With $x_0 = A(T_1)e^{iT_0} + \text{c.c.}$, write $\dot{x}_0 = iAe^{iT_0} + \text{c.c.}$ and cube it; the resonant (e^{iT_0}) part of $-\varepsilon\dot{x}_0^3$ combines with the $-2\partial_{T_0}\partial_{T_1}x_0$ term to give the solvability condition $2iA' + 3i|A|^2A \cdot (\text{const}) = 0$. In amplitude–phase form $A = \frac{1}{2}Re^{i\phi}$ this yields $\phi' = 0$ (no frequency shift) and

$$R' = -\tfrac{3}{8}R^3.$$

Separating, R^{-2} grows linearly: $R^{-2} = R_0^{-2} + \frac{3}{4}\varepsilon t$, so $R \sim (\frac{3}{4}\varepsilon t)^{-1/2} \propto t^{-1/2}$ at long times — algebraic, not exponential. Physically, cubic damping $\propto \dot{x}^3$ weakens fast as the oscillation shrinks (a small velocity cubed is very small), so it cannot sustain the constant *fractional* decay rate that linear damping gives; the envelope therefore trails off far more slowly. (A numerical check at $\varepsilon = 0.05$, $R_0 = 2$ matches $R^{-2} = \frac{1}{4} + \frac{3}{4}\varepsilon t$ to three figures.)

Ex. 2. With $x_0 = Ae^{iT_0} + \text{c.c.}$ the resonant coefficient of $-\varepsilon(x_0^2 - 1)\dot{x}_0$ is $\varepsilon\big(A - |A|^2A\big) \cdot (\text{factor})$; the solvability condition gives, in real form, $R' = \frac{1}{2}R(1 - \frac{1}{4}R^2)$. The right-hand side vanishes at $R = 0$ (unstable) and $R = 2$ (stable, since the bracket changes from $+$ to $-$ there), so every nonzero amplitude is drawn to $R = 2$. That single attracting amplitude, approached from both larger and smaller R, is a *limit cycle*: an isolated periodic orbit that neighbouring trajectories spiral onto. Contrast the Duffing case, where $R' = 0$ left the amplitude undetermined by the dynamics (a whole family of orbits, set by the initial data); here the nonlinearity *selects* one.

Ex. 3. Outer: drop $\varepsilon y''$ to get $y' + y = 0$, $y_{\text{out}} = Ce^{-x}$; the layer is at $x = 0$ (the coefficient of y' is positive), so y_{out} takes the right condition, $C = e$, $y_{\text{out}} = e^{1-x}$. Inner: $x = \varepsilon\xi$ gives $Y'' + Y' = 0$, $Y(0) = 0$, so $Y = A(1 - e^{-\xi})$; matching $Y(\infty) = y_{\text{out}}(0) = e$ fixes $A = e$. Composite $= e^{1-x} + e(1 - e^{-x/\varepsilon}) - e = e^{1-x} - e^{1-x/\varepsilon}$. Numerically at $\varepsilon = 0.05$ this matches the exact solution to a few percent (the error is $\mathcal{O}(\varepsilon)$, as expected for a leading-order composite); halving to $\varepsilon = 0.01$ visibly narrows the layer by a factor of five, confirming width $\propto \varepsilon$.

Ex. 4. Manifold: $y - x = 0$, so $x = h(y) = y$. QSS: $\dot{y} = -x = -y$, hence $y \propto e^{-t}$ — slow decay at rate 1. Exact system $\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -1/\varepsilon & 1/\varepsilon \\ -1 & 0 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$ has eigenvalues $\lambda = \big(-1 \pm \sqrt{1-4\varepsilon}\big)/(2\varepsilon)$, i.e. $\lambda_{\text{slow}} \approx -1$ and $\lambda_{\text{fast}} \approx -1/\varepsilon$. QSS keeps the slow eigenvalue -1 (matching $\dot{y} = -y$) and discards the fast one $-1/\varepsilon$, which governs the initial layer in which x collapses onto y. It is valid because $\partial_x f = \partial_x(y - x) = -1 < 0$: the manifold is attracting. (For $\varepsilon = 0.05$ the exact eigenvalues are -1.056 and -18.94.)

Ex. 5. The complex c is fast, since ε multiplies $\dot{c}$. Setting $\varepsilon = 0$: $s - (s + \kappa)c = 0$, so the manifold is $c = h(s) = \dfrac{s}{s + \kappa}$. Substituting into $\dot{s}$:

$$\dot{s} = -s + (s + \kappa - \lambda)\frac{s}{s + \kappa} = \frac{-s(s + \kappa) + (s + \kappa - \lambda)s}{s + \kappa} = -\frac{\lambda s}{s + \kappa},$$

the Michaelis–Menten law. The manifold is attracting: $\partial_c\big[s - (s + \kappa)c\big] = -(s + \kappa) < 0$. The omitted initial layer is the brief $\mathcal{O}(\varepsilon)$ phase in which c rises from its initial value (typically $c(0) = 0$, no complex yet) up to $h(s_0)$; only afterwards does the reduced law apply. This is the calculation Segel and others made rigorous, and it is why enzyme kinetics is usually written as a single rate law rather than a pair of equations.

Ex. 6. Numerical. Expect: for smaller ε the trajectory hugs the outer branches of the cubic more tightly and the jumps at the folds become sharper and faster; $x(t)$ looks increasingly like a square-ish wave (long slow segments, near-vertical transitions). Starting off the manifold (say $x(0)$ well away from a branch at some $y(0)$) shows a near-horizontal fast dash onto the nearest attracting branch, over a time $\mathcal{O}(\varepsilon)$ — the initial layer of Theorem 5.8(i). Near a fold, QSS fails because the attracting branch is ending: $\partial_x f = 1 - x^2 \to 0$ as $x \to 1$, so the fast relaxation that QSS relies on becomes arbitrarily slow just before it reverses into the jump — the hypothesis of the theorem ($\partial_x f < 0$, bounded away from 0) breaks down exactly there.

