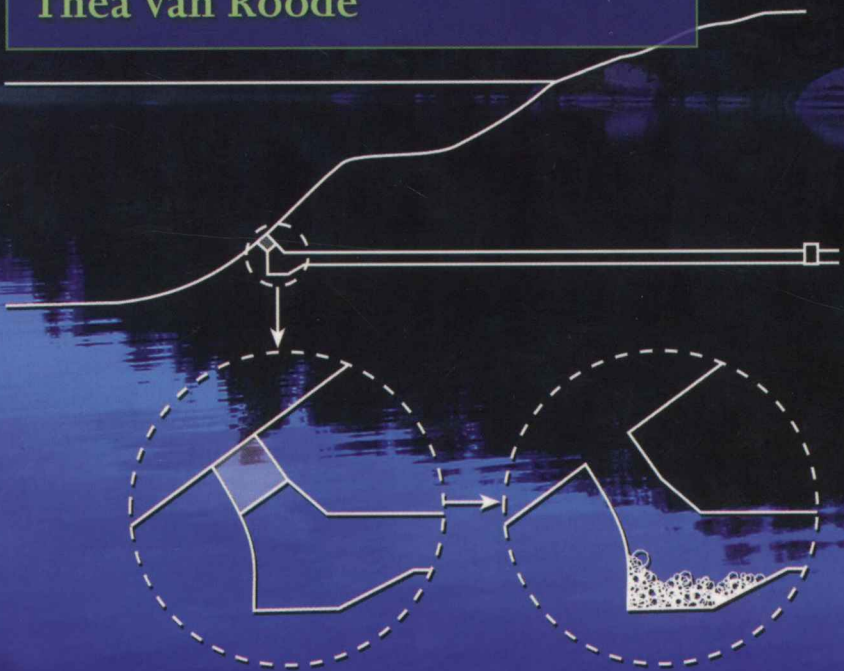
STUDENT MATHEMATICAL LIBRARY
Volume 27

Mathematical Modelling

A Case Studies Approach

Reinhard Illner
C. Sean Bohun
Samantha McCollum
Thea van Roode



AMS
American Mathematical Society

Mathematical Modelling

A Case Studies Approach

This page intentionally left blank

STUDENT MATHEMATICAL LIBRARY
Volume 27

Mathematical Modelling

A Case Studies Approach

Reinhard Illner
C. Sean Bohun
Samantha McCollum
Thea van Roode

AMERICAN MATHEMATICAL SOCIETY

2000 *Mathematics Subject Classification.* Primary 00–01, 00A69; Secondary 00A71, 93A30.

For additional information and updates on this book, visit
www.ams.org/bookpages/stml-27

Library of Congress Cataloging-in-Publication Data

Mathematical modelling : a case studies approach / Reinhard Illner... [et al.].
p. cm. — (Student mathematical library, ISSN 1520-9121 ; v. 27)
Includes bibliographical references.
ISBN 0-8218-3650-1 (alk. paper)
1. Mathematical models. I. Illner, Reinhard. II. Series.

QA401.C352 2005
511′.8—dc22 2004046225

---

10 9 8 7 6 5 4 3 2 10

Dedicated to those who brought us these projects.
Their imagination was the seed for this book.

This page intentionally left blank

Contents

Preface

Mathematical modelling is a subject without boundaries in every conceivable sense. Wherever mathematics is applied to another science or sector of life, the modelling process enters in a conscious or subconscious way. Significantly, this is the process in which much of mathematics was originally started, even in the geometric problems encountered by the Greeks so many centuries ago.

The axiomatic method in mathematics, while putting much of mathematics on solid ground, can push this modelling history of mathematics to some degree out of view; rather than focus on questions of measurement or prediction, one worries about existence and uniqueness, about basic structures, or about more generality. In a significant fraction of the mathematical community, those who actually compute real life phenomena are labelled as "engineers" or "physicists", though they may need just as much mathematical skill as those who spend their time with rigourous proofs. We contend that mathematical modelling deserves attention by all scientists as it serves science and engineering from many points of view:

- It occupies a middle ground between mathematics and most other science and engineering disciplines.

- It is the primary testing and development ground for the power of mathematical language as applied to real life problems.
- It provides avenues of high sophistication and motivation to abstract mathematics; this is of particular value for students who have mathematical skills but lack motivation (yes, that breed exists).
- It serves society.

This list is easily amended.

Mathematical modelling problems are implicitly found in most textbooks in physics, engineering, chemistry, computer science, biology, and even in such subjects as psychology or sociology. In recent years, more emphasis has been placed on a complete description of the modelling cycle, and accordingly there are a number of textbooks for courses at all levels. One could easily teach a modelling course even at the first year university level, and certain modelling problems are clearly part of the standard calculus sequence (constrained optimization problems, solving linear differential equations, computing moments of inertia, etc.). More serious attempts at systematic modelling are left until the third year, when students have acquired sufficient skills in basic mathematical disciplines like calculus, linear algebra, discrete mathematics, and probability and statistics.

The problem then is not when to start modelling, but what to choose. The amount of material is unlimited, and one could easily fill an entire term with applications of systems of ordinary differential equations, or Markov processes, or linear programming, or ... Wait! We just listed subjects that contain mathematical modelling, but which offer enough material to be treated in their own right. One can, of course, extract modelling problems from these subjects, spell out the modelling cycle, and present the material from a slightly different point of view.

This is the philosophy of most textbooks on mathematical modelling, and this text is really no exception. Why, you may ask, did we write another book while there are already numerous texts from which to choose? Three key words should answer this question: *size*,

level, and *fun*. First, we wanted to produce a text limited to what can comfortably be covered in one term. In fact, the text emerged from lectures which one author (R.I.) has given at the University of Victoria during the years 1997–1999. Certain material was chosen from original papers as well as from texts at both the undergraduate and graduate level. These texts were invariably found to contain more material than could be covered within one term. In addition, some of the graduate material was too demanding for a third year course. Eventually, we focussed on nine topics deemed appropriate for a third year audience, and easily complemented by student presentations on various other modelling problems. Selection of the topics for the book was based on the variety of mathematics the students would be exposed to, and (more importantly) the motivational value each topic would have.

This text is written for students and partly by students. Thea van Roode and Samantha McCollum took the course in the spring of 1999 and spent the summer compiling the material into a LaTeX file. This file was recently reworked by Sean Bohun and Reinhard Illner to produce what you now hold in your hand.

The following is a brief synopsis of the subjects covered in the text, noting the origins of the material.

Chapter 1: Models of crystallization are well understood in physical chemistry, and the classical Kolmogorov–Avrami model is found in many textbooks. We derive it here using probabilistic arguments and introducing the concept of a spatial Poisson process. A recent experiment indicates that the model is ill suited to describe incomplete crystallization processes, and an alternative for such cases based on recent original research is presented. We would like to thank Terry Gough for permission to use this problem in the text.

Chapter 2: This is probably the most original of all of the modelling problems in the book. Ten years ago a retired engineer from B.C. Hydro (known to the authors) knocked at Reinhard Illner's door and presented this strange problem that he had solved approximately in 1961 with only his slide rule and a healthy analytic mind. The problem of filling a closed tunnel

with water leads, as will be seen, to a formally very singular initial value problem (not really, of course!) that can be analysed as an excellent example to get a feeling for the concept of well-posedness, numerical solvability, and stability. Due to the intuitive mathematical skill of the B.C. Hydro engineer, a catastrophe was averted; in Chapter 2, we revisit his problem.

Chapter 3: If you like money, this chapter is for you. Starting from scratch, we develop the basic formulas for interest with yearly, monthly, daily, ..., continuous compounding. The formulas for mortgages, loans, and variable interest rates are covered and then combined with a little bit of risk theory to compute the present values of annuities. While it's not a course on financial mathematics, you should be able to argue with your bank after reading this chapter. We are grateful to Bill Reed, as much of this material was introduced by him into an earlier version of the course.

Chapter 4: This topic is more standard fare in mathematical modelling courses. Everybody knows a bit about dimensional analysis, but is it clear to you that dimensional analysis is really linear algebra? And what do nuclear explosions have to do with linear algebra? Well, this chapter starts with an exciting story on the first nuclear explosion on Earth—the Los Alamos bomb. We saw this first in the beautiful (but more advanced) text by Fred Wan [**U**], quoted in the chapter.

Chapter 5: Predator-prey systems are very much standard examples in courses on ordinary differential equations, and they are a must in modelling texts. The famous story of the Adriatic Fishery after the First World War makes for a fascinating introduction to the Lotka–Volterra model. From there, one stumbles naturally onto the concept of phase plane, stationary points, periodic orbits, linearization, and the classification of equilibria.

Chapter 6: Optimizing a fishery: When should a fleet of constant size be allowed to resume fishing a stock that is depleted at the present time? This is a disturbingly familiar problem in places like British Columbia, where salmon stocks that seemed inexhaustible thirty years ago are at risk.

Chapter 7: Functional equations are an ancient mathematical subject with applications in strange places. The problem of formal justice, addressed in this chapter, was originally brought to us by David Gartrell from the Department of Sociology at the University of Victoria, and the equations were supplied by a thick volume on formal justice written by the sociologist Sołtan [**O**]. However, no mathematical analysis of these equations was done in that volume; the analysis was done by one of the authors in [**H**] and is reproduced here. The tools which emerge here as useful are dimensional analysis (again) and separation of variables.

Chapter 8: Traffic flow problems produce a wealth of good mathematics and offer such good motivation. Who hasn't been stuck in a traffic jam? This topic is naturally divided into two parts: Chapter 8 keeps track of individual cars, leading to systems of differential-delay equations. Essentially, we follow the treatment in [**K**]. The key part of this treatment is the transition from a microscopic to macroscopic description presented in Chapter 9, which can be used in equilibrium situations.

Chapter 9: This is a continuation of the theory developed in Chapter 8. A macroscopic model is developed by appealing to the microscopic theory of the previous chapter. Students learn the significance of continuity assumptions, and the emergence of a scalar conservation law is the reward. An introductory discussion on the nature and properties of such conservation laws, including shock and rarefaction waves, concludes the chapter.

We find time to do additional material in class, and such material is done partly by student presentations and partly in the form of a computer lab. In the computer lab, students are introduced to MAPLE and its power to compute integrals, solve differential equations, produce least-square fits, or draw phase diagrams. One of the authors (C.S.B.) has successfully used MATLAB equipped with the symbolic toolbox for the same purpose. All of these tasks are related to the modelling problems listed above. In the student presentations, we extend the basic theory covered in class. Typical problems discussed are the use of dimensional analysis for finding a reasonable

relationship between the weight and the cooking time of a turkey, the search for structurally stable modifications of the Lotka–Volterra system, the use of systems of ordinary differential equations for the design of rock-climbing cam devices, the application of graph theory to city planning, and others.

We owe big thanks to Rod Edwards and Denton Hewgill, who used preliminary versions, caught many mistakes and weaknesses, and made many constructive suggestions. We also thank Anton Arnold, Mike Doncheski, Deborah Mirdamadi, and Holger Teismann for their input.

The question of purpose should never arise in a mathematical modelling course, and indeed, we haven't encountered it.

Victoria, June 2004

Chapter 1

Crystallization Dynamics

Concepts and Tools: Elementary probability theory, calculus, Poisson process, Voronoi diagrams

Consider a substance that crystallizes about crystallization nuclei. The microcrystals accumulating about the nuclei will grow until they impinge upon one another, eventually resulting in the crystallization of the entire volume of the substance. Figure 1 illustrates such a crystallization process. The relevant question and focus of this section is to determine the fraction of the substance that has crystallized by some time t for crystallization phenomena of this type.

In 1997, chemists at the University of Victoria conducted the following experiment: A mixture of carbon dioxide, CO_2, and acetylene, C_2H_2, was sprayed at 90 degrees Kelvin on a metal sheet. A spectroscopic analysis confirmed the formation of a hitherto unknown crystalline binary phase $CO_2 \cdot C_2H_2$. This binary phase was found to be metastable and, over a period of approximately five hours, gave way to the formation of CO_2 crystals embedded in a matrix of amorphous C_2H_2. Figure 2 displays this decomposition. Spectroscopy can be used to measure the amount of the formed CO_2 with good reliability but no model was available that could accurately predict these results. We will develop and discuss mathematical models that predict the amount of crystallized CO_2 at any given time.

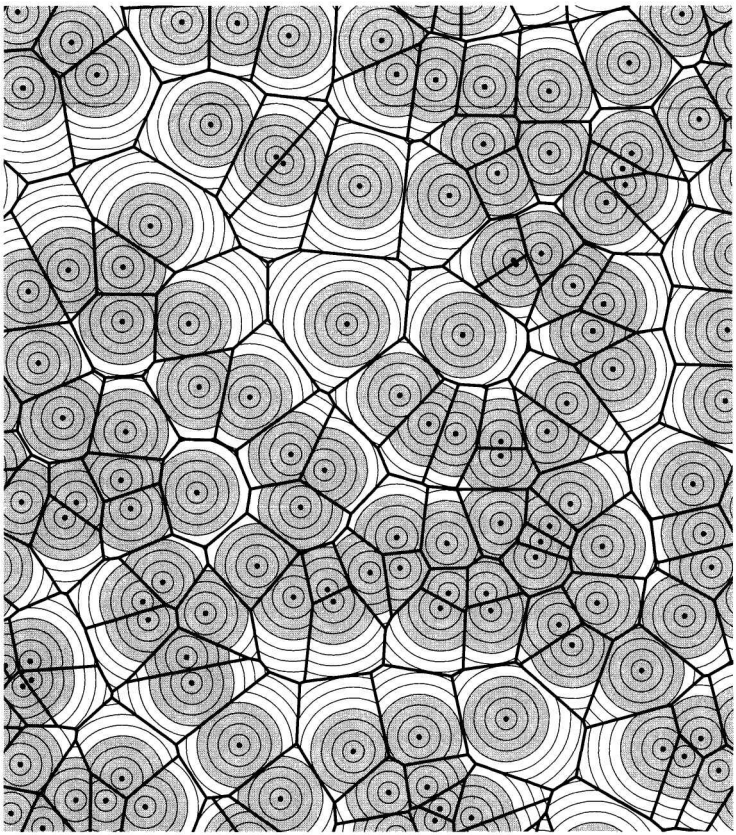


Figure 1. A Voronoi diagram illustrating the crystallization process: crystallization nuclei grow until they impinge upon one another.

The so-called Kolmogorov–Avrami model (from here on referred to as the K-A model) is a well-known classical tool for the prediction of growth curves such as these; however, in this case it did not yield accurate results and a new model needed to be developed. Before we can derive this new model, we need to introduce the K-A model and analyse why it fails in this particular instance.

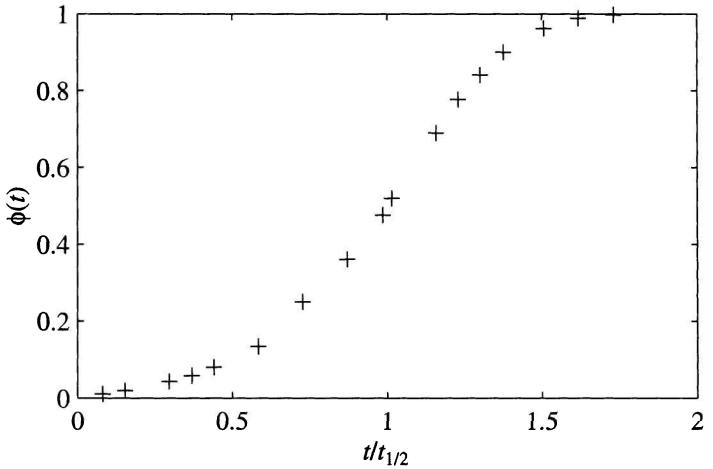


Figure 2. $\varphi(t)$ denotes the fractional decomposition of $CO_2 \cdot C_2H_{2(s)}$ to $CO_{2(s)}$ as a function of time t. The time axis is scaled so that one time unit corresponds to one half of the CO_2 being crystallized. In reality, the complete crystallization process took approximately five hours. The data for the curve was extracted from absorption curves similar to those described by Figure 3.

1.1. Derivation of the K-A Model

We consider a large (macroscopic) volume V in which N impurities (which act as nucleation sites) are equidistributed independently of each other. Let Q denote a particular nucleation site and P be a fixed but arbitrary point within V. To remove any boundary effects we assume that the distance from P to Q, denoted as $|PQ|$, is sufficiently small so as to ensure that the sphere centred at P with radius $|PQ|$ lies entirely within V. Figure 4(a) illustrates this situation. For $a > 0$

$$\text{Prob}\,(|PQ| \leq a) \approx \frac{4\pi a^3}{3V},$$

which implies that

$$\text{Prob}\,(|PQ| > a) \approx 1 - \frac{4\pi a^3}{3V}.$$

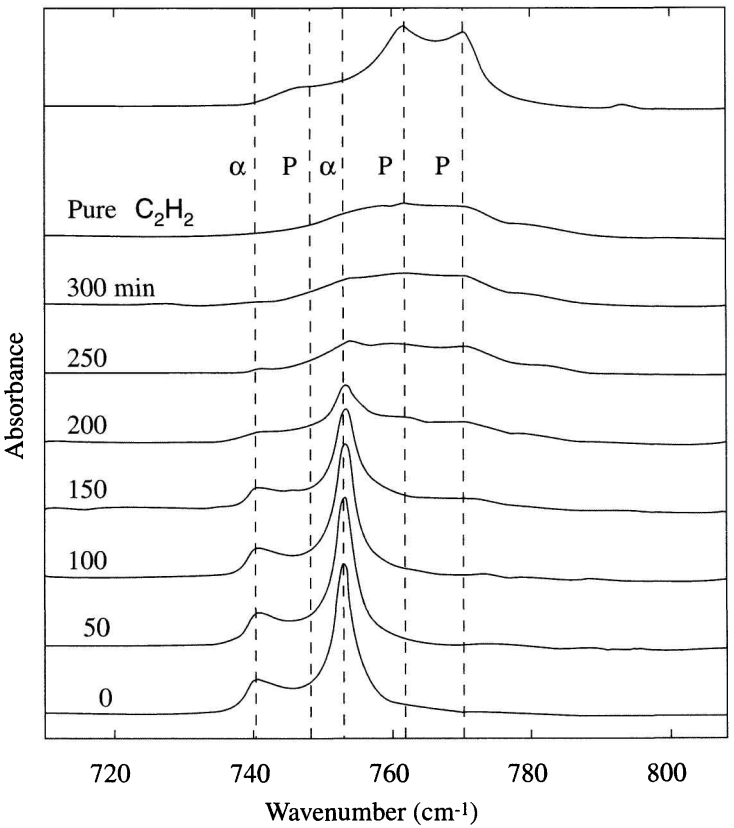


Figure 3. Time dependence of the infrared spectra recorded during the decomposition of $CO_2 \cdot C_2H_2$ at a temperature of 90 K. Shown are the bending vibrations of C_2H_2. Transitions of the reactant $CO_2 \cdot C_2H_2$ are labelled α while transitions of the products are labelled P.

Now assume that N independent nuclei are equidistributed in V as shown in Figure 4(b). By independence, we have

$$\text{Prob}\left(\min_{i\in\{1,\ldots,N\}} |PQ_i| > a\right) = \left(1 - \frac{4\pi a^3}{3V}\right)^N.$$

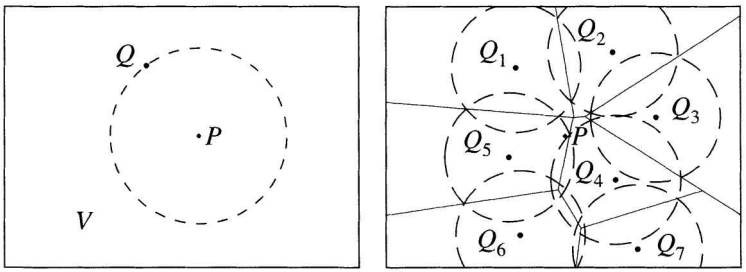


Figure 4. (a) An impurity Q within a volume V is depicted. P is an arbitrary but fixed point within the domain. (b) Likewise, this figure shows a similar situation with N impurities denoted by Q_i with $i \in \{1, \ldots, N\}$. P remains fixed but arbitrary. Also shown are the boundaries of the Voronoi cells.

If we denote by X the distance from P to the nearest Q_i, this becomes

$$\text{Prob}\,(X > a) = \left(1 - \frac{4\pi a^3}{3V}\right)^N. \tag{1.1}$$

If we denote by $F_N(x)$ the cumulative distribution function of the random variable X associated with x, equation (1.1) implies that

$$F_N(x) = \text{Prob}\,(X \leq x) = 1 - \left(1 - \frac{4\pi x^3}{3V}\right)^N. \tag{1.2}$$

Imagine now that we have many microscopic nuclei in the comparatively large volume V. To this end, we investigate the limit of (1.2) as both N and V become arbitrarily large, but such that there is a fixed number of λ nuclei per unit volume in V. Consequently, let $N = \lambda V$ so that equation (1.2) gives

$$F_N(x) = 1 - \left(1 - \frac{4\pi\lambda x^3}{3N}\right)^N,$$

and taking the limit as $N \to \infty$ with fixed λ yields

$$F_N(x) \longrightarrow F(x) = 1 - e^{-4\pi\lambda x^3/3}.$$

Assuming further that the radii of the crystallizing globules expand at speed v, it is now easy to see that the fraction of the material

crystallized by time t, denoted by $\varphi(t)$, is given by

$$\varphi(t) = \text{Prob}\,(X < vt) = F(vt) = 1 - e^{-4\pi\lambda v^3 t^3/3}. \tag{1.3}$$

The preceding derivation assumed that this crystallization process is three dimensional, but a variant of the argument also holds in two dimensions. Indeed chemists typically generalize the model to

$$\varphi(t) = 1 - e^{-kt^n}, \tag{1.4}$$

where k and n are fitted constants. Equation (1.4) with $n = 2, 3$ is known as the K-A model. Allowing n to take on any positive real value gives the generalized K-A model. See Exercise 3.

The process we followed in the above derivation gives rise to a so-called Poisson distribution with intensity λ. We explain this in detail:

1.2. Emergence of the Poisson Distribution from the Binomial Distribution

Suppose that N nuclei are equidistributed, independently of each other, in a volume V. Let $\Omega \subset V$ and let $p = |\Omega|/|V|$. $p \in [0, 1]$ is the probability that a randomly chosen nucleus is in Ω. The probability of finding k nuclei in Ω is then

$$P_k^N = \binom{N}{k} p^k (1-p)^{N-k}.$$

This is the binomial distribution. It is exact, but very cumbersome to work with when N is large. It is well approximated by a *Poisson distribution with intensity* I if $N \to \infty$ and $|V| \to \infty$ such that $I = N/|V|$ is kept fixed. We think of Ω as fixed in the process. Then, writing $|V| = N/I$,

$$\begin{aligned} P_k^N &= \binom{N}{k} \left(\frac{\Omega I}{N}\right)^k \left(1 - \frac{\Omega I}{N}\right)^{N-k} \\ &= \frac{(\Omega I)^k}{k!} \left(1 - \frac{\Omega I}{N}\right)^N \left(1 - \frac{\Omega I}{N}\right)^{-k} \frac{(N-k+1)}{N} \cdots \frac{N}{N}. \end{aligned}$$

In the limit where $N \to \infty$ (but k is fixed) we find

$$P_k^N \to P_k = \frac{(\Omega I)^k}{k!} e^{-\Omega I}$$

because

$$\lim_{N\to\infty}\left(1-\frac{\Omega I}{N}\right)^{-k}=1$$

and

$$\lim_{N\to\infty}\frac{(N-k+1)\cdots N}{N^k}=1.$$

The probability distribution

$$P_k=\frac{(\Omega I)^k}{k!}e^{-\Omega I},\qquad k=0,1,2,\ldots,$$

derived here is known as the Poisson distribution, used in many applications. If X denotes the number of nucleation sites in Ω, then $P(X=k)=P_k$, and an easy exercise shows that

$$E(X)=\Omega I.$$

The limit in which the Poisson distribution arises is the same as the limit for which we derived (1.3), the K-A model.

1.3. Testing the K-A Model

Now that we have derived the K-A model, let us determine whether it is an effective theoretical predictor for the situation at hand. More specifically: Can the generalized K-A model explain the data for the $CO_2 \cdot C_2H_2$ conversion? The empirical data and the theoretical predictions from the K-A model are best compared if both are transformed and plotted on a ln–ln graph. We convert $\varphi(t)=1-e^{-kt^n}$ from the K-A model as follows. Isolating the exponential gives

$$e^{-kt^n}=1-\varphi(t),$$

and taking the logarithm of both sides yields $kt^n=-\ln(1-\varphi)$. We take the logarithm once again to obtain

$$\ln k+n\ln t=\ln\left[-\ln(1-\varphi)\right]. \tag{1.5}$$

Relabelling the right-hand side of equation (1.5) as $f(t)$ gives

$$\ln k+n\ln t=f(t)$$

which, since $\ln k$ and n are constants, is linear in $\ln t$. If the K-A model were applicable to our crystallization process, performing a similar conversion on the experimental data by replacing $\varphi(t)$ in equation (1.5) with the empirical values would result in a straight line.

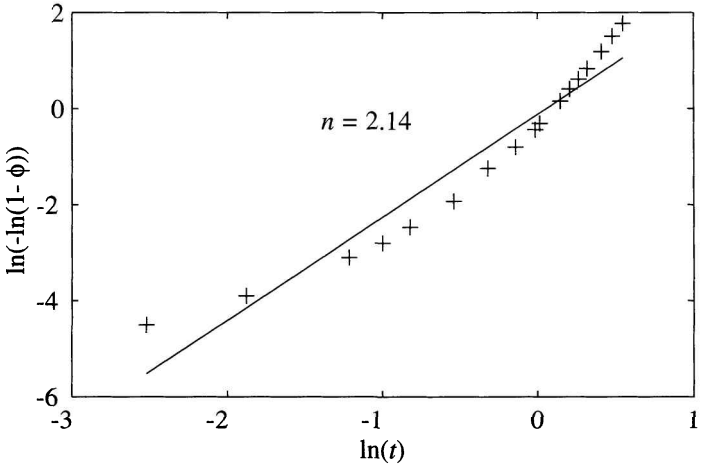
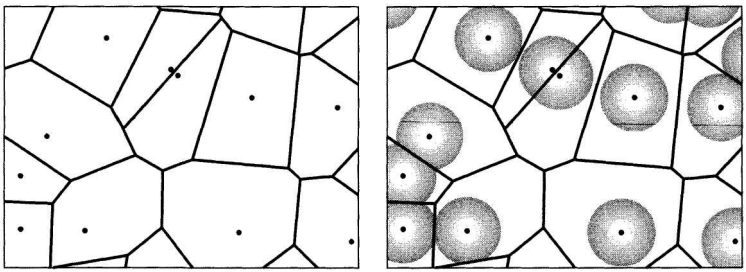


Figure 5. This graph plots the transformed empirical data from the $CO_2 \cdot C_2H_2$ conversion. Because the plot of the data is not linear, it is apparent that the K-A model is not an accurate theoretical predictor for this process.

Figure 5 shows the representation of this experimental data. The logarithmic conversion of the data points clearly does not lie on a straight line. In fact, the effective value of n increases from a value of approximately 1 to 4 as the reaction progresses. An increasing value of n has only rarely been reported [V], and even in this case n decreased at the end of the reaction. These observations suggest that the K-A model is inadequate for the theoretical prediction of the conversion of $CO_2 \cdot C_2H_2$ into CO_2 (crystalline) and C_2H_2 (amorphous waste). While the K-A model is known to work quite well for most crystallization phenomena, it was conjectured that it failed to give accurate predictions in this instance because of the presence of an amorphous waste product in the crystallization process. A new model incorporating the presence of C_2H_2 needed to be developed.

1.4. A New Model

Our new model will be based on empirical approximations to the distribution densities of the volumes of three-dimensional Voronoi cells.



(a) (b)

Figure 6. (a) Given several nuclei, a Voronoi diagram is one that partitions the plane into sections in such a way that a partitioning line will be equidistant between any two points (nuclei) that it separates. (b) Voronoi diagram depicting the growth of crystallization nuclei.

These cells are defined by Poisson generated crystallization nuclei which grow until the cell under consideration is filled by CO_2 crystals and C_2H_2 amorphous waste. The cumulative growth curve is then computed by averaging the individual growth curves of each cell with respect to the distribution density of the volumes of the cells.

A Voronoi diagram associated with a set of points $(Q_1, \ldots, Q_N)$ in the plane (or in space) is the partition of the plane (or space) which associates to each Q_i all the points P in the plane that are closer to Q_i than to any other Q_j. In a crystallization process as discussed here, such a point P would become part of the crystalline globule growing about Q_i. Hence, Voronoi diagrams are of obvious relevance to our problem.

Consider a Voronoi diagram as shown in Figure 6, and assume that the points are the nuclei for the crystallization of CO_2. We will assume a simple growth curve for an individual globule

$$g(c,t) = \begin{cases} kt^3 & \text{for } t \leq b, \\ c & \text{for } t > b, \end{cases}$$

where $g(t)$ is the volume of the globule at time t and b, c, k are positive constants chosen so that $g(t)$ is continuous. The idea is that this growth is cubic until the Voronoi cell is full (with the crystallized globule and the waste product).

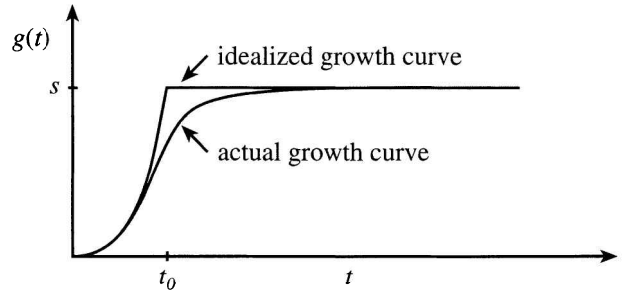


Figure 7. The ideal growth curve is cubic until it reaches $t_0 = (s/k)^{1/3}$ and then remains constant as depicted. In reality, once one or more boundaries of the cell are reached, the growth will slow down and behave more like the actual growth curve.

Let s be half the volume of the Voronoi cell in question (here, we incorporate the 1 to 1 stoichiometry of the $CO_2 \cdot C_2H_2$ mixture). As $g(t)$ reaches s, the growth should stop; therefore,

$$g(s,t) = \begin{cases} kt^3 \text{ for } t \leq \left(\frac{s}{k}\right)^{1/3}, \\ s \quad \text{for } t > \left(\frac{s}{k}\right)^{1/3}, \end{cases} \tag{1.6}$$

and, of course, $g((s/k)^{1/3}) = s$. Figure 7 displays this idealized growth curve. In reality, the growth will not be cubic after one or more boundaries of the Voronoi cell are reached; at this point the growth will slow down and eventually stop once all the corners of the cell are filled. Figure 7 suggests a more realistic growth curve in which the detailed shape will depend on the individual cell.

For simplicity we will use the idealized growth curves of the Voronoi cells as in Figure 7 (i.e., the assumption that growth proceeds as kt^3 until the cell is half full), where $g(s,t)$ satisfies (1.6), and average them over the volumes of the Voronoi cells emerging from a Poisson process. In this case $s > 0$, and it is equal to half the volume of the Voronoi cell.

1.5. The Averaging Process

Assume that a distribution function of the volumes of the Voronoi cells $f(s)$, with $s > 0$, is known. Then the probability that the volume of a cell is contained in the interval $\beta \leq s \leq \gamma$ is given by

$$\text{Prob}(\beta < s < \gamma) = \int_{\beta}^{\gamma} f(s)\, ds. \tag{1.7}$$

The experimentally observed fraction of CO_2 that has crystallized by time t is obtained in this model by averaging the growth curve $g(s,t)$ over the set of possible volumes. Since the probability distribution function for the volume is given by $f(s)$, we have

$$\varphi(t) = E(g(\cdot,t)) = \int_{0}^{\infty} g(s,t) f(s)\, ds. \tag{1.8}$$

For $s \in [0, kt^3)$, $t > (s/k)^{1/3}$, implying that $g = s$. In a similar fashion we see that $g = kt^3$ for $s \in [kt^3, \infty)$, and expression (1.8) can be written as

$$\varphi(t) = \int_{0}^{kt^3} s f(s)\, ds + kt^3 \int_{kt^3}^{\infty} f(s)\, ds, \tag{1.9}$$

which is the basis of our new model.

1.6. Choosing the Probability Density f

In order to use the result from Section 1.5, we need the probability density f. Although the actual distribution of f is unknown (this is a subject of active research in computational geometry), we will examine a few choices.

