Part one

Introduction

1

Mathematical modeling

1.1 What is a model?

Mathematical modeling is a subject that is difficult to teach. It is what applied mathematics (or, to be precise, physical applied mathematics) is all about, and yet there are few texts that approach the subject in a serious way. Partly, this is because one learns it by practice: there are no set rules, and an understanding of the 'right' way to model can only be reached by familiarity with a wealth of examples. That is what this book aims to provide.

A model is a representation of a process. Usually, a *mathematical* model takes the form of a set of equations describing a number of variables, and we distinguish between continuous models, in which the variables vary continuously in space and time, and discrete models, whose variables vary discontinuously. Examples of discrete models are nonlinear recurrence equations for population size in nonoverlapping generations (for example, the well-known logistic equation $x_{n+1} = \lambda x_n(1 - x_n)$) or probability distributions for Markov processes. Another example would be ARMA (auto-regressive moving average) models for the prediction of stochastic time series.

In this book, we are exclusively concerned with continuous models, and in practice that means models formulated as differential equations, both ordinary and partial. Other types of continuous model give rise to integro-differential equations (e.g., age-dependent population growth, nucleation and kinetics of crystal growth) or delay-differential equations (e.g., in ring-cavity lasers and models of cell maturation and growth).

Applied mathematicians have a procedure, almost a philosophy, that they apply when building models. First, there is a phenomenon of interest that one wants to describe or, more importantly, explain. Observations of the phenomenon lead, sometimes after a great deal of effort, to a hypothetical mechanism that can explain the phenomenon. The purpose of a model is then to formulate a description of the mechanism in quantitative terms, and the analysis of the resulting model leads to results that can be tested against the observations. Ideally, the model also leads to predictions which, if verified, lend authenticity to the model. It is important to realize that all models are idealizations and are limited in their applicability. In fact, one usually *aims* to over-simplify; the idea is that if a model is basically right, then it can subsequently be made more complicated, but the analysis of it is facilitated by having treated a simpler version first.

In formulating continuous models, there are three main ways of prescribing governing equations. The classical procedure is to formulate exact conservation laws. The laws of mass, momentum, and energy in fluid mechanics are obvious examples of these. In certain situations, conservation laws involve empiricism. For example, momentum conservation in a turbulent fluid motion may be represented by the friction correlation $\tau_w = f\rho u^2$, where τ_w is the wall shear stress, ρ is density, u is velocity, and f is a friction factor that is determined from experiment. Such 'laws' may depend on the precise physical constitution of the fluid, and may not be uniquely determined. Lastly, there are what might be termed 'hypothetical' laws, based on qualitative reasoning in the absence of precise rules. For example, in the Lotka–Volterra model of interacting predator and prey populations, the death rate of the prey is supposed to be proportional to the product of each population. This is a phenomenological assumption that is plausibly akin to the law of mass action in chemical reactions but which nevertheless has no quantitative basis. In this case, the usefulness of the model is in explaining the mechanism whereby interacting populations can oscillate.

1.2 The procedure of modeling

Problem identification

Mathematical modeling begins with the identification of a problem. There is something we don't understand, a phenomenon that requires explanation, and we begin by trying to identify a plausible mechanism. Sometimes this is quite straightforward. For example, consider the exasperating effect on the motorist of traffic jams on motorways. You drive along until quite suddenly you hit a slow-moving wall of traffic. You then move at a snail's pace until all of a sudden, the road is clear and you can drive freely. You wonder why, if the road is clear, couldn't the drivers ahead get on with it?

How should we model the density of traffic flow? First, we need to define a model context, and we need to know what appropriate variables are. Traffic density (cars per unit length of road) is one such variable and traffic speed is another, and we will want these to be functions of time and distance. We idealize the real situation by supposing that traffic travels in a single lane and also that the variables may be represented as continuous (indeed, differentiable) functions of space and time: This is essentially the *continuum approximation*, familiar in the modeling of fluids and other continua.

To formulate a model, we require laws (often of conservation type) and constitutive relations between variables, which may be based on experiment or empirical reasoning. For example, if the density of cars in a (single) line of traffic is $\rho(x, t)$, where x is distance along the traffic lane and t is time, the conservation of cars requires (subscripts denote partial derivatives)

$$\rho_t + q_x = 0, \tag{1.1}$$

where q is the car flux, that is, $q = \rho v$, and v is the car velocity. A simple phenomenological assumption is that car speed is determined entirely by the density, that is, $v = v(\rho)$, and v decreases as ρ increases. For example, take $v = 1 - \rho$, where car density is measured by its ratio to that of the maximum (bumper to bumper) density.