Recommended reading

Chapter 3 of Holmes [2] and also Chapter 2.4 to 2.5 of Holmes [1] remain the natural companion: it carefully and with many worked examples develops the methods of multiple scales, boundary layers, and matched asymptotics, and is the standard reference for the material in Sections 5.1 and 5.2. For the slow–fast material of Section 5.3, Strogatz [3] gives the clearest elementary account of relaxation oscillations and the Van der Pol geometry, with the phase-plane pictures drawn out in full. The quasi-steady-state reduction is also treated in Chapter 6.4 of Holmes [2].

References for this chapter

[1] M. H. Holmes. *Introduction to the Foundations of Applied Mathematics*. Vol. 56. Texts in Applied Mathematics. New York: Springer, 2009.

[2] M. H. Holmes. *Introduction to Perturbation Methods*. 2nd ed. New York: Springer, 2013.

[3] S. H. Strogatz. *Nonlinear Dynamics and Chaos*. 2nd ed. Boulder, CO: Westview Press, 2015.

Chapter 6

Constructing models: balance laws and closures

So far we have taken models as given and learned to simplify and analyse them. We now step back and ask where the equations come from. The answer, for an enormous range of models, has a fixed two-part structure: an exact *balance law* expressing conservation, and an empirical *constitutive closure* expressing the specific behaviour of the materials or mechanisms involved. Learning to see this structure — and to locate the closure, because that is the part that is really an assumption — is the key both to building models and to criticising them. We then reinterpret bifurcations, which your Dynamical Systems course computes, as *predictions* a model makes.

6.1 The balance law

Principle 6.1 (Balance / conservation)**.** *For any conserved quantity Q within a fixed region (a* control volume *or* compartment*),*

$$\frac{\mathrm{d}Q}{\mathrm{d}t} = (\textit{rate in}) - (\textit{rate out}) + (\textit{rate of generation}).$$

This bookkeeping is exact; it makes no assumption about mechanism. Applied to mass, momentum, energy, charge, or number of individuals, and to a small control volume in the continuum limit, it produces most of the governing equations of applied mathematics. The simplest instance is a well-mixed compartment.

Example 6.2 (A mixing tank)**.** A tank holds volume V of brine at concentration $c(t)$. Brine of concentration c_{in} flows in at volumetric rate q, and well-mixed brine flows out at the same rate. Balancing the mass of salt,

$$\frac{\mathrm{d}(Vc)}{\mathrm{d}t} = q\,c_{\mathrm{in}} - q\,c \quad \Rightarrow \quad \dot{c} = \frac{q}{V}\,(c_{\mathrm{in}} - c).$$

Nondimensionalising time by V/q gives $\dot{\hat{c}} = c_{\mathrm{in}} - \hat{c}$: the tank relaxes to c_{in} on the residence-time scale V/q. Everything here came from bookkeeping alone.

Balance laws are not confined to mass in tanks: the same bookkeeping governs momentum, energy, and angular momentum. Chaining compartments together — tanks in series, and more generally a directed graph of compartments with conserved flow along the edges — recurs across pharmacokinetics, ecology, and engineering (an exercise below). But to see the balance principle in a quite different guise, consider a swinging rigid body.

Example 6.3 (A rigid-body (compound) pendulum)**.** A rigid body of mass m is pivoted at a point P and swings in a vertical plane; let θ be the angle between the downward vertical and the

line from P to the centre of mass, which lies a distance R from P. The relevant balance is now for *angular momentum* about P: its rate of change equals the applied torque. Gravity, acting at the centre of mass, supplies a restoring torque $-mgR\sin\theta$, so

$$I\,\ddot{\theta} = -\,mgR\sin\theta,$$

where I is the moment of inertia of the body about P. This is an exact balance law with no empirical closure, and it has exactly the form of the simple pendulum. For small oscillations $\sin\theta \approx \theta$, giving $\ddot{\theta} + (mgR/I)\theta = 0$ and period

$$T = 2\pi\sqrt{\frac{I}{mgR}} = 2\pi\sqrt{\frac{L_{\text{eff}}}{g}}, \qquad L_{\text{eff}} := \frac{I}{mR}.$$

So *any* rigid-body pendulum swings like a simple pendulum of length L_{eff}: the entire extended body collapses to a single effective length. (A point mass at distance L has $I = mL^2$ and $R = L$, hence $L_{\text{eff}} = L$, as it must.) All the geometry enters only through I, obtained from a mass integral, and the dimensional prediction $T \propto \sqrt{L_{\text{eff}}/g}$ of Chapter 2 reappears — with the effective length now made explicit.

6.2 Constitutive closures

The balance law is never enough by itself: it relates rates to fluxes, but says nothing about *what the fluxes are*. That gap is filled by a *constitutive relation* (or *closure*), an empirical statement about the specific system.

Definition 6.4 (Constitutive closure)**.** A *constitutive relation* expresses a flux or source in terms of the state variables. Unlike a balance law it is not exact; it is a modelling choice, validated by experiment within a limited range.