1.6.1. An Empirical Choice for f. Empirical studies suggest that a reasonable guess for the true probability density is

$$f(s) = \beta s^2 e^{-\gamma s^2}, \qquad s \geq 0, \tag{1.10}$$

where β and γ are parameters and γ relates to the variance. Since $f(s)$ is a probability density β and γ are not completely free but must be chosen so that

$$\int_{0}^{\infty} f(s)\, ds = 1. \tag{1.11}$$

To investigate this further, for any $n \in \mathbb{N} \cup \{0\}$ let

$$I_n = \int_0^\infty x^n e^{-x^2}\, dx.$$

Integrating by parts we see that for $n \geq 1$,

$$I_{n+1} = \int_0^\infty x^{n+1} e^{-x^2}\, dx = -\frac{x^n}{2} e^{-x^2}\Big|_0^\infty + \frac{n}{2}\int_0^\infty x^{n-1} e^{-x^2}\, dx = \frac{n}{2} I_{n-1}.$$

In addition,

$$I_0 = \int_0^\infty e^{-x^2}\, dx = \frac{\sqrt{\pi}}{2}, \qquad I_1 = \int_0^\infty x e^{-x^2}\, dx = \frac{1}{2},$$

and

$$\int_0^\infty \beta s^n e^{-\gamma s^2}\, ds = \beta \gamma^{-(n+1)/2} I_n.$$

With these observations expression (1.11) gives the condition $\beta = 4\gamma^{3/2}/\sqrt{\pi}$.

Another condition on β and γ can be determined by considering equation (1.9) in the limit as $t \to \infty$. Substituting (1.10) into (1.9) gives

$$\varphi(t) = \beta \int_0^{kt^3} s^3 e^{-\gamma s^2}\, ds + \beta k t^3 \int_{kt^3}^\infty s^2 e^{-\gamma s^2}\, ds. \tag{1.12}$$

Since in the limit as $t \to \infty$ all of the CO_2 becomes crystallized, we have the secondary condition that

$$\beta \int_0^\infty s^3 e^{-\gamma s^2}\, ds = 1. \tag{1.13}$$

A graphical representation of function (1.12) that satisfies condition (1.13) can be computed using any numerical computation package such as MAPLE or MATLAB. Figure 8 illustrates that this empirical model, with the value $k = 0.3193$, gives an excellent match to the experimental data.

1.6.2. Systematic Derivation of f: An Attempt. The f used in the previous calculation was a purely empirical choice, suggested by numerical experiments [**L**]. Let us see what happens if we make another choice. First, we will derive the probability density of the volume of the largest sphere fitting in the Voronoi cell as in the derivation

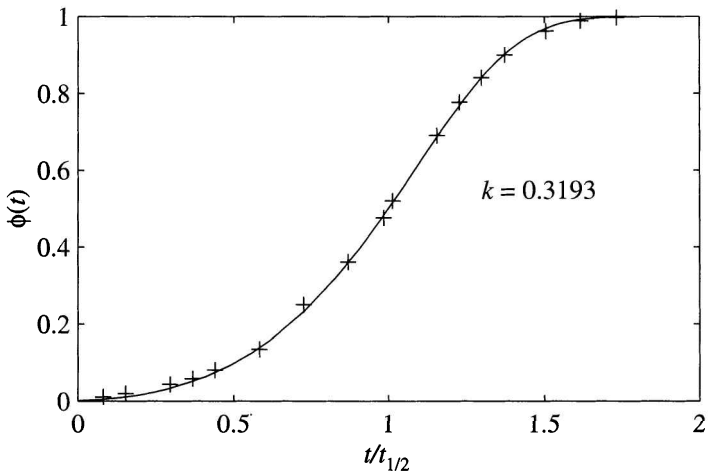


Figure 8. The resulting curve as defined in Section 1.6.1, with $k = 0.3193$ and shifted left by a factor of 0.1742, closely approximates the experimental data points.

of the K-A model in Section 1.1. This appears to be a reasonable approach since f is the probability distribution of the volume of a cell.

Assume that N nuclei are Poisson distributed within a volume V, and choose one of these nuclei to be arbitrary but fixed. If we let X denote the distance from this nucleus to its nearest neighbour, then

$$\text{Prob}\,(X > x) = \left(1 - \frac{4\pi x^3}{3V}\right)^{N-1}.$$

Therefore, the cumulative distribution function of the radius of the cell, $R_N(x)$, is given by

$$R_N(x) = \text{Prob}\ (X \leq x) = 1 - \left(1 - \frac{4\pi x^3}{3V}\right)^{N-1} \tag{1.14}$$

and

$$r_N(x) = R_N'(x) = \frac{4\pi(N-1)}{V} x^2 \left(1 - \frac{4\pi x^3}{3V}\right)^{N-2}.$$

Examining the limit as $N \to \infty$ and $V \to \infty$ such that $N = \lambda V$ with $\lambda > 0$, where λ is the intensity of the Poisson process, we find,

$$R_N(x) \to R(x) = 1 - e^{-4\pi\lambda x^3/3} \tag{1.15}$$

and

$$r(x) = R'(x) = 4\pi\lambda x^2 e^{-4\pi\lambda x^3/3}. \tag{1.16}$$

We note that the convergences in (1.2), (1.15), and (1.16) are simple consequences of l'Hôpital's rule.

We now use (1.16) and (1.15) to estimate the probability density of the volume of the largest sphere that will fit into the Voronoi cell centred at the chosen nucleus. Let S be the volume of this sphere which has radius $X/2$, then

$$\text{Prob}\,(S \le s) = \text{Prob}\left(\frac{4\pi}{3}\left(\frac{X}{2}\right)^3 \le s\right).$$

Simplifying and solving inside the brackets for X, we have

$$\text{Prob}\,(S \le s) = \text{Prob}\left(X \le \left(\frac{6s}{\pi}\right)^{1/3}\right),$$

and, using equation (1.15), the cumulative probability distribution of S is given by

$$F(s) := R\left(\left(\frac{6s}{\pi}\right)^{1/3}\right) = 1 - e^{-8\lambda s}. \tag{1.17}$$

If we denote the probability density associated with $F(s)$ as $f(s)$, then from (1.17) we obtain

$$f(s) = \frac{dF}{ds} = \frac{d}{ds}\left(1 - e^{-8\lambda s}\right) = 8\lambda e^{-8\lambda s}. \tag{1.18}$$

As a final approximation to the density f of volume distribution of Voronoi cells in (1.9), f seems a reasonable guess. Let us see what happens if we do this.

Replacing the f in equation (1.9) with (1.18) we have

$$\varphi(t) = 8\lambda \int_0^{kt^3} s e^{-8\lambda s}\, ds + 8\lambda k t^3 \int_{kt^3}^{\infty} e^{-8\lambda s}\, ds.$$

Evaluating the integrals gives

$$\varphi(t) = \frac{1}{8\lambda}\left(1 - e^{-8\lambda k t^3}\right).$$

In other words, our new theory is a genuine extension of the K-A theory!

1.7. Why Did the K-A Model Fail?

The K-A model itself is correct when the crystals fill the total volume. As well, the empirical choice for the distribution function illustrates that if the volume distribution function $f(s)$ is chosen appropriately, then the experimental data can be reproduced, despite the approximation involved in ignoring boundary effects. The attempt at a systematic derivation of $f(s)$ resulted in an extension of the K-A model, yet this extended model was no better at predicting the experimental data than the original K-A model. The problem here is that the distribution of Voronoi cell volumes is not well captured by spheres whose radii are proportional to the distance to the nearest other nuclei. Indeed, if one computes the expected volume of all of the Voronoi cells, we have

$$E(NS) = N \int_0^\infty s f(s)\, ds = 8N\lambda \int_0^\infty s e^{-8\lambda s}\, ds, = \frac{N}{8\lambda} = \frac{V}{8}$$

recalling that $N = \lambda V$. If we had used spheres of radius $X/2^{1/3}$ rather than $X/2$, we would have ended up with $E(NS) = V/2$ as we should have, but in doing this we still do not represent the distribution of Voronoi cell volumes well. It is the presence of the waste product that forces us to consider the distribution of individual cell volumes, as it stops the growth of the crystal before the space is filled, and the time at which the growth stops depends on the size of the cell.

To be completely accurate we need to take the shapes of the Voronoi cells into account, but this is too hard. Rather, any improvements to the model must explicitly account for the location of the waste product. Clearly, it is the presence of the waste product that causes the deviation of our crystallization process from the K-A model.

Exercises

(1) (The raindrop problem) At time $t = 0$, rain starts to fall at an even and steady rate of I^* droplets per unit time per unit area over a large pond (imagine an ocean). Each droplet creates a

wave which spreads outward at a constant velocity $v > 0$ radially from the point of impact.

Let P be an arbitrary but fixed point on the pond. The probability that N waves have passed P at time t is given by a Poisson distribution,

$$\text{Prob}\{N \text{ waves have passed through } P\} = e^{-E(t)} \frac{E(t)^N}{N!},$$

where $E(t)$ is the expected number of waves that have passed through P by time t.

(a) Compute $E(t)$. (Assume there is no interaction between the waves.)

(b) Find the probability that at least one wave has crossed P by time t. Compare this probability to the crystallization rate that we computed from the K-A model.

(2) A two-dimensional crystal growth process is modelled by an Avrami equation $\varphi(t) = 1 - e^{-kt^2}$, where $k > 0$.

(a) Sketch $\varphi(t)$.

(b) Experimental observations suggest that the inflection point of φ occurs at $t = 2$ hours. Find the dimension and a numerical value for k.

(c) Use a Taylor expansion for e^x to show that for $t \ll 1$, $\varphi(t) \sim kt^2$. Explain why this initial behaviour for φ must be expected.

(d) A chemist asks you whether knowledge of k suffices to compute the intensity (number/unit volume) of the crystallization nuclei. What is your answer?

(3) Suppose that the intensity of the Poisson process describing the crystallization nuclei is time dependent and given by $\lambda + g(t)$, where $g(0) = 0$ and g is continuous and monotonically increasing (take $g(t) = \epsilon t$ as an example). Follow the method from Exercise 1 to derive a reasonable K-A model for this scenario.

(4) Compute the growth curve

$$\varphi(t) = \int_0^{kt^3} x f(x)\, dx + kt^3 \int_{kt^3}^{\infty} f(x)\, dx$$

for the following cases:

(a) $f(x) = \frac{1}{A}\chi_{[0,A]}(x)$, where $A > 0$ is fixed. Here

$$\chi_{[0,A]}(x) = \begin{cases} 1 \text{ for } x \in [0, A], \\ 0 \text{ for } x \notin [0, A]. \end{cases}$$

(b) $f(x) = \lambda e^{-\lambda x}$, $\lambda > 0$. (This density was derived as the probability density of the volumes of the largest spheres that fit inside the Voronoi cells generated by a Poisson process.)

(5) The Voronoi diagram associated with three points in a plane always consists of three lines meeting at one point.
   (a) What is the geometric meaning of this point?
   (b) Can you construct a Voronoi diagram associated with N points in a plane such that all of its defining lines meet at one point?

(6) It can be proved that $\int_0^\infty e^{-x^2}\,dx$ exists. Proceeding under the assumption that it does exist and that the other integrals used exist, show that $\int_0^\infty e^{-x^2}\,dx = \sqrt{\pi}/2$ as follows:
   (a) Show that

$$\left(\int_0^\infty e^{-x^2}\,dx\right)\left(\int_0^\infty e^{-y^2}\,dy\right) = \int_0^\infty \int_0^\infty e^{-(x^2+y^2)}\,dx\,dy. \tag{1.19}$$

   (b) Express (1.19) as an integral of the form $\int_0^\infty \int_0^{\pi/2} f(r, \theta)\,d\theta\,dr$, and evaluate it over the unbounded region.
   (c) Combine the results of (a) and (b).

(7) In this question we explore in detail the properties of the empirical distribution chosen in Section 1.6.1.
   (a) Determine the values of β and γ as explicitly as possible.
   (b) Using the results derived in Section 1.6.1, find the expected value of S in terms of γ and β. Then find the variance of S, $\text{var}(S) = E(S^2) - [E(S)]^2$.
   (c) The paper by Gough et al. [**G**] states that the ratio of the standard deviation to the mean has a value of 0.422. Verify that the exact value is $\sqrt{3\pi/8 - 1}$ and that it depends solely on the normalization

$$\int_0^\infty f(s)\,ds = 1.$$

Table 1. Experimental data for Exercise 9.

n	t_n	y_n	n	t_n	y_n	n	t_n	y_n
1	0.0802	0.011	7	0.7259	0.250	13	1.3000	0.841
2	0.1522	0.020	8	0.8696	0.361	14	1.3741	0.900
3	0.2955	0.044	9	0.9842	0.476	15	1.5049	0.962
4	0.3674	0.059	10	1.0132	0.520	16	1.6158	0.989
5	0.4391	0.081	11	1.1564	0.690	17	1.7305	0.9972
6	0.5827	0.135	12	1.2284	0.777			

(8) Suppose that the distribution of volumes is not continuous but rather can take on only two values. In particular suppose that

x	1	2
$p(x)$	p	$1-p$
$s(x)$	s_1	s_2

where $0 \leq p \leq 1$ are the probabilities and $0 \leq s_1 < s_2$ are the respective volumes. Using the discrete version of (1.9)

$$\varphi(t) = \sum_{\{x:s(x)\leq kt^3\}} s(x)p(x) + kt^3 \sum_{\{x:s(x)>kt^3\}} p(x),$$

determine the resulting growth curve $\varphi(t)$.

(9) CLASS PROJECT

First verify that under the substitution $\lambda = kt^3$ that $\varphi(t)$ is a solution of the coupled system

$$\begin{aligned} \frac{du}{d\lambda} &= -f(u), \qquad u(0) = 1, \\ \frac{dv}{d\lambda} &= u, \qquad v(0) = 0, \end{aligned}$$

where $v(\lambda) = \varphi((\lambda/k)^{1/3})$. By solving the system with $f(u)$ as determined by (1.10), determine the value of λ where $v = 1/2$. You may want to use some type of iterative technique. Denote this value as $\lambda_{1/2}$.

The experimental data points shown in Figure 2 are listed in Table 1 where the t_n have been scaled so that $t = 1$ corresponds to a value of $y = 1/2$. For the second part of this project use the

data points find the value of k that minimizes the error

$$e(k) = \sum_{n=1}^{17} (y_n - v(\lambda_n))^2 ,$$

where $\lambda_n = k[t_n - 1 + (\lambda_{1/2}/k)^{1/3}]^3$. In particular, verify that $k = 0.3193$ and that for this value of k the curve should be shifted to the left by an amount of 0.1742. For this part of the project you may want to investigate various numerical minimization algorithms such as the bisection method, quadratic interpolation, or the golden section search method.

NOTES

This problem arose in chemistry research at the University of Victoria, published in [**F, G**]. The Kolmogorov–Avrami model is well known in physical chemistry and may be found in standard reference books (e.g., [**V**]). A good source for the statistical properties of Poisson generated Voronoi diagrams is [**B**].

Chapter 2

Will the Valve Hold?

Concepts and Tools: Ordinary differential equations, numerical analysis

The generation of hydroelectric power involves a potentially destructive environmental force that must be carefully managed. Should, for any reason, a proposed scheme of power generation be flawed by design, a great deal of damage could be inflicted on neighbouring areas. This is the responsibility faced by B.C. Hydro in 1961 when attempting to tap a water reservoir by drilling a tunnel through a mountainside. Figure 1 illustrates this situation.

This tunnel was drilled so that a rock plug between the tunnel and the lake was left intact. At the opposite end a valve had been installed for later control of the water flow. These control measures are depicted in Figure 1. The control valve was shut so that when the rock plug, which had been charged with explosives, was blown clear, water would rush into the tunnel to fill it and, in the process, compress the air that had become trapped. The valve had been designed to withstand twice the hydrostatic pressure of the lake. However, the evening before the rock plug was to be removed, it was pointed out that the compressed air bubble would have to absorb all of the dynamic energy of the water rushing in and filling the tunnel, and that

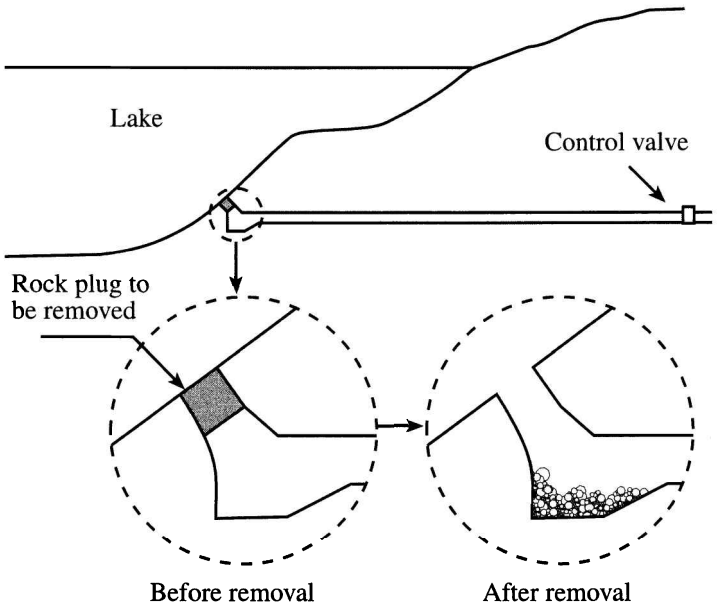


Figure 1. This figure illustrates the general approach: a tunnel cut through a mountain in order to tap a water reservoir. The tunnel is initially sealed with a rock plug at one end and a control valve at the other end.

the pressure might temporarily exceed the safe threshold. B.C. Hydro was faced with a pressing issue: When the rock plug was blown, would the valve be able to withstand the maximum pressure?

The responsible engineer had a slide rule and 24 hours to solve this problem. He knew that the valve had been designed to withstand twice the hydrostatic pressure of the lake, and predicted after a sleepless night of calculations that the maximum pressure would be about 180% of the lake's hydrostatic pressure. This indicated that the valve would indeed withstand the maximum pressure. When the valve was blown, the pressure was in fact measured to be approximately 175%; he had been incredibly accurate using only primitive computational tools. In this chapter we analyse this problem, revisit his method of accurately predicting the maximum pressure, and discuss it from a modern analytical and computational point of view.

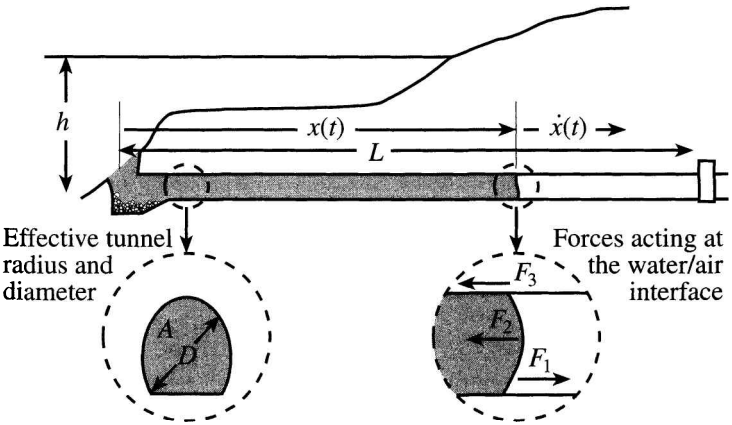


Figure 2. Shown are the parameters h, $x(t)$, $\dot{x}(t)$, and L. Here $\dot{x}(t)$ denotes the velocity of the water. The inset to the left illustrates a typical tunnel cross-section with area A and an effective diameter of D. To the right are the forces that act on the water/air interface.

2.1. Terminology

We will begin by introducing some of the terms that will be used in the following sections. They are as follows:

Process parameters		Tunnel parameters	
$x(t)$	position of the water/air interface at time t	A	cross-sectional area of the tunnel
ρ	density of water	L	length of the tunnel
P_o	atmospheric pressure	D	diameter of the tunnel
h	depth of the lake	f	coefficient of friction between the water and the wall of the tunnel
g	acceleration due to gravity		

The graphical description of the parameters is shown in Figure 2. Throughout this chapter, we may suppress t for clarity, but x should always be considered a function of time.

2.2. The Relevant Forces

In order to determine the maximum pressure, denoted by $P_{\max}$, that will occur in the air bubble in the tunnel, we will examine the forces which affect the motion of the water. (In general forces are denoted as vectors; however, because we are only dealing with one dimension in this problem, we will drop the vector notation.) Using the fact that force is equal to pressure times area, we formulate three forces which are relevant to the determination of the maximum pressure:

- F_1 := the hydrostatic force,
- F_2 := the force due to adiabatic compression,
- F_3 := the force exerted by friction.

2.2.1. F_1: The Hydrostatic Force. The first force affecting the motion of the water is determined by the hydrostatic pressure: the pressure exerted by the water reservoir which is being tapped. This hydrostatic pressure, denoted by P_{hs}, is given by

$$P_{hs} = \rho g h + P_o;$$

therefore, F_1 becomes

$$F_1 = P_{hs} A = (\rho g h + P_o)\, A, \tag{2.1}$$

where we recall that A is the cross-sectional area of the tunnel and h is kept constant, a reasonable assumption.

2.2.2. F_2: The Force Due to Adiabatic Compression. The second force is a reaction force of the compressed air. If we assume that there is no change in heat in the gas while it is compressed (*adiabatic compression*), we can use the adiabatic equation of state

$$PV^\gamma = \text{ constant}, \tag{2.2}$$

where $\gamma > 0$ is a parameter of the gas ($\gamma = 1.4$ for air). If we take the volume as $V = A(L - x)$ and denote $P(0)$ as P_o, then the constant of (2.2) becomes $P_o A^\gamma L^\gamma$. (This follows by setting $t = 0$ in (2.2).) Using this fact allows equation (2.2) to be expressed as

$$P(t) A^\gamma (L - x)^\gamma = P_o A^\gamma L^\gamma,$$

which, upon solving for $P(t)$, yields

$$P(t) = P_o \left(\frac{L}{L - x} \right)^{\gamma}. \tag{2.3}$$

Consequently, the subsequent force, F_2, is given by

$$F_2(t) = -P(t)A = -AP_o \left(\frac{L}{L - x} \right)^{\gamma} \tag{2.4}$$

which is negative because the force exerted by the compressed air is in opposition to the hydrostatic force, as shown by Figure 2.

2.2.3. F_3: The Force Exerted by Friction. In order to determine the force exerted by friction, we will assume that the water is in *fully turbulent flow*. In this case, a known (semi-empirical) formula for the associated pressure loss ΔP is

$$\Delta P(t) = \frac{\rho f l}{2D} \dot{x}^2,$$

where l is the length of water flow in contact with the wall of the tunnel and f is a dimensionless friction coefficient.[1] If we identify l with $x(t)$, then the force due to friction, F_3, becomes

$$F_3(t) = -\Delta P(t)A = -\frac{\rho f A}{2D} x\dot{x}^2 \tag{2.5}$$

which, like F_2, also acts against F_1, as illustrated by Figure 2. In actual fact, the direction of the frictional force should oppose the direction of motion. Because of this, the sign of F_3 should be $-\text{sgn}(\dot{x})$ where

$$\text{sgn}(a) = \begin{cases} 1 \text{ for } a > 0, \\ 0 \text{ for } a = 0, \\ -1 \text{ for } a < 0. \end{cases}$$

Therefore, equation (2.5) becomes

$$F_3(t) = -\text{sgn}(\dot{x}) \frac{\rho f A}{2D} x\dot{x}^2. \tag{2.6}$$

While the tunnel is being filled, we expect $\dot{x}(t) > 0$ and $\text{sgn}(\dot{x}(t)) = 1$. While the water flow into the tunnel is not likely to reverse itself in practice, it is nevertheless useful to model F_3 such that this contingency is covered. After all, when the bubble reaches the maximum

[1] This semi-empirical expression is known as the Darcy–Weisbach equation.

pressure, we expect $\dot{x}(t)$ to approach zero, and air and water will mix turbulently at this time. Our model will then lose validity.

2.3. The Equation of Motion

According to the first law of thermodynamics, the time rate of change of a material's internal and kinetic energy is equal to the rate of change of the heat transferred to the material region less the rate of work done by the material region. By neglecting the internal energy of the fluid and noting that the fluid in the tunnel is not heated, we have

$$\frac{dE}{dt} = -\frac{dW}{dt}, \tag{2.7}$$

where E is the kinetic energy of the material region given by

$$E = \frac{1}{2}mv^2 = \frac{1}{2}\rho A x \dot{x}^2. \tag{2.8}$$

Here m is the mass, given by $m = \rho A x(t)$, and the velocity $v = \dot{x}$.

Now that we have determined the relevant forces, we are able to find the equation of motion for the interface. This hinges on the fact that for a given force F acting *on* the fluid, the corresponding rate of work done *by* the fluid is $-dW/dt = F\dot{x}$ so that the right-hand side of (2.7) is

$$\begin{aligned} -\frac{dW}{dt} &= \dot{x}\sum_i F_i \\ &= \dot{x}\left[(\rho g h + P_o)A - AP_o\left(\frac{L}{L-x}\right)^{\gamma} - \operatorname{sgn}(\dot{x})\frac{\rho f A}{2D}x\dot{x}^2\right] \end{aligned} \tag{2.9}$$

using expressions (2.1), (2.4), and (2.6). By using expression (2.8), the left-hand side of (2.7) becomes

$$\frac{dE}{dt} = \rho A \dot{x}\left(\frac{1}{2}\dot{x}^2 + x\ddot{x}\right), \tag{2.10}$$

and combining equations (2.9) and (2.10), we have

$$\begin{aligned} &\rho A\dot{x}\left(\frac{1}{2}\dot{x}^2 + x\ddot{x}\right) \\ &\quad = \dot{x}\left[(\rho g h + P_o)A - AP_o\left(\frac{L}{L-x}\right)^{\gamma} - \operatorname{sgn}(\dot{x})\frac{\rho f A}{2D}x\dot{x}^2\right] \end{aligned} \tag{2.11}$$

or

$$x\ddot{x} = gh + \frac{P_o}{\rho} - \frac{P_o}{\rho}\left(\frac{L}{L-x}\right)^{\gamma} - \operatorname{sgn}(\dot{x})\frac{f}{2D}x\dot{x}^2 - \frac{1}{2}\dot{x}^2, \tag{2.12}$$

a second-order ordinary differential equation to be solved numerically for suitable initial conditions.

2.4. Analysis: Is the Initial Value Problem Well-Posed?

Having developed an equation to model the motion of water, we need to determine whether the initial value problem is solvable. In order to do this, we first need to determine the initial conditions for equation (2.12). Because the rock plug remains intact at time $t = 0$, the water (having not yet entered the tunnel) will not be moving; this suggests that the equation should have initial conditions $x(0) = \dot{x}(0) = 0$. As we shall see, these initial conditions lead to an ill-posed problem with no solution.

The right-hand side of equation (2.12) is a function of x and $\dot{x}$, which we will abbreviate by $G(x, \dot{x})$:

$$G(x,\dot{x}) = gh + \frac{P_o}{\rho} - \frac{P_o}{\rho}\left(\frac{L}{L-x}\right)^{\gamma} - \operatorname{sgn}(\dot{x})\frac{f}{2D}x\dot{x}^2 - \frac{1}{2}\dot{x}^2, \tag{2.13}$$

which implies

$$x\ddot{x} = G(x,\dot{x}) \tag{2.14}$$

taken with initial conditions $x(0) = \dot{x}(0) = 0$. Hence, evaluating (2.13) at time $t = 0$ gives

$$G(0,0) = gh > 0,$$

which must be positive to ensure that the water flows into the tunnel. This implies that, at the same initial time, the left-hand side of equation (2.14) should be a positive constant as well. However, $x\ddot{x}(0) = 0$, which hints at a contradiction. We realize that equation (2.14) cannot hold at $t = 0$ if $x(0) = \dot{x}(0) = 0$. However, this alone does not invalidate our modelling; the equation may possibly hold for $t > 0$, and there may be a solution which assumes the initial values. Next we show that there is no such solution.

2.4.1. A Class of Ill-Posed Problems. For a better understanding of the difficulty, we examine the less complex equation

$$x\ddot{x} = C, \qquad x(0) = \dot{x}(0) = 0, \tag{2.15}$$

where C is a positive constant, and now prove by contradiction that problem (2.15) has no solution. It will be transparent from the proof that this *ill-posedness* also applies to our problem (2.14) with zero initial data.

Assume that a solution $x(t)$ does exist for equation (2.15) with $x(t) > 0$ for $t > 0$. Then solving for $\ddot{x}$ and multiplying by $\dot{x}$ yields

$$\dot{x}\ddot{x} = C\frac{\dot{x}}{x}.$$

Considering solutions on the interval $[\delta, t]$ with $\delta > 0$, and integrating both sides with respect to t, we obtain

$$\int_\delta^t \dot{x}\ddot{x}\, dt = C\int_\delta^t \frac{\dot{x}}{x}\, dt,$$

which becomes

$$\frac{1}{2}\left[\dot{x}(t)\right]^2 - \frac{1}{2}\left[\dot{x}(\delta)\right]^2 = C\left[\ln x(t) - \ln x(\delta)\right]. \tag{2.16}$$

Fixing t and taking the limit as $\delta \to 0^+$, the left-hand side of equation (2.16) gives

$$\frac{1}{2}\left[\dot{x}(t)\right]^2 = b,$$

where b is some fixed value. However, the right-hand side yields

$$\lim_{\delta \to 0^+} C\left[\ln x(t) - \ln x(\delta)\right] \longrightarrow \infty,$$

which is a contradiction; therefore, there is no solution to equation (2.15). We say that the initial value problem is ill-posed.

2.5. Revising the Model

It is apparent from our analysis of equations (2.12) and (2.15) that problems of the type

$$\ddot{x} = \frac{G(x, \dot{x})}{x}, \qquad x(0) = \dot{x}(0) = 0, \tag{2.17}$$

are not well-posed and therefore will not sufficiently model the situation at hand. Consequently, a modification must be made to correct

for the discontinuity that occurs in the model with respect to the initial conditions. In order to better understand how to effectively revise the model, we need a bit of theory of initial value problems for ordinary differential equations.

2.5.1. Second-Order Equations as Systems of First-Order Equations. Initial value problems for second-order differential equations (such as equation (2.17)) of the type

$$\ddot{x}(t) = f(x, \dot{x}, t), \qquad x(0) = a,\ \dot{x}(0) = b, \tag{2.18}$$

may be rewritten as systems of first-order ordinary differential equations. To do this set $x_1(t) = x(t)$ and $x_2(t) = \dot{x}(t)$, such that equation (2.18) becomes

$$\dot{x}_2(t) = \ddot{x}(t) = f(x_1, x_2, t), \qquad x_1(0) = a,\ x_2(0) = b.$$

Now define the vector $\mathbf{x}(t)$ as $\mathbf{x}(t) := (x_1(t), x_2(t))$ and let

$$\dot{\mathbf{x}}(t) = F(\mathbf{x}, t), \tag{2.19}$$

where $F(\mathbf{x}, t) = (x_2(t), f(x_1, x_2, t))$. This first-order system (2.19) with initial condition $\mathbf{x}(0) = (a, b)$ is equivalent to (2.18).

2.5.2. Unique and Local Solvability. Critical to the analysis and revision of the model is an understanding of the conditions which are sufficient for the unique local solvability of equation (2.19). Such conditions, as presented in intermediate texts on ODEs are:

(1) $F(\mathbf{x}, t)$ is continuous in an open neighbourhood of $t = 0$, $\mathbf{x} = (a, b)$.

(2) $\nabla_{\mathbf{x}} F$ is bounded in this neighbourhood. (This entails the so-called Lipschitz continuity of F in the neighbourhood.)

Remark 2.1. These conditions are sufficient but *not* necessary. See the examples below.

Example 2.2. Consider the system

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= x_1 x_2 - t^2 x_1^2, \end{aligned}$$

with initial conditions $x_1(0) = 5$, $x_2(0) = 2$.

F is continuous in any neighbourhood of $t = 0$, $x_1 = 5$, $x_2 = 2$, and its derivatives are bounded and continuous in this same neighbourhood. Therefore, we conclude that a unique local solution exists. It may only be defined locally, i.e., for a sufficiently small time interval.

Example 2.3. In this next example, we examine the system equivalent to (2.15)

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= \frac{c}{x_1}, \end{aligned}$$

with the initial conditions $x_1(0) = x_2(0) = 0$. Note that $\dot{x}_2$ is not continuous at $t = 0$, nor is the partial derivative of c/x_1 bounded (or even defined!) there. Indeed, we have already seen that this problem has no solution.