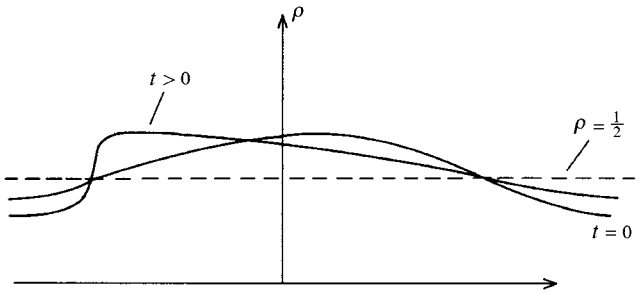

Fig. 1.1. Traffic jam formation: If a uniform density stream of traffic (with $\rho < \frac{1}{2}$) is perturbed so that $\rho_{\max} > \frac{1}{2}$, a jam forms, which is experienced by the oncoming driver as a shock (sudden deceleration) followed by a rarefaction wave (gradual acceleration)

Then $q = \rho(1-\rho)$, and the local wave speed is $q'(\rho) = 1 - 2\rho$. Thus $q' > 0$ for $\rho < 1/2$, $q' < 0$ for $\rho > 1/2$. An arbitrary initial disturbance to a uniform density state $\rho = \rho_0$ evolves as shown in Fig. 1.1 if the maximum value of ρ at $t = 0$, $\rho_{\max}$, is greater than $1/2$. Values of $\rho > 1/2$ move backward and form a shock that moves with speed

$$v_d = 1 - (\rho_+ + \rho_-), \tag{1.2}$$

where $\rho_\pm$ are the values just in front and just behind the shock. Because $v_- > v_d$, cars approach the shock and suffer a drop of speed as they pass through it. Further, because $v > q'$, they pass through the jam and eventually emerge in the undisturbed region again.

Through this simple model, we gain insight. We see that sudden changes in car speed can be associated with passage through a shock wave; the traffic flow is essentially that of a compressible medium. However, we also see that the model has important limitations. For example, one expects traffic jams to form because of, for example, lane closures on motorways, or where traffic lanes merge. Simple models to represent such phenomena can be based on Eq. (1.1) but require more realistic governing equations. Another limitation is that shocks (crashes) are usually avoided, and this may be represented by allowing traffic speed to depend on the spatial derivative of the density, which introduces a crucial diffusive term.

Thus we see the resolution (to some extent) of the problem in terms of a *mechanism*: the dependence of car speed on density and the resultant variation of the kinematic wave speed $q'(\rho)$. More generally, one often seeks a mechanism for the problem as specified. Often, this will be a verbal description only, which one then endeavors to translate into a mathematical formulation. For instance, consider the formation of underground cave systems, particularly in karst regions such as those near Postojna in Slovenia or near Doolin, County Clare, in the west of Ireland. Underground rivers are formed by the dissolution of the alkaline rocks by meteoric[1] groundwater that seeps through the porous fissured rock. There is an obvious mechanism here, because if the porosity (i.e., void space) is larger at some location, then fluid flow will be larger there, and hence also the rate of dissolution. This gives a positive feedback

[1] Meaning derived from the atmosphere, i.e., rainwater, which tends to be slightly acid.

mechanism whereby channels can form through increased amounts of dissolution at greater flow rates.

Model formulation

Once a problem is identified and a mechanism proposed, then one must formulate it mathematically. Often the difficulty lies in the choice of complexity: one wants the ease of a simpler model, but on the other hand one should include every relevant process. Different modelers will differ on what is important, and there is no unique 'right answer.' Formulation involves equations and boundary conditions, and if the problem is a sensible representation of the physics, it will usually (though not always) be well-posed. Mathematical analysts have as their program the establishment of well-posedness of a model, with the view that such results help design and validate suitable numerical solution procedures. In this book, we will not emphasize this approach, but rather will emphasize that of the applied analyst, whose business is to find the actual solutions.

Reduction

Solution of the proposed model now proceeds differently according to the modeler's background. Often (and this is largely the case for engineers and applied scientists) a model is a numerical model, and a solution means a numerical solution. There are two levels of difficulty with this. At the primitive level, direct numerical computations can founder because of ill-posedness or stiffness of the equations. More seriously, computation can limit insight, because of an inability to pose questions properly.

The first difficulty is aided by some pretreatment of the governing equations. It is, for example, often unhelpful to solve problems unless they have been nondimensionalized. When this is done properly (and this forms the focal point of this book), then the presence of small or large dimensionless parameters can be an indicator of singular perturbations, and thus stiffness. In many cases (most, in fact), this numeral inconvenience is an aid to analysis, facilitating the use of perturbation methods that can be used to gain insight into the solutions.