Familiar closures include:

- *Newton's law of cooling*: heat flux $\propto$ temperature difference, $\dot{T} = -k(T - T_a)$;
- *Fick's law*: diffusive flux $\propto -$concentration gradient;
- *Ohm's law*: current $\propto$ voltage;
- *the law of mass action*: reaction rate $\propto$ product of concentrations;
- *a traffic fundamental diagram*: vehicle flux as an empirical function of density;
- *a nonlinear circuit element*: current as a prescribed function of voltage (Chapter 7).

Principle 6.5 (Anatomy of a model)**.** *model = balance law (exact)+constitutive closure (empirical). When reading or criticising a model,* find the closure*: it is where the assumptions — and the vulnerabilities — live.*

Example 6.6 (Car-following: a model and its closure)**.** Model a line of cars, each reacting to the one ahead. For car i, let h_i be its *headway* (the gap to the car in front) and v_i its velocity. Two ingredients build the model. The first is exact bookkeeping: the headway grows at the rate the car ahead pulls away,

$$\dot{h}_i = v_{i+1} - v_i,$$

a kinematic balance that is not negotiable. The second is the *closure* — a model of the driver. A common choice is that each driver accelerates towards a preferred speed set by the gap ahead,

$$\dot{v}_i = \alpha\big(V(h_i) - v_i\big),$$

with sensitivity α and an *optimal-velocity function* $V(h)$ giving the speed a driver chooses at headway h. This V is the constitutive closure, and it is a pure modelling choice: it should vanish at small gaps (stop when too close), rise smoothly, and saturate at the speed limit V_t for large gaps. A convenient family is

$$V(h) = V_t \frac{[(h - h_{\min})/s]^k}{1 + [(h - h_{\min})/s]^k} \quad (h > h_{\min}), \qquad V = 0 \ \ (h \le h_{\min}),$$

whose shape is tuned by the parameters s and k. All the driver behaviour lives in this closure; the bookkeeping equation is fixed, but V is exactly the part one argues about, refines, and tests — as we do next.

6.3 Bifurcations as model predictions

Your Dynamical Systems course teaches how to *find* bifurcations. In modelling, the point of a bifurcation is different: it is a *falsifiable prediction* that the system undergoes a qualitative change — switching on, buckling, beginning to oscillate — as a parameter crosses a threshold. Computing where the bifurcation lies is computing a prediction one can test.

We recall the three basic local bifurcations only as *predictions* (normal forms; the machinery is elsewhere):

- *Saddle–node* $\dot{x} = r - x^2$: two equilibria for $r > 0$, none for $r < 0$ — predicts the abrupt *appearance or disappearance* of a steady state (a tipping point).
- *Transcritical* $\dot{x} = rx - x^2$: two equilibria exchange stability at $r = 0$ — predicts an *onset threshold* beyond which a previously-zero state becomes active.
- *Hopf* $\dot{r} = \alpha r - r^3$ (in polar form for a planar system) at parameter crossing $\alpha = 0$: a stable equilibrium gives way to a small-amplitude limit cycle — predicts the *onset of spontaneous oscillation.*

Example 6.7 (The onset of traffic jams)**.** Put n cars on a ring road of circumference L, each following the one ahead as in Example 6.6 (the last car follows the first). The model always admits a *uniform-flow* state: every car equally spaced at headway $h^* = L/n$, all moving at the same speed $v^* = V(h^*)$. The modelling question is whether this smooth flow is stable — and this is where a bifurcation makes a prediction. Linearising about the uniform state (the computation, using Appendix A, is pursued in the tutorials) shows that uniform flow is stable precisely when

$$V'(h^*) < \tfrac{1}{2}\,\alpha,$$

and unstable when $V'(h^*) > \frac{1}{2}\alpha$. The prediction is sharp and testable: when drivers are too sluggish (α small) or the density sits on the steep part of the optimal-velocity curve (V' large), smooth flow spontaneously breaks into stop-and-go waves — *phantom traffic jams* with no external cause. Increasing the density through the critical value crosses this stability boundary exactly as one crosses a bifurcation; and in the dense limit $h^* \le h_{\min}$, where $V(h^*) = 0$, the flow locks up completely into gridlock. The optimal-velocity closure we chose in Example 6.6 thus predicts both the jamming instability and its threshold — and a different closure would move that threshold, which is the sense in which the prediction tests the model.

Remark 6.8. Buckling of a loaded column (a saddle-node/pitchfork as load increases) and the onset of oscillation in a feedback system (a Hopf as gain increases) are read the same way: each bifurcation is a boundary in parameter space across which behaviour changes qualitatively, and the model's job is to predict where that boundary lies.

6.4 Sanity-checking a model

A model that has just been written down should be interrogated before it is trusted.

- *Dimensional check*: is every term dimensionally homogeneous (Chapter 2)?
- *Limiting cases*: do known simple behaviours emerge when a parameter is set to zero or infinity?
- *Conservation check*: does the model conserve what it should (total mass, probability, energy in the absence of sources)?
- *Sign and monotonicity*: do the signs of terms encode the right physical tendencies (does adding damping remove energy)?
- *Qualitative behaviour*: does the model at least have the right kind of solutions (bounded, oscillatory, monotone) as physical intuition demands?

A model that fails any of these is telling you something before you spend effort solving it.

Exercises

Exercise 6.1. A lake of volume V receives a pollutant at constant rate S and loses it by outflow q and by first-order decay at rate λ. Write the balance law and the closures, form the model, nondimensionalise, and find the steady-state concentration. Identify which parts are exact and which are constitutive.

Exercise 6.2. Assemble a model for a falling body subject to *quadratic* drag by combining Newton's second law (balance) with the closure $F_{\text{drag}} = \frac{1}{2}C_D\rho A v^2$. Find the terminal velocity and compare, as a modelling choice, with the linear-drag closure of Chapter 3.