Example 2.4. As a third example, consider the equation

$$\dot{x} = \sqrt{|x|}, \qquad x(0) = 0,$$

with solutions $x(t) = 0$ and

$$x(t) = \begin{cases} 0 & \text{for } t \leq a, \\ \dfrac{(t-a)^2}{4} & \text{for } t > a, \end{cases}$$

with $a > 0$. By observing that

$$\frac{d}{dt}\left[\frac{(t-a)^2}{4}\right] = \frac{t-a}{2} = \sqrt{\frac{|t-a|^2}{4}},$$

we verify that, for any value of a, $x(t)$ is in fact a solution. However, if we examine the derivative, we see that $\frac{d}{dx}(\sqrt{|x|})$ is not bounded as $x \to 0$. Thus, solutions do exist for this function as verified; however, they may not be unique because the boundedness condition has been violated. Figure 3 displays various solution curves.

Example 2.5. Finally, consider

$$\sqrt{|x|}\ddot{x} = 1, \qquad x(0) = \dot{x}(0) = 0.$$

Here, both conditions are violated, yet it is easy to determine constants b and $r > 0$, such that bt^r solves the initial value problem for

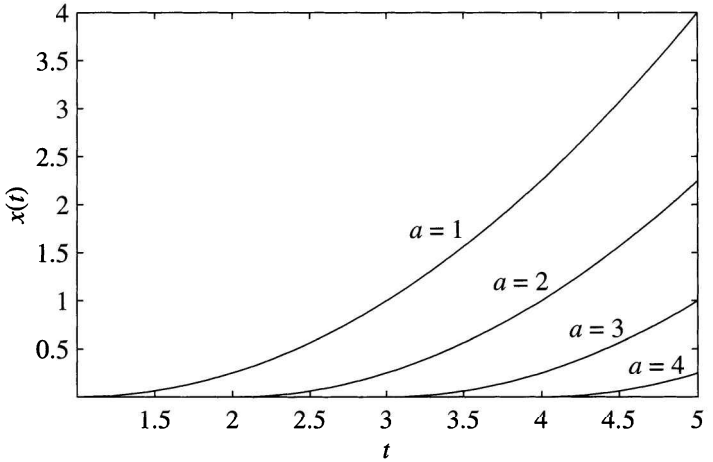


Figure 3. The solution to the differential equation for this example is not unique. The illustration shows the solution curves for various values of a.

$t > 0$. This example shows that the conditions for unique solvability are sufficient but not necessary.

We are now able to examine some possible revisions of our model. On physical grounds we expect that we should revise the model such that we obtain a unique solution. Moreover, the solution should depend continuously on the given data.

2.6. Revision 1: Adding a Reference Distance

In this first method of revision, we observe that some water in the lake will be displaced during this process. To correct for this, a reference distance will be added to $x(t)$ as depicted in Figure 4(a) when defining the mass. We denote this reference distance as d, and redefine the mass to be $m = \rho A\,(x + d)$ accordingly. Therefore, (2.10) becomes

$$\frac{dE}{dt} = \rho A \dot{x} \left[\frac{1}{2} \dot{x}^2 + (x + d) \ddot{x} \right],$$

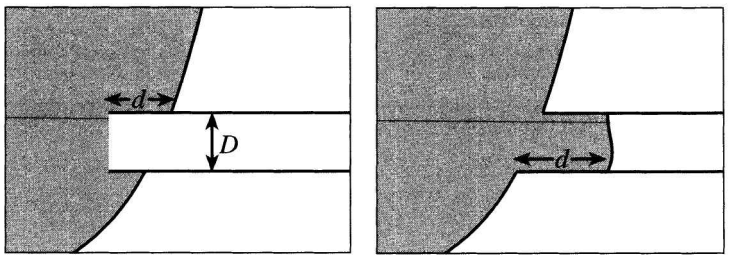


Figure 4. (a) Revision 1: A reference distance d is added to account for the water in the lake which is displaced. (b) Revision 2: The reference distance d is considered to be inside the tunnel. This second revision will alter the initial conditions.

and (2.12), the equation of motion, is similarly modified yielding

$$(2.20)\quad (x+d)\,\ddot{x} = gh + \frac{P_o}{\rho} - \frac{P_o}{\rho}\left(\frac{L}{L-x}\right)^{\gamma} - \operatorname{sgn}(\dot{x})\frac{f}{2D}x\dot{x}^2 - \frac{1}{2}\dot{x}^2.$$

Notice that the reference distance does not contribute to the frictional term or the pressure term. Denoting the right-hand side by $G(x,\dot{x})$ as in Section 2.4 we have

$$(x+d)\,\ddot{x} = G(x,\dot{x}),$$

which gives the revised equation of motion as

$$\ddot{x} = \frac{G(x,\dot{x})}{x+d}, \qquad x(0) = \dot{x}(0) = 0.$$

This modified model, unlike the previous one represented by (2.12), is continuous at time $t = 0$ and thus has a unique local solution that will depend on the chosen value for d. Table 1 displays results obtained for various choices. This method was the one used by the B.C. Hydro engineer. He had only one shot at the calculation. He used $d = D$.

2.7. Revision 2: Changing the Initial Conditions

Another simple way to modify (2.12) to correct for the discontinuity that occurs at $t = 0$ is to change the initial conditions. To do so, we assume that some water has already entered the tunnel at time

Table 1. Maximal pressures obtained for the two revised models. There are three nondimensional quantities that determine the motion. Variation in the first three columns are a result of choosing $L = 500, 1000, 2000$ m, $h = 50, 100, 200$ m, $f = 0, 0.05, 0.1$, and $d = 0.1, 1, 2$ m. Nominal values of the parameters are $d = 2$ m, $L = 1000$ m, $f = 0.05$, $D = 2$ m, $P_o = 1$ atm $= 101325$ Pa, $\rho = 1000$ kg m^{-3}, $g = 9.8$ m s^{-2} and $h = 100$ m.

d/L	fL/D	$P_o/\rho gh$	$P_{\max}^{\text{rev } 1}$(atm)	$P_{\max}^{\text{rev } 2}$(atm)
$L = 500, 1000, 2000$ m				
0.004	12.5	0.10339	17.78	17.74
0.002	25.0	0.10339	14.02	14.01
0.001	50.0	0.10339	12.31	12.31
$h = 50, 100, 200$ m				
0.002	25.0	0.20679	7.009	7.006
0.002	25.0	0.10339	14.02	14.01
0.002	25.0	0.05170	30.79	30.76
$f = 0, 0.05, 0.1$				
0.002	0	0.10339	320.5	320.5
0.002	25.0	0.10339	14.02	14.01
0.002	50.0	0.10339	12.31	12.31
$d = 0.1, 1, 2$ m				
0.0001	25.0	0.10339	14.01	14.01
0.001	25.0	0.10339	14.02	14.01
0.002	25.0	0.10339	14.02	14.01

$t = 0$ (i.e., take the reference distance d from Section 2.6 as inside the tunnel as shown by Figure 4(b)) which gives the initial condition $x(0) = d > 0$.

Now assume that the velocity of the water is given by $\dot{x}(0) = \sqrt{2gh}$, recalling that h is the depth of the lake and g the acceleration due to gravity. This velocity was determined using Toricelli's law for the speed at which water pours out of a beaker at equilibrium, while disregarding friction (which will not have had a significant impact at this point). Figure 5 illustrates the logic behind this choice for $\dot{x}$.

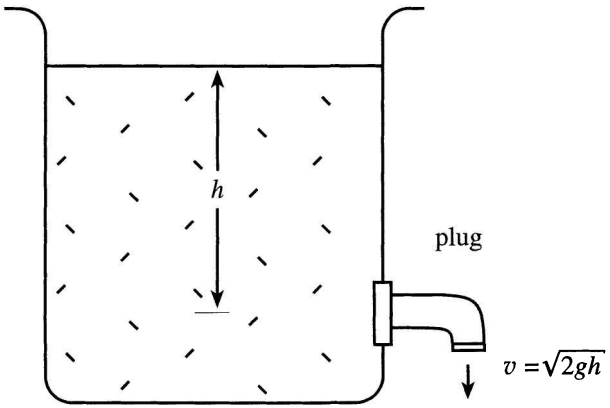


Figure 5. Toricelli's law dictates that the speed at which a liquid leaves a hole in the bottom of a tank of depth h is $v = \sqrt{2gh}$. Notice that this is equal to the speed acquired by a body falling freely through a vertical distance h.

Using these new initial conditions in our original equation of motion

$$x\ddot{x} = gh + \frac{P_o}{\rho} - \frac{P_o}{\rho}\left(\frac{L}{L-x}\right)^{\gamma} - \operatorname{sgn}(\dot{x})\frac{f}{2D}x\dot{x}^2 - \frac{1}{2}\dot{x}^2,$$

$x(0) = d$, $\dot{x}(0) = \sqrt{2gh}$ yields results which are consistent with those from Section 2.6. Both methods of revising the model remove the discontinuity that occurred and give consistent results. In both revisions the maximal pressure proved to be very robust with respect to d except when d becomes too small.

2.8. $P_{\max}$: The Maximal Pressure

We now return to the purpose of this chapter, which is to estimate the maximum pressure $P_{\max}$ that results when the rock plug is blown and the water enters the tunnel. We determine this pressure by numerically integrating the model derived in this chapter. In particular, we solve the nondimensionalized version of Revision 1,

$$2\left(x + \frac{d}{L}\right)\ddot{x} = 1 + \frac{P_o}{\rho g h}\left[1 - \left(\frac{1}{1-x}\right)^{\gamma}\right] - \operatorname{sgn}(\dot{x})\frac{fL}{D}x\dot{x}^2 - \dot{x}^2,$$

$x(0) = \dot{x}(0) = 0$, where fL/D and $P_o/\rho gh$ are varied while holding the ratio $d/L = 0.002$. The contour lines correspond to the maximal

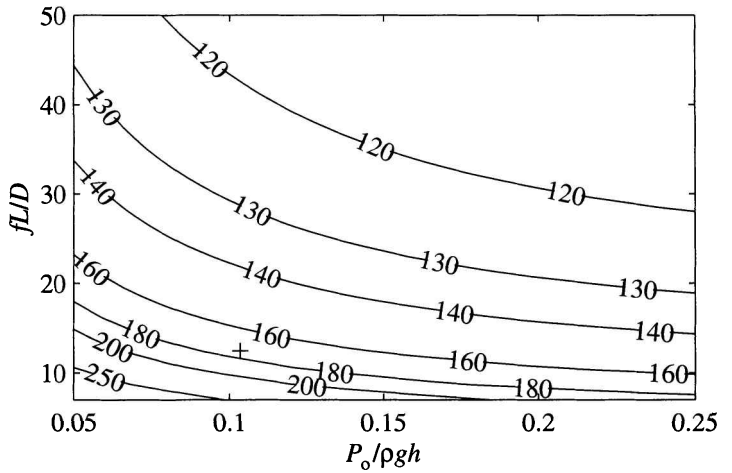


Figure 6. Displayed are the contours of equal maximal pressure corresponding to Revision 1 with $d/L = 0.002$ and various values of the nondimensional parameters fL/D and $P_o/\rho gh$. For a given maximal pressure, the value assigned to the contour is the differential pressure on the valve expressed as a percentage of the hydrostatic pressure. The cross corresponds to $f = 0.025$, $h = 100$ m, and $L = 1000$ m with $(P_{\max} - P_o)/\rho gh = 1.73$.

differential pressure experienced by the valve which is expressed in terms of the hydrodynamic pressure ρgh.

From Figure 6 it is apparent that the maximal pressure is inversely proportional to the friction factor, f, the tunnel length, L, and the atmospheric pressure, P_o, while it is directly proportional to the depth of the lake, h, and the diameter of the tunnel, D. Assuming that the tunnel has walls of roughened concrete, we use a value of $f = 0.025$ [**R**]. This value together with $h = 100$ m and $L = 1000$ m results in a differential pressure of 173% of the hydrostatic pressure and is indicated by the cross on the figure. Additional parameters chosen were $P_o = 1$ atm $= 101325$ Pa, $\rho = 1000$ kg m^{-3}, $g = 9.8$ m s^{-2}, and $\gamma = 1.4$. Table 1 illustrates that changes in the reference distance have little effect on $P_{\max}$. This is significant because in reality it is unknown how the flow at the bottom of the lake behaves.

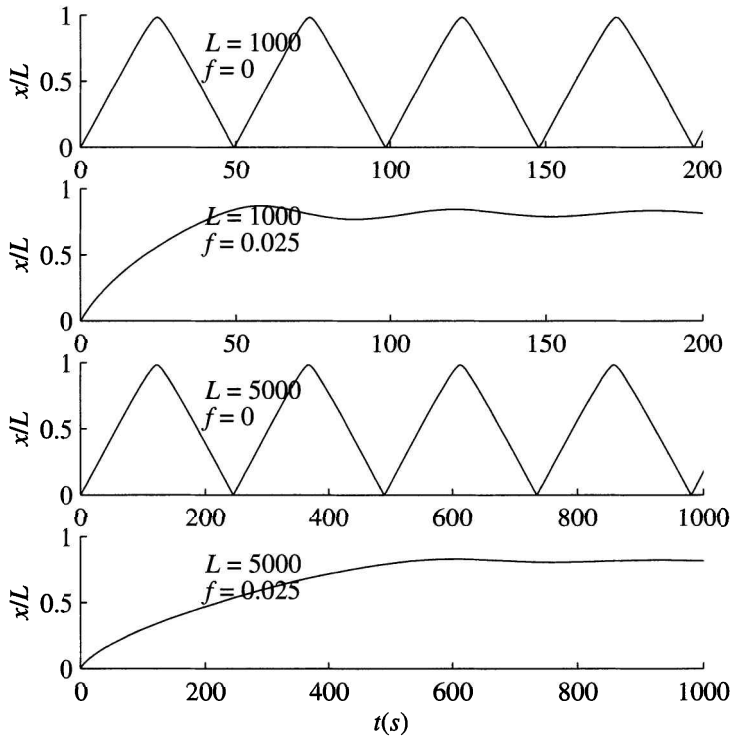


Figure 7. For each graph $P_o = 1$ atm, $\rho = 1000$ kg m^{-3}, $g = 9.8$ m s^{-2}, $\gamma = 1.4$, $h = 100$ m, $D = 2$ m, and $d = 2$ m. Time dependence of the interface position $x(t)$ is as determined by equation (2.20).

We conclude with Figures 7 and 8 which show the behaviour of $x(t)$ and $P(t)$ under different circumstances.

It should be evident that the model developed in this chapter is only valid up to the point where the maximum pressure is reached. After that point, the interface between the water and the air bubble will cease to exist, and the air will escape as bubbles through the tunnel into the lake. Hence, our numerical calculations are meaningless past the point of maximum pressure.

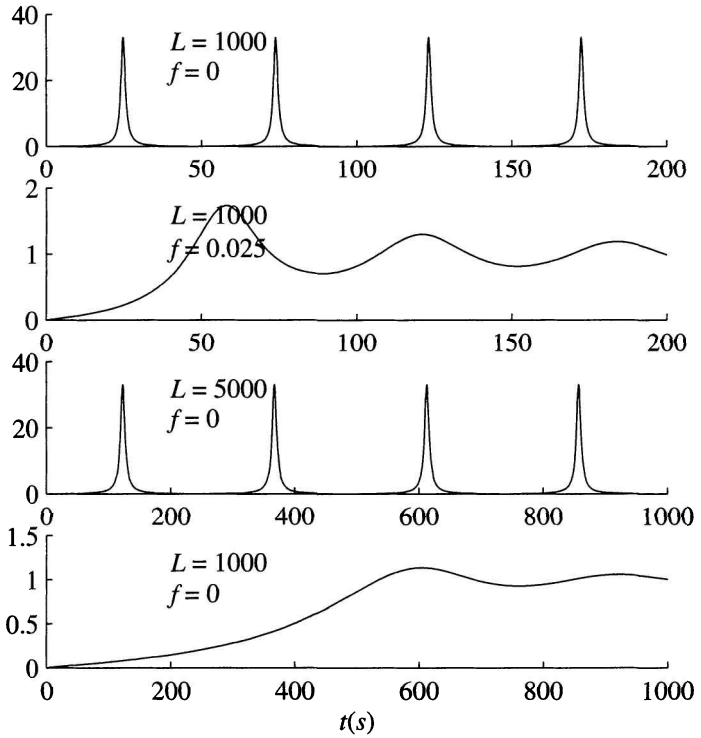


Figure 8. Pressure dependence corresponding to Figure 7. The ordinate is the nondimensional quantity $[P(t) - P_0]/\rho gh$ with $P(t)$ given by equation (2.3).

Exercises

(1) Check that the model for adiabatic air compression in a tunnel is dimensionally consistent (i.e., verify that all the terms in the model have the same dimension). Use the [m, kg, s] system. For

example, the term $x\ddot{x}$ has dimension $\mathrm{m}^2\mathrm{s}^{-2}$. Verify in particular that the coefficient of friction f is dimensionless.

(2) What additional term must be added to the model equations in the air compression problem if the tunnel slopes downward at an angle $\alpha > 0$?

(3) Give a derivation of Toricelli's law.

(4) Find $b > 0$ and $r > 0$ such that $x(t) = bt^r$ solves the initial value problem from Example 2.5.

(5) Use expression (2.20) to determine the equilibrium position of the water/air interface. Find the equilibrium pressure of the air in the tunnel.

(6) Reproduce the numerical results in Table 1. Investigate how the pressure depends on the friction coefficient f for a tunnel of length 5000 m. What happens for $f = 0$? Use a numerical ODE solver of your choice (MAPLE, MATLAB, etc.).

(7) CLASS PROJECT
Investigate the unique solvability of $\dot{x} = f(x,t)$, $x(0) = x_0$ by consulting an intermediate or advanced text on ordinary differential equations. Explore the concept of Lipschitz continuity of a function, investigate its past in the classical (Picard-Lindelöf) existence and uniqueness theorem. Look up the Peano existence theorem. See how these theorems relate to the examples given in Section 2.5.

(8) CLASS PROJECT
Consider the situation when the path of the tunnel is given by the monotonically decreasing function $y = \eta(x)$ with $\eta(0) = h$, the depth of the lake at the tunnel entrance. Using the derivation of the differential equation (2.12) as a template, find an equation for x, the displacement of the water at time t. Ensure that the ODE (2.12) is recovered in the case of a constant tunnel depth, $\eta(x) \equiv h$.

NOTES

In 1961, Bill H. was responsible for making the decision whether or not to ignite the explosives in the rock plug. The day before the explosion, it was brought to his attention that the compressed air in the tunnel would have to absorb all the dynamic energy of the invading water; however, it was not evident whether twice the hydrostatic pressure of the lake would suffice to do so. Bill made up and used a slide rule to approximately solve (overnight) the model presented in this chapter. It was intuitively clear to him that it was necessary to add a reference distance (in this case one tunnel diameter) in order to make the problem well-posed. His prediction was an impressive achievement, in particular given that he had only one chance to do the calculations. Bill brought this problem to the Department of Mathematics and Statistics at the University of Victoria in 1989, requesting that a computer simulation be done.

This page intentionally left blank

Chapter 3

How Much Will that Annuity Cost Me?

Concepts and Tools: Probability, calculus

The mathematics of finance includes such concepts as compounding interest, mortgages, loans, and annuities. In order to establish some basic principles, we first revisit the compounding of interest at various rates. These basic principles are then used to develop the more complex processes arising in mortgages, loans, and annuities. In the exploration of these topics, discrete and continuous processes are examined and derived.

3.1. Interest Basics

We will begin by establishing the basic principles regarding interest. Let P_o denote the original investment and r be the nominal interest rate per year. If $P(t)$ denotes the accumulated amount after t years, then investing P_o for one year at interest rate r yields

$$P(1) = P_o + rP_o = P_o(1 + r). \tag{3.1}$$

After two years this becomes

$$P(2) = P(1) + rP(1) = P_o(1+r) + rP_o(1+r) = P_o(1+r)^2.$$

Therefore, $P(t)$, the amount after a duration of t years, is given by

$$P(t) = P_o(1+r)^t, \tag{3.2}$$

an exponential function of t. Here interest was compounded on a yearly basis with each subsequent year compounding interest on both the original amount P_o, and the interest earned to that point. Because interest is earned not only on P_o but also on the interest, it is more profitable to compound interest on a more frequent basis. For instance, compounding could take place on a semiannual, monthly, daily, or even on a continuous basis.

3.1.1. Semiannual, Monthly, and Daily Compounding. In order to compound on any basis other than yearly, we need to modify (3.2) slightly. Because the interest is set at a rate which is to be applied annually, if we wish to compound on a more frequent basis we must correct for this. To do so, we divide the annual rate, r, by the number of conversion periods per year. For example, if we compound semiannually, equation (3.2) becomes

$$P(t) = P_o(1+i)^n \tag{3.3}$$

where $i = r/2$ for two conversion periods and $n = 2t$, since t is in units of years. For a one-year period this gives

$$P(1) = P_o\left(1 + \frac{r}{2}\right)^2 = P_o\left(1 + r + \frac{r^2}{4}\right).$$

Comparing this to (3.1), it is clear that compounding semiannually is more profitable than compounding annually.

A similar adjustment is made to equation (3.2) when compounding monthly. Proceeding according to equation (3.3), we divide r by 12 (the number of months in a year) so that $i = r/12$ and set $n = 12t$. This yields

$$P(t) = P_o\left(1 + \frac{r}{12}\right)^{12t},$$

and, likewise, for compounding daily over t years we have

$$P(t) = P_o\left(1 + \frac{r}{365}\right)^{365t}.$$

Table 1. Return on $P_o = \$1000$ if compounding m times a year.

Rate	Conversion periods(m)	$P_m(10)$	% of continuous return
Annually	1	\$1628.89	98.800%
Monthly	12	\$1647.01	99.900%
Daily	365	\$1648.66	99.996%
Continuously	–	\$1648.72	100.000%

Using this basic approach, we arrive at a general formula for compounding interest,

$$P_m(t) = P_o\left(1 + \frac{r}{m}\right)^{mt}, \tag{3.4}$$

where t is the number of years for which P_o is invested and m denotes the number of conversion periods during that time.

3.1.2. Continuous Compounding. As stated in Section 3.1, the more frequently the interest is compounded, the greater the return. Therefore, it is desirable to be able to compound on a continuous basis in order to maximize the interest earned. To do so, we examine the limit of equation (3.4) as $m \to \infty$ which gives

$$P_c(t) = \lim_{m\to\infty} P_m(t) = \lim_{m\to\infty} P_o\left(1 + \frac{r}{m}\right)^{mt} = P_o e^{rt}, \tag{3.5}$$

which is the well-known formula for compounding continuously.

Having developed equations for different rates of compounding, a numerical example is presented to illustrate the difference in return.

Example 3.1. If we let $P_o = 1000$ and use a 5% interest rate ($r = 0.05$), then Table 1 shows the result of compounding annually, monthly, daily, and continuously for a period of ten years. Clearly, the more often the interest is compounded, the greater the return.

3.2. Mortgages

Now that we have established this background, we are able to examine mortgages and determine how fixed periodic payments are derived. As before, P_o will denote the initial principal and r the nominal interest rate per year. As well, denote the amortization time (the length

of time over which the mortgage will be paid) by T, and the fixed periodic payments on the mortgage by x. These payments can be made on any basis: annual, monthly, biweekly, etc. Computing x such that the mortgage will be paid off after T years (the case where $P(T) = 0$) is the objective of this section.

To begin we consider the case where the payments are made on a yearly basis. If we let $P(i)$ be the remaining debt after i years, and $P(0) = P_o$, then

$$P(i+1) = P(i)(1+r) - x, \qquad 0, 1, 2, \ldots .$$

Therefore, letting $i = 0$ gives

$$P(1) = P(0)(1+r) - x = P_o(1+r) - x, \tag{3.6}$$

and for $i = 1$ we have

$$P(2) = P(1)(1+r) - x = P_o(1+r)^2 - x(1+r) - x.$$

This generalizes recursively to a formula for $P(T)$ the accumulated debt after T years given in Theorem 3.2.

Theorem 3.2. *For any* $T \in \mathbb{N}$

$$P(T) = P_o(1+r)^T - x\left[\sum_{i=0}^{T-1}(1+r)^i\right]. \tag{3.7}$$

Proof. We use the principle of mathematical induction. Referring back to (3.6), we have established that equation (3.7) holds for $T = 1$. Assume that it also holds for $T = n$. Let $T = n + 1$, then

$$P(n+1) = P(n)(1+r) - x,$$

which, after using expression (3.7) for $P(n)$, yields

$$P(n+1) = \left\{P_o(1+r)^n - x\left[\sum_{i=0}^{n-1}(1+r)^i\right]\right\}(1+r) - x.$$

Simplifying, this becomes

$$P(n+1) = P_o\,(1+r)^{n+1} - x\left[\sum_{i=0}^{n}(1+r)^i\right],$$

which is in the required form. Therefore, by the principle of mathematical induction, equation (3.7) is true for all $T \in \mathbb{N}$. □

3.2.1. The Mortgage Payments. Using equation (3.7), we are able to determine x, the fixed periodic mortgage payments. Recall that for a finite geometric sum

$$\sum_{i=0}^{T-1} q^i = \begin{cases} \dfrac{q^T - 1}{q - 1} & \text{for } q \neq 1, \\ T & \text{for } q = 1, \end{cases}$$

and if we set $q = 1 + r$, we have

$$\sum_{i=0}^{T-1} (1+r)^i = \frac{(1+r)^T - 1}{(1+r) - 1} = \frac{(1+r)^T - 1}{r},$$

where $r \neq 0$. Therefore, (3.7) simplifies further to

$$P(T) = P_o(1+r)^T - \left[(1+r)^T - 1\right] \frac{x}{r}. \tag{3.8}$$

To determine how large the fixed periodic payments need to be in order to pay off the mortgage in T years, we set $P(T) = 0$, and solve for x in (3.8). This yields

$$x_1 = \frac{P_o r(1+r)^T}{(1+r)^T - 1} \tag{3.9}$$

as the annual mortgage payment.

For payments that are made m times a year we have

$$x_m = \frac{P_o \frac{r}{m} \left(1 + \frac{r}{m}\right)^{mT}}{\left(1 + \frac{r}{m}\right)^{mT} - 1}. \tag{3.10}$$

Payments made on a monthly basis are calculated in with $m = 12$. The following example assumes monthly payments.

Example 3.3. A Canadian real estate agency distributes a flyer quoting monthly payments of \$610.39 for twenty-five years on a property valued at \$100,000. The nominal interest rate is 5.5%. However, calculating the monthly payments using (3.10), we have

$$x_{12} = \frac{100000 \left(\frac{0.055}{12}\right) \left(1 + \frac{0.055}{12}\right)^{12 \cdot 25}}{\left(1 + \frac{0.055}{12}\right)^{12 \cdot 25} - 1} = 614.09,$$

which is slightly in excess of the advertised payments. The reason for this is a small but subtle point which we have so far ignored.

Table 2. Nominal versus effective interest rates for daily compounding.

r	r_{eff}	% change
0.050	0.0513	2.53%
0.075	0.0779	3.83%
0.100	0.1052	5.16%
0.125	0.1331	6.50%
0.150	0.1618	7.87%

3.2.2. Nominal and Effective Interest Rates. The discrepancy which occurs in Example 3.3 relates to the use of nominal versus effective interest rates. The effective yearly rate, computed only on the original principal, is that rate which produces the same interest as the nominal rate compounded m times per year. We denote the effective interest rate as r_{eff}, and observe that our definition is equivalent to the relationship

$$\left(1+\frac{r}{m}\right)^m = 1 + r_{eff},$$

where m is the number of conversion periods per year. As Table 2 illustrates for daily compounding, the effective rate is greater than the nominal rate; therefore, frequent payments (with formulas using x derived from the nominal rate) would increase the effective interest rate. To remove this variability in the value of r_{eff} used to compute mortgage payments, by Canadian law, all mortgages are *compounded semi-annually and not in advance* (CSNA). Under these guidelines, for monthly payments, the relationship between the nominal and effective interest rates becomes

$$\text{CSNA}: \quad \left(1+\frac{r}{2}\right)^2 = \left(1+\frac{r_{eff}}{12}\right)^{12}.$$

Note that here the published rate is the one used for semi-annual compounding, and the effective rate is calculated depending on the frequency of payments (monthly in the example). Substituting in the value $r = 0.055$, this gives $r_{eff} \approx 0.05438018 < r$. Now returning to Example 3.3, we can re-evaluate the monthly mortgage payment using this CSNA effective interest rate. This yields $x_{12} = \$610.391527$ which is the quoted rate of \$610.39.

3.2.3. Continuous Compounding of Mortgages. To compound mortgages on a continuous basis we use the relationship established by (3.5) in Section 3.1.2:

$$P(t) = P_o\, e^{\rho t}, \tag{3.11}$$

where ρ denotes the interest rate corresponding to an annual rate r_{eff} by

$$e^{\rho} = 1 + r_{eff}.$$

Assume that the mortgage is paid off continuously at a rate $x > 0$ where $P(t)$ is the principal. Then for a small $\Delta t > 0$,

$$P(t + \Delta t) \approx P(t) + [\rho P(t) - x]\, \Delta t \tag{3.12}$$

or

$$\frac{P(t + \Delta t) - P(t)}{\Delta t} \approx \rho P(t) - x.$$

Taking the limit as $\Delta t \to 0$ and denoting the derivative as $P'(t)$, we have the linear ordinary differential equation

$$P'(t) - \rho P(t) = -x, \qquad P(0) = P_o.$$

In order to solve for $P(t)$, we multiply through by an integrating factor $e^{-\rho t}$ that yields

$$[P'(t) - \rho P(t)]\, e^{-\rho t} = \frac{d}{dt}\left[P(t)e^{-\rho t}\right] = -xe^{-\rho t}.$$

Integrating both sides from 0 to T gives

$$\int_0^T \frac{d}{dt}\left[P(t)e^{-\rho t}\right] dt = -x \int_0^T e^{-\rho t}\, dt$$

and, upon evaluating both integrals, this becomes

$$P(T)e^{-\rho T} - P_o = \frac{x}{\rho}\left(e^{-\rho T} - 1\right).$$

Now solving for $P(T)$ we have

$$P(T) = P_o e^{\rho T} - \frac{x}{\rho}\left(e^{\rho T} - 1\right),$$

which is the remaining debt at time T. To determine the payments, x, which will pay off the mortgage in T years, we set $P(T) = 0$ and

solve for x. Therefore, the formula for continuous payments is given by

$$x_c = \frac{P_o \rho \, e^{\rho T}}{e^{\rho T} - 1}, \tag{3.13}$$

which is also the continuous limit of the discrete formula (3.10). See Exercise 1.

We now have enough information to determine the ρ that would be used when payments are made on a continuous basis. Returning to the situation presented in Example 3.3, ρ is found by using the relation

$$e^{\rho} = \left(1 + \frac{r}{2}\right)^2,$$

which follows the CSNA guideline of compounding semi-annually. As $r = 0.055$ in this example, one has $\rho \approx 0.05425$. Comparing this result to the one obtained for the effective monthly rate at the end of Section 3.2.2 ($r_{eff} = 0.05438018$), it is seen that the effective rate for continuous compounding is less than that reported if compounding is on a monthly basis.

3.2.4. Variable Interest Rates. The preceding section assumed that ρ was a constant rate; however, in many cases ρ will actually vary as a function of time. In order to retain the set amortization time, the rate at which payments are made must also be time dependent; therefore, $x = x(t)$. If this adjustment is not made, there exists the possibility that the debt will not be paid off within the amortization time.

This adjusted rate with variable interest rate $\rho(t)$ is given by equation (3.13) as

$$x(t) = P(t)\frac{\rho(t)e^{\rho(t)(T-t)}}{e^{\rho(t)(T-t)} - 1}, \qquad 0 \le t < T, \tag{3.14}$$

where $x(t)$ is the instantaneous rate of payment that would satisfy the debt if we assume that the interest rate does not change after time t. We now re-establish the relationship presented in (3.13) using a variable interest rate $\rho(t)$. Substituting (3.14) for x and moving it to the left-hand side we have

$$P'(t) - \rho(t)P(t) + P(t)\frac{\rho(t)}{1 - e^{-\rho(t)(T-t)}} = 0.$$

Grouping like terms and rearranging yields

$$P'(t) + \rho(t)\frac{e^{-\rho(t)(T-t)}}{1 - e^{-\rho(t)(T-t)}}P(t) = 0, \tag{3.15}$$

which can be expressed as

$$P'(t) + \alpha(t)P(t) = 0, \tag{3.16}$$

where α is the appropriate substitution.