The second difficulty resides in the ability to use big computers to solve problems directly. Here is an example. Computation of convective flow in the Earth's mantle is a problem that realistically depends on a number of different dimensionless parameters: a Rayleigh number Ra, a viscosity number ε, an activation volume number μ, a dissipation number D, an internal heating number H, and so on. One aim of solving the equations representing this flow is to determine the dependence of the dimensionless heat flux, given by the Nusselt number Nu, on these various parameters: $Nu = Nu[Ra, \varepsilon, \mu, \ldots]$. The simplest situation, that of constant viscosity and no internal heating, is well understood both analytically and numerically. However, even with two independent parameters (Ra, ε), results are confusing and inconclusive. The problem is that the kinds of value of relevance to the Earth ($Ra = 10^7$, $\varepsilon = 1/40$, for example) have been, in the past, inaccessible to computer simulation–the asymptotic limits are too severe. One therefore has to extrapolate results at smaller Ra and/or

larger ε to more extreme values: but, really, to do this sensibly, one needs a theoretical understanding of the correct limiting behavior, and although this has been done for Ra and ε, it has not for combinations of more parameters.

What happens is that the original question, what is $Nu[Ra, \varepsilon, \ldots]$ when $Ra = 10^7$, $\varepsilon = 1/40$, etc., is replaced by the question, what is $Nu[Ra, \varepsilon, \ldots]$ when, say, $Ra = 10^6$, $\varepsilon = 1/10$; the answer is then extrapolated to the more extreme parameter values. In this situation, the correct extrapolative procedure is to analyze the problem at extreme values and use the numerical results as a test for the predictions: each should complement the other. Successful mathematical modeling needs to combine different approaches, rather than elevate any one in a misplaced ascendancy.

The first thing we wish to do with a continuous model is to nondimensionalize it. It is then possible to identify in a rational way whether different terms are large or small. If the latter, they can in some (but not all) circumstances be ignored. One is thus led to a reduced model, which is a simplification of the original problem but not significantly less accurate. One must keep in mind what the question is. If one seeks not so much a quantitative simulation as a theoretical insight, then it may be judicious to simplify further. For example, detailed simulation of turbulent fluid flow probably requires use of averaged models such as k-ε models, but in some circumstances, Bernoulli's law may be sufficient and even Laplace's equation for the velocity potential![2]

Analysis

Reduction is the process whereby a model is simplified, most often by the neglect of various small terms. In some cases, their neglect leads to what are called *singular perturbations* (for which, see Chapter 4), whose effect can be understood by using the method of *matched asymptotic expansions*. In particular, it is often possible to break down a complicated model into simpler constituent processes, which, for example, operate on different space and time scales. Analytic dissection in this way leads to an overall understanding that is not so simply available through straightforward numerical computations, and indeed it also provides a methodology for simplifying such computations.

In analyzing a model, one is often led through a sequence of similar types of calculation: the existence and nature of steady solutions; their stability and instability, and consequent bifurcations to oscillations and traveling waves; hysteresis and the associated phenomenon of blow-up; secondary instabilities, and the occurrence of chaotic behavior. By studying a variety of models, as we do here, one thus 'learns' modeling by seeing the same sequence of processes carried out on a wide variety of different systems.

Computation

At some judicious point, numerical results need to be obtained. The use of these may be complementary: to obtain quantitative results in parametric regions where analysis is impossible. Or they may be validatory: they provide an independent confirmation

[2] For those unfamiliar with these fluid mechanical terms, further discussion is given in Chapter 6.

of analytic results. Sometimes, scientists think of analysis as confirming numerical results; in reality, one often needs to *design* numerical experiments to complement analytic results: straightforward but unthinking approaches often lead to apparent contradictions where a more carefully designed computation would remove these.

In problems where analytic progress is possible, the eventual problem to be solved numerically is often relatively simple, and we do not dwell on this aspect of modeling in this book; it should, however, be emphasized that it is an important component. Sometimes, indeed, it is *simpler* to solve the original problem numerically *rather* than the simplified model!

Model validation

Ideally, a mathematical model ends by returning to its origin. We look to see whether the model and its analysis explains the phenomenon we are interested in. Does the predicted curve fit the experimental data? Does the predicted stability curve agree with the experimentally determined values? The whole art of mathematical modeling lies in its self-consistency. It is an inexact science that derives its justification from the fact that apparently arbitrary assumptions are seen to work. And ultimately, this is the justification for a model: it helps us to understand an experimental observation. There is no unique or 'correct' model; but there are good models and bad models. The skill of modeling lies in being able to judge which is which.