Exercise 6.3. For the car-following model of Example 6.6 on a ring road, uniform flow is stable if and only if $V'(h^*) < \frac{1}{2}\alpha$. Using the optimal-velocity family given there with $k = 3$, $s = 1$, plot $V(h)$ and $V'(h)$, find the headway at which V' is largest, and hence identify the range of densities most prone to spontaneous jams. What does increasing the sensitivity α do to that range?

Exercise 6.4. Two well-mixed tanks in series, each of volume V with throughflow q, obey $\dot{c}_1 = \frac{q}{V}(c_{\text{in}} - c_1)$ and $\dot{c}_2 = \frac{q}{V}(c_1 - c_2)$. Derive these balances, and show that if $c_{\text{in}} = 0$ and the outflow of the second tank is returned to the first (a closed loop), the total salt is conserved. What does it tell you if your equations fail this conservation check?

Exercise 6.5. A uniform rod of mass m and length L is pivoted at one end. Given that its moment of inertia about that end is $I = \frac{1}{3}mL^2$ and its centre of mass sits at $R = L/2$, use Example 6.3 to find its effective length L_{eff} and its small-oscillation period, and compare with a simple pendulum of length L.

Exercise 6.6 (†). Choose a phenomenon with an obvious threshold (a heated wire that suddenly glows, a population that suddenly persists, an amplifier that suddenly whistles). Propose a minimal model as balance-plus-closure, identify the bifurcation that would produce the threshold, and state what quantitative prediction the model makes about *where* the threshold sits. Say which part of your model you are least sure of, and why it is the closure.

Chapter 7

Reading and improving a model

The final chapter adds the last skill a modeller needs. So far we have *built* models; but just as often one is handed a model already made — by a colleague, in a report, in a research paper — and the task is to understand it, judge it, and improve it. That is the discipline here: *reconstruct* a model from its assumptions, *locate the closure*, *criticise* it, and *extend* it so that it makes a new, checkable prediction. We work it through in full on the car-following model of traffic (Chapter 6), which is transparent enough to rebuild in a line yet rich enough to criticise and extend; a second example, the Van der Pol oscillator, shows the same steps applied elsewhere. Everything is self-contained — no outside reading is required — but a classic paper is suggested at the end for anyone who would like to see these ideas in an original source.

7.1 How to read a model

Whether a model reaches you as a paragraph of assumptions or a published paper, the way to read it is the same, and it is not from top to bottom. Read instead for structure:

- read first for the *model*: what are the state variables, what is conserved (the balance law), and what is the constitutive assumption (the closure)?
- separate the *stated* assumptions from the *unstated* ones — the latter are usually the interesting ones;
- locate the *closure* (Principle 6.5): it is the negotiable core of the model, and the first place to push;
- ask what the model *predicts* that could be wrong, and under what circumstances it would fail.

Rewriting the model in your own notation, from its assumptions, is the single most valuable step; only then are you in a position to improve it.

7.2 Reconstructing a model: car-following

We met the car-following model in Chapter 6; here we rebuild it from its assumptions, as an exercise in reconstruction. Picture a single lane of cars, numbered along the road, each driver watching only the car ahead. Two assumptions suffice.

Bookkeeping (the balance law). If h_i is the headway of car i — the gap to the car in front — then the gap grows exactly as fast as the car ahead pulls away:

$$\dot{h}_i = v_{i+1} - v_i.$$

This is exact; it assumes nothing about behaviour.

The driver (the closure). Each driver has a preferred speed that depends on the gap, and accelerates towards it:

$$\dot{v}_i = \alpha\big(V(h_i) - v_i\big),$$

with sensitivity α and an *optimal-velocity function* $V(h)$ — zero at small gaps (stop when too close), rising smoothly, saturating at the speed limit for large gaps (the family with shape parameters k, s was given in Chapter 6).

That is the whole model. Its two parts are now plain: the headway equation is fixed bookkeeping, while $V(h)$ and the sensitivity α carry every assumption about how people drive. *The closure is $V(h)$.*

7.3 Criticising the model

With the structure exposed, we interrogate the assumptions — aiming, as always, at the closure and the omissions rather than at surface detail.

- *Instantaneous reaction.* Real drivers respond to the gap only after a reaction-time delay; the model assumes none. This is the most consequential omission.
- *The shape of $V(h)$.* The optimal-velocity curve is fitted to typical behaviour, not derived; its steepness (the parameters k, s) is the softest, most negotiable part of the model.
- *Identical drivers.* Every car shares the same α and V; real traffic mixes cautious and aggressive drivers.
- *Determinism, single lane.* There are no fluctuations and no overtaking.

Each is a place the model might fail. The reaction delay and the closure $V(h)$ are the weakest, and so the first candidates for improvement.

7.4 Improving the model as a prediction

The modelling discipline of Chapter 6 insists that an improvement *earn its keep by making a prediction* — a falsifiable, qualitative claim, ideally a transition as some parameter is varied.

The base model already makes one such prediction, derived in Chapter 6: on a ring road the uniform-flow state loses stability, breaking into stop-and-go waves, exactly when

$$V'(h^*) > \tfrac{1}{2}\alpha.$$

Crossing this threshold — by raising the density, so h^* falls onto the steep part of V, or by lowering the sensitivity α — is a bifurcation, and the resulting "phantom jams" are the model earning its keep.

Now improve it, and watch the prediction move:

- *Add a reaction delay τ*, so a driver responds to the gap as it was a moment earlier. The prediction: the delay is destabilising, so the jamming threshold shifts and jams appear at *lower* sensitivity or density than before — a definite, testable change.
- *Make one driver slow* (a smaller V or α). The prediction: a persistent bottleneck forms behind that car, with a standing queue upstream.