To solve for $P(t)$, we multiply through by the appropriate integrating factor, $e^{\int_0^t \alpha(\tau)\,d\tau}$, and integrate. This yields

$$\int_0^t \frac{d}{dt}\left[P(t)e^{\int_0^t \alpha(\tau)\,d\tau}\right] dt = 0.$$

Evaluating the integral from 0 to t gives

$$P(t)e^{\int_0^t \alpha(\tau)\,d\tau} = P_o,$$

which, upon solving for $P(t)$ and replacing α with the appropriate quantity from expression (3.15), becomes

$$P(t) = P_o \exp\left[-\int_0^t \frac{\rho(\tau)e^{-\rho(\tau)(T-\tau)}}{1 - e^{-\rho(\tau)(T-\tau)}}d\tau\right].$$

By construction $P(T) = 0$, which means that the mortgage has been repaid.

3.3. Loan Repayment

Determining when a loan will be repaid follows a derivation similar to that of mortgages in the previous section. In this case, we will assume a constant rate of payment, x, and a changing interest rate, $\rho(t)$, so that we begin with the expression

$$P'(t) - \rho(t)P(t) = -x,$$

which becomes

$$P(t)e^{-\int_0^t \rho(\tau)\,d\tau} = P_o - x\int_0^t e^{-\int_0^\tau \rho(\sigma)\,d\sigma}d\tau,$$

if we use the appropriate integrating factor and integrate from 0 to t as in previous sections. The loan will be repaid when $P(T) = 0$ for

some T; therefore, we have

$$P_o = x \int_0^T e^{-\int_0^\tau \rho(\sigma)\, d\sigma} d\tau,$$

a transcendental equation for T that can be solved numerically if ρ is known. The easiest case, as before, is when ρ is constant.

3.4. Present Value

The present value means how much a payment expected t years in the future is worth now. For example, for monthly compounding, the present value, PV, is found by solving equation (3.4) for P_o which gives

$$\text{PV} = P_o = \frac{P_{12}(t)}{(1+i)^{12t}},$$

where $i = r/12$. Similarly, to find the present value of an investment which is compounded continuously with a constant interest rate, equation (3.11) gives

$$\text{PV} = P(t)e^{-\rho t}.$$

For a variable interest rate where $P'(t) - \rho(t)P(t) = 0$, we solve for $P(t)$ as in Section 3.2.4. This yields

$$\text{PV} = P(t)e^{-\int_0^t \rho(\tau)\, d\tau}.$$

When computing mortgages, the present value of all the future payments should equal the principal, P_o. See Exercise 4.

Example 3.4. Assuming continuous compounding with a constant interest rate of $\rho = 0.05$, the present value of \$10,000 expected in ten years will be

$$\text{PV} = 10,000e^{-(0.05)(10)} \approx \$6065.31.$$

3.5. Annuities

Mortgage and loan repayments are both forms of annuities; they are sequences of payments that are made on a structured basis. In this section we will examine another type of annuity where a fixed amount of money is invested (with an insurance company) in return for a constant cash flow at a rate x until death. These *income annuities*

differ from mortgages and loans in the respect that life expectancy is a random variable (not everyone lives to the same age).

In order to derive a formula for these payments, we further develop the concept of present value. As stated, income annuities are sold for a fixed amount, K_o, such that $K_o \geq \text{PV}$, the present value. In this case,

$$K_o \geq \text{PV} = \int_0^T x e^{-\rho t}\, dt = \frac{x}{\rho}\left(1 - e^{-\rho T}\right)$$

where T is the time of death and ρ is the interest rate here compounded on a continuous basis.

Example 3.5. An insurance policy was purchased fifteen years ago. In return for the original investment, the insurance company issued continuous payments of \$1500 a month (i.e., \$18,000 per year). This was compounded continuously at a 5% nominal interest rate. Therefore, the original investment was

$$\text{PV} = \frac{18,000}{0.05}\left(1 - e^{-0.05 \cdot 15}\right) = \$189,948.04.$$

Two problems arise with such a representation. The first relates to the fact that, in reality, the interest rate is not generally held constant. The second is that T itself is a random variable since, as previously mentioned, life expectancy is a random variable. The latter problem is addressed in the following section by the use of hazard rate functions. To do so we again need some probability theory.

3.6. Hazard Rate Functions

Hazard rate functions address the problem of variable life spans for certain types of annuities. The hazard rate function attempts to assess the expected lifetime of an individual. If we let t measure the age of a randomly chosen individual (including deceased individuals), then the hazard rate function, $h(t)$, is given by

$$h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \text{Prob}\left\{A \,\middle|\, B\right\}, \tag{3.17}$$

where A is the event that the individual will die within the time period $(t, t + \Delta t)$, B is the event that the individual is alive at time t, and

Δt represents a time increment. The probability that an individual is alive at time t is called the survivor function, denoted as $S(t)$, such that

$$S(t+\Delta t) = \text{Prob}\{\text{individual is alive at } t+\Delta t\},$$

which is equivalent to

$$S(t+\Delta t) = \text{Prob}\left\{\overline{A} \text{ and } B\right\}, \tag{3.18}$$

where we denote by $\overline{A}$ the event that the individual is alive at $t+\Delta t$. Consequently, by the definition of conditional probabilities, equation (3.18) becomes

$$S(t+\Delta t) = \text{Prob}\left\{\overline{A} \mid B\right\} \cdot \text{Prob}\{B\}. \tag{3.19}$$

3.6.1. Linking Survivor and Hazard Rate Functions. The survivor function is linked with the hazard function by the relationship

$$S(t+\Delta t) - S(t) = \text{Prob}\left\{\overline{A} \mid B\right\} \cdot \text{Prob}\{B\} - \text{Prob}\{B\},$$

where $S(t+\Delta t)$ is as defined by equation (3.19). Grouping like terms and identifying $\text{Prob}\{B\}$ with $S(t)$ yields

$$S(t+\Delta t) - S(t) = \left(\text{Prob}\left\{\overline{A} \mid B\right\} - 1\right) S(t) = -\left(\text{Prob}\left\{A \mid B\right\}\right) S(t),$$

which, upon substituting $\text{Prob}\left\{A \mid B\right\}$ from (3.17), becomes

$$S(t+\Delta t) - S(t) = (-h(t)\Delta t)S(t) + o(\Delta t).$$

Rearranging, we have

$$\frac{S(t+\Delta t) - S(t)}{S(t)\Delta t} = -h(t) + \frac{o(\Delta t)}{\Delta t}.$$

By taking the limit as $\Delta t \to 0$, we find that

$$\frac{S'(t)}{S(t)} = -h(t). \tag{3.20}$$

Now integrating from 0 to t in order to solve for $S(t)$ yields

$$\ln S(t) - \ln S(0) = -\int_0^t h(\tau)\, d\tau. \tag{3.21}$$

Clearly, $S(0) = 1$; the probability of being alive at birth is 1. Therefore, solving equation (3.21) for $S(t)$ gives

$$S(t) = \exp\left[-\int_0^t h(\tau)\, d\tau\right]. \tag{3.22}$$

This now enables us to determine the cumulative distribution function and the density distribution function of life expectancy. Since $S(t)$ is the probability that the individual is alive at t, this implies that

$$1 - S(t) = \text{Prob}\{\text{individual dies before } t\}.$$

If T is the life expectancy of an individual, then the cumulative distribution function of T, denoted as $F(t)$, is given as

$$F(t) = \text{Prob}\{T \leq t\} = 1 - S(t). \tag{3.23}$$

Consequently, the density distribution function of T, $f(t)$, is

$$f(t) = F'(t) = -S'(t), \tag{3.24}$$

which, upon replacing $S'(t)$ as given by (3.20), becomes

$$f(t) = h(t)S(t). \tag{3.25}$$

3.6.2. Examples of Hazard Rate Functions. The following are examples of different hazard rate functions with the survivor function, cumulative distribution function, and density distribution function as defined by equations (3.22), (3.23), and (3.25), respectively.

Example 3.6. The constant hazard rate function is $h(t) := \lambda$ where λ, is a constant. The survivor function in this case is

$$S(t) = \exp\left[-\int_0^t h(\tau)\, d\tau\right] = e^{-\lambda t}.$$

Therefore, this implies that the cumulative distribution function $F(t)$ is

$$F(t) = 1 - S(t) = 1 - e^{-\lambda t}$$

and that the density distribution function is given by

$$f(t) = F'(t) = h(t)S(t) = \lambda e^{-\lambda t}. \tag{3.26}$$

A constant hazard rate function is an unrealistic choice, and thus leads to unrealistic predictions; however, it is useful for the purpose of analysis. The following functions are more commonly used.

Example 3.7. The Weibull hazard rate function is defined as $h(t) := \alpha\beta t^{\beta-1}$ for $\alpha > 0$ and $\beta > 0$. The h is unbounded when $\beta > 1$. Using

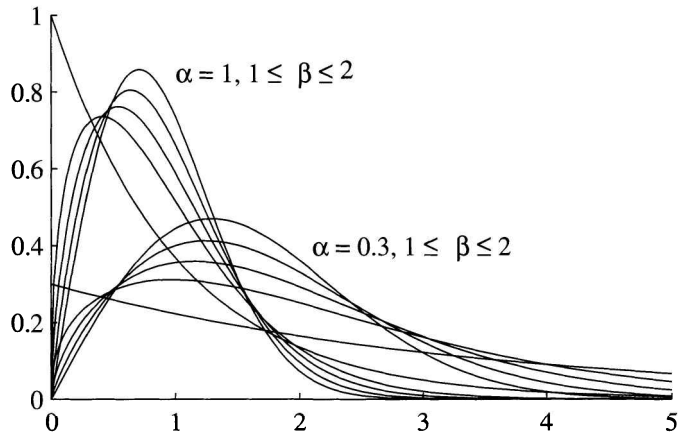


Figure 1. The Weibull distribution $f(t) = \alpha\beta t^{\beta-1}e^{-\alpha t^{\beta}}$ for $\alpha = 0.3$ and $\alpha = 1$ with $1 \leq \beta \leq 2$.

this distribution function $S(t) = e^{-\alpha t^{\beta}}$ and $F(t) = 1 - e^{-\alpha t^{\beta}}$. This implies that the density function is

$$f(t) = \alpha\beta t^{\beta-1}e^{-\alpha t^{\beta}}.$$

Figure 1 displays the density distribution function of life expectancy for the Weibull hazard function where $\beta = 5$ and $\alpha = 1$.

Example 3.8. The linear hazard rate function is defined to be $h(t) := \alpha + \beta t$, where $\beta > 0$. Therefore,

$$S(t) = e^{-t(\alpha+\beta t/2)},$$

implying that $F(t) = 1 - e^{-t(\alpha+\beta t/2)}$ and $f(t) = (\alpha + \beta t)\, e^{-t(\alpha+\beta t/2)}$.

Example 3.9. The Gompertz hazard rate function is defined to be $h(t) := e^{\alpha}e^{\beta t}$ with parameters $\alpha, \beta \in \mathbb{R}$ (typically, $\beta > 0$). Hence,

$$S(t) = \exp\left[\frac{e^{\alpha}}{\beta}\left(1 - e^{\beta t}\right)\right],$$

and $F(t)$ and $f(t)$ can be determined as in the previous examples. Notice that $\log h(t) = \alpha + \beta t$, i.e., $\log h(t)$ is an affine linear function.

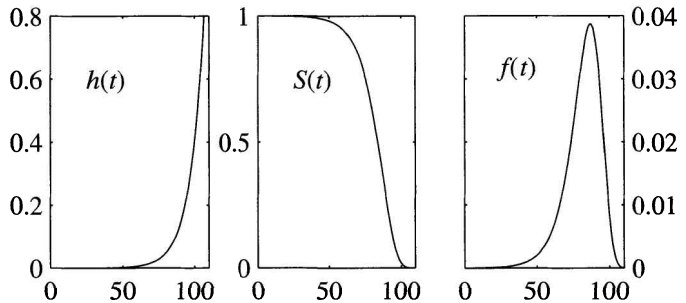


Figure 2. Displayed are the hazard rate function $h(t)$, survivor function $S(t)$, and probability density $f(t)$ for the Gompertz hazard with $\alpha = -11.43$ and $\beta = 0.1053$.

Figure 2 displays the Gompertz hazard rate function for $\alpha = -11.43$ and $\beta = 0.1053$. Also displayed are the corresponding survivor function and probability density.

Example 3.10. In Figure 3 we display statistically observed log hazard functions for humans, separated into male and female data sets.

The data employed for Figure 3 are taken from the 1982–1988 Mortality Table published by the Canadian Institute of Actuaries, and may be found in [**W**]. Here Q_x denotes the (observed) probabilities that an individual of age x will die before age $x+1$. The straight line portion of the curve beyond age 35 was used to estimate values for α and β in Example 3.9.

Three things are prominently visible:

(1) $\log h$ is well approximated by an affine linear function for ages above 35.

(2) Except for very young children, hazard rates for females are *always* lower than males.

(3) Hazard rates drop fairly steeply from birth to approximately the age of 6 (presumably explained by the risk of infant death and early childhood diseases or accidents), then level off until there is a prominent bulge, much more pronounced

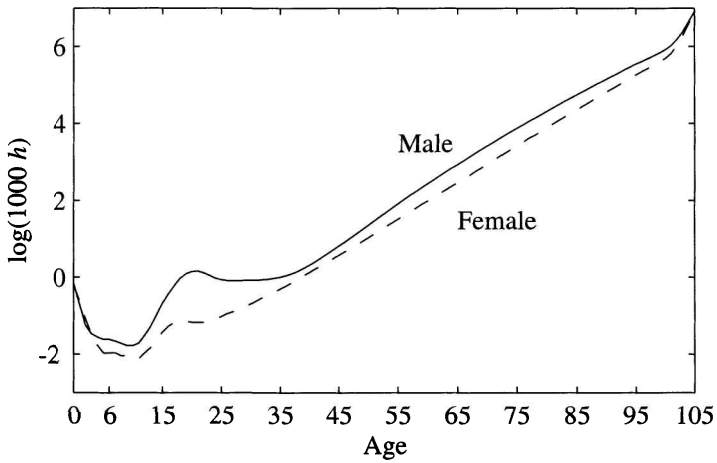


Figure 3. Statistically observed log hazard function for males and females for the years 1982–1988.

for males, between the ages of 15 and 25. We leave it to the reader's imagination to explain this bulge.

3.7. Expected Lifetime

We are now able to deal with the problem arising in life insurance policies with respect to variable life spans. Using the hazard rate function and the relevant functions derived from it, we first compute the expected lifetime of an individual. This is given as

$$E(T) = \int_0^\infty t f(t)\, dt,$$

where $E(T)$ denotes the expected lifetime, and T, the random variable *life expectancy*, has a density distribution $f(t)$. Using equation (3.24), the above becomes

$$E(T) = -\int_0^\infty tS'(t)\, dt,$$

which is integrated by parts to yield

$$E(T) = [-tS(t)]|_0^\infty + \int_0^\infty S(t)\, dt.$$

Assume that $S(t)$ is such that

$$\lim_{t\to\infty} tS(t) = 0$$

(i.e., that no one lives forever); the expected lifetime becomes

$$E(T) = \int_0^\infty S(t)\,dt.$$

Our theory still includes the possibility that the expected lifetime of an individual is unbounded; an individual could live forever. This will occur if the hazard function is such that $\int_0^\infty h(\tau)\,d\tau < \infty$. In this case, $\lim_{t\to\infty} S(t) > 0$ and necessarily $E(T) = \infty$. The previous calculations do not apply to this case as they depend on some assumptions which are violated here (mainly the assumption that everyone eventually dies). Rather, one has to use (3.23) and proceed from there. Of course for realistic choices (h increasing) these problems do not arise. See Exercise 6.

3.8. An Annuity Problem

At this point, we return to examine annuities further. If we denote the present value of the annuity by Y, then from Section 3.5,

$$Y = \frac{x}{\rho}\left(1 - e^{-\rho T}\right), \tag{3.27}$$

where x/ρ is the present value of an annuity paid in perpetuity. (Clearly $Y \le x/\rho$.) Examining the probability that the present value is less than or equal to a value y, we have

$$\text{Prob}\{Y \le y\} = \text{Prob}\left\{\frac{x}{\rho}\left(1 - e^{-\rho T}\right) \le y\right\},$$

or equivalently

$$\text{Prob}\{Y \le y\} = \text{Prob}\left\{T \le -\frac{1}{\rho}\ln\left(1 - \frac{\rho y}{x}\right)\right\}.$$

The cumulative distribution function, $F_Y(y)$, is therefore given by

$$F_Y(y) = \text{Prob}\{Y \le y\} = \begin{cases} 1 - S\left(-\frac{1}{\rho}\ln\left(1 - \frac{\rho y}{x}\right)\right) & \text{for } \rho y \le x, \\ 1 & \text{for } \rho y > x, \end{cases}$$

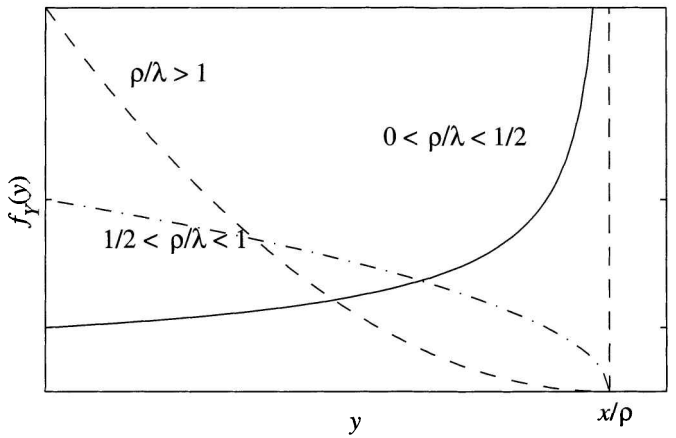


Figure 4. The probability distribution of the present value of the annuity given by (3.29) is illustrated above. There are three possible shapes of this distribution. The particular shape is determined by the ratio of the interest rate to the hazard rate, ρ/λ. For this illustration, x and ρ were set to one and λ was varied.

and the density distribution function, $f_Y(y)$, is

$$f_Y(y) = \begin{cases} -\frac{1}{x-\rho y} S'\left(-\frac{1}{\rho}\ln\left(1-\frac{\rho y}{x}\right)\right) & \text{for } \rho y \le x, \\ 0 & \text{for } \rho y > x. \end{cases}$$

Recalling from equation (3.24) that $f(t) = -S'(t)$, this implies that

$$f_Y(y) = \begin{cases} \frac{1}{x-\rho y} f\left(-\frac{1}{\rho}\ln\left(1-\frac{\rho y}{x}\right)\right) & \text{for } \rho y \le x, \\ 0 & \text{for } \rho y > x. \end{cases} \tag{3.28}$$

If we focus on a constant hazard $h(t) = \lambda > 0$, as in Example 3.6, and substitute the appropriate density distribution function $f(y) = \lambda e^{-\lambda y}$, then, after simplification, equation (3.28) becomes

$$f_Y(y) = \begin{cases} \frac{\lambda}{x}\left(1-\frac{\rho y}{x}\right)^{\frac{\lambda}{\rho}-1} & \text{for } \rho y \le x, \\ 0 & \text{for } \rho y > x. \end{cases} \tag{3.29}$$

Figure 4 displays this density distribution function $f_Y(y)$ for various scenarios involving the interest rate ρ.

Using this relation, the expected value of the present value of the annuity, $E(Y)$, is given as

$$E(Y) = \frac{\lambda}{x} \int_0^{\frac{x}{\rho}} y \left(1 - \frac{\rho y}{x}\right)^{\frac{\lambda}{\rho}-1} dy, \tag{3.30}$$

which, by substituting $z = \rho y/x$ and integrating by parts, yields

$$E(Y) = \frac{x}{\lambda + \rho}. \tag{3.31}$$

It should not come as a surprise that the expected value of the annuity is maximized when the probability that the individual eventually dies approaches zero ($h(t) \equiv 0$). In this case $E(Y) = x/\rho$, which is the present value of an annuity paid in perpetuity. In the general case, the hazard rate can just be added to the interest rate. In Section 3.9.1 below we reproduce (3.31) with an alternative method.

3.9. $V(Y)$: How the Expected Value of the Annuity Varies

Knowledge of the variance of the expected lifetime is important to assess the risk an insurance company accepts in selling annuities. From probability theory,

$$V(Y) = E\left([Y - E(Y)]^2\right) = E(Y^2) - [E(Y)]^2,$$

which is nonnegative by construction. Using the expected value function for the constant hazard rate defined by relations (3.30) and (3.31), the variance is given by

$$V(Y) = \int_0^{\frac{x}{\rho}} y^2 f_Y(y)\, dy - \left(\frac{x}{\lambda + \rho}\right)^2,$$

which, upon substituting the appropriate value for $f_Y(y)$ from (3.29) and integrating by parts, yields a variance of

$$V(Y) = \frac{\lambda x^2}{(\lambda + \rho)^2 (\lambda + 2\rho)}. \tag{3.32}$$

3.9.1. Alternative Methods of Computing $E(Y)$ and $V(Y)$. It is possible to compute the expected lifetime, $E(Y)$, and the variance, $V(Y)$, in another way. If we denote the present value, Y, as $g(T)$, then from (3.27)

$$g(T) = Y = \frac{x}{\rho}\left(1 - e^{-\rho T}\right),$$

and the expected value can be written as

$$E\left(g(T)\right) = \int_0^\infty g(\tau) f(\tau)\, d\tau = \frac{x}{\rho}\int_0^\infty \left(1 - e^{-\rho\tau}\right) f(\tau)\, d\tau.$$

Using the density distribution function for a constant hazard function from (3.26), this becomes

$$E\left(g(T)\right) = \frac{x}{\rho}\int_0^\infty \left(1 - e^{-\rho\tau}\right) \lambda e^{-\lambda\tau}\, d\tau,$$

which, upon evaluation, yields

$$E\left(g(T)\right) = \frac{x}{\lambda + \rho}.$$

This result for the expected lifetime is consistent with the result in Section 3.8. The variance can be found in a similar manner.

Exercises

(1) A loan (or mortgage) with principle P_o and interest rate r is repaid with monthly payments of size x.
   (a) Find the time T needed to pay off the loan as a function of P_o, r, and x. Assume the interest is compounded monthly.
   (b) Repeat part (a) but assume that interest is compounded continuously and that payments are made continuously at rate x.

(2) Mr. Smith takes out a mortgage for \$120,000. The agreed amortization time is twenty-five years and the negotiated initial interest rate is 7.5%, fixed for five years. During this time, Mr. Smith makes monthly payments based on this rate. Assume the interest is compounded monthly.

At the end of the five years, the bank offers a reduced interest rate of 6%, fixed for another five years. Mr. Smith accepts these conditions and chooses to make payments on a weekly basis. How large are these payments?

Remark 3.11. Do your calculations using 7.5% and 6% as yearly nominal interest rates (i.e., ignore the fact that Canadian mortgages are CSNA).

(3) Verify that the formula (3.13)

$$x_c = \frac{P_o \rho \, e^{\rho T}}{e^{\rho T} - 1}$$

arises as a limit from the formula (3.10)

$$x_m = \frac{P_o \frac{r}{m} \left(1 + \frac{r}{m}\right)^{mT}}{\left(1 + \frac{r}{m}\right)^{mT} - 1}.$$

Explain the limit procedure in detail.

(4) Suppose that an interest rate is constant at 4% this year, 6% next year, and 5% for the following year. Assuming continuous compounding, what is the present value of \$10,000? (You expect to have \$10,000 at the end of the three years.)

(5) For a mortgage of size P_o, it was demonstrated that in order to pay off the mortgage after T years, continuous payments at the rate

$$x_c = \frac{P_o \rho e^{\rho T}}{e^{\rho T} - 1}$$

are necessary.

(a) Show that the present value of these payments is P_o when the interest rate ρ is assumed fixed.

(b) Repeat part (a), but assume annual compounding at a constant interest rate $r > 0$ with annual payments

$$x_1 = \frac{P_o r \, (1 + r)^T}{(1 + r)^T - 1}.$$

(6) If $s := \lim_{t \to \infty} S(t) > 0$, some individuals will never die. What modifications to the formula relating h, S, f, and T are required in this scenario?

(7) Assume a Gompertz hazard $h(t) = e^{\alpha+\beta t}$, and a constant interest rate $\delta > 0$.
   (a) Find an integral expression for the expected value of an annuity of $\$x$ per year paid out continuously. Simplify as much as you can, but do not evaluate (it is not possible to do this in explicit terms).
   Hint: You do not need to compute the distribution density of Y.
   (b) Find an expression involving integrals for the variance of Y as in part (a).

(8) Derive the formula (3.32) of the variance of an annuity for a constant hazard λ by using the alternative method from Section 3.9.1.

(9) Consider a hazard rate function

$$h(t) = \begin{cases} \lambda > 0 & \text{for } 0 \le t \le 50, \\ \lambda + \alpha(t-50) & \text{for } t > 50 \end{cases}$$

(constant hazard to age 50, linearly increasing hazard after 50). Find the survivor function $S(t)$ from $S'/S = -h$ (distinguish $t \le 50$ and $t > 50$).
Hint: First solve $S'/S = -h$ on $t \in [0, 50]$, where $S(0) = 1$. Compute $S(50)$ and use this as the initial value for $S'/S = -h$ on $t \in (50, \infty)$.

(10) CLASS PROJECT
A very unrealistic assumption in our analysis is that the interest rate ρ remains constant. Suppose the value of ρ is allowed to change at fixed time intervals $n\Delta t$, $n = 0, 1, 2, \ldots$, such that at each time $n\Delta t$, $\rho_{\text{new}} = \rho_{\text{old}} + \Delta\rho$, where $\Delta\rho$ is a random fluctuation. Write an essay in which you discuss the necessary adjustments to the analysis of this chapter.

Chapter 4

Dimensional Analysis

Concepts and Tools: Linear algebra, basis, Gaussian elimination

When the United States exploded the nuclear bomb "Trinity" at the Los Alamos test site in 1945, both the data and the motion picture footage from the explosion were classified. Two years later, although the data was still classified, the government released a movie of the explosion. Sir G.I. Taylor (a British physicist, 1886–1975) managed to determine the energy released during the explosion using only the radius of the expanding blast wave at time t and the principles of dimensional analysis. When the data were later unclassified, his calculations proved quite accurate. Among other examples, we will retrace Taylor's calculations in this chapter and develop the principles of dimensional analysis.

4.1. A Classical Example: The Pendulum

Dimensional analysis provides a powerful method for analysing physical events and relationships without requiring much knowledge of the underlying physics or ordinary differential equations. For example, consider the motion of a pendulum as depicted in Figure 1. Here, m denotes the mass, l the length of the pendulum, $x(t)$ the arclength, and θ the angle of deflection from the vertical (measured in radians). One standard method to determine the period of this pendulum is to set up and solve the differential equation that models the dynamics.

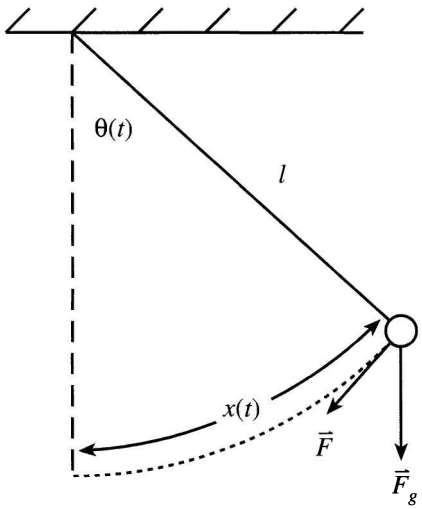


Figure 1. The above figure models the motion of a pendulum where m denotes the mass and l denotes the length of the pendulum. The radial position is $x(t)$, and θ is the angle the pendulum makes with the vertical (measured in radians). $\vec{F}_g$ is the force due to gravity and $\vec{F}$ is the force parallel to the direction of the motion.

This is given by $m\ddot{x}(t) = -mg\sin\theta(t)$, where $x = l\theta$ and $\ddot{x}$ denotes its second derivative with respect to time. (Figure 2 clarifies how this relationship is derived.) Replacing $x(t)$ as defined and simplifying this equation we have

$$l\ddot{\theta}(t) = -g\sin\theta(t), \tag{4.1}$$

where $\ddot{\theta}$ denotes the second derivative of θ and g is the acceleration due to gravity. When θ is small, $\sin\theta \approx \theta$ and formally equation (4.1) simplifies to

$$\ddot{\theta}(t) + \frac{g}{l}\theta(t) = 0.$$

The above equation has solutions $\theta(t)$ such that

$$\theta(t) = A\cos\omega t + B\sin\omega t,$$

where $\omega^2 = g/l$ and A and B are constants depending on the initial conditions. Denoting the period as T and recalling that both $\sin\theta$ and

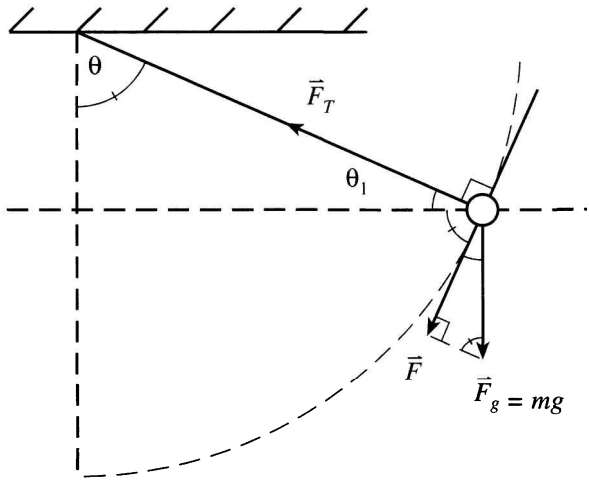


Figure 2. Here $\theta_1 = \frac{\pi}{2} - \theta$. $\vec{F} = -mg\sin\theta$ is the force parallel to the direction of motion, while $\vec{F}_T$ is the tension force perpendicular to the direction of motion.

$\cos\theta$ have a period of 2π implies that $\omega T = 2\pi$. Thus the pendulum has period

$$T = \frac{2\pi}{\omega} = 2\pi\sqrt{\frac{l}{g}}. \tag{4.2}$$

Remarkably, we can derive the essential dependence of T and l and g without setting up and solving the differential equation by the process of dimensional analysis. Assume that there exists a function f such that

$$T = f(l, g, \theta, m),$$

and denote the dimensions of the physical quantities T, l, g, θ, and m by brackets. If, for example, we set the units of these field variables such that

$$[T] = \mathrm{s}, \qquad [g] = \mathrm{m\,s^{-2}}, \qquad [l] = \mathrm{m},$$
$$[m] = \mathrm{kg}, \qquad [\theta] = \text{dimensionless},$$

then the period of the pendulum, T, can be determined by combining l, g, θ, and m in such a way that the resulting function has the same

dimensions as T (the case where $[m^\alpha l^\beta g^\gamma]$ has dimensions 's'). Thus we are interested in determining α, β, and γ such that the expression

$$\mathrm{s} = (\mathrm{kg})^\alpha(\mathrm{m})^{\beta+\gamma}(\mathrm{s})^{-2\gamma}$$

is satisfied. Notice that θ is ignored since it carries no dimensions. Because the left-hand side does not include the dimensions of kilograms or meters, it is clear that $\alpha = 0$ and that $\beta + \gamma = 0$. The remaining relationship, $-2\gamma = 1$, yields $\gamma = -1/2$ and $\beta = 1/2$. Consequently,

$$\left[m^0\, l^{1/2}\, g^{-1/2}\right] = (\mathrm{kg})^0(\mathrm{m})^{1/2}\left(\mathrm{m\ s}^{-2}\right)^{-1/2} = \mathrm{s},$$

as required. Therefore, if T is divided by a factor of $m^\alpha l^\beta g^\gamma = (l/g)^{1/2}$, we have

$$\sqrt{\frac{g}{l}}T = \sqrt{\frac{g}{l}}f(l, g, m, \theta), \tag{4.3}$$

and both sides of this equation are dimensionless. The factor $(g/l)^{1/2}$ can be absorbed into the (so far undetermined) function f, allowing (4.3) to be rewritten as

$$\sqrt{\frac{g}{l}}T = F(l, g, m, \theta). \tag{4.4}$$

Because changing the unit of measurement for m, l, or g will change the value of that variable but will not affect the left-hand side, the function F represented by (4.4) must be independent of m, g, and l. Consequently, equation (4.4) is given by

$$\sqrt{\frac{g}{l}}T = F(\theta), \tag{4.5}$$

implying that the period is

$$T = \sqrt{\frac{l}{g}}F(\theta),$$

which is consistent with relation (4.2).