1.3 Choosing the model

Consider, for example, the problem of modeling the climate. The weather is determined by heat and mass transfer in the atmosphere, which is (more or less) a blanket of air some ten kilometres thick that shrouds the planet (it extends above this *troposphere*, but the processes above the *tropopause* at ten kilometres mainly concern radiative heat exchange and absorption). The basic process of the weather is convection of the atmosphere driven by the differential heat input due to solar radiation between the equator and the poles. This heat imbalance causes a poleward heat flux by various convective processes both in the atmosphere and in the oceans. The vigorous rotation of the Earth causes this slow poleward circulation of the atmosphere to be distorted to a primarily *azimuthal* flow (the zonal wind), which is itself *baroclinically* unstable, and leads to the characteristic large-scale feature of the circulation, a wave-like undulation in the latitude of isobars.

Predicting the weather

Meteorologists predict the weather by writing fluid dynamic equations describing the convective motion of the atmosphere and seeking to solve them numerically. Predictions can be made for periods on the order of days, but it is generally thought that forecasting is impossible beyond about a week because of the chaotic nature of the atmospheric flow. One might infer from this that describing atmospheric dynamics on longer timescales is impossible, but this is not so – it is a question of choosing an appropriate model.

Ice ages

For example, the regular occurrence of ice ages at intervals of about a hundred thousand years can be predicted in some models of climate dynamics. Typically, these models describe the heat and mass transport in atmosphere and oceans by using semiempirical constitutive laws (for the average transport rates), with variable coefficients that depend on the size of the ice sheet cover. The ice sheets wax and wane over thousands of years, so that short term variability on a timescale of days, weeks, years, or even decades is unimportant and can be *averaged*. Thus the short-term unpredictability of the weather is irrelevant to the problem of modeling the longer term evolution of the climate, *provided* the more rapid fluctuations of the various atmospheric transports can be suitably parameterized. The type of model one obtains is very different and can be dramatically simpler. The simplest (energy balance) models are zeroth-order (i.e., algebraic)!

The smaller scale

But equally, we can go to the opposite extreme. One often observes, from aircraft, beautiful roll-like patterns in cloud formations. These are due to local convective processes, and they, too, can be modeled by fluid dynamic equations: However, the space and time scales are much smaller than those involved in weather prediction, and the fundamentally important larger scale effects of rotation are irrelevant on such smaller scales. Therefore, although the relevant model to describe these convective flow patterns is based on fluid equations, it will be rather different to a weather prediction model: For example, phase change effects are important (owing to the presence of water droplets in clouds) but rotation is not.

In summary, modeling is a subjective pursuit, and the nature of the problems dictates the type of model that is relevant. The same process may require different models, depending on the question that is of interest.

1.4 Some examples

In this section, we discuss further some examples of the mathematical models that were mentioned above.

Age-dependent population growth Suppose a population has a size distribution $\phi(a, t)$, where a is age and t is time: $\phi\delta a$ is the number between ages a and $a + \delta a$. The birth rate $b(a)$ depends on age, as does the mortality rate $m(a)$. These may also depend on time through social effects, for example, the baby boom of the sixties. A cohort of individuals age at a constant rate, $\dot{a} = 1$, while their numbers decline at a rate $m(a)$, whence $\dot{\phi} = -m\phi$. These are the characteristic equations for the partial differential equation

$$\phi_t + \phi_a = -m\phi, \tag{1.3}$$

known as the Von Foerster equation, and the birth rate appears in the boundary

condition

$$\phi(0,t) = \int_0^\infty b(a)\phi(a,t)\,da. \tag{1.4}$$

Further discussion of age-structured population models can be found in the books by Murray (1989) and Hoppensteadt (1975).

Nucleation and kinetics of crystal growth The classic theory of phase change kinetics is given in two papers by Avrami (1939, 1940). Suppose crystals are nucleated and grow from a melt, such that at time t, they occupy a volume $V(t)$ per unit volume. We also define $V'(t)$ to be the volume fraction the crystals *would* have had if different crystals did not meet. We define Y to be the rate of growth of crystal interfaces and I to be the rate of nucleation of new crystals. Both Y and I depend on temperature. In a time interval $(\tau, \tau+\delta\tau)$, $\delta N' = I(\tau)\,\delta\tau$ new crystals are created, and each of these attains a volume $v = a\{\int_\tau^t Y(\theta)\,d\theta\}^3$ at time t, where a is a shape factor ($a = 4\pi/3$ for spheres, $a = 8$ for cubes, for example). It follows that

$$V' = a\int_0^t I(\tau)\left\{\int_\tau^t Y(\theta)\,d\theta\right\}^3 d\tau \tag{1.5}$$

is the *fictive* volume, and

$$\frac{\partial V'}{\partial t} = 3a\int_0^t I(\tau)Y(t)\left\{\int_\tau^t Y(\theta)\,d\theta\right\}^2 d\tau. \tag{1.6}$$