And two further predictions of the base model, from Chapter 6, are worth reproducing: the flow rate $r(h) = V(h)/h$ is maximal at the *capacity* headway where $V'(h^*) = V(h^*)/h^*$, and in the dense limit $h^* \le h_{\min}$ the flow locks up into gridlock.

Reproducing the base prediction (ring-road jams) numerically, then implementing one of these improvements and testing the shift it predicts, is precisely the capstone project — and, in miniature, what research modelling is: take a transparent model, find where it is thin, and extend it so that it makes a new claim you can check.

7.5 The same discipline elsewhere: the Van der Pol oscillator

The steps just taken — reconstruct, find the closure, improve-as-prediction — are not special to traffic. As a second illustration, consider a model built by *analogy*: an electrical circuit whose behaviour mimics a self-sustaining oscillation. It contains an inductor L, a capacitor C, and a nonlinear element whose current–voltage characteristic $i = f(V)$ has a region of *negative* slope near the origin.

Reconstruction. The balance law is Kirchhoff's current law at the node, $C\dot{V} + I_L + f(V) = 0$ with $L\dot{I}_L = V$; differentiating,

$$C\ddot{V} + f'(V)\dot{V} + \frac{V}{L} = 0.$$

The closure is the characteristic f. Van der Pol's cubic choice, $f(V) = -\alpha V + \frac{1}{3}\beta V^3$ so that $f'(V) = -\alpha + \beta V^2$, is what makes it oscillate: near the origin the effective damping is negative and *feeds* energy in, while for large V it dissipates, driving the amplitude to a fixed value. Scaling voltage by $\sqrt{\alpha/\beta}$ and time by $\sqrt{LC}$ collapses every constant into one dimensionless parameter,

$$\boxed{\ddot{v} - \mu(1 - v^2)\dot{v} + v = 0}, \qquad \mu = \alpha\sqrt{L/C}. \tag{7.1}$$

This is the Van der Pol equation; as with traffic, the interesting behaviour lives entirely in the closure. Its large-μ motion is the relaxation oscillation analysed in Chapter 5 (Example 5.9): a slow build-up and a fast discharge, repeating.

Improvement-as-prediction. Two natural extensions each make a checkable claim. Forcing the oscillator periodically, $\ddot{v} - \mu(1 - v^2)\dot{v} + v = A\cos\Omega t$, predicts *frequency locking*: the oscillator entrains to rational ratios $p{:}q$ of the drive over intervals of Ω (*Arnold tongues*), with the locking ratio moving through a *devil's staircase* as Ω is swept — a genuinely nonlinear effect no linear system can produce. Coupling two such oscillators predicts *mutual entrainment* above a threshold coupling strength, and its loss below — again a bifurcation. The discipline is identical to the traffic case; only the model has changed.

Further reading (optional)

The continuous cousin of our car-following model — traffic treated as a fluid, with a density-dependent flux — was introduced in a classic pair of short papers by Lighthill and Whitham (1955) and Richards (1956), the origin of the *fundamental diagram* relating flow to density. They make a rewarding first research paper to read: brief, self-contained, and squarely on the topic of this chapter. Reading them is entirely optional — nothing above depends on it — but anyone wishing to see these ideas in an original source will find them a good place to start.

Exercises

Exercise 7.1. From the two assumptions of Section 7.2 — headway bookkeeping and a preferred-speed closure — reconstruct the car-following equations in your own notation, and state clearly which part is the exact balance law and which is the constitutive closure.

Exercise 7.2. Propose one improvement to the car-following model (a reaction-time delay, a heterogeneous slow driver, or a different closure $V(h)$), and state the *prediction* it makes: a qualitative transition, and what one would observe on each side of it.

Exercise 7.3. Carry out in full the nondimensionalisation leading from $C\ddot{V} + (-\alpha + \beta V^2)\dot{V} + V/L = 0$ to (7.1), and confirm $\mu = \alpha\sqrt{L/C}$.

Exercise 7.4. Verify the Liénard transformation used in Example 5.9: starting from (7.1), show that $y = \dot{v}/\mu + \frac{1}{3}v^3 - v$ leads to the slow–fast system quoted there, and identify the small parameter.

Exercise 7.5 (†). Force a single Van der Pol oscillator, $\ddot{v} - \mu(1 - v^2)\dot{v} + v = A\cos(\Omega t)$. Fixing μ and A, sweep Ω and record the locking ratio (output period over forcing period). Plot the resulting devil's staircase and interpret each plateau as a frequency-locked state.

Appendix A

Linear stability and the phase plane

Several tutorials and assignments ask you to find the equilibria of a model, decide whether they are stable, and sketch the resulting dynamics. These are the standard tools of the qualitative theory of differential equations — the subject of a typical Dynamical Systems course — and we collect here, in the compact form we need for modelling, the few facts that make those problems self-contained. Nothing is proved in full.

A.1 Equilibria and linearisation

Consider an autonomous system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^n.$$

An *equilibrium* (fixed point, steady state) is a point $\mathbf{x}^*$ with $\mathbf{f}(\mathbf{x}^*) = \mathbf{0}$: started there, the system stays there. To see what happens *near* $\mathbf{x}^*$, put $\mathbf{x} = \mathbf{x}^* + \boldsymbol{\xi}$ with $\boldsymbol{\xi}$ small and Taylor-expand,

$$\dot{\boldsymbol{\xi}} = \mathbf{f}(\mathbf{x}^* + \boldsymbol{\xi}) = J\boldsymbol{\xi} + \mathcal{O}(|\boldsymbol{\xi}|^2), \qquad J = D\mathbf{f}(\mathbf{x}^*), \quad J_{ij} = \left.\frac{\partial f_i}{\partial x_j}\right|_{\mathbf{x}^*},$$

where J is the *Jacobian* at the equilibrium. Discarding the quadratic remainder leaves the *linearisation* $\dot{\boldsymbol{\xi}} = J\boldsymbol{\xi}$.