In view of (4.5), $F(\theta)$ is an even function of θ; assuming that F is smooth, it is possible to expand F in a Taylor series as

$$F(\theta) = F(0) + F'(0)\theta + \frac{1}{2}F''(0)\theta^2 + \cdots.$$

Because F is even, it follows that $F'(0) = F^{(3)}(0) = F^{(5)}(0) = \cdots = 0$ (the proof of this is left to the reader in Exercise 1), thus

$$F(\theta) = F(0) + \frac{1}{2}F''(0)\theta^2 + \frac{1}{4!}F^{(4)}(0)\theta^4 + \cdots,$$

which indicates that to first order in θ, $F(\theta)$ is well approximated by $F(0)$.

4.2. Dimensional Analysis: The General Procedure

Having taken a cursory look at how to approach the analysis of phenomena using dimensional analysis, we now establish the procedure more rigourously. The phenomena of interest involve measurable scalar quantities $\{u, W_1, W_2, \ldots, W_n\}$ with dimensions $\{[u], [W_i]\}$. If we assume that there exists an unknown function f such that

$$u = f(W_1, \ldots, W_n), \tag{4.6}$$

then it is possible to analyse this function using fundamental dimensional units $(L_1, L_2, \ldots, L_m)$, where $m \leq n$. In the SI system[1] there are seven fundamental dimensional units: mass (M), length (L), time (T), electric current (A), thermodynamic temperature (K), amount of substance (mol) and luminous intensity (cd). Because of this, practitioners assign $L_1 = M$, $L_2 = L$, $L_3 = T$, $L_4 = A$, $L_5 = K$, $L_6 = \text{mol}$, $L_7 = \text{cd}$. This assignment will be made in Section 4.3. In the case of the pendulum,

$$u = T, \qquad W_1 = M, \qquad W_2 = l, \qquad W_3 = g,$$

so that

$$[u] = \text{s}, \qquad [W_1] = \text{kg}, \qquad [W_2] = \text{m}, \qquad [W_3] = \text{m s}^{-2}.$$

Therefore, the fundamental dimensional units are chosen to be $L_1 = \text{kg}$, $L_2 = \text{m}$, and $L_3 = \text{s}$.

We postulate that all descriptive quantities in mathematical models have dimensions that are products of powers of these fundamental dimensional units. (This is the axiomatic assumption which makes

[1]The International System of Units (SI) is the modern form of the metric system.

dimensional analysis possible). For example, the dimensions of energy are

$$[\text{Energy}] = \frac{\text{kg m}^2}{\text{s}^2} = L_1 L_2^2 L_3^{-2},$$

where $L_1 = \text{kg}$, $L_2 = \text{m}$ and $L_3 = \text{s}$. In general, if Z is some descriptive quantity, then its dimensions may be written in terms of fundamental dimensional units as

$$[Z] = L_1^{\alpha_1} \cdots L_m^{\alpha_m}.$$

For instance, $m = 3$ in the above energy example. The number of fundamental dimensional units for the phenomena of interest determines the number of primary and secondary quantities, which we define next.

4.2.1. Primary and Secondary W_i. For the function represented by (4.6), it is necessary to separate the W_i as primary and secondary quantities. In order to do so, we choose a subset $\{P_1, \ldots, P_m\}$ of the set $\{W_1, \ldots, W_n\}$ such that the number of elements in the subset corresponds to the number of fundamental dimensional units and the $\{P_1, \ldots, P_m\}$ are linearly independent in the sense that their dimensions are independent. Specifically, if $L_1, \ldots, L_m$ are the fundamental dimensional units and $[P_i] = L_1^{\alpha_{i,1}} L_2^{\alpha_{i,2}} \cdots L_m^{\alpha_{i,m}}$, the vectors $(\alpha_{i,1}, \ldots, \alpha_{i,m}) \in \mathbb{R}^m$ have to form a linearly independent set (a basis) of $\mathbb{R}^m$. With this correspondence, the fundamental dimensional unit L_i corresponds to the ith canonical basis vector given by

$$e_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T,$$

where the 1 is in position i. This subset $\{P_i\}$ of $\{W_i\}$ chosen in this way but otherwise arbitrary, contains what we call the primary quantities. Of course, the dimension of the left-hand side of (4.6) must also be a combination of the L_i, $i = 1, \ldots, m$.

We leave it to the reader to consider the situation where no set of primary quantities can be found (the case where the dimensions of the space spanned by the $(\alpha_{i,1}, \ldots, \alpha_{i,m})$ corresponding to all the $\{W_1, \ldots, W_n\}$ is strictly less than m).

If we assume that there exist m such primary quantities, then the remaining quantities

$$\{W_1, \ldots, W_n\} - \{P_1, \ldots, P_m\} = \{S_1, \ldots, S_{n-m}\}$$

are termed secondary and are expressible as linear combinations of the primary ones. Therefore, returning to (4.6), we make the distinction

$$u = f(P_1, \ldots, P_m, S_1, \ldots, S_{n-m}), \tag{4.7}$$

where u and $\{S_j : j = 1, \ldots, n-m\}$ are secondary quantities and $\{P_i : i = 1, \ldots, m\}$ are linearly independent, primary quantities. The dimensions of the secondary quantities are, by definition, expressible in terms of the dimensions of the P_i. In this sense u and the S_j are linearly dependent on the primary quantities.

4.2.2. Constructing the Quantities D and Π. In order to proceed, we form a quantity D such that $[D] = [u]$ where

$$D = P_1^{\alpha_1} \cdots P_m^{\alpha_m}.$$

Note that D is a combination of the primary quantities and that the $\alpha_1, \ldots, \alpha_m$ are uniquely defined (Exercise 2). Now if we let Π be the dimensionless function formed by $\Pi = u/D$, then (4.7) becomes

$$\Pi = \frac{u}{D} = \frac{1}{D} f(P_1, \ldots, P_m, S_1, \ldots, S_{n-m}). \tag{4.8}$$

Because the S_j are dependent on the P_i, it is possible to find a set of powers $\alpha_{j,1}, \ldots, \alpha_{j,m}$ such that

$$[S_j] = [P_1]^{\alpha_{j,1}} \cdots [P_m]^{\alpha_{j,m}}$$

(the secondary quantities are expressed as combinations of the primary quantities). For each S_j we form the quantity D_j such that

$$D_j = P_1^{\alpha_{j,1}} \cdots P_m^{\alpha_{j,m}},$$

where $[D_j] = [S_j]$, and we let Π_j be the dimensionless quantity $\Pi_j = S_j/D_j$. Thus, $S_j = D_j \Pi_j$ and equation (4.8) becomes

$$\Pi = \frac{1}{D} f(P_1, \ldots, P_m, D_1\Pi_1, \ldots, D_{n-m}\Pi_{n-m}),$$

which, by absorbing D and the D_j into the function f, is written as

$$\Pi = F(P_1, \ldots, P_m, \Pi_1, \ldots, \Pi_{n-m}), \tag{4.9}$$

where Π and Π_j are dimensionless. F must effectively be independent of the primary quantities; otherwise, we could rescale the fundamental dimensional units in such a way that, for example, only the value of P_1 would change (see Exercise 3). As none of the dimensionless quantities would be affected by this change, Π would not vary; therefore, F has to be independent of P_i.

4.2.3. The Buckingham Pi Theorem. We are now able to solve for u in equation (4.9) by multiplying through by D. This gives

$$u = \Pi D = DF(W_1, \ldots, W_m, \Pi_1, \ldots, \Pi_{n-m}).$$

We summarize our findings as a theorem which is known as the Buckingham Pi Theorem.

Theorem 4.1.

(1) *A relation* $u = f(W_1, \ldots, W_n)$ *in a mathematical model, which is unchanged for any system of measurement units, can be written as a relation among dimensionless combinations of the original quantities.*

(2) *The number of independent dimensionless combinations involved is equal to the difference between the number of original quantities* (n) *and the number of fundamental dimensional units* (m).

In effect,

$$u = DF(\Pi_1, \ldots, \Pi_{n-m}) = DF\left(\frac{S_1}{D_1}, \ldots, \frac{S_{n-m}}{D_{n-m}}\right). \tag{4.10}$$

Notice that the explicit dependence of the right-hand side on the primary quantities disappears. The power of this theorem becomes apparent in its applications. For a case in point, let us return to the nuclear explosion of 1945.

4.3. The Energy Released by a Nuclear Bomb

We now use the method presented in the previous sections to reproduce the results which Sir G.I. Taylor obtained when he determined

the energy released by the explosion. We can take the radius of the expanding shock wave as the dependent variable and assume

$$r = f(t, E, \rho_o, P_o),$$

where E is the released energy, t denotes the time, and ρ_o and P_o denote the ambient air density and pressure.

There are three fundamental dimensional units: L_1 represents mass (in kilograms), L_2 length (in meters), and L_3 time (in seconds). Therefore, the dimensions of the above variables are

$$[r] = L_2, \qquad [E] = L_1 L_2^2 L_3^{-2},$$
$$[t] = L_3, \qquad [P_o] = L_1 L_2^{-1} L_3^{-2}, \qquad [\rho_o] = L_1 L_2^{-3}.$$

Because there are three fundamental dimensional units, there will be three primary variables. Here we arbitrarily choose t, E, and ρ_o as primary and express them as combinations of the fundamental dimensional units. This results in three linearly independent vectors

$$[t] = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad [E] = \begin{pmatrix} 1 \\ 2 \\ -2 \end{pmatrix}, \quad [\rho_o] = \begin{pmatrix} 1 \\ -3 \\ 0 \end{pmatrix},$$

such that the three rows correspond to the three fundamental dimensional units L_1, L_2, and L_3, respectively.

Now, as in Section 4.2.2, we form D, a combination of the primary quantities t, E, and ρ_o, such that $[D] = [r]$. This is accomplished by solving

$$[r] = [t]^\alpha [E]^\beta [\rho_o]^\gamma$$

for α, β, and γ. Expressing this in terms of the fundamental dimensional units gives

$$L_2 = L_3^\alpha \left(L_1 L_2^2 L_3^{-2}\right)^\beta \left(L_1 L_2^{-3}\right)^\gamma,$$

or in matrix form

$$\begin{pmatrix} 0 & 1 & 1 \\ 0 & 2 & -3 \\ 1 & -2 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

Augmenting the matrix allows us to solve by the process of Gaussian elimination as follows. Switching the first and the last row, then

multiplying the second row by $1/2$ and subtracting it from the third row yields

$$\left(\begin{array}{ccc|c} 0 & 1 & 1 & 0 \\ 0 & 2 & -3 & 1 \\ 1 & -2 & 0 & 0 \end{array}\right) \rightarrow \left(\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 2 & -3 & 1 \\ 0 & 1 & 1 & 0 \end{array}\right)$$
$$\rightarrow \left(\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 1 & -3/2 & 1/2 \\ 0 & 0 & 5 & -1 \end{array}\right).$$

This is now in a form which allows us to easily solve for γ. We find $\alpha = 2/5$, $\beta = 1/5$, and $\gamma = -1/5$ so that

$$D = [r] = \left(\frac{t^2 E}{\rho_o}\right)^{1/5},$$

which has the same dimensions as r by construction. Returning to the procedure set out in Section 4.2.2, we let Π be the function formed by $\Pi = r/D$. This gives the dimensionless

$$\Pi = r\left(\frac{\rho_o}{t^2 E}\right)^{1/5}.$$

The next step is to establish the D_j and Π_j for the remaining secondary quantities. In this case there is only one: P_o. As with D, D_1 associated with P_o is formed by combining the primary quantities t, E, and ρ_o, in such a way that $[D_1] = [P_o]$. We find

$$D_1 = \left(\frac{E^2 \rho_o^3}{t^6}\right)^{1/5}$$

by following the same method used to construct D. Consequently,

$$\Pi_1 = P_o\left(\frac{t^6}{E^2 \rho_o^3}\right)^{1/5}.$$

Therefore, using equation (4.10) and making the appropriate substitutions for D and Π_1, we have

$$r = DF(\Pi_1) = \left(\frac{t^2 E}{\rho_o}\right)^{1/5} F\left(P_o\left(\frac{t^6}{E^2 \rho_o^3}\right)^{1/5}\right) \tag{4.11}$$

as the relationship for r. The function F is still unknown.

4.3.1. Calculating the Released Energy. To estimate the released energy, the function F in equation (4.11) must be approximated. By examining the relationship $P_o\left(t^6E^{-2}\rho_o^{-3}\right)^{1/5}$, it is clear that for a nuclear bomb, the time elapsed will be relatively short while the energy expended will be extremely large. Because of this, the argument of F will be very small for the duration of the explosion. Therefore, approximating $F(x)$ with its value at $x = 0$, equation (4.11) becomes

$$r \approx \left(\frac{t^2E}{\rho_o}\right)^{1/5} F(0).$$

Through experiments with small explosives, Taylor suggested that $F(0) \approx 1$. This implies that, to a good approximation,

$$r = \left(\frac{t^2E}{\rho_o}\right)^{1/5},$$

which we now convert to a linear function in order to compute the energy. Taking the logarithm of both sides produces

$$\frac{5}{2}\log_{10} r = \log_{10} t + \frac{1}{2}\log_{10}\left(\frac{E}{\rho_o}\right), \tag{4.12}$$

where $\rho_o = 1.25$ kg m^{-3}. Figure 3 illustrates the transformed data from Taylor's original values given in Table 1. A least-squares fit of this data gives an estimate of $1/2\log_{10}(E/\rho_o) \simeq 6.90$ so that $E = 8.05 \times 10^{13}$ Joules. Using the conversion factor of 1 kiloton $= 4.186 \times 10^{12}$ Joules gives the strength of Trinity as 19.2 kilotons. It was later revealed that the actual strength of Trinity was 21 kilotons. This demonstrates the predictive power of dimensional analysis.

Table 1. Taylor's original data with the time t measured in milliseconds and the radius r in meters.

t	$r(t)$	t	$r(t)$	t	$r(t)$	t	$r(t)$	t	$r(t)$
0.10	11.1	0.80	34.2	1.50	44.4	3.53	61.1	15.0	106.5
0.24	19.9	0.94	36.3	1.65	46.0	3.80	62.9	25.0	130.0
0.38	25.4	1.08	38.9	1.79	46.9	4.07	64.3	34.0	145.0
0.52	28.8	1.22	41.0	1.93	48.7	4.34	65.6	53.0	175.0
0.66	31.9	1.36	42.8	3.26	59.0	4.61	67.3	62.0	185.0

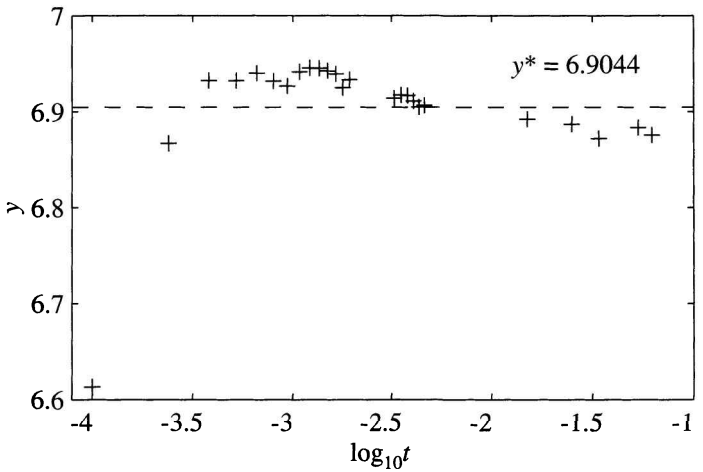


Figure 3. The graph of $y = 5/2 \log_{10} r - \log_{10} t$ is identified with the quantity $1/2 \log_{10} (E/\rho_o)$. A least-squares fit yields $y^* = 6.90$ and because ρ_o is known, it is possible to solve for E, the energy released during the explosion.

4.4. Exploration: How to Cook a Turkey

The process of cooking a turkey seems a fairly straightforward and simple matter. A basic rule of thumb suggests preheating the oven to 400°F, and then cooking it for twenty minutes per pound at 350°F. (Actually, cookbooks like Betty Crocker's are a little more sophisticated. They lengthen the cooking time with the weight of the turkey, while suggesting reduced temperatures for larger turkeys. Eons of kitchen experience and thousands of burnt turkeys must have contributed to this amount of detail.)

According to the rule of thumb, the time required to cook a turkey is directly proportional to and depends only upon the mass. Although this procedure has the advantages of being easy to remember and implement, it seems too simplistic. First, it is unlikely that mass is the *only* variable that will affect the time needed to cook the turkey. Second, and most importantly, why do some people who adhere to this rule serve turkey that is dry and tough rather than moist and tender? It would appear that further investigation is required.

4.4.1. Considering Other Variables. As stated, our traditional model only takes into consideration the mass of the turkey when determining the cooking time. However, many factors play a part here. By only using the mass of the turkey, much of the variability that exists in nature is ignored. For example, two turkeys with the same mass and different diameters would cook at different rates: the thicker the bird, the more insulation, the more time necessary. Thus, the diameter is important in determining cooking time, and, is a good description of the size of the turkey. To determine the mass of the turkey using the diameter, we need to know the density of the bird. Since bone and tissue have different densities, this variable will mean the effective mean density of the bird.

Oven temperature is clearly another variable as indicated clearly by Betty Crocker's recipes. Intuitively, the higher the temperature, the less time should be required. However, we recognize that different foods cook at different rates; therefore, a fourth variable, the thermal conductivity of the turkey, is needed to account for the rate at which heat energy is absorbed.

Considering these four variables, the diameter, l, the oven temperature, T, the average density, ρ, and the thermal conductivity, κ, we assume a function

$$t = f(l, \rho, T, \kappa), \tag{4.13}$$

where t denotes the cooking time. With this established, we conduct dimensional analysis to evaluate the accuracy of the traditional rule of thumb for cooking a turkey.

4.4.2. Dimensional Analysis. Following the procedure laid out in the text, we examine the dimensions of the variables in (4.13) which are

$$[t] = \text{s}, \quad [l] = \text{m}, \quad [\rho] = \text{kg m}^{-3}, \quad [T] = \text{kg m}^{-1}\text{s}^{-2}, \quad [\kappa] = \text{m}^2\text{s}^{-1}.$$

(Here temperature is defined as energy per volume, and the thermal conductivity is measured as energy times the length divided by the product of the area, the time, and the temperature. Consult textbooks on physics for the background on this.) We have chosen the fundamental dimensional units as $L_1 = \text{kg}$, $L_2 = \text{m}$, and $L_3 = \text{s}$.

Expressing the above quantities in terms of these fundamental dimensional units yields

$$[t] = L_3, \qquad [l] = L_2,$$
$$[T] = L_1 L_2^{-1} L_3^{-2}, \qquad [\kappa] = L_2^2 L_3^{-1}, \qquad [\rho] = L_1 L_2^{-3}.$$

As there are three fundamental dimensional units, there will be three primary quantities which can be chosen arbitrarily from the four variables l, ρ, T, κ in our function. Choosing l, ρ, and κ as our primary variables results in three linearly independent vectors

$$[l] = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad [\rho] = \begin{pmatrix} 1 \\ -3 \\ 0 \end{pmatrix}, \qquad [\kappa] = \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix},$$

where the three rows correspond to L_1, L_2, and L_3, respectively.

As in the text, we now want to form D, a combination of the primary quantities, such that $[D] = [t]$. To do this, we solve

$$[t] = [l]^\alpha [\rho]^\beta [\kappa]^\gamma,$$

which has the matrix form

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & -3 & 2 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

for α, β, and γ. This gives $\alpha = 2$, $\beta = 0$, and $\gamma = -1$. Thus,

$$D = \frac{l^2}{\kappa},$$

which has by construction the same dimensions of t. We find the function Π such that $\Pi = t/D$, which gives

$$\Pi = t\left(\frac{\kappa}{l^2}\right).$$

D_1 and Π_1 for the secondary quantity T are found in a similar manner.

This produces

$$D_1 = \frac{\rho\kappa^2}{l^2},$$

which implies that

$$\Pi_1 = T\left(\frac{l^2}{\rho\kappa^2}\right),$$

and we obtain the dependency

$$t = DF(\Pi_1) = \frac{l^2}{\kappa} F\left(\frac{Tl^2}{\rho\kappa^2}\right) \tag{4.14}$$

with a still unknown function F.

4.4.2.1. *Simplifying the Function.* The power of dimensionless analysis is exhausted at this point. As in the case of an explosion discussed in the chapter, we have to invoke experimental data in order to obtain further information on F. To simplify the burden and just test the mentioned rule of thumb, let us first assume that we will keep the temperature constant (at, say, 350°F), so that this variability will not matter for our argument. Also, it is reasonable to assume that the thermal conductivity and average density of turkeys does not vary much from one bird to another, as all turkeys are made of similar tissue and bones. It then follows that the mass of the turkey, denoted as M, is simply proportional to the volume of the turkey. Applying these generalizations to (4.14) and substituting $M^{1/3}$ for l yields

$$t = M^{2/3}\,\tilde{F}(M^{2/3}), \tag{4.15}$$

where proportionality constants have been absorbed into this new function $\tilde{F}$.

The new relationship (4.15) seems rather useless; all we really see is that it is reasonable to expect that t should be a function of M. However, if we revisit (4.14) for a moment, we remember that the F there is dimensionless; it is this dimensionless function we need to determine, and it depends on only one variable. As temperature, density, and thermal conductivity are kept fixed, only the variable $M^{2/3}$ remains, and we have to proceed by testing simple hypotheses regarding $\tilde{F}$.

The simplest hypothesis is that $\tilde{F}$ is constant. If so, cooking time is just proportional to $M^{2/3}$ as opposed to being a linear function as represented by the rule of thumb. However, keep in mind that $\tilde{F}(M^{2/3})$ is not *a priori* known to be constant. If, for example, $\tilde{F}(x) = Cx^{1/2}$ where C is some constant, the resulting t dependence on the mass M would in fact be linear.

4.4.2.2. *Alternate Functions.* In addition to the consideration mentioned in Section 4.4.2.1, there are two other criticisms of the expression (4.15). First, the choice of T as a secondary quantity is completely arbitrary. Had we chosen κ as the secondary quantity instead, our results would look entirely different. Specifically, this approach yields

$$t = DF(\Pi_1) = l\sqrt{\frac{\rho}{T}}G\left(\frac{\kappa}{l}\sqrt{\frac{\rho}{T}}\right). \tag{4.16}$$

Proceeding as in Section 4.4.2.1 to (4.16) we have a second function for the cooking time given by

$$t = M^{1/3}\tilde{G}(M^{1/3}). \tag{4.17}$$

Using just the mass instead of the density and the length, and performing a dimensional analysis, yields the function $t = h(M, \kappa, T)$ such that

$$t = \left(\frac{M^2}{\kappa T^2}\right)^{1/5} H(0),$$

which is simplified to

$$t = M^{2/5}\tilde{H}(0). \tag{4.18}$$

This result looks appealing as it lies between (4.15) and (4.17) and does not include a secondary quantity dependent on the mass. However, we have discarded information by reducing the number of independent variables.

4.4.3. Evaluating the Functions. We now have three alternative functions to describe the cooking time of a turkey, in addition to the general rule of thumb that we set out to evaluate. In order to determine which of these options yields the best results, it is helpful to analyse some data. In discussing this problem, Giordano and Weir [**E**] included the actual results of cooking turkeys of various sizes which (assuming adequate cooking skills) offer some insight. Table 2 displays these results with values from other sources. A graph of this data can be used to give some indication which of these functions is the most appropriate model. Because we do not have any information regarding the functions $\tilde{F}(M^{2/3})$, $\tilde{G}(M^{1/3})$, and $\tilde{H}(0)$, we will test the models under the assumptions that these functions are constant.

Table 2. Cooking time required for turkeys of various weights.

M: weight (lbs.)	5	6	10	12	15	20	25	30
t: time (hours)	2	2.5	3.4	3.5	4.5	5.4	5.8	7

Figure 4 illustrates the dependence of t/M^p against the weight M where $p = 2/3, 2/5$, and $1/3$, respectively. The data points and the resulting best fit line for each function are displayed. As can be seen, $\tilde{F}(M^{2/3})$, $\tilde{G}(M^{1/3})$, and $\tilde{H}(0)$ are all well represented by a linear function of M; however, only the $p = 2/3$ model ($\tilde{F}(M^{2/3})$) is essentially constant. Thus, it is apparent that the cooking time is best represented by function (4.15) where we assume that the proper equation is $t = Cw^{2/3}$ where C is some constant.

4.4.3.1. *A New Rule.* The constant C in the equation $t = CM^{2/3}$ is easily determined from the slope of the best fit line in Figure 4 to be 0.7185 hours/lbs$^{2/3}$ or 45 min/lbs$^{2/3}$ to within a couple of minutes. Thus our new rule is given by

$$t = 45M^{2/3} \tag{4.19}$$

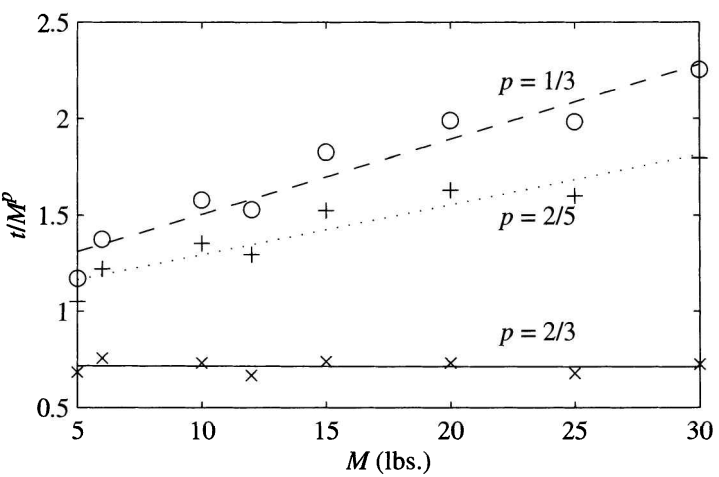


Figure 4. The cooking time, t, for turkeys of various weights, w, are presented when $p = 2/3$, $p = 2/5$ and $p = 1/3$.

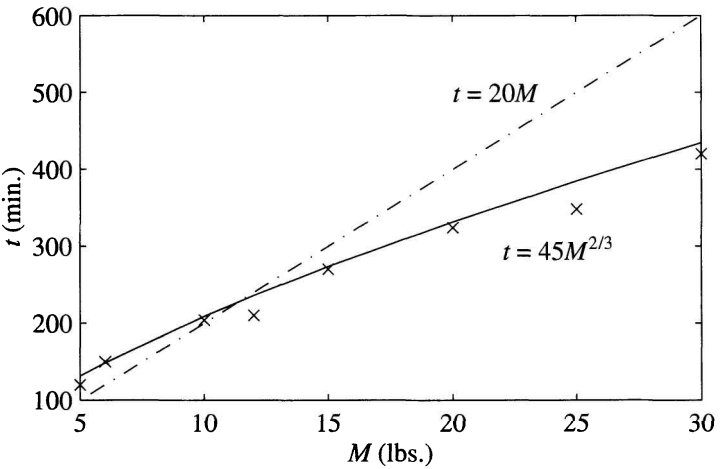


Figure 5. The old rule of thumb $t = 20M$ is displayed with the new rule $t = 45M^{2/3}$. The data points from Table 2 have also been plotted. As is evident from the graph, the new rule (4.19) is a much better predictor of the time required to cook a turkey.

where M is weight in pounds and time t is in minutes. Figure 5 graphs our new rule (4.19) with the traditional rule $t = 20M$.

Plotting the actual cooking times (as in Figure 5) from Table 2 shows that our new rule is indeed a much better model. Even when a fifteen minute error is allowed, the old rule still only predicts well for turkeys within the range of nine to fourteen pounds.

4.4.4. Further Considerations. It is possible to take the analysis of this problem even further by including other variables such as air

Table 3. Additional functions

Variables	Function	Simplified equation
M, l, P, T	$t = f_1(M, l, P, T)$	$t \propto M^{1/3} F_1(0)$
l, κ, P, T	$t = f_2(l, \kappa, P, T)$	$t \propto M^{2/3} F_2(0)$
M, κ, P, T	$t = f_3(M, \kappa, P, T)$	$t \propto M^{2/5} F_3(0)$
M, l, P, T, c	$t = f_4(M, l, P, T, c)$	$t \propto M^{1/3} F_4(0, 0)$

pressure (which would be relevant if a pressure cooker was used) or specific heat. However, the addition of more variables significantly complicates the analysis (as there are currently only three primary variables), and does not necessarily ensure a better cooked bird than the rule we developed in Section 4.4.3.1. A few additional formulas have been included in Table 3 which incorporate the cooking pressure P (which has dimensions $[P] = \text{kg m}^{-1}\text{s}^{-2}$) and specific heat c (which has dimensions $[c] = \text{kg}^{-1}\text{m}^3$) for completeness. All other variables are defined as in previous sections. These functions can be verified using dimensional analysis and applying the assumptions made in Section 4.4.2.1.

Exercises

(1) Suppose that $f = f(x)$ is a smooth, even function; i.e., that $f(x) = f(-x)$ for all $x \in \mathbb{R}$ and that the derivatives of any order exist. Show that $f'(x)$ is odd ($f'(x) = -f'(-x)$), $f''(x)$ is even, $f^{(3)}(x)$ is odd, etc. Use this to prove that $f'(0) = 0$, $f^{(3)}(0) = 0$, $f^{(5)}(0) = 0, \ldots$.

(2) Prove that the $\alpha_1, \ldots, \alpha_m$ in the definition of D (Section 4.2.2) are unique.

(3) (a) Find an estimate for the order of magnitude of the dimensionless quantity $P_o(t^6 E^{-2} \rho_o^{-3})^{-1/5}$. Choose $t = 5$ s, P_o and ρ_o ambient pressure, and air density in the kg·m·s - system, and make a guess for E (for example, $E = 10^3$ Joules, or 10^9 Joules, or 10^{12} Joules).

(b) Compute the energy E from (4.12) by assuming $t = 5$ s, $r = 1000$ m.

(4) Suppose that $M_1, \ldots, M_m$ are primary variables in a mathematical model, and that $L_1, \ldots, L_m$ measure fundamental dimensional units. A rescaling of the fundamental dimensional units implies that L_j is converted to $\tilde{L}_j$, where $L_j = \beta_j \tilde{L}_j$ for some $\beta_j > 0$. (For example, if $L_j = \text{m}$ and $\tilde{L}_j = \text{cm}$, then $\beta_j = 100$.)

Let M_j be the value of the primary variable if the fundamental dimensional units are measured as $L_1, \ldots, L_m$, and let

$\tilde{M}_j$ be the value of the primary variable if they are measured as $\tilde{L}_1, \ldots, \tilde{L}_m$. Show that for any set of positive constants $k_1, \ldots, k_m$ there exist scale factors β_j, $j = 1, 2, \ldots, m$, such that

$$\tilde{M}_1 = k_1 M_1, \quad \tilde{M}_2 = k_2 M_2, \quad \ldots, \quad \tilde{M}_m = k_m M_m.$$

(5) A spherical raindrop falls from a motionless cloud. Its terminal velocity v is expected to depend on its size (characterized by its radius r), its density ρ, the gravitational acceleration g, and the viscosity of air μ, where $[\mu] = \text{kg m}^{-1}\text{s}^{-1}$. Determine this dependence as explicitly as possible, ignoring other factors.

(6) Repeat Sir Taylor's problem of the radius of the shock wave in an explosion as a function of time, released energy, ambient air density, and pressure, choosing time, energy, and pressure as the primary variables. What functional relationship do you find?

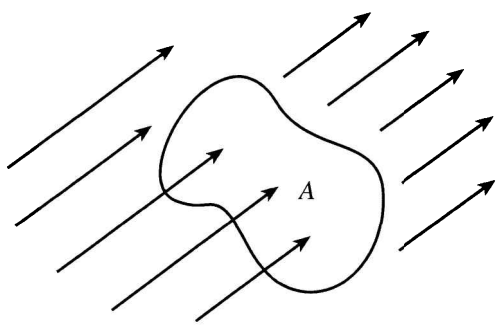
(7) The force K experienced by an object with a cross-sectional area A is a fluid with density ρ and speed u (see picture). It is assumed to be a function of ρ, A and u:

$$K = f(\rho, A, u).$$

Show that there is a constant c such that

$$K = c\rho u^2 A.$$

Hint: Choose kg, m, s as fundamental dimensional units. Note that f depends on three primary quantities and on *no* secondary quantities.



Chapter 5

Predator-Prey Systems

Concepts and Tools: Systems of ordinary differential equations

When Italy declared war on Austria in 1915, both sides feared an invasion of their ports. Mines were set in many of the seaports in the Adriatic Sea, preventing fishing for the duration of the war. When the war ended three years later and the mines were removed, it was expected that the fisherman would have a better than usual catch since the fish stocks had had three years to replenish. Surprisingly, the opposite was true.