Now we suppose that at time t, the fraction of the fictive crystal surface that lies inside the actual crystals should be V, and thus of the fictive growth $\delta V'$ in the interval δt, only the fraction $1-V$ contributes to actual growth. Thus $\delta V = (1-V)\,\delta V'$, so that

$$\frac{\partial V}{\partial t} = 3a(1-V)\int_0^t I(\tau)Y(t)\left\{\int_\tau^t Y(\theta)\,d\theta\right\}^2 d\tau. \tag{1.7}$$

Recently, this theory (used in the study of solid-solid phase transitions in metallurgy) has been applied to the solidification of magma chambers (Brandeis, Jaupart, and Allegre, 1984).

Delay differential equations These occur in a number of different applications, particularly medical, where the delay may be due to finite maturation time (for example in white blood cell populations (Mackey and Glass, 1977) or in the humoral immune response (Dibrov, Livshits, and Volkenstein, 1977a,b), or to a finite transport time (hence the delay in respiratory control models due to transport of blood gases in the arteries). A common form of such equations is the *delay-recruitment* equation

$$\varepsilon\dot{x} = -x + f(x_1), \tag{1.8}$$

where $x_1 = x(t-1)$, which can be derived in respiratory control models (Fowler, Kalamangalam, and Kember, 1993), blood cell populations, ring cavity lasers (Ikeda and Matsumoto 1987) and population biology (May, 1980; Gurney, Blythe, and Nisbet, 1980). In the last case, the decay term $-x$ is the mortality rate, whereas the delay term is the regeneration rate (or recruitment rate), taken as a nonlinear function of the population size at an earlier time (the delay here is the gestation time).

Lotka–Volterra equations The simplest model of interacting populations was proposed by Volterra (1926). The same model was used by Lotka (1920) to illustrate the phenomenon of undamped oscillations in a model chemical reaction. If x and y are the predator and prey populations, then these equations are

$$\begin{aligned} \dot{x} &= \alpha xy - \beta x, \\ \dot{y} &= \gamma y - \delta xy, \end{aligned} \tag{1.9}$$

representing constant specific birth and mortality rates for prey and predator, respectively. The predators' specific growth rate αy depends on availability of the prey as food source, whereas the prey death rate δx depends on the number of predators. These equations have oscillatory solutions but have the unsatisfactory feature of forming a conservative system, and oscillations of any magnitude are possible. More realistic versions of the model can remove this degeneracy (see Murray 1989).

Traffic flow modeling Modeling traffic flow was used by Whitham and Lighthill as an example of their 'kinematic wave theory.' There is a good discussion in Whitham's (1974) book. In general, shocks will form, and as with shock waves in gas dynamics, this suggests hunting for a physically plausible mechanism that can prevent shock collision. One such mechanism is the realization that if there is a change in density, this also affects drivers' reactions. We might expect that v is lower if the density increases ahead of the driver, thus $\partial v/\partial \rho_x < 0$, and a simple model is then

$$v = 1 - \rho - \delta \rho_x; \tag{1.10}$$

this leads to the nonlinear diffusion equation

$$\rho_t + (1 - 2\rho)\rho_x = \delta(\rho \rho_x)_x, \tag{1.11}$$

which allows a diffusive shock structure of width $O(\delta^{1/2})$.

Formation of cave systems A recent model analyzing the onset of cave systems in limestone regions is that of Groves and Howard (1994). As with many applied problems, one finds similar phenomena in a wide variety of different contexts. For example, meltwater at the surface of a valley glacier finds its way through crevasses to the bed, where it drains along the valley floor through a network of channels. The development of a channeled flow occurs in a similar way to that in limestone, except that melting plays the part of erosion. The basic theory is given by Röthlisberger (1972). Similar erosive/melting instabilities are the cause of channel formation in dendritically solidifying alloys (Copley *et al.*, 1970) and lie at the heart of the erosional formation of river drainage networks (Willgoose, Bras, and Rodriguez-Iturbe, 1991a,b; Smith and Bretherton, 1972; Kramer and Marder, 1992). The web that a particular mathematical model can weave through the different sciences is part of the fun of applying mathematics.