Principle A.1 (Linearised stability)**.** *If every eigenvalue of J has* negative *real part, the equilibrium is asymptotically stable (nearby solutions return to it); if some eigenvalue has* positive *real part, it is unstable. When no eigenvalue has zero real part — the* hyperbolic *case — the nonlinear flow near $\mathbf{x}^*$ looks qualitatively like the linear flow $\dot{\boldsymbol{\xi}} = J\boldsymbol{\xi}$ (the Hartman–Grobman theorem). A zero real part is borderline: the linearisation does not decide, and the neglected nonlinear terms take over.*

This turns "is this equilibrium stable?" into "compute the eigenvalues of a matrix."

A.2 The planar case: trace and determinant

For a planar system ($n = 2$) the eigenvalues of J satisfy the characteristic equation

$$\lambda^2 - \tau\lambda + \Delta = 0, \qquad \tau = \operatorname{tr} J, \quad \Delta = \det J,$$

so $\lambda_\pm = \frac{1}{2}\big(\tau \pm \sqrt{\tau^2 - 4\Delta}\,\big)$, and the whole classification follows from the two numbers τ and Δ (Figure A.1):

- $\Delta < 0$: real eigenvalues of opposite sign — a *saddle* (always unstable); its eigenvectors give the incoming (stable) and outgoing (unstable) directions.

- $\Delta > 0,\ \tau^2 > 4\Delta$: real eigenvalues of the same sign — a *node*, stable if $\tau < 0$, unstable if $\tau > 0$.
- $\Delta > 0,\ \tau^2 < 4\Delta$: complex conjugate eigenvalues — a *spiral* (focus), stable if $\tau < 0$, unstable if $\tau > 0$.
- $\tau = 0,\ \Delta > 0$: pure imaginary eigenvalues — a *centre* (borderline; nonlinear terms decide).

In short: $\Delta > 0$ **together with** $\tau < 0$ **is the condition for linear stability**; $\Delta < 0$ is a saddle; and the sign of $\tau^2 - 4\Delta$ separates nodes from spirals.

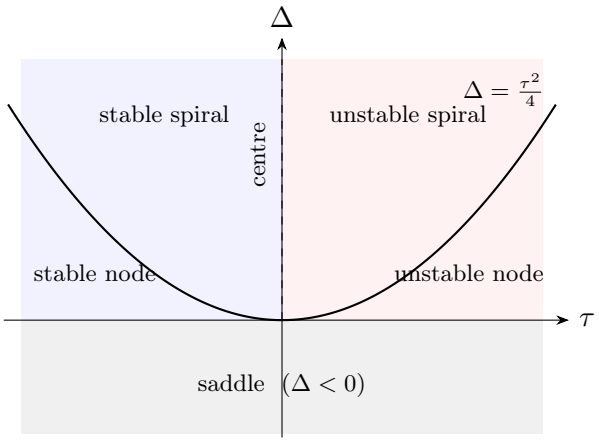


Figure A.1: The trace–determinant plane. Stability lives in the upper-left quadrant ($\Delta > 0$, $\tau < 0$); the parabola $\Delta = \tau^2/4$ separates nodes (below) from spirals (above); everything below the τ-axis is a saddle.

Remark A.2 (Stability regions in a parameter plane). When a model carries parameters, τ and Δ depend on them, and the two inequalities $\Delta > 0$, $\tau < 0$ carve out the region of parameter space in which the equilibrium is stable. Finding such a region — for example in the control gains (a, b) of a balancing problem — is exactly a matter of writing down τ and Δ and imposing these inequalities.

A.3 A worked example: the pendulum

The nondimensional damped pendulum of Chapter 3, written as a first-order system, is

$$\dot{x} = y, \qquad \dot{y} = -\sin x - c\,y, \qquad c \geq 0.$$

Equilibria require $y = 0$ and $\sin x = 0$, i.e. $(x, y) = (n\pi, 0)$: the bob hanging straight down ($x = 0$) or balanced straight up ($x = \pi$), and their 2π-translates. The Jacobian is

$$J(x, y) = \begin{pmatrix} 0 & 1 \\ -\cos x & -c \end{pmatrix}.$$

Downward equilibrium $(0, 0)$: here $J = \left(\begin{smallmatrix} 0 & 1 \\ -1 & -c \end{smallmatrix}\right)$, so $\tau = -c$ and $\Delta = 1$. Since $\Delta = 1 > 0$, the point is stable for every $c > 0$; and $\tau^2 - 4\Delta = c^2 - 4$ sorts the behaviour into the three damping regimes: a *centre* when $c = 0$ (undamped — perpetual oscillation), a *stable spiral* for $0 < c < 2$ (under-damped — decaying oscillation), and a *stable node* for $c > 2$ (over-damped — no oscillation).

Upward equilibrium $(\pi, 0)$: now $\cos\pi = -1$, so $J = \left(\begin{smallmatrix} 0 & 1 \\ 1 & -c \end{smallmatrix}\right)$ and $\Delta = -1 < 0$: a *saddle*, for every c. Balancing a pendulum upright is unstable no matter how it is damped — which is why it takes active control. Its eigenvectors, from $(J - \lambda I)\mathbf{v} = \mathbf{0}$, are $\mathbf{v}_\pm = (1, \lambda_\pm)$ with $\lambda_\pm = \frac{1}{2}\left(-c \pm \sqrt{c^2 + 4}\right)$: the stable direction ($\lambda_- < 0$) and the unstable direction ($\lambda_+ > 0$).