A possible explanation of this phenomenon was later given by Volterra (1926) in a study of predator-prey systems. Lotka had independently derived the same model in the context of a chemistry application in 1920. This chapter examines the Lotka–Volterra model, its relevance to this situation, and more general dynamical systems described by systems of *autonomous* ordinary differential equations.

5.1. The Lotka–Volterra Model

To understand the basis of this model, we will begin by considering a system in which there are only two species: the number of predators at time t, $P(t)$, and the number of its prey, $N(t)$. The nature of the

interaction between these two species is given by the following system of differential equations: The rate of change of the prey population is given as

$$\frac{d}{dt}N(t) = aN(t) - bN(t)P(t), \tag{5.1}$$

where the growth rate is represented by the term $aN(t)$, and the effects of predation by the term $-bN(t)P(t)$ (a and b are positive constants). Losses due to predation are proportional to both the number of predators and prey. For the predator, the corresponding rate is

$$\frac{d}{dt}P(t) = -dP(t) + cN(t)P(t), \tag{5.2}$$

where $-dP(t)$ is the effect of mortality, and $cN(t)P(t)$ is the growth from predation (c and d are positive constants). Denoting the derivatives of N and P as N' and P', suppressing the t, and grouping like terms allows (5.1) and (5.2) to be expressed as

$$N' = N(a - bP), \tag{5.3}$$

$$P' = P(-d + cN). \tag{5.4}$$

These equations are known as the Lotka–Volterra system; as we will see, it offers a simple explanation of how predator and prey populations fluctuate. Notice that the right-hand sides of the equations do not depend explicitly on time. Systems with this property are called *autonomous.*

To gain a better understanding of this interaction we examine a phase portrait for this system. For this reason, we rescale (5.3) and (5.4) and introduce a vector field which allows us to determine and analyse the curves of the phase portrait.

5.1.1. Rescaling the System. Equations (5.3) and (5.4) seem to contain four *free* parameters. However, only one parameter ($\alpha := d/a$) is essential: the others can be *removed* by rescaling, without changing the structure of the phase diagram. To this end, let

$$\tau = at, \qquad u(\tau) = \frac{c}{d}N(t), \qquad v(\tau) = \frac{b}{a}P(t).$$

After some calculation (see Exercise 1) one arrives at the rescaled Lotka–Volterra system

$$\frac{du}{d\tau} = u(\tau)\,[1 - v(\tau)]\,, \tag{5.5}$$

$$\frac{dv}{d\tau} = \alpha v(\tau)\,[u(\tau) - 1]\,, \tag{5.6}$$

where $\alpha = d/a$.

5.1.2. The Vector Field. Consideration of the vector field $\vec{F}(u, v)$ from the right-hand sides of equations (5.5) and (5.6),

$$\vec{F}(u, v) = \langle u(1 - v), \alpha v(-1 + u) \rangle\,,$$

allows us to study (5.5) and (5.6) in the phase plane as displayed by Figure 1. There is a steady state solution at $u = 1$ and $v = 1$ found by setting both (5.5) and (5.6) equal to zero. Otherwise, as demonstrated in the figure, the predator and prey populations oscillate. This occurs when an increase in predators eventually leads to a decrease in prey which in turn eventually leads to a decrease in predators, and so forth. On the basis of this very elementary observation, we expect counterclockwise oscillations in phase space.

If we abbreviate $\vec{U}(\tau) = \langle u(\tau), v(\tau) \rangle$, equations (5.5), (5.6) become

$$\frac{d\vec{U}}{d\tau} = \vec{F}\left(\vec{U}(\tau)\right). \tag{5.7}$$

It is evident that $\vec{F}(\vec{U})$ defines the direction field in the phase plane, and any solution $\vec{U}(\tau)$ to equation (5.7) is a curve in the plane with the property that the tangent vector of the position at any time $\tau = \tau_o$ is given by the vector field $\vec{F}(\vec{U}(0))$.

Example 5.1. Consider the vector $\vec{g}(t) = \langle g_1(t), g_2(t) \rangle$, where the derivatives of g_1 and g_2, denoted as g_1' and g_2', are given as

$$g_1' = g_1^2 - g_1 g_2 e^t \quad \text{and} \quad g_2' = g_2^2 - g_1 g_2 e^{-t}.$$

Therefore, the vector field $\vec{G}$ is given by

$$\vec{G}\left(\vec{g}(t), t\right) = \left\langle g_1^2 - g_1 g_2 e^t, g_2^2 - g_1 g_2 e^{-t} \right\rangle,$$

where $\vec{G}$ is both explicitly and implicitly dependent on t.

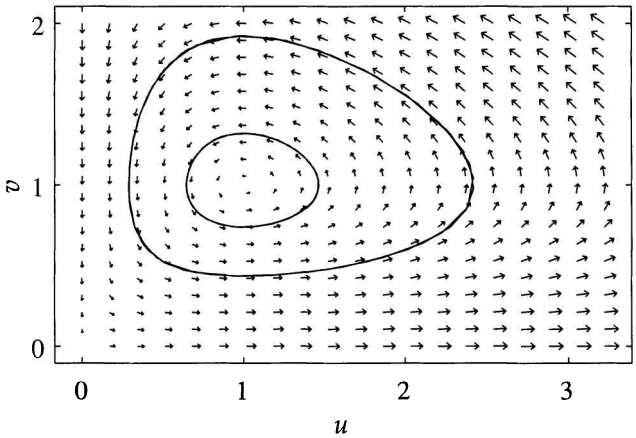


Figure 1. At $(u, v) = (1, 1)$, $\vec{F}(1, 1) = \vec{0}$ and the system is at equilibrium. All other points in the phase plane are associated with the tangential direction illustrated.

5.1.3. Autonomy. Included in the general theory of systems of differential equations are

$$\frac{d\vec{U}}{d\tau} = \vec{F}\langle \vec{U}(\tau), \tau \rangle,$$

where $\vec{F}$ depends on τ both implicitly and explicitly.

Systems where $\vec{F} = \vec{F}\left(\vec{U}(\tau)\right)$ (i.e., the right-hand side does not depend explicitly on τ) are called autonomous. As mentioned earlier, the Lotka–Volterra system is autonomous. This implies in particular that the phase portrait will be time invariant, or, more precisely, a phase portrait is possible.

5.1.4. The Level Curves. Next we analyse the phase paths so that an accurate phase portrait for the predator and prey population fluctuations may be established.

Theorem 5.2. *Let $u(0) > 0$ and $v(0) > 0$. The solution trajectory $\tau \mapsto (u(\tau), v(\tau))$ is periodic; i.e., there exists $T = T(u(0), v(0))$ such that $u(T) = u(0)$ and $v(T) = v(0)$.*

Proof. Assuming that $v \neq 1$ and $u > 0$, we divide equation (5.6) by (5.5) such that

$$\frac{dv}{du} = \alpha \frac{v(u-1)}{u(1-v)},$$

which eliminates τ. Separating the variables and integrating yields

$$\int \frac{1-v}{v}\, dv = \alpha \int \frac{u-1}{u}\, du, \qquad u > 0,\ v > 0,$$

which, upon evaluation, becomes

$$\ln v - v + C_0 = \alpha\,(u - \ln u)\,, \tag{5.8}$$

where C_0 is a constant determined because the point $(u(0), v(0))$ must lie on the curve. By rearranging equation (5.8) we obtain

$$\alpha u + v - \ln\,(u^{\alpha} v) = C_0$$

and denoting the left-hand side by $g(u, v)$ gives the expression

$$g(u, v) = C_0. \tag{5.9}$$

To see how this function behaves, we take the limit of (5.9) as $u, v \to \infty$ and $u, v \to 0$. Clearly the function $g(u, v)$ gets arbitrarily large when $u \to \infty$, $v \to \infty$, $u \to 0$, or $v \to 0$. It is an easy exercise to show that g has a global minimum at the point $(1, 1)$ and, provided $C_0 \geq \alpha - 1$, the level curves exist and are closed. The solution trajectories correspond to these closed level curves of $g(u, v)$. □

We are now going to introduce external perturbations such as fishing.

5.2. The Effect of Interference on the System

The system of equations developed in Section 5.1 does not account for any influences on population growth other than those of the predator and prey relationship. We modify the system in order to include the external influence of fishing. For simplicity, let us assume that fishing will extract the same fraction, δ, per number of predators and prey;

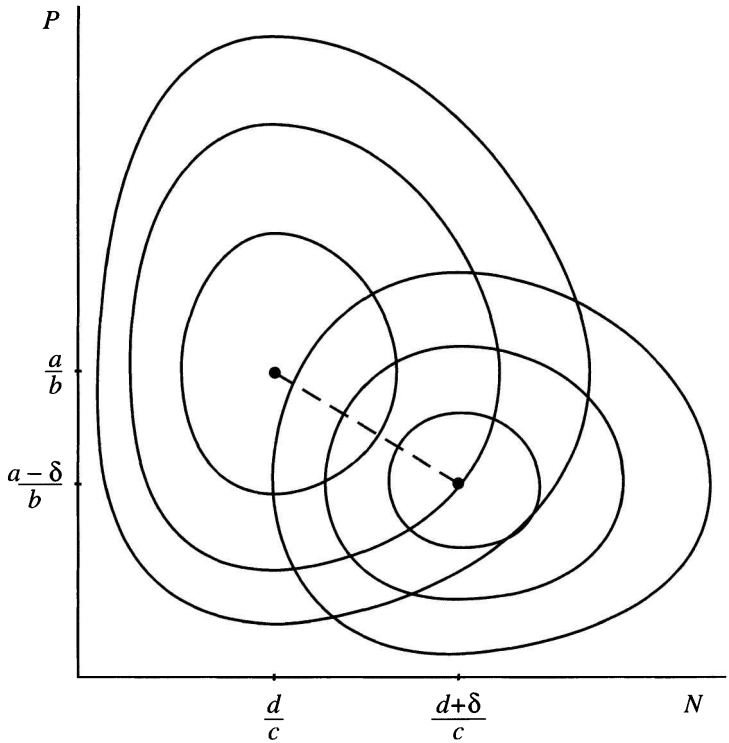


Figure 2. Without fishing, the system has an equilibrium point at $(d/c, a/b)$. When fishing resumes, the system shifts to a new equilibrium point $(d + \delta/c, a - \delta/b)$.

therefore, to model the effect of fishing, the unscaled equations (5.3) and (5.4) are modified to

$$\begin{aligned} N' &= N(a - bP) - \delta N = N(a - bP - \delta), \\ P' &= P(cN - d) - \delta P = P(cN - d - \delta). \end{aligned}$$

Figure 2 compares the phase portraits of predator and prey populations with and without fishing. As is apparent, the equilibrium point for the system shifts with fishing. We see that the new system will be at equilibrium when $P = P_{\text{eq}} = (a - \delta)/b$ and $N = N_{\text{eq}} = (d + \delta)/c$.

5.2.1. The Catch. If the predator and prey populations were to remain at equilibrium while fishing occurs, the catch would be given by

$$\delta(N_{\text{eq}} + P_{\text{eq}}) = \delta\left(\frac{d+\delta}{c} + \frac{a-\delta}{b}\right) = \delta\left(\frac{d}{c} + \frac{a}{b}\right) + \delta\left(\frac{\delta}{c} - \frac{\delta}{b}\right).$$

However, as was the case during the war, when fishing ceases, the system with $\delta = 0$ will take over and the evolution of the numbers will shift to a nonequilibrium phase path, as shown by Figure 2. To determine the average catch after the war when fishing resumed, a good estimate for both fish populations after 1915 is necessary. This seems a hopeless task given the lack of information about parameters. Remarkably, though, the Lotka–Volterra system has the property that the average populations

$$\bar{N} := \frac{1}{T}\int_0^T N(t)\,dt \quad \text{and} \quad \bar{P} := \frac{1}{T}\int_0^T P(t)\,dt$$

over a cycle with period T are identical to the equilibrium values. We formulate this as a theorem.

Theorem 5.3. *If the population model is given by the system represented by* (5.3) *and* (5.4), *then* $\bar{N} = d/c$ *and* $\bar{P} = a/b$.

Proof. By definition $T > 0$ is such that $N(t+T) = N(T)$ and $P(t+T) = P(T)$ for all values of t.

Consider the original system without fishing represented by equations (5.3) and (5.4). Focusing on equation (5.3) and separating the variables yields

$$\frac{N'(t)}{N(t)} = a - bP(t)$$

so that by integrating from $t = 0$ to $t = T$ one has

$$\ln\frac{N(T)}{N(0)} = \int_0^T [a - bP(t)]\,dt.$$

Using the fact that $N(T) = N(0)$, this can be arranged to yield

$$\bar{P} = \frac{1}{T}\int_0^T P(t)\,dt =: \frac{a}{b},$$

where $\bar{P}$ denotes the average value of $P(t)$ or the interval $[0, T]$. The average prey population is found similarly. □

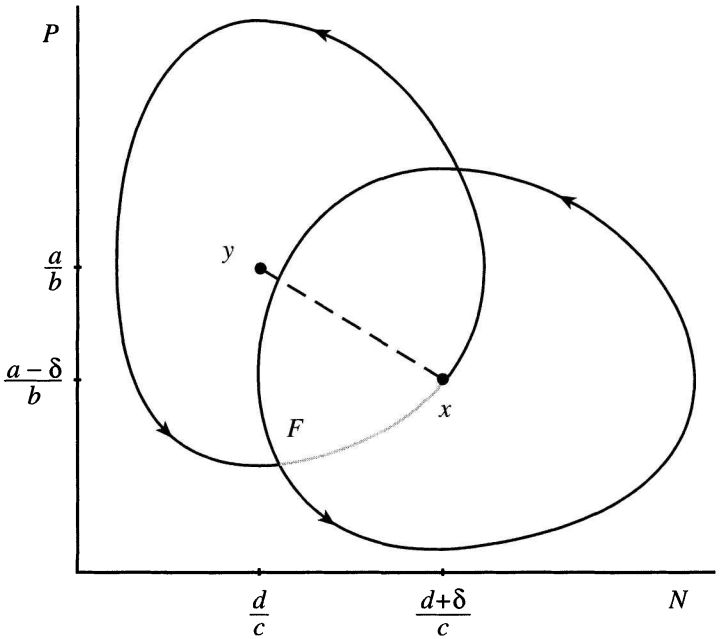


Figure 3. As shown, x is the equilibrium point of the system with fishing, and y is the equilibrium point of the system without. When fishing ceases, the phase path will follow that of the system without fishing, and, as the populations will no longer be at equilibrium, it will oscillate. When fishing is resumed, the population will not return to the previous equilibrium. For example, resuming fishing at the point F shown will initially result in a decrease in catch despite the fact that the fish stocks were given a chance to replenish.

The size of the fish population, and thus the size of the catch, will of course depend on the stage of the periodic cycle. For example, Figure 3 illustrates a possible scenario that could explain the decreased catch when fishing is resumed after an interruption.

A disturbance such as fishing perturbs a periodic system. When that disturbance is removed (as when fishing was stopped due to the war), the system will return to one of the unperturbed oscillating states. If the fishing is then resumed at a time when both predator and

prey populations are depleted due to the natural cycle, the resulting catch will be less than was usual in the past.

This explanation, while simple and mathematically rigourous, is open to many criticisms. First and foremost, the model is far too simplistic; its appeal lies mostly in its ability to offer a possible explanation for an unexpected phenomenon. Apart from the biological shortcomings of the model, it also suffers from a mathematical degeneracy: As we will discuss, the equilibrium belonging to Lotka–Volterra is a *centre* and hence structurally unstable. Modifications of Lotka–Volterra which possess structurally stable phase portraits are considered to be more realistic, and we show an example of such a modification in a project at the end of this chapter.

For now, however, we use the Lotka–Volterra model as a case study to introduce more qualitative tools for the analysis of dynamical systems.

5.2.2. Behaviour near the Equilibrium Point. To gain additional information about the system, we study its linearized version near the equilibrium points. To this end we first revisit the rescaled Lotka–Volterra system

$$u' = u(1-v), \tag{5.10}$$

$$v' = \alpha v(u-1), \tag{5.11}$$

which has an equilibrium point at $u = v = 1$. Let us assume that $u(t)$ and $v(t)$ are close to this equilibrium point, and shift the phase plane coordinate system so that it is centred at the equilibrium point. Let $U(t) := u(t) - 1$ and $V(t) = v(t) - 1$ so that

$$\begin{aligned} u(t) &= U(t) + 1, \\ v(t) &= V(t) + 1. \end{aligned}$$

Differentiating and substituting the above into (5.10) and (5.11) for u and v, we have

$$U' = u' = (U+1)(-V) = -V - UV$$

and

$$V' = v' = \alpha(V+1)U = \alpha U + \alpha UV.$$

Since quadratic terms are much smaller than linear terms for small values of $|U|, |V|$, we omit the quadratic terms to find $U' = -V$ and $V' = \alpha U$. Differentiating U once again yields

$$U'' = -V' = -\alpha U,$$

which is equivalent to $U'' + \alpha U = 0$. This linear simplified system has solutions

$$\begin{aligned} U(t) &= A\cos\left(\sqrt{\alpha}t\right) + B\sin\left(\sqrt{\alpha}t\right), \\ V(t) &= A\sqrt{\alpha}\sin\left(\sqrt{\alpha}t\right) - B\sqrt{\alpha}\cos\left(\sqrt{\alpha}t\right), \end{aligned}$$

which parameterizes an elliptical orbit. Their period is $T = 2\pi/\sqrt{\alpha}$.

5.3. Linearization: The General Procedure

Consider now a general autonomous system of ordinary differential equations for $\vec{x}(t) \in \mathbb{R}^2$, written as $x'(t) = \vec{F}(\vec{x}(t))$. The vector field $\vec{F}$ is assumed to have a zero (an equilibrium for the system) at $\vec{x}_o$ so that $\vec{F}(\vec{x}_o) = \vec{0}$. Then the linearized system about $\vec{x}_o$ is defined to be

$$\vec{X}'(t) := D\vec{F}(\vec{x}_o)\vec{X}(t), \tag{5.12}$$

where D is the matrix of partial derivatives. (Effectively, we are replacing $\vec{F}$ near $\vec{x}_o$ by its linear approximation. Section 5.3.1 provides the rationale for this step.)

In the case of the two dependent variables that we are mostly concerned with, $\vec{x}(t) = \langle x(t), y(t)\rangle$, with $x' = f(x, y)$ and $y' = g(x, y)$. Since $\vec{F}(\vec{x}_o) = \vec{0}$, f and g both vanish at $\vec{x}_o = \langle x_o, y_o\rangle$. By setting $X(t) = x(t) - x_o$ and $Y(t) = y(t) - y_o$, we can write the linearized system (as defined by equation (5.12) for this two-dimensional system) as

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{pmatrix} f_x(x_o, y_o) & f_y(x_o, y_o) \\ g_x(x_o, y_o) & g_y(x_o, y_o) \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix},$$

where X' and Y' denote the derivatives of X and Y. The partial derivatives of the functions f and g (denoted as f_x, f_y, g_x, and g_y) are evaluated at the equilibrium point (x_o, y_o). The linearization shifts the equilibrium to the point zero in the X, Y coordinates.

As a specific example we revisit the Lotka–Volterra system. Recall the system of equations (5.10) and (5.11)

$$\begin{aligned} u' &= f(u,v) := u(1-v), \\ v' &= g(u,v) := \alpha v(u-1). \end{aligned}$$

An equilibrium point is found at $u_o = v_o = 1$ as before, and as above we set $U = u - 1$ and $V = v - 1$. The linearized system is then

$$\begin{pmatrix} U' \\ V' \end{pmatrix} = \begin{pmatrix} f_u(u_o, v_o) & f_v(u_o, v_o) \\ g_u(u_o, v_o) & g_v(u_o, v_o) \end{pmatrix} \begin{pmatrix} U \\ V \end{pmatrix}.$$

Substituting the partial derivatives of f and g into the matrix gives

$$\begin{pmatrix} U' \\ V' \end{pmatrix} = \begin{pmatrix} 1 - v\big|_{(1,1)} & -u\big|_{(1,1)} \\ \alpha v\big|_{(1,1)} & \alpha(u-1)\big|_{(1,1)} \end{pmatrix} \begin{pmatrix} U \\ V \end{pmatrix},$$

and evaluating these at (u_o, v_o) yields

$$\begin{pmatrix} U' \\ V' \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ \alpha & 0 \end{pmatrix} \begin{pmatrix} U \\ V \end{pmatrix},$$

the linearized form of the Lotka–Volterra model about the equilibrium point $(1,1)$. (Writing this as a system of equations yields $U' = -V$ and $V' = \alpha U$, consistent with our results from Section 5.2.2.)

It must be emphasized that the linearized system is a different system than the original system, and it can provide qualitative information about the latter only in small neighbourhoods of the equilibrium point. Before we explore this further, we justify the approximation process using a Taylor expansion.

5.3.1. Justification of the Linearization. This linear approximation method may be justified using Taylor's theorem. If the system is at equilibrium at $\vec{x} = \vec{x}_o$, then from $\vec{X}(t) = \vec{x}(t) - \vec{x}_o$ it is apparent that $\vec{x}(t) = \vec{X}(t) + \vec{x}_o$. Therefore, using the fact that $\vec{X} + \vec{x}_o - \vec{x}_o = \vec{X}$, the vector field $\vec{F}(\vec{x}) = \vec{F}(\vec{X} + \vec{x}_o)$ will be given by

$$\begin{aligned} \vec{F}(\vec{X} + \vec{x}_o) &= \vec{F}(\vec{x}_o) - \vec{F}(\vec{x}_o) + \vec{F}(\vec{X} + \vec{x}_o) \\ &= \vec{F}(\vec{x}_o) + \int_0^1 \frac{d}{d\tau} \vec{F}(\tau \vec{X} + \vec{x}_o)\, d\tau. \end{aligned} \tag{5.13}$$

Inside the integral the chain rule yields

$$\frac{d}{d\tau}\vec{F}(\tau\vec{X}+\vec{x}_o) = D\vec{F}(\tau\vec{X}+\vec{x}_o)\vec{X},$$

where D is the matrix of partial derivatives. Expanding this about the point $\tau = 0$ and using Taylor's theorem, we have

$$\begin{aligned}\frac{d}{d\tau}\vec{F}(\tau\vec{X}+\vec{x}_o) &= D\vec{F}(\vec{x}_o)\vec{X} + \frac{d^2}{d\tau^2}\vec{F}(\tau\vec{X}+\vec{x}_o)\Big|_{\tau=0}\tau \\ &\quad + \frac{1}{2}\frac{d^3}{d\tau^3}\vec{F}(\tau\vec{X}+\vec{x}_o)\Big|_{\tau=0}\tau^2 + \cdots\end{aligned}$$

so that

$$\frac{d}{d\tau}\vec{F}(\tau\vec{X}+\vec{x}_o) \approx D\vec{F}(\vec{x}_o)\vec{X}$$

since the higher order terms will be small relative to $D\vec{F}(\vec{x}_o)\vec{X}$ when $\|\vec{X}\|$ is small. As $\vec{F}(\vec{x}_o) = \vec{0}$, equation (5.13) becomes

$$\vec{F}(\vec{X}+\vec{x}_o) \simeq \int_0^1 D\vec{F}(\vec{x}_o)\vec{X}\,d\tau = D\vec{F}(\vec{x}_o)\vec{X},$$

which provides the desired explanation for the linear approximation method.

Example 5.4. Consider the vector $\vec{x}(t)$ such that

$$\begin{aligned}x' &= (x-1)\sqrt{y} = f(x,y), \\ y' &= (y-4)e^x = g(x,y).\end{aligned}$$

Setting $f(x,y) = g(x,y) = 0$ yields $(x_o, y_o) = (1,4)$ as an equilibrium point; therefore, we let $X = x-1$ and $Y = y-4$. Consequently, the linearized system in matrix form is given by

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{pmatrix} \sqrt{y}\Big|_{(1,4)} & \dfrac{x-1}{2\sqrt{y}}\Big|_{(1,4)} \\ (y-4)e^x\Big|_{(1,4)} & e^x\Big|_{(1,4)} \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}.$$

Evaluating this, we have the result

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & e \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix},$$

which may also be expressed as the system of equations $X' = 2X$ and $Y' = eY$.

5.4. Solving Linear Systems

The linearization procedure yields a system of linear ordinary differential equations with constant coefficients. In matrix form, such systems can be written as

$$\vec{x}' = A\vec{x}, \tag{5.14}$$

where A is a $n \times n$ constant real-valued coefficient matrix if we have n dependent variables. We will be mostly concerned with the case $n = 2$.

Linear systems with a constant coefficient matrix are solvable and can be used to classify the behaviour of the original system near the equilibrium point (x_o, y_o). (For the linearized system which we have been working with, $(0, 0)$ will be the equilibrium point since under the transformation $X = x - x_o$ and $Y = y - y_o$ we have $(x_o, y_o) \mapsto (0, 0)$.)

To solve (5.14) we make an ansatz $\vec{x}(t) = \vec{v}e^{\lambda t}$; therefore,

$$\vec{x}'(t) = \lambda \vec{v} e^{\lambda t}$$

and

$$A\vec{x}(t) = A\vec{v}e^{\lambda t},$$

which implies that for (5.14) to hold, we have to have $A\vec{v} = \lambda\vec{v}$. So $\vec{v}$ must be an eigenvector of A associated with the eigenvalue λ. This leads to a universal solution procedure.

We only discuss the case of two equations, and assume that the matrix A has two distinct eigenvalues. The analysis generalizes to arbitrary numbers of equations, and to the case of repeated eigenvalues. We refer to intermediate texts on ordinary differential equations for the general solution procedure.

5.4.1. The Case of Two Distinct Eigenvalues. Assume that for the linear system $\vec{x}' = A\vec{x}$, there exist two eigenvalues $\lambda_1 \neq \lambda_2$ for the 2×2 matrix A. It is possible that these eigenvalues are complex conjugates. We prove that then there exist two linearly independent, possibly complex, eigenvectors $\vec{v}_1$ and $\vec{v}_2$ associated with λ_1 and λ_2. (Recall that two vectors $\vec{v}_1$ and $\vec{v}_2$ are said to be linearly independent if the identity

$$a\vec{v}_1 + b\vec{v}_2 = \vec{0}$$

holds only for $a = b = 0$.)

Theorem 5.5. *Suppose $\vec{v}_1$ and $\vec{v}_2$ are eigenvectors belonging to distinct eigenvalues λ_1 and λ_2. Then $\vec{v}_1$ and $\vec{v}_2$ are linearly independent.*

Proof. Assume that

$$a\vec{v}_1 + b\vec{v}_2 = \vec{0}.$$

Applying the matrix A, it is clear that

$$a\lambda_1\vec{v}_1 + b\lambda_2\vec{v}_2 = \vec{0},$$

and substituting $b\vec{v}_2 = -a\vec{v}_1$ gives

$$a(\lambda_1 - \lambda_2)\vec{v}_1 = \vec{0}.$$

As $\lambda_1 \neq \lambda_2$ and $\vec{v}_1 \neq 0$, this implies that $a = 0$. The fact that $b = 0$ follows immediately. □

If there are two different eigenvalues, the general solution for systems of the form (5.14) is then

$$\vec{x}(t) = C_1 e^{\lambda_1 t}\vec{v}_1 + C_2 e^{\lambda_2 t}\vec{v}_2, \tag{5.15}$$

where C_1 and C_2 are constants. It is transparent from (5.15) that the behaviour of the solution as $t \to \infty$ is determined by the real parts of the eigenvalues λ_1 and λ_2. For example, if both real parts are negative, all solutions will go to zero. We defer a discussion of the various cases until later.

5.4.2. Determining the Eigenvalues and Eigenvectors. We return to the system represented by (5.14). In order to find the eigenvalues λ_1 and λ_2 of the constant coefficient matrix A, we denote the determinant of A as $|A|$ and set

$$|A - \lambda I| = 0, \tag{5.16}$$

where I is the identity matrix. The eigenvalues are found by solving the resulting quadratic equation for λ (or, for more than two dependent variables, finding the roots of the polynomial $|A - \lambda I|$, known as the *characteristic polynomial* of A).

Example 5.6. Consider the system

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{pmatrix} -3 & 1 \\ -2 & -1 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}, \tag{5.17}$$

where A is the constant coefficient matrix. To find the eigenvalues of this system we set $|A - \lambda I| = 0$, which gives

$$\begin{vmatrix} -3-\lambda & 1 \\ -2 & -1-\lambda \end{vmatrix} = 0.$$

This yields $\lambda^2 + 4\lambda + 5 = 0$; using the quadratic formula we find the complex conjugate eigenvalues $\lambda_1 = -2 + i$ and $\lambda_2 = -2 - i$. The implications of this with respect to the behaviour of the system will be discussed later.

Having found the eigenvalues of the matrix A, we next compute the associated eigenvectors $\vec{v}_1$ and $\vec{v}_2$. They are found by solving

$$(A - \lambda_i I)\,\vec{v}_i = \vec{0} \tag{5.18}$$

for $\vec{v}_i$; $i = 1, 2$. Let us return to the system presented in Example 5.6.

Example 5.7. To find the eigenvector $\vec{v}_1 = (r_1, s_1)^T$ associated with $\lambda_1 = -2 + i$ for the system represented by (5.17), we make the appropriate substitutions into (5.18) which gives

$$\begin{pmatrix} -1-i & 1 \\ -2 & 1-i \end{pmatrix} \begin{pmatrix} r_1 \\ s_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The two equations are identical (as can be seen by multiplying row one by $1 - i$). If we choose $r_1 = 1$, we find the eigenvector

$$\vec{v}_1 = \begin{pmatrix} r_1 \\ s_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1+i \end{pmatrix}.$$

The eigenvector associated with λ_2 can be found in a similar manner or by noting that since λ_2 is the complex conjugate of λ_1 and A is real, $\vec{v}_2$ will be the complex conjugate of $\vec{v}_1$. Thus

$$\vec{v}_2 = \begin{pmatrix} r_2 \\ s_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1-i \end{pmatrix}.$$

The general solution of Example 5.6 is therefore given by

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = C_1 e^{(-2+i)t} \begin{pmatrix} 1 \\ 1+i \end{pmatrix} + C_2 e^{(-2-i)t} \begin{pmatrix} 1 \\ 1-i \end{pmatrix}. \tag{5.19}$$

To eliminate the complex numbers and find the real solutions, let $C_1 = (c_1 + ic_2)/2$ and $C_2 = (c_1 - ic_2)/2$, where c_1, c_2 are arbitrary but real constants. Equivalently, just keep C_1 in (5.19) complex, set $C_2 = 0$, and take the real part of the solution given in (5.19). This real part is still a solution.

Focusing on $x(t)$, this process yields

$$\begin{aligned} x(t) &= \frac{c_1 + ic_2}{2} e^{(-2+i)t} + \frac{c_1 - ic_2}{2} e^{(-2-i)t} \\ &= e^{-2t} \left[\frac{c_1}{2}(e^{it} + e^{-it}) + \frac{ic_2}{2}(e^{it} - e^{-it}) \right] \end{aligned}$$

or

$$x(t) = e^{-2t}(c_1 \cos t - c_2 \sin t);$$

$y(t)$ is found in a similar manner. Summarizing, we obtain from (5.19) the real solutions

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = e^{-2t} \cos t \begin{pmatrix} c_1 \\ -(c_1 - c_2) \end{pmatrix} + e^{-2t} \sin t \begin{pmatrix} -c_2 \\ -(c_1 + c_2) \end{pmatrix}.$$

The constants c_1 and c_2 are determined by the initial conditions.

5.5. Classification of the Equilibria

For linear systems with general solutions of the type

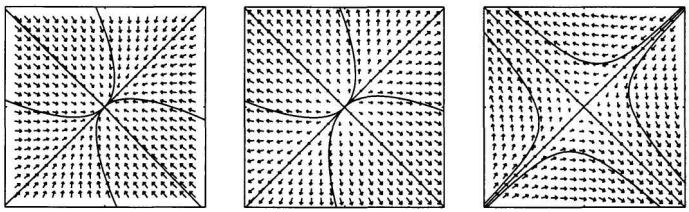
$$\vec{x}(t) = c_1 e^{\lambda_1 t} \vec{v}_1 + c_2 e^{\lambda_2 t} \vec{v}_2,$$

the eigenvalues do in fact determine the behaviour of the system near the zero equilibrium point; in particular, the real parts of these eigenvalues reveal the long term behaviour. Equilibrium points of general nonlinear systems can in this way be classified as stable and unstable nodes, saddle points, attractive or repulsive spirals, and centres. Among other factors, this classification will depend on whether the eigenvalues are real or complex conjugates.