A.4 Phase portraits

A *phase portrait* is a sketch of representative trajectories in the state space. A workable recipe: mark the equilibria and their local types (from the Jacobian); draw the *nullclines*, the curves $\dot{x} = 0$ and $\dot{y} = 0$ on which trajectories move purely vertically or horizontally; use the eigenvector directions at each saddle to draw the incoming and outgoing separatrices; then fill in the flow consistently between them. For the pendulum this yields the familiar picture: closed orbits circling each downward equilibrium, saddles at the upright positions, and — once damping is present — spirals winding into the downward rest state.

Exercises

Exercise A.1. Classify the equilibrium at the origin of $\dot{\mathbf{x}} = A\mathbf{x}$ for $A = \left(\begin{smallmatrix} -1 & 2 \\ 0 & -3 \end{smallmatrix}\right)$, $A = \left(\begin{smallmatrix} 0 & 1 \\ -4 & 0 \end{smallmatrix}\right)$, and $A = \left(\begin{smallmatrix} 1 & 2 \\ 2 & 1 \end{smallmatrix}\right)$, using τ and Δ in each case.

Exercise A.2. Show that $\dot{\boldsymbol{\xi}} = \left(\begin{smallmatrix} 0 & 1 \\ \alpha & \beta \end{smallmatrix}\right)\boldsymbol{\xi}$ is asymptotically stable if and only if $\alpha < 0$ and $\beta < 0$. Interpreting $-\alpha$ as an effective stiffness and $-\beta$ as an effective damping, explain why stabilising an inverted pendulum by feedback requires the controller to supply *both* a restoring force and damping.

Exercise A.3 (†). A model has an equilibrium whose Jacobian is $J = \left(\begin{smallmatrix} -1 & 1 \\ \gamma & -1 \end{smallmatrix}\right)$, with γ a parameter. Find the value of γ at which the equilibrium changes stability, identify what type it is on either side, and describe what happens to nearby trajectories as γ crosses that value.

Appendix B

Partial derivatives and partial differential equations

Two chapters of these notes reach, briefly, beyond ordinary differential equations. Chapter 5 treats the solution of an oscillator as a function of *two* time variables and differentiates with respect to each; Chapters 6 and 7 describe quantities that vary in *space as well as time*, governed by balance laws in continuum form. Neither uses the machinery of a full course on partial differential equations. This appendix collects the small amount that is actually needed, so that the notes stay self-contained. It assumes only what a second calculus course provides: partial derivatives and the multivariable chain rule.

B.1 Partial derivatives and the chain rule

A function of several variables, say $u(x,t)$, has a *partial derivative* with respect to each. The symbol $\partial_x u \equiv \partial u/\partial x$ means: differentiate u with respect to x, holding every other variable (t here) fixed. It is an ordinary derivative in disguise — the disguise being that the other variables are frozen. Higher and mixed partials follow the same rule; for the smooth functions we deal with, the order of mixed differentiation does not matter, $\partial_x \partial_t u = \partial_t \partial_x u$.

The one tool we lean on is the *chain rule*. If a quantity depends on t through several intermediate variables, its rate of change collects one contribution from each. For $u\big(a(t), b(t)\big)$,

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \frac{\partial u}{\partial a}\frac{\mathrm{d}a}{\mathrm{d}t} + \frac{\partial u}{\partial b}\frac{\mathrm{d}b}{\mathrm{d}t}.$$

Everything in this appendix is an application of these two facts.

B.2 The two-scale chain rule (Chapter 5)

The method of multiple scales writes the solution as a function of a fast time $T_0 = t$ and a slow time $T_1 = \varepsilon t$, treated as independent, and then needs $\mathrm{d}/\mathrm{d}t$ and $\mathrm{d}^2/\mathrm{d}t^2$ expressed through ∂_{T_0} and ∂_{T_1}. This is exactly the chain rule above, with $a = T_0$ and $b = T_1$.

Since $T_0 = t$ gives $\mathrm{d}T_0/\mathrm{d}t = 1$, and $T_1 = \varepsilon t$ gives $\mathrm{d}T_1/\mathrm{d}t = \varepsilon$, the chain rule reads

$$\frac{\mathrm{d}}{\mathrm{d}t} = \frac{\mathrm{d}T_0}{\mathrm{d}t}\,\partial_{T_0} + \frac{\mathrm{d}T_1}{\mathrm{d}t}\,\partial_{T_1} = \partial_{T_0} + \varepsilon\,\partial_{T_1}.$$

That is the operator identity the recipe uses. The second derivative is obtained by applying it twice:

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} = \big(\partial_{T_0} + \varepsilon\,\partial_{T_1}\big)^2 = \partial_{T_0}^2 + 2\varepsilon\,\partial_{T_0}\partial_{T_1} + \varepsilon^2\,\partial_{T_1}^2,$$

where the cross term carries a factor 2 because $\partial_{T_0}\partial_{T_1} = \partial_{T_1}\partial_{T_0}$ (equality of mixed partials), and the operator is squared as an ordinary binomial.