5.5.1. Real vs. Complex Conjugate Eigenvalues. We begin by examining linear systems with two real eigenvalues.

Case 1: Real Eigenvalues: $\lambda_1, \lambda_2 < 0$

The equilibrium is a *stable node.* As λ_1 and λ_2 are both negative, the exponentials $\exp(-\lambda_i t)$ in the general solution



(a) (b) (c)

Figure 4. (a) Case 1: A stable node. This occurs when $\lambda_1, \lambda_2 < 0$. (b) Case 2: Here $\lambda_1, \lambda_2 > 0$, which is an unstable node. (c) Case 3: When $\lambda_1 < 0 < \lambda_2$, or vice versa, the equilibrium is a saddle point. Note that the direction field is given by the arrows in the background of each graph.

will force all solutions of the system to decay to 0. However, as the eigenvalues are not necessarily equal, this decay will occur at different rates in different directions. Figure 4(a) displays a typical phase portrait of such systems.

Case 2: Real Eigenvalues: $\lambda_1, \lambda_2 > 0$

The equilibrium is an *unstable node.* As shown by Figure 4(b), solutions of the system will grow to infinity since, in this case, the exponentials will be to a positive power. The eigenvector associated with the eigenvalue of greater absolute value will exert a stronger "push" on the trajectory.

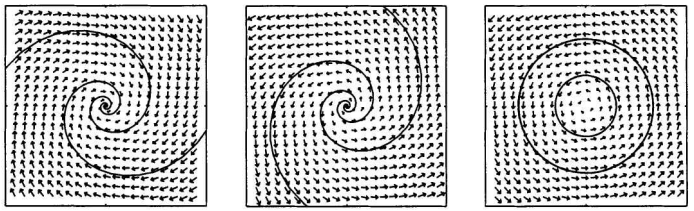
Case 3: Real Eigenvalues: $\lambda_1 < 0 < \lambda_2$

The equilibrium is a *saddle point* near the equilibrium. The system decays toward the origin on multiples of the eigenvector associated with the negative eigenvalue and is pulled to infinity otherwise. Figure 4(c) depicts a saddle point. Saddle points are unstable equilibria.

The remaining cases, namely stable and unstable spirals, and centres occur in systems with complex conjugate eigenvalues. The eigenvalues here are $\lambda_1 = \alpha + \beta i$ and $\lambda_2 = \alpha - \beta i$, with real α and β, which determine the qualitative behaviour.

Case 4: Complex Conjugate Eigenvalues: $\alpha < 0$

When the real part of the eigenvalue, α, is negative the



(a) (b) (c)

Figure 5. (a) Case 4: Because $\alpha < 0$ for $\lambda = \alpha + \beta i$, the equilibrium is a stable spiral. (b) Case 5: Here $\alpha > 0$, which is an unstable spiral. (c) Case 6: The graph depicts a centre which occurs when $\alpha = 0$. The arrows in the background indicate the direction field. The direction of the rotation of this vector field will be determined by β. In the first figure, $\beta < 0$ and thus the rotation is clockwise. In the latter two figures, $\beta > 0$ so the rotation is counterclockwise as shown.

equilibrium is a *stable spiral* with solutions that decay towards the origin. Whether these solutions spiral clockwise or counterclockwise will depend on the value of β. Figure 5(a) displays a stable spiral where $\beta < 0$, and the phase path spirals clockwise.

Case 5: Complex Conjugate Eigenvalues: $\alpha > 0$

Systems which have complex conjugate eigenvalues, where $\alpha > 0$, display *unstable spirals*. As shown by Figure 5(b), the solutions will spiral outwards away from the equilibrium. The motion is counterclockwise in this figure since $\beta > 0$ for the system represented.

Case 6: Complex Conjugate Eigenvalues: $\alpha = 0$

Systems with eigenvalues of the form $\lambda = \pm i\beta$ are called *centres* with solutions that form concentric ellipses shown in Figure 5(c). As before, β will determine the direction of the motion.

Remark 5.8. The attentive reader will have noticed that we have ignored the case where $\lambda_1 = \lambda_2 \in \mathbb{R}$. This case requires a slightly more sophisticated treatment involving the concepts of generalized eigenvectors and the Jordan normal form. We refer to standard intermediate texts on ordinary differential equations.

5.6. The Phase Paths

The preceding sections provide useful guidelines on how various linear and linearized systems will behave. The restriction to two different eigenvalues also allows a drastic simplification of the system by employing suitable linear transformations. This also enables us to find explicit representations of the phase paths.

To begin, consider systems with two real eigenvalues. Let $\vec{u}(t) = S\vec{x}(t)$ where S is a constant invertible matrix. Taking the derivative of $\vec{u}(t)$ and using equation (5.14) gives

$$\vec{u}'(t) = S\vec{x}'(t) = SA\vec{x}(t),$$

where we recall that A is a constant coefficient matrix with real eigenvalues $\lambda_1 \neq \lambda_2$. Solving the original system $\vec{u}(t) = S\vec{x}(t)$ for $\vec{x}$ and substituting this into the above equation yields

$$\vec{u}'(t) = SAS^{-1}\vec{u}(t), \tag{5.20}$$

where SAS^{-1} is a new matrix.

The matrix S should now be chosen so that the resulting system is simpler than the original system. To do this, let $T = (\vec{v}_1, \vec{v}_2)$ where $\vec{v}_1$ and $\vec{v}_2$, the eigenvectors of A, are column vectors. (Recall from Section 5.4.1 that these vectors are independent.) Multiplying this matrix by A we have that

$$AT = \begin{pmatrix} \lambda_1\vec{v}_1, & \lambda_2\vec{v}_2 \end{pmatrix} = T\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = T\Lambda, \tag{5.21}$$

where Λ denotes the matrix which has eigenvalues λ_1 and λ_2 in the diagonal. Because T has independent columns, the determinant of T is not zero. This implies that T is invertible; therefore, $\Lambda = T^{-1}AT$ and indeed the new system will be simpler if we choose $S = T^{-1}$.

Substituting this choice for S into equation (5.20) and using the result from (5.21) gives

$$\vec{u}'(t) = T^{-1}AT\vec{u}(t) = T^{-1}T\Lambda\vec{u}(t) = \Lambda\vec{u}(t).$$

Therefore, $\vec{u}(t)$ satisfies $\vec{v}'(t) = \Lambda\vec{u}(t)$ and is expressed as

$$u_1' = \lambda_1 u_1, \tag{5.22}$$

$$u_2' = \lambda_2 u_2, \tag{5.23}$$

so the equations are effectively decoupled. Dividing (5.23) by (5.22) yields

$$\frac{du_2}{du_1} = \frac{\lambda_2}{\lambda_1}\frac{u_2}{u_1}$$

or

$$\int \frac{du_2}{u_2} = \frac{\lambda_2}{\lambda_1}\int \frac{du_1}{u_1}.$$

Evaluating the integrals we have

$$\ln|u_2| = \frac{\lambda_2}{\lambda_1}\ln|u_1| + C,$$

where C is some constant. Solving for u_2 gives the result

$$u_2 = \pm C_2 |u_1|^{\lambda_2/\lambda_1}, \tag{5.24}$$

which is the equation for the phase path of systems if both eigenvalues are real.

Using equation (5.24), we are now better able to interpret the phase portraits. In fact, the fraction λ_2/λ_1 will determine the behaviour of the system. When $\lambda_2/\lambda_1 > 0$ (as in cases 1 and 2 in Section 5.5.1), the system will have a stable or unstable node. When $\lambda_2/\lambda_1 < 0$ (as in case 3), the system will have a saddle point.

5.6.1. Phase Path Equations when Eigenvalues are Complex. We move on to systems with complex conjugate eigenvalues (the cases where $\lambda_1 = \alpha + \beta i$). As with the real case, it is possible to find a matrix S (in this case complex) such that $\vec{u}(t) = S\vec{x}(t)$, which satisfies

$$u_1'(t) = \alpha u_1(t) - \beta u_2(t), \tag{5.25}$$

$$u_2'(t) = \beta u_1(t) + \alpha u_2(t). \tag{5.26}$$

To solve this, let $z(t) = u_1(t) + iu_2(t)$, with polar representation $z(t) = r(t)e^{i\theta(t)}$. Differentiating $z(t)$ and using both (5.25) and (5.26) yields

$$z'(t) = u_1'(t) + iu_2'(t) = \alpha u_1(t) - \beta u_2(t) + i\beta u_1(t) + i\alpha u_2(t).$$

After grouping like terms, we find that

$$z'(t) = \alpha\left[u_1(t) + iu_2(t)\right] + i\beta\left[u_1(t) + iu_2(t)\right] = (\alpha + i\beta)z(t). \tag{5.27}$$

Differentiating the polar representation of $z(t)$ and using (5.27) we have

$$z'(t) = r'(t)e^{i\theta(t)} + ir(t)\theta'(t)e^{i\theta(t)} = (\alpha + i\beta)z(t),$$

where $z(t)$ has polar form $z(t) = r(t)e^{i\theta(t)}$. This implies that

$$r' + ir\theta' = \alpha r + i\beta r,$$

therefore, equating real and imaginary parts, $r'(t) = \alpha r(t)$ and $\theta'(t) = \beta$. So α determines the stability of the system, and β indicates the direction of the rotation as noted in cases 4, 5, and 6.

Example 5.9. The linearized Lotka–Volterra system is of the above type. Recall that

$$\begin{pmatrix} U' \\ V' \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ \alpha_1 & 0 \end{pmatrix} \begin{pmatrix} U \\ V \end{pmatrix}$$

(we have called the parameter in the system α_1 to distinguish it from the α in the general theory). If we set $U_1(t) = U(t)$ and $V_1(t) = \frac{1}{\alpha_1}V(t)$, the system for (U_1, V_1) will be of the above type with $\alpha = 0, \beta = \alpha_1$. By our classification scheme we see again that the equilibrium point is a centre.

5.6.2. Structurally Stable and Unstable Systems: A Criticism. Phase diagrams in the neighbourhood of a steady state where $\operatorname{Re}\lambda_i \neq 0 \forall i$ have the property that, with respect to small perturbations in the system, they are stable; i.e., if we add a small perturbation to the right-hand side of the system,

$$\vec{x}' = \vec{F}(\vec{x}) + \epsilon\vec{g}(\vec{x}), \qquad 0 < \epsilon \ll 1,$$

the type of equilibrium will remain unchanged for a small enough ϵ even though the actual equilibrium point may have shifted slightly. Figure 6 illustrates the effect of a slight shift in the equilibrium of a stable system. A major criticism of the Lotka–Volterra model is that the phase diagram associated with a centre is structurally unstable: as the complex conjugate pair of eigenvalues of the linearized system will move with the perturbation, small perturbations will typically change the centre to an attractive or repulsive spiral. It is therefore unrealistic to expect that a biological system, which cannot persist in isolation from its environment, is accurately described by a dynamical system with a centre.

In Exploration A (see Section 5.8) we present a natural modification of the Lotka–Volterra system addressing this weakness.

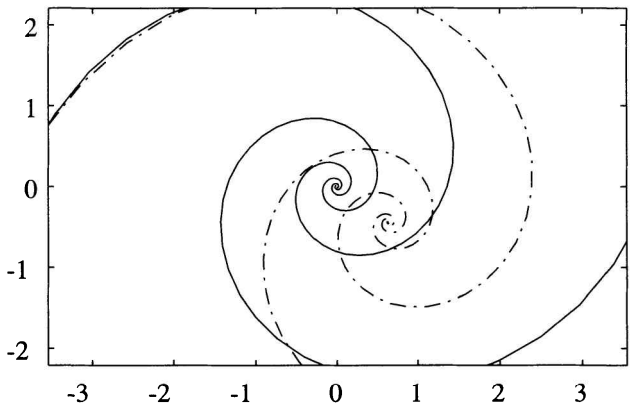


Figure 6. Stable systems such as this stable spiral are not significantly affected by a slight shift in the equilibrium point.

5.7. Multiple Species

The Lotka–Volterra model may readily be extended to include multiple species which all interact with one another within the system. For example, suppose we examine the system with four species: N_1, N_2, N_3, and N_4 such that N_2 preys on N_1 and N_3 preys on both N_1 and N_2. Assume that N_4, in this instance, has a symbiotic relationship with N_2 and is neutral with respect to the other two. The resulting system of equations would look like

$$\begin{aligned} N_1' &= c_1 N_1 \left(1 - \frac{N_1}{K}\right) - d_1 N_1 N_2 - e_1 N_1 N_3, \\ N_2' &= c_2 N_1 N_2 - d_2 N_2 + e_2 N_2 N_4 - f_2 N_2 N_3, \\ N_3' &= c_3 N_3 N_1 + d_3 N_3 N_2 - e_3 N_3, \\ N_4' &= c_4 N_2 N_4 - d_4 N_4. \end{aligned}$$

In the above system, K is the environmental carrying capacity and the c_i, d_i, e_i, and f_i are real valued nonnegative constants.

5.8. Exploration A: Structural Stability

5.8.1. Improved Predator-Prey Models. In this chapter we saw that predator-prey populations can and do exhibit oscillatory behaviour. An observed example of such oscillations are the fluctuations of the lynx and snowshoe hare populations in the northern coniferous forests of the United States, whose populations have been observed to oscillate with an approximate period of ten years [**K**]. However, this example is not consistent with the Lotka–Volterra model, because the peaks of the lynx populations seem to always precede the peaks of the hare populations, while in the Lotka–Volterra model the predator oscillation always lags behind the prey population.

Another unrealistic feature of the Lotka–Volterra model is that it is structurally unstable in the sense that small perturbations will change the character of the equilibrium. We discuss this in some detail.

5.8.2. The Lotka–Volterra Model: Its Limitations. As before let $N(t)$ and $P(t)$ represent the prey and predator populations respectively. We review the model: the interaction between the two populations is given by the system of equations

$$N' = N(a - bP), \tag{5.28}$$

$$P' = P(-d + cN), \tag{5.29}$$

where a, b, c, and d are positive valued constants. There is a steady state solution at the equilibrium point $(N, P) = (d/c, a/b)$. In the absence of outside interference, the populations will remain steady at this equilibrium, while other initial values will lead to sustained oscillations. In particular, the equilibrium is stable but not asymptotically stable.

Although this system of nonlinear differential equations has no closed form solution, it is easily solved numerically. Figure 7 displays the behaviour of this system when $a = 4$, $b = 2$, $c = 3/2$, and $d = 3$. In particular the closed phase paths imply that both populations oscillate with the same period.

It is clear that this model is exceedingly simplistic. In this exploration we investigate what happens if we make a more realistic

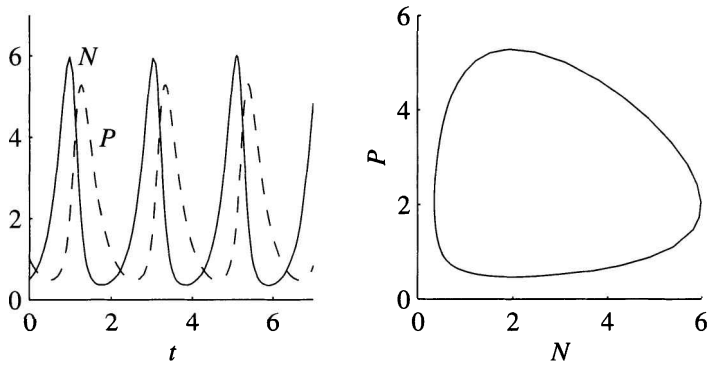


Figure 7. The individual components and phase trajectory for the Lotka–Volterra model with the parameter set $a = 4$, $b = 2$, $c = 3/2$, and $d = 3$.

assumption regarding the growth rate of the prey population. Recall that in the absence of a predatory influence, $P(t)$ grows exponentially without bounds, a clearly unrealistic model.

5.8.3. The Logistic Growth Law: Revising the Model. In reality, populations do not grow without bound, but have a finite carrying capacity. This capacity reflects limiting factors such as disease and the availability of resources which necessarily keep a population at a level which can be supported by its environment. To include this property, the Lotka–Volterra system requires some adjustment.

To this end, suppose that the prey population will in fact obey the logistic growth law when undisturbed by predators [**K**]. In this scenario, an undisturbed prey population, $N(t)$, would now be described by the equation

$$\frac{dN}{dt} = aN\,(1 - \epsilon N)\,, \tag{5.30}$$

where $\epsilon = 1/K$ is a small positive parameter which prevents $N(t)$ from growing without bounds. The number K is the carrying capacity of the environment. By separation of variables it is easy to solve (5.30). The solution is

$$N(t) = \frac{N(0)\,e^{at}}{1 + \epsilon\,N(0)\,(e^{at} - 1)}.$$

As $t \to \infty$, clearly $N(t) \to K$.

With this established, we replace the original model (5.28–5.29) with the revised system

$$N' = N(a - bP - \epsilon aN), \tag{5.31}$$

$$P' = P(-d + cN). \tag{5.32}$$

In most cases, $\epsilon \ll 1$ since the maximum possible population of a prey species tends to be large. Furthermore, the size of the prey population at equilibrium will certainly not exceed the maximum prey population; therefore, it is reasonable to assume that $\epsilon \ll c/d$.

5.8.3.1. *The Qualitative Behaviour of the New System.* We next determine the nature of the equilibrium point. The equilibria are found (as discussed in the text) by setting both (5.31) and (5.32) to zero and solving for N and P. This yields two equilibria points: $(0, 0)$ and, of greater interest, the nontrivial point

$$(N_o, P_o) = \left(\frac{d}{c}, \frac{a}{bc}(c - \epsilon d)\right).$$

The linearized system in matrix form about the equilibrium point of interest is given by

$$\begin{pmatrix} N' \\ P' \end{pmatrix} = \begin{pmatrix} a - bP - 2\epsilon aN\big|_{(N_o,P_o)} & -bN\big|_{(N_o,P_o)} \\ cP\big|_{(N_o,P_o)} & -d + cN\big|_{(N_o,P_o)} \end{pmatrix} \begin{pmatrix} N \\ P \end{pmatrix},$$

which simplifies to

$$\begin{pmatrix} N' \\ P' \end{pmatrix} = \begin{pmatrix} -\dfrac{\epsilon ad}{c} & -\dfrac{bd}{c} \\ \dfrac{a}{b}(c - \epsilon d) & 0 \end{pmatrix} \begin{pmatrix} N \\ P \end{pmatrix}.$$

We will abbreviate the constant coefficient matrix by A. Setting $|A - \lambda I| = 0$ to determine the eigenvalues of this linearized system gives

$$\begin{vmatrix} -\dfrac{\epsilon ad}{c} - \lambda & -\dfrac{bd}{c} \\ \dfrac{a}{b}(c - \epsilon d) & -\lambda \end{vmatrix} = \lambda^2 + \frac{\epsilon ad}{c}\lambda + \frac{ad}{c}(c - \epsilon d) = 0,$$

and the eigenvalues are

$$\lambda_{1,2} = -\frac{\epsilon ad}{2c} \pm \sqrt{\left(\frac{\epsilon ad}{2c}\right)^2 - \frac{ad}{c}(c - \epsilon d)}. \tag{5.33}$$

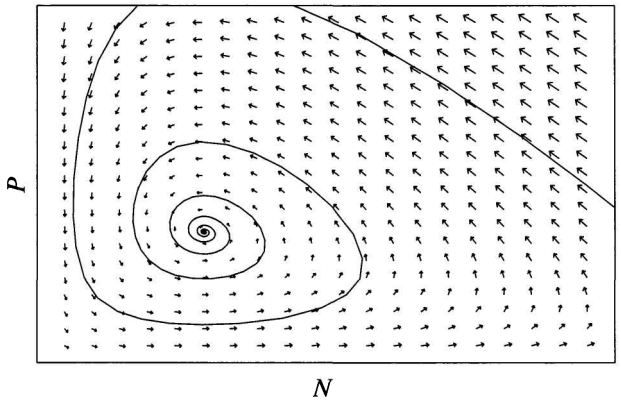


Figure 8. A typical phase trajectory for the revised model is displayed when $a = 4$, $b = 2$, $c = 3/2$, $d = 3$ and $\epsilon = 0.1$. The result is the stable spiral shown.

Consider the discriminant of (5.33). As $\epsilon \ll c/d$ and $(\epsilon ad)^2/(2c)^2 > 0$, we have $c - \epsilon d > 0$ and

$$-\frac{ad}{c}(c - \epsilon d) < 0.$$

We next examine the relative magnitudes of these terms to determine whether or not the eigenvalues are complex. Since $\epsilon \ll c/d$, we can set $\epsilon = \delta c/d$ where $0 < \delta \ll 1$. Now let C_1 and C_2 be positive constants, arbitrary but fixed. Then, if δ is small enough, we have

$$C_1(1 - \delta) \gg C_2\delta^2.$$

Therefore, substituting ϵ with $C_1 = ad$ and $C_2 = a^2/4$ into the above relation, it follows that for small ϵ

$$\frac{ad}{c}(c - \epsilon d) \gg \left(\frac{\epsilon ad}{2c}\right)^2,$$

and thus the discriminant of equation (5.33) is negative. Therefore, λ_1 and λ_2 are complex conjugate eigenvalues with a negative real part. Hence, for small values of $\epsilon > 0$, the system will behave like a stable spiral near this equilibrium point with solutions that decay toward the stable equilibrium as shown in Figure 8. Notice from the figure

that regardless of the initial conditions, all trajectories in the phase plane approach the same equilibrium point. Hence, this modified system does not explain the behaviour of predator-prey populations that exhibit oscillatory behaviour. It does, however, have the advantage of structural stability over the original model.

5.8.4. Structural Stability. The concept of structural stability (as we understand it here for our purposes) refers to the type of behaviour a dynamical system displays in a neighbourhood of a fixed point. Thus structural stability is not a property of the system alone; it is a property that a system will possess (or not possess) in a neighbourhood of a given steady state. This idea certainly generalizes to other types of invariant sets (for example, to limit cycles) but we will not explore such generalizations in detail.

A nonlinear system

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

is said to be *structurally stable* in a neighbourhood of a steady state (a,b) if for bounded and smooth functions $s(x,y), t(x,y)$ and sufficiently small $\epsilon > 0$ the perturbed system

$$\frac{dx}{dt} = f(x,y) + \epsilon\, s(x,y), \qquad \frac{dy}{dt} = g(x,y) + \epsilon\, t(x,y)$$

will possess a steady state (a_ϵ, b_ϵ) near the steady state (a,b), and the type of equilibrium is unchanged for small enough ϵ. By this definition the original Lotka–Volterra model is structurally unstable.

Why is this important for biological models? With any biological model, there are always factors that are ignored, yet influence the behaviour of the system. Hence a model can only be considered realistic if it does not change its character in the presence of small perturbations. The original Lotka–Volterra model breaks down in the face of this dynamic interplay.

The revised Lotka–Volterra model is structurally stable near the relevant equilibrium. Although it fails to properly account for the oscillations observed in predator-prey ecosystems, it does maintain its phase diagram when the system is exposed to random factors that cause small changes to the ecosystem.

5.8.5. Possible Improvements to the Model. Limiting the prey population using the logistic growth model is only one possible improvement that can be made to the Lotka–Volterra model. The original system contains many assumptions which, if called into question, reveal inadequacies: for example, considerations such as a limited food supply and alternate food sources may necessitate the modelling of hierarchical feeding structures involving multiple species.

It is clear that we desire a model whose solutions form cycles that do not collapse when subtly perturbed by random factors. Limit cycles, i.e., periodic solutions which attract nearby states and persist under small perturbations of the system, are the type of cycle one expects to observe in nature. Though stable enough to handle fluctuations, our revised model fails in that it does not produce cycles at all; instead, all trajectories in the phase plane spiral towards the same equilibrium point. While the revised model does not form cycles, the original Lotka–Volterra model could not handle random fluctuations to the system.

Advanced texts on ordinary differential equations explain how limit cycles arise under so-called *Hopf bifurcations.*

5.9. Exploration B: The Lorenz Attractor

The method of analysing a general system of first-order ordinary differential equations by identifying steady states and using local linearization works in any number of dimensions. As an instructive and inspiring example we discuss the Lorenz system of equations, derived by Edward Lorenz 40 years ago in the context of atmospheric science, and famous because of the presence of a strange attractor in its three-dimensional phase space. The equations involve three nonnegative parameters σ, b and $r \geq 1$, and three dependent variables, x, y, z, whose dynamics is driven by the system

$$\begin{aligned} \dot{x} &= \sigma(y - x), \\ \dot{y} &= rx - y - xz, \\ \dot{z} &= xy - bz. \end{aligned} \tag{5.34}$$

Steady solutions are given as the critical points of the vector field on the right and are easily found to be $P_1 = (0, 0, 0)$ and the nontrivial

pair

$$P_2 = \left(\sqrt{b(r-1)}, \sqrt{b(r-1)}, r-1\right),$$
$$P_3 = \left(-\sqrt{b(r-1)}, -\sqrt{b(r-1)}, r-1\right).$$

The Lorenz system is one of the most studied dynamical systems because of the intriguing behaviour of the typical trajectory. The book by Sparrow [**P**] deals exclusively with the Lorenz system and contains everything from explanations of the meaning of the parameters to extensive numerical simulations.

Our goal here is much more modest. We will choose parameters inside the well-known interesting range and analyse the linear stability properties of the three steady states computed earlier. The student is then challenged to explain how these stability properties can be consistent with the typical trajectory approaching the famous Lorenz attractor.

We will fix $\sigma = 10$, $b = 8/3$, and vary r from 1 to, say, about 28. The sample trajectory depicted below in Figure 9 was produced by choosing $r = 28$ and starting with the initial value $Q = (1, 1, 1)$.

Linearization of (5.34) near P_1 produces the system

$$\frac{d}{dt}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} -10 & 10 & 0 \\ r & -1 & 0 \\ 0 & 0 & -\frac{8}{3} \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix}. \tag{5.35}$$

The eigenvalues of the matrix on the right are found from the characteristic polynomial $-(\lambda + 8/3)[\lambda^2 + 11\lambda - 10(r-1)] = 0$, so

$$\lambda_1 = -\frac{8}{3}, \qquad \lambda_{2,3} = -\frac{11}{2} \pm \frac{1}{2}\sqrt{81 + 40r}.$$

Notice that the fixed points P_2 and P_3 are complex valued for $r < 1$ (i.e., they are not really present in phase space until $r > 1$) and all three eigenvalues computed above are negative and real while $r < 1$. In this parameter range the fixed point P_1 is asymptotically stable, there are no other real fixed points, and P_1 is in fact globally attracting, i.e., every trajectory will converge to P_1 (our linear stability analysis does not prove this in full generality, but it is true).

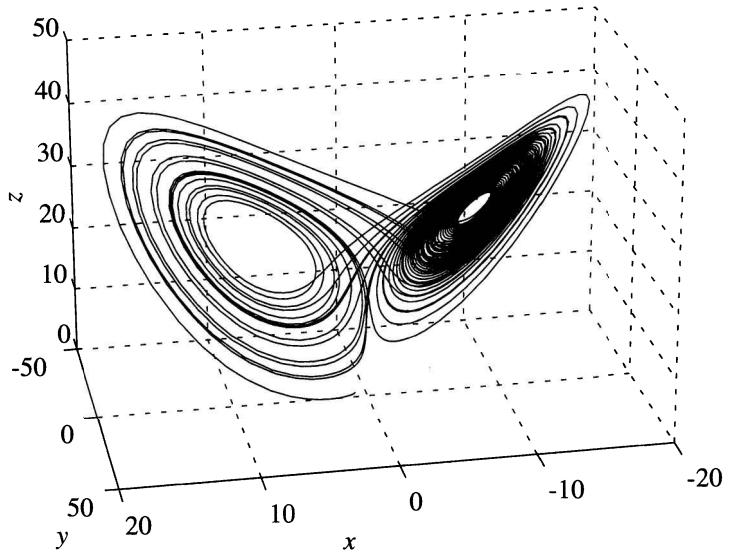


Figure 9. The trajectories of the Lorenz system are obtained by choosing $\sigma = 10, b = 8/3, r = 28$. The saddle between the two "wings of the butterfly" lies at the origin.

As r crosses the threshold $r = 1$, the two additional fixed points P_2 and P_3 emerge (in the language of the fluid dynamics at the basis of the Lorenz system, this is the onset of convective flow). Notice that λ_3 becomes positive while λ_1 and λ_2 remain negative. In the language of dynamical systems theory, this implies that there is a *stable manifold* through P_1, which is a two-dimensional surface immersed in phase space and tangent (at P_1) to the eigenvectors belonging to λ_1 and λ_2.

We next linearize the system about P_2 (linearization about P_3 is completely analogous, so we will not bother with that). At P_2 the linearized system is

(5.36)

$$\frac{d}{dt}\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} -10 & 10 & 0 \\ 1 & -1 & \sqrt{\frac{8}{3}(r-1)} \\ \sqrt{\frac{8}{3}(r-1)} & \sqrt{\frac{8}{3}(r-1)} & -\frac{8}{3} \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix}.$$

(The calculation is left as an exercise; we use u, v, w as symbols for the variables to emphasize that these variables are deviations from the coordinates of the fixed point P_2.)

The characteristic polynomial of the matrix in (5.36) is

$$p(\lambda) = 3\lambda^3 + 41\lambda^2 + 8(r + 10)\lambda + 160(r - 1).$$

Use MAPLE to verify the following assertions concerning the roots of $p(\lambda)$:

(1) For $1 < r < r_1 \approx 1.3456$, there are three real negative roots, and so P_2 is a locally attractive node (a sink). Two of these eigenvalues meet for $r = r_1$.

(2) For $r_1 < r < r_2 \approx 24.737$, there is one real negative eigenvalue, and there is a pair of complex conjugate eigenvalues $\lambda_2 = \overline{\lambda_3}$ with $\operatorname{Re}\lambda_2 < 0$. The phase diagram near P_2 is a stable spiral. Each trajectory will be attracted rapidly towards a plane transversal to the eigenvector associated with λ_1, and then spiral in tangentially to this plane towards P_2.

(3) Finally, as $r > r_2$, we retain the real eigenvalue $\lambda_1 < 0$, but for the complex conjugate pair λ_2, λ_3, we then have $\operatorname{Re}\lambda_2 > 0$, so P_2 is no longer a stable equilibrium (the phase portrait near P_2 is an unstable spiral). In this range (in which our reference value of $r = 28$ resides) there are no stable fixed points left, so the typical trajectory will not end up at a fixed point.

What happens to the typical trajectory? It is easily seen that it cannot run off to infinity. In fact, if we consider the function of the dependent variables

$$V(x, y, z) := rx^2 + \sigma y^2 + \sigma(z - 2r)^2,$$

an elementary calculation (Exercise!) gives that along trajectories

$$\frac{dV}{dt} = -2\sigma\left[rx^2 + y^2 + b(z - r)^2 - br^2\right],$$

and it is not hard to conclude from this that trajectories will stay inside sufficiently large domains $\{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 \leq C\}$ in phase space. (In fact, more is true: there is a sphere of this type

such that every trajectory, no matter haw far from 0 it starts, will eventually enter and stay inside that sphere.)

Finally, the dynamics given by the Lorenz system contracts the three-dimensional volume in phase space, meaning that the volume of a small cube anchored at x, y, z and with side lengths Δx, Δy, Δz, will shrink exponentially fast. In fact, if we denote the vector field on the right-hand side of (5.34) by $\vec{F}(x, y, z)$, then an elementary calculation shows that $\nabla \cdot \vec{F}(x, y, z) = -\sigma - 1 - b$, and in conjunction with the divergence theorem this implies the claimed contraction (in fact, at a very fast rate for our choice of parameters).

So the dynamics must lead us onto a *thin* yet bounded set, and this set is the famous Lorenz attractor, a butterfly-type infinite *string* with fractal cross-sections, delicately suspended between the stable and unstable manifolds associated with the three unstable fixed points P_1, P_2, and P_3. The linear stability analysis we learned and used in the course of this chapter allows us to make statements about the system near each one of these points, but it leaves enough room in phase space for the existence of such a strange object.

Exercises

(1) Show that the system

$$\begin{aligned} N' &= N\,(a - bP)\,, \\ P' &= P\,(-d + cN) \end{aligned}$$

under the substitutions

$$\tau = at, \qquad u(\tau) = \frac{c}{d}N(t), \qquad v(\tau) = \frac{b}{a}P(t),$$

can be written in the form

$$\begin{aligned} \frac{du}{d\tau} &= u(\tau)\,[1 - v(\tau)]\,, \\ \frac{dv}{d\tau} &= \alpha v(\tau)\,[u(\tau) - 1]\,, \end{aligned}$$

where $\alpha = d/a$.