Why this is not really a partial differential equation. At each order in ε, the multiple-scales calculation produces an equation such as $\partial_{T_0}^2 x_0 + x_0 = 0$. Although $x_0(T_0, T_1)$ is a function of two variables, this is solved as an *ordinary* differential equation in T_0 alone — its solution is $\cos T_0$ and $\sin T_0$ — with the only novelty that the "constants" of integration may depend on the other variable T_1. The equation that then fixes that dependence, the solvability condition, is likewise an *ordinary* differential equation, in T_1. So multiple scales is two ordinary differential equations stitched together by the chain rule above; it borrows the notation of partial derivatives but none of the theory of partial differential equations. That theory is the subject of the next sections, and is needed only for the balance-law chapters.

B.3 What a partial differential equation is

A *partial differential equation* (PDE) is an equation relating the partial derivatives of an unknown function of two or more variables. Where an ordinary differential equation determines a function of one variable — a trajectory $x(t)$ — a PDE determines a *field*: a quantity like a density $\rho(x,t)$ or a temperature $u(x,t)$ spread over space and evolving in time. The balance-law chapters need only the simplest and most physically transparent PDE, the one that expresses conservation, together with two standard examples built from it. We do not solve these equations here; we only need to recognise them and know where they come from.

B.4 The continuum balance law (Chapters 6, 7)

Chapter 6 states the balance principle for a well-mixed compartment: the rate of change of the amount of some conserved quantity equals what flows in minus what flows out. When the quantity is instead spread continuously along a line — cars on a road, heat in a rod, a pollutant in a river — the same bookkeeping produces a PDE.

Let $\rho(x,t)$ be the *density* of the conserved quantity (amount per unit length) and $q(x,t)$ its *flux* (amount crossing a point per unit time, in the $+x$ direction). Fix any interval $[a,b]$. The amount inside it is $\int_a^b \rho\,\mathrm{d}x$, and by the balance principle its rate of change equals the flux in at a minus the flux out at b:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_a^b \rho(x,t)\,\mathrm{d}x = q(a,t) - q(b,t).$$

Write the right-hand side as an integral, $q(a,t) - q(b,t) = -\int_a^b \partial_x q\,\mathrm{d}x$, and move the time derivative inside the left-hand integral:

$$\int_a^b \big(\partial_t \rho + \partial_x q\big)\mathrm{d}x = 0.$$

This holds for *every* interval $[a,b]$, which is possible only if the integrand itself vanishes everywhere. Hence the *continuum balance law*

$$\boxed{\partial_t \rho + \partial_x q = 0}$$

(or $\partial_t \rho + \partial_x q = S$ if there is a source of strength S per unit length). It is the field version of the compartment balance of Chapter 6, and like that balance it is exact: it says only that the quantity is conserved. As always, it is not closed until a *constitutive relation* supplies the flux q in terms of ρ — exactly the closure step of that chapter.

B.5 Two equations worth recognising

Different closures turn the balance law into different PDEs. Two recur so often that they are worth knowing by sight.

The transport (advection) equation. If the conserved stuff is simply carried along at speed c, the flux is $q = c\rho$, and the balance law becomes

$$\partial_t \rho + c\,\partial_x \rho = 0.$$

Its solutions are travelling waves: any initial profile $\rho(x,0) = \rho_0(x)$ slides rightward unchanged, $\rho(x,t) = \rho_0(x - ct)$. The continuum traffic model mentioned in Chapter 7 is of this type, with a flux $q = Q(\rho)$ set by an empirical fundamental diagram, so that the signal speed $Q'(\rho)$ depends on the density — which is what allows the profile to steepen into a jam.

The diffusion equation. If instead the flux runs *down the gradient* of ρ — Fick's law, $q = -D\,\partial_x \rho$, one of the closures listed in Chapter 6 — the balance law becomes

$$\partial_t \rho = D\,\partial_{xx} \rho,$$

the *diffusion equation*, with $[\,D\,] = L^2 T^{-1}$. It spreads any concentration out, smoothing sharp features over time. This is the equation behind the self-similar $\sqrt{Dt}$ spreading law of Chapter 2: with no imposed length in the problem, dimensional analysis alone forced the width of a diffusing front to grow like $\sqrt{Dt}$, and it is this equation that the front is obeying.

Recognising which of these a model reduces to — carried along, or spread out — is usually enough to anticipate its qualitative behaviour, in the same spirit as reading the dynamics off a critical manifold in Chapter 5.

Bibliography

[1] G. I. Barenblatt. *Scaling, Self-Similarity, and Intermediate Asymptotics*. New York: Cambridge University Press, 1996.

[2] Clive L. Dym. *Principles of Mathematical Modeling*. 2nd ed. Amsterdam and Boston: Elsevier Academic Press, 2004. ISBN: 9780122265518.

[3] A. C. Fowler. *Mathematical Models in the Applied Sciences*. Cambridge: Cambridge University Press, 1997.

[4] M. H. Holmes. *Introduction to the Foundations of Applied Mathematics*. Vol. 56. Texts in Applied Mathematics. New York: Springer, 2009.

[5] M. H. Holmes. *Introduction to Perturbation Methods*. 2nd ed. New York: Springer, 2013.

[6] S. D. Howison. *Practical Applied Mathematics: Modelling, Analysis, Approximation*. Cambridge: Cambridge University Press, 2005.

[7] C. C. Lin and L. A. Segel. *Mathematics Applied to Deterministic Problems in the Natural Sciences*. Classics in Applied Mathematics. Philadelphia: Society for Industrial and Applied Mathematics, 1988.

[8] J. D. Logan. *Applied Mathematics*. 4th ed. Hoboken, NJ: Wiley, 2013.

[9] S. H. Strogatz. *Nonlinear Dynamics and Chaos*. 2nd ed. Boulder, CO: Westview Press, 2015.