(2) (a) Use MAPLE to solve for x, y, and z in the following system:

$$\begin{aligned} ax + 3y + 3z &= 10, \\ x - y + az &= 2, \\ 3x - 2y + z &= 6. \end{aligned}$$

There are several ways to do this. You can use the `eqns` and `solve` commands; or you can use the `linalg` package. Try larger systems of equations (for your own benefit).

(b) Evaluate the integral

$$\int_0^\infty e^{-x^2} \ln x \, dx.$$

(c) Find the eigenvalues of the matrix

$$\begin{bmatrix} a & b & c \\ b & a & b \\ c & b & a \end{bmatrix}.$$

Hint: Use `linalg` and `eigenvals`.

(3) Show that all phase paths for the system

$$\begin{aligned} x' &= y^5, \\ y' &= -x^3 \end{aligned}$$

are closed curves about $(0, 0)$.

(4) Sketch the phase diagrams for the following systems of differential equations:

(a)

$$\begin{aligned} x' &= x - y, \\ y' &= x + y - 2xy, \end{aligned}$$

(b)

$$x'' + x - x^3 = 0.$$

You may use MAPLE or MATLAB. First find all equilibria points, then sketch the direction field and a few typical phase paths.

Hint: Set $y = x'$ in part (b).

(c) Determine the linear approximations to the system in (a) at all equilibria.

(5) Assume that a function $F(x, y)$ has partial derivatives of second order. Use the chain rule to evaluate

$$\frac{d^2}{dt^2}\left[F(tx, ty)\right]\Big|_{t=0},$$

and conclude that this term is of quadratic order in x and y.

(6) Classify the equilibrium point $(0, 0)$ as being either a saddle point, stable or unstable node, stable or unstable spiral, or centre for the systems of differential equations of the form

$$\frac{d}{dt}\vec{x} = A\vec{x},$$

where

(a)

$$A = \begin{bmatrix} -1 & 3 \\ -2 & 2 \end{bmatrix},$$

(b)

$$A = \begin{bmatrix} -1 & 3 \\ -\frac{1}{2} & 2 \end{bmatrix},$$

(c)

$$A = \begin{bmatrix} -1 & 3 \\ 1 & 2 \end{bmatrix}.$$

(7) Solve the initial value problem

$$\begin{aligned} x' &= 2x + y, \qquad x(0) = 1, \\ y' &= 2x + 3y, \qquad y(0) = 3, \end{aligned}$$

explicitly.

(8) The nonlinear first-order system

$$\begin{aligned} x' &= a_1 x - b_1 xy, \\ y' &= a_2 y - b_2 xy, \end{aligned}$$

where a_1, a_2, b_1, and b_2 are positive constants, models the time evolution of two species competing for the same food supply.

(a) Find the equilibria.

(b) Classify the equilibria (linearize!).

(c) Sketch a phase portrait.

(9) Suppose the predator-prey (fish) species described by the Lotka–Volterra model are subject to selective fishing such that only the prey population is fished at rate $\delta > 0$. Describe the effect of this fishing on the phase diagram.

This page intentionally left blank

Chapter 6

A Control Problem in Fishery Management

Concepts and Tools: Ordinary differential equations, control theory

In this chapter we are concerned with the optimal harvesting of a renewable resource like, say, a fish population. Obviously, one has to take into account matters of sustainability and profitability. We shall talk about a fishery but the analysis applies with little change to other scenarios, for example forestry.

The key questions we are going to address are as follows:

- Given a reasonable growth model for a fish population, what is the maximal sustainable catch?
- Given certain economic parameters, such as interest rates, prices, overhead costs etc., what is the maximal sustainable profit?
- Suppose that a fleet has fished a certain homogeneous fish population $x(t)$ to a level of depletion which necessitates that fishing cease for a time, and assume that the fish population will recover if given the chance. At what time and rate should fishing resume to maximize the long term profit?

6.1. Variables and Parameters

The following table defines the quantities for the next few sections

Variables	
$x(t)$	the size of the fish population at time t
$b(t)$	the number of boats operational at time t
$h(t)$	the harvest rate (units of fish caught per unit time)
Parameters	
c_B	the overhead cost of maintaining one boat per unit of time
n	the mean number of fishers per boat
p	the selling price per unit of fish
w	one fisher's wage per unit of time

In reality, p, w, n, and c_B are not constant; however, we will assume for simplicity that they are constant in some of this chapter. Also, we shall mostly be interested in a fishing fleet of constant size $b(t) := U > 0$. The harvest rate $h(t)$ is usually taken as $h(t) = qx(t)b(t)$, i.e., proportional to the fish population and the number of boats with a proportionality constant q known as *catchability*. For the constant fishing fleet size, therefore,

$$h(t) = qUx(t).$$

6.2. The Logistic Growth Model

We make the simple assumption that the fish population under consideration grows according to logistic growth models if left alone (implicitly this rules out oscillations due to predation, as discussed in the previous chapter). The validity of this assumption is certainly questionable, as the history of the Adriatic fishery described earlier so vividly demonstrates.

By this assumption, the population $x(t)$ will satisfy the ordinary differential equation

$$x'(t) = Rx(t)\left[1 - \frac{x(t)}{K}\right], \tag{6.1}$$

where R is the constant, intrinsic rate of increase in the population and K is the carrying capacity of the environment. This equation is known as the logistic growth model. If we include fishing in this

model (say that the fishing fleet becomes active at a time $s > 0$), the equation (6.1) changes to

$$x'(t) = Rx(t)\left[1 - \frac{x(t)}{K}\right] - qUx(t), \qquad t > s,$$

which is satisfied for times $t > s$. After some rearrangement one has

$$x' = \begin{cases} Rx\left(1 - \dfrac{x}{K}\right) & \text{for } t \leq s, \\ Rx\left(1 - \dfrac{x}{K} - \dfrac{qU}{R}\right) & \text{for } t > s, \end{cases} \tag{6.2}$$

where $s > 0$ is the time when fishing commences. The solution to (6.2) will be denoted as $x_1(t)$ or $x_2(t)$ for $t \leq s$ and $t > s$, respectively.

It is convenient to express the initial fish population as a fraction of the carrying capacity. Hence we write $x(0) = K/N$ where $N > 1$. Solving for $x_1(t)$ we have

$$x_1(t) = \frac{K}{1 + (N-1)e^{-Rt}}. \tag{6.3}$$

(The proof of this is left as an exercise. Equation (6.1) is a separable first-order equation and easily integrated.)

When fishing starts, we aim to fish in such a way as to keep the fish population steady for all of the future (this is sustainability). Therefore, to have a stable fishery when $t > s$, the steady solution x_2 of equation (6.2) is required. To achieve this steady state, x_2 is chosen to satisfy

$$1 - \frac{x_2}{K} - \frac{qU}{R} = 0,$$

which yields the solution

$$x_2 = \left(1 - \frac{qU}{R}\right)K, \tag{6.4}$$

which is valid for $t > s$. If fishing starts at time $s > 0$, the fish population $x(t)$ satisfies (6.1) for $t < s$, and for $t > s$ we insist that $x(t) = x_2$. Of course, $x(t)$ must be continuous at $t = s$ and therefore

$$x_1(s) = \frac{K}{1 + (N-1)e^{-Rs}} = \left(1 - \frac{qU}{R}\right)K = x_2(s).$$

Isolating the exponential gives

$$e^{-Rs} = \frac{1}{N-1}\left(\frac{R}{R-qU} - 1\right),$$

and taking the logarithm of both sides then solving for s yields

$$s = \frac{1}{R}\ln\left[(N-1)\left(\frac{R}{qU} - 1\right)\right], \tag{6.5}$$

which gives s as a function of the fishing fleet U. This formula for s indicates when fishing should be resumed. However, this calculation so far gives no information about maximal sustainable catches or profits. We next turn our attention to these matters.

6.3. Maximizing the Sustainable Catch

6.3.1. Continuous Fishing. It is remarkably easy to find the maximal sustainable catch from the analysis presented in Section 6.2. A steady fish population with a constant fishing fleet is given by expression (6.4) as

$$x_2 = \left(1 - \frac{qU}{R}\right)K.$$

The catch associated with this population level is of course

$$H(U) := qx_2U = qUK\left(1 - \frac{qU}{R}\right),$$

and we wish to choose U such that $H(U)$ becomes maximal. To this end, just compute the derivative

$$H'(U) = qK - \frac{2q^2KU}{R}$$

and solve the equation $H'(U) = 0$. As a result, $U^* = R/2q$, which determines the optimal number of boats. The equilibrium population x_2 is then $x_2 = K/2$, and the catch itself is $H(U^*) = RK/4$.

6.3.2. Seasonal Fishing. Many, if not most renewable natural resources are harvested on a seasonal basis, so that there is a period (typically the better part of a year) where the resource can grow undisturbed, and then there is a (usually rather short) harvesting

season. In this section we present a very elementary method to analyse such scenarios. Standard examples are again fisheries and logging operations. We will continue to use the fishery example.

Let x_n be the fish population in the nth year. In the absence of fishing, the discrete model under consideration is of the type

$$x_n = x_{n-1} + R_1 x_{n-1}\left(1 - \frac{x_{n-1}}{K}\right), \tag{6.6}$$

where x_{n-1} is the size of the fish population from the previous year and the latter half of the equation represents the growth of the population using the logistic growth model given in (6.1). Note that we are now talking about growth rather than growth rates as in the differential equations model. Effectively, the present model arises from equation (6.1) via the standard Euler approximation where we take $R_1 = R\Delta t$. The discrete dynamical system (6.6) is therefore quite a rough approximation of the logistic growth model; however, it offers an acceptable estimate of next season's population in terms of this year's, an estimate which can be used to predict maximal sustainable catch, optimal fleet size, etc.

Our objective is again to maintain a fish population which may be fished with optimal results from year to year; therefore, to maximize the sustainable catch it is only necessary to consider the growth portion of the model. The sustainable catch will be maximized for a population x^* for which the growth term is maximized. Hence we solve

$$\frac{d}{dx}\left[R_1 x\left(1 - \frac{x}{K}\right)\right]\Bigg|_{x^*} = 0,$$

which occurs when $x^* = K/2$. This is identical to the population x_2 computed in the previous section. The maximal possible harvest is then determined by substituting x^* back into the growth model:

$$R_1 x^*\left(1 - \frac{x^*}{K}\right) = \frac{R_1 K}{2}\left(1 - \frac{1}{2}\right) = \frac{R_1 K}{4}.$$

This result is identical to the one from the previous calculation, except that R has been replaced by R_1.

Now it seems natural to expect that the size of the fishing fleet required to harvest the maximal sustainable catch is $R_1/2q_1$, consistent with the result obtained from the differential equations model.

However, this is not quite accurate, because of a subtle point that we have so far ignored.

The point is that we have not yet defined whether x_n is the fish population before or after the fishing season. The two will obviously not be exactly the same. Assuming that the fishing season is relatively short with respect to the closure period, the difference between the two will just equal the catch.

Let us agree that x_n is the fish population in year n *after* the fishing season. We then can write the complete model, including fishing, as

$$\begin{aligned} x_n^* &= x_{n-1} + R_1 x_{n-1}\left(1 - \frac{x_{n-1}}{K}\right), \\ x_n &= x_n^* - q_1 U x_n^*. \end{aligned}$$

Here the constant q_1 should be thought of as $q_1 = q\Delta t$. x_n^* is the fish population in year n before fishing, and x_n is the population after fishing. In particular, if $0 < x_{n-1} < K$, then $x_n^* > x_{n-1}$. So the fish population isn't really steady, and fleet sizes have to take this into account.

Substituting $x^* = K/2$ for x_n and x_{n-1} and equating

$$q_1\left[x^* + R_1 x^*\left(1 - \frac{x^*}{K}\right)\right]U = R_1 x^*\left(1 - \frac{x^*}{K}\right)$$

yields

$$U^* = \frac{R_1}{q_1(2 + R_1)}$$

(verify this as an exercise). Please compare this with the optimal fleet size predicted by continuous fishing.

Note that the arguments in this analysis depend on only three inputs: the size of the fish population at some point in time, the carrying capacity of the environment, and the constant, intrinsic rate of increase in the population. All three can be estimated from current or historical fishing statistics.

The method generalizes to other growth models in which the growth function, g, is a smooth, univalent, nonnegative function of the population. More specifically, that means $g = g(x)$ such that $g(0) = 0$, there is a carrying capacity K such that $g(K) = 0$, and g

assumes a unique maximum at $x_o \in [0, K]$. The model in question is then $x' = g(x)$, and much of the previous analysis carries over. We refer to [**C**] for a more comprehensive and advanced treatment.

6.4. Maximizing the Profit

6.4.1. The Objective Functional. We now include the additional complexity of economic parameters such as wages, prices, overhead costs and interest rates. First we will set up an objective functional which will represent that sustainable profit. Profit is revenue less total costs. The revenue per unit time, denoted as $P_{\text{rev}}(t)$, is $P_{\text{rev}}(t) = ph(t)$. The total cost per boat, c, is given as $c = c_B + nw$; and the cost per unit time, $cb(t)$, is

$$cb(t) = c_B b(t) + nwb(t).$$

This gives the profit per unit time, denoted as $P(t)$, as

$$P(t) = ph(t) - cb(t). \tag{6.7}$$

Let us include an interest (discount) rate $\delta > 0$, which we assume to be constant. Using equation (6.7), the present value of the expected profit at some time t, $E(P(t))$, is given by

$$E(P(t)) = e^{-\delta t}\left[ph(t) - cb(t)\right].$$

Integrating $E(P(t))$, the true return or total profit rate in present dollars, denoted as J, is expressed as

$$J = \int_0^\infty e^{-\delta t}\left[ph(t) - cb(t)\right] dt. \tag{6.8}$$

This is the objective functional to be maximized.

A severe weakness in this model is that the interest rate, δ, is assumed to apply to everything, yet c_B, p, and w are assumed to be constant. In reality, p will probably grow with or even faster than general inflation, and c_B and w are hard to predict since they are influenced by factors such as technological advancements, union negotiations, government policies, and taxation rates. It is possible to include stochastic fluctuations in numerical simulations in order to arrive at more realistic predictions. However, for a first analysis we will proceed with the given unrealistic assumptions.

As before, we take the harvest function as $h(t) = qx(t)b(t)$ where q is catchability. The functional (6.8) becomes

$$J(b) = \int_0^\infty e^{-\delta t} b(t) \left[pqx(t) - c\right] dt, \tag{6.9}$$

which is a functional of the fleet size $b = b(t)$.

It is easy to include a modification where the selling price of fish and the cost per boat, p and c, both increase with time. Suppose for example that they both grow at the same rate α such that

$$p = p(t) = p_o e^{\alpha t}, \tag{6.10}$$

$$c = c(t) = c_o e^{\alpha t}. \tag{6.11}$$

We then obtain an objective functional

$$J(b) = \int_0^\infty e^{(\alpha-\delta)t} b(t) \left[p_o q x(t) - c_0\right] dt,$$

which is well defined if $\delta > \alpha$ (if $\delta \leq \alpha$, the functional will in general no longer be finite). In the rest of this section we will act as if p and c are constant. The case where they grow, as in (6.10) and (6.11) with $\delta > \alpha$, follows by replacing δ by $\delta - \alpha$.

We can now answer the question at what time $t = s > 0$ a fleet of constant size $b(t) = U$ should resume fishing such that $J(b)$ is maximized.

6.4.2. Finding the Optimal Fleet Size. Having found a formula for s (the time when fishing should resume), as a function of U in (6.5), the profit functional (which in (6.9) is a function of $b(\cdot)$, i.e., of U and s) is really a function of U alone.

Substituting $x(t) = x_2(t)$ from (6.4) and s as given in (6.5) reduces (6.9) to

$$J(U) = U \int_{s(U)}^\infty e^{-\delta t} \left[pqK\left(1 - \frac{qU}{R}\right) - c\right] dt,$$

where, for a given fleet size U, everything except the exponent is constant. Integration gives

$$J(U) = U \left[pqK\left(1 - \frac{qU}{R}\right) - c\right] \frac{e^{-\delta s(U)}}{\delta},$$

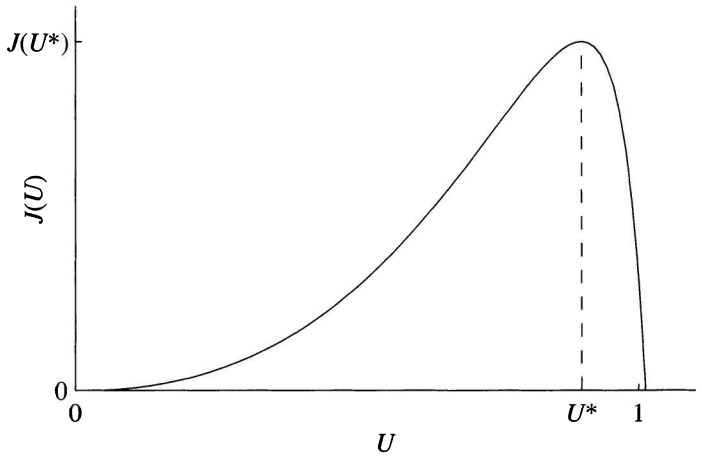


Figure 1. Displayed is the viable range for the solutions of $J'(U) = 0$. The profit function $J(U)$ has a maximum at U^* as shown. Parameters used in this plot are: $p = 2$, $q = 0.5$, $K = 50$, $\delta = 0.7$, $R = 0.55$, $N = 25$ and $c = 4$. As a result, $\beta = 0.08$, the viable range is $0 \leq U \leq 1.1$, and $U^* = 0.8992$.

and after making the substitution $\beta = c/pqK$ this becomes

$$J(U) = \frac{pqKU}{\delta}\left(1 - \frac{qU}{R} - \beta\right)e^{-\delta s(U)}.$$

This profit function will be negative if $1 - qU/R - \beta < 0$. This indicates conditions under which it is no longer profitable to fish. Otherwise, when the profit function is positive, $J(U)$ will behave as shown by Figure 1.

The profit will be at a maximum at the point U^* is shown. To find U^*, we solve $J'(U) = 0$. Choosing the solution U^*, which lies in the viable range displayed by Figure 1, we find that the profit function $J(U)$ will be maximized at the point U^* such that

$$U^* = \frac{R}{4q}\left[3 - \beta + \frac{\delta}{R} - \sqrt{\left(1 + \beta - \frac{\delta}{R}\right)^2 + \frac{8\beta\delta}{R}}\right]$$

(see Exercise 2). Fishing should resume at the time $s^* = s(U^*)$ in order to maximize the profit.

6.4.3. When Costs Rise Faster than Inflation: Maximal Sustainable Profit. Suppose that, with the notation of the previous section, $\alpha \geq \delta$, such that the discount rate is smaller than the rate at which prices and costs rise. The objective functional used in the previous section (the present value of the total profit over all future time for the fishery) will in this case be divergent, so it makes no sense to maximize it. However, we can still try to maximize the rate of profit per unit time, $P(t) = ph(t) - cb(t)$, in an equilibrium situation. This would be called a maximum sustainable profit model. In this case we take $b(t) = U$ (constant), so $h(t) = qUx(t)$, and $P(t) = U[pqx(t) - c]$, with $x(t) = K(1 - qU/R)$ at equilibrium. See expression (6.4). The rate of profit $P(t)$ will therefore be a simple function of U, namely

$$J(U) = U\left[pqK\left(1 - \frac{qU}{R}\right) - c\right], \tag{6.12}$$

which is maximized for $U^* = R(1-\beta)/2q$, with $\beta = c/pqK$, as before.

The associated equilibrium fish population is then

$$x_{\text{eq}} = \frac{K}{2}(1 + \beta), \tag{6.13}$$

the catch rate is

$$h^* = qx_{\text{eq}}U^* = \frac{RK}{4}(1 - \beta^2), \tag{6.14}$$

and the sustained profit rate P^* is

$$P^* = U^*(pqx_{\text{eq}} - c) = \frac{pRK}{4}(1 - \beta)^2. \tag{6.15}$$

If we compute the profit rate for the maximization of the sustainable catch from Section 6.3 (where costs were ignored), we find that

$$P^* = \frac{pRK}{4} - \frac{cR}{2q} = \frac{pRK}{4}(1 - 2\beta), \tag{6.16}$$

where the definition of β has been used.

This is slightly smaller than the value from (6.15), and the equilibrium in (6.13) is slightly larger than the one obtained earlier. So what do these calculations teach us? What would you tell the government about how to run a fishery?

Exercises

(1) Verify that the function

$$x(t) = \frac{K}{1 + (N-1)e^{-Rt}}$$

solves the initial value problem

$$x'(t) = Rx\left(1 - \frac{x}{K}\right), \qquad x(0) = \frac{K}{N}.$$

(2) The purpose of this problem is to verify some of the identities that arose in the fisheries control problem. (This is a rather mechanical problem and a little laborious.)

Consider the function derived in class for $J(U)$ given by

$$J(U) = \frac{pqKU}{\delta}\left(1 - \frac{Uq}{R} - \frac{c}{pqK}\right)e^{-\delta s(U)},$$

where

$$s(U) = \frac{1}{R}\ln\left[(N-1)\left(\frac{R}{qU} - 1\right)\right].$$

Verify that the profit function $J(U)$ is maximized for

$$U^* = \frac{R}{4q}\left[3 - \beta + \frac{\delta}{R} - \sqrt{\left(1 + \beta - \frac{\delta}{R}\right)^2 + \frac{8\beta\delta}{R}}\right],$$

where $\beta = c/pqK$. Thus $s(U^*)$ is the optimal time to start fishing.

Observe that for $\delta = 0$ the expression for U^* reduces to the fleet size obtained in Section 6.4.3.

(3) Verify identities (6.13), (6.14), (6.15), and (6.16).

This page intentionally left blank

Chapter 7

Formal Justice

Concepts and Tools: Functional equations

The concept of formal justice relates to the idea that there should be a relationship between the qualifications of professionals and the compensation received for their work. These qualifications contain positive criteria: level of education, skill, and seniority, and may as well include negative criteria. For example, the number and severity of mistakes committed on the job.

As discussed in sociology texts like, for example, "The Causal Theory of Justice" [**O**], an individual's compensation (their wage or punishment) should bear a reasonable relationship to these qualifications in order to establish a fair wage scale. This chapter introduces and solves functional equations for this problem.

7.1. The Basic Functional Equation

The concept of formal justice in the creation of fair wage scales is not a new one; it was Aristotle who first considered the problem more than 2300 years ago when he suggested proportional justice. The word *proportion* suggests immediately that qualification as well as compensation must be measurable on a numerical scale. This is easy

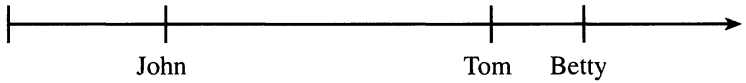


Figure 1. Three professionals are evaluated depending on their qualifications which are measured by a number $x \in \mathbb{R}$.

for the compensation, but a rather delicate task for the qualification. To proceed, it will simply be assumed that it is possible to measure the qualification in question by a number $x \in \mathbb{R}^+$ as indicated by Figure 1. If we consider just one qualification measured by $x \in \mathbb{R}^+$, then the compensation should be a function of x, denoted by $m(x)$. Formal justice should then be expressed in terms of properties of m. For example, Aristotle's proportional justice idea states that the qualification and the wage should be proportional, i.e., $m(x)$ should satisfy the relationship

$$\frac{m(x)}{m(y)} = \frac{x}{y}, \qquad x, y \in \mathbb{R}^+.$$

For $y = 1$ this gives $m(x)/m(1) = x$ and hence $m(x) = m(1)x$. Therefore, $m(x) = cx$ where c is some constant. A professional who has twice the qualifications than another is entitled to twice the wage. This wage scale is limited as it only allows for linear relationships. This is clearly too restrictive.

In [**O**] the concept of formal justice is generalized to mean that $m : \mathbb{R}^+ \to \mathbb{R}^+$ should be a homomorphism with respect to the ratio scale; i.e., that

$$m\left(\frac{x}{y}\right) = \frac{m(x)}{m(y)}, \qquad x, y \in \mathbb{R}^+. \tag{7.1}$$

This functional equation is our basic object of study in this chapter.

7.1.1. Solving the Functional Equation.

Theorem 7.1. *Assume that m is continuous at some point $x_o \in \mathbb{R}^+$ and satisfies*

$$m\left(\frac{x}{y}\right) = \frac{m(x)}{m(y)}, \qquad x, y \in \mathbb{R}^+.$$

Then m is of the form $m(x) = x^p$ for some $p \in \mathbb{R}$.

Proof. It is immediate that $m(x) = x^p$ solves the functional equation. The nontrivial part is to show that these are all the solutions. We do this in some detail.

Step 1: Using the homomorphism, one has

$$m(xy) = m\left(\frac{x}{y^{-1}}\right) = \frac{m(x)}{m\left(\frac{1}{y}\right)} = \frac{m(x)m(y)}{m(1)},$$

which, since

$$m(1) = m\left(\frac{x}{x}\right) = \frac{m(x)}{m(x)} = 1,$$

gives the result

$$m(xy) = m(x)m(y), \qquad x, y \in \mathbb{R}^+. \tag{7.2}$$

Thus this equivalent form of (7.1) may be used.

Step 2: Let $h : \mathbb{R} \to \mathbb{R}$ with $h(t) = \ln[m(e^t)]$. The function $h(t)$ satisfies *Cauchy's functional equation*, i.e., $h(t+s) = h(t)+h(s)$ for all $s, t \in \mathbb{R}$. In fact, from the definition of h,

$$h(t+s) = \ln[m(e^{t+s})] = \ln[m(e^t e^s)], \qquad t, s \in \mathbb{R},$$

and using Step 1 and the rules for logarithms, this becomes

$$h(t+s) = \ln[m(e^t)m(e^s)] = \ln[m(e^t)] + \ln[m(e^s)] = h(t) + h(s)$$

as claimed.

Step 3: We next derive basic properties of h. As $h(t) = h(t+0) = h(t) + h(0)$, it is clear that $h(0) = 0$. Consequently,

$$h(0) = h(t-t) = h[t + (-t)] = h(t) + h(-t) = 0,$$

and thus $-h(t) = h(-t)$. Therefore,

$$h(t-s) = h(t) + h(-s) = h(t) - h(s).$$

Step 4: We establish that $h(nt) = nh(t)$ for all $t \in \mathbb{R}$ and $n \in \mathbb{Z}$. Cauchy's equation for the case where $n = 2$ gives

$$h(2t) = h(t+t) = h(t) + h(t) = 2h(t),$$

and by induction

$$h(n) = h(\underbrace{1 + 1 + \cdots + 1}_{n \text{ times}}) = nh(1), \qquad n \in \mathbb{N}.$$

This argument can be combined with Step 3 to establish the assertion for the cases $n = 0$ and $-n \in \mathbb{N}$ as well.

Step 5: Let $p := h(1)$. We can now show that $h(r) = pr$ for all $r \in \mathbb{Q}$. From Step 4 $h(n) = nh(1) = np$ for all $n \in \mathbb{N}$. Similarly,

$$p = h(1) = h\left(n\frac{1}{n}\right) = nh\left(\frac{1}{n}\right)$$

and, dividing both sides by n, $h(1/n) = p/n$. If $r \in \mathbb{Q}$ so that $r = a/b$ with $a, b \in \mathbb{Z}$ and $b \neq 0$, then

$$h(r) = h\left(\frac{a}{b}\right) = ah\left(\frac{1}{b}\right) = \frac{a}{b}p;$$

that is, $h(r) = pr$ for all $r \in \mathbb{Q}$.

Step 6: The last identity can be proved to hold for all $r \in \mathbb{R}$ provided that h is continuous at every point. To this end, we first show that h is continuous at $t = 0$. As m is continuous at a point $x_o \in \mathbb{R}^+$, $h(t) = \ln(m(t))$ is continuous at $e^t = x_o$. In mathematical terminology,

$$\lim_{\epsilon \to 0} [h(x_o + \epsilon) - h(x_o)] = 0$$

and from Step 3

$$\lim_{\epsilon \to 0} h(x_o + \epsilon - x_o) = \lim_{\epsilon \to 0} h(\epsilon) = 0.$$

As we already know that $h(0) = 0$, it is clear that $h(t)$ is continuous at $t = 0$.

Step 7: The continuity at $t = 0$ is sufficient to show that the function $h(t)$ is continuous at every $t \in \mathbb{R}$. Using Step 3 gives

$$\lim_{\epsilon \to 0} [h(t + \epsilon) - h(t)] = \lim_{\epsilon \to 0} h(t + \epsilon - t) = \lim_{\epsilon \to 0} h(\epsilon) = 0.$$

Thus $h(t)$ is continuous everywhere.

Step 8: Using Step 7, it is now possible to show $h(t) = pt$ for all $t \in \mathbb{R}$. Let $t \in \mathbb{R}$ and choose a sequence $\{t_i\}_{i \in \mathbb{N}}$ such that $t_i \in \mathbb{Q}$ for all $i \in \mathbb{N}$ and that $t_i \to t$ as $i \to \infty$. Using the continuity of $h(t)$ established by Step 7,

$$h(t) = h\left(\lim_{i \to \infty} t_i\right) = \lim_{i \to \infty} h(t_i),$$

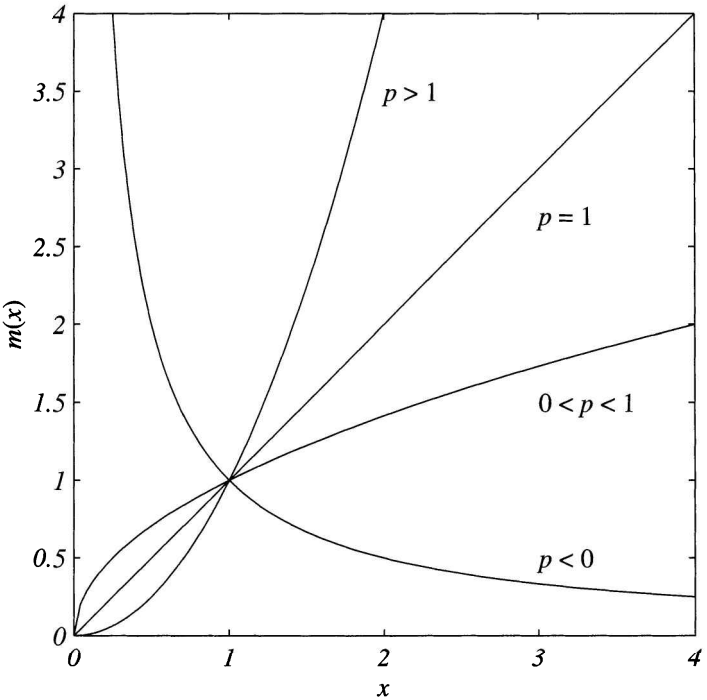


Figure 2. Sołtan's model allows for both linear and nonlinear relationships. Observe that all curves pass through the point $(1, 1)$. Aristotle's model is the case where $p = 1$.

but from Step 5, $h(t_i) = pt_i$ since $t_i \in \mathbb{Q}$, so that

$$h(t) = \lim_{i \to \infty} pt_i = pt.$$

Therefore, $h(t) = pt$ for all $t \in \mathbb{R}$.

Step 9: Having established the form of $h(t)$, we finally compute $m(t)$. Using the definition of $h(t)$ we have

$$h(t) = pt = \ln[m(e^t)]$$

so that $m(e^t) = (e^t)^p$. Thus if we let $x = e^t$, we see that m is of the form $m(x) = x^p$, as required. □

Figure 2 displays compensation scales for this model of formal justice. Aristotle's model is a special case (the case where $p = 1$). All compensation curves must go through the point $(1, 1)$ because $m(1) = 1$ is independent of p. The cases where $p > 0$ are realistic compensation scales for determining salaries based on positive qualifications, such as level of education. If $p < 0$, this implies that salary should decrease with qualification which is highly unrealistic. However, this compensation scale may be of some significance in situations where negative qualifications (e.g., criminal records) are taken into account.

7.1.2. A Criticism. The previous model, although broader in scope than Aristotle's, contains a serious shortcoming. The first indication that a problem exists is that all solutions of the basic functional equation pass through the point $(1, 1)$. This property is clearly not invariant under scale changes for the measurements of either qualification or compensation. For example, consider a scenario where the wage will be paid in units of cents rather than dollars; then $m^*(x) = 100m(x)$. Recalling that $m(x/y) = m(x)/m(y)$ we find that

$$m^*\left(\frac{x}{y}\right) = 100m\left(\frac{x}{y}\right) = 100\frac{m(x)}{m(y)},$$

and using the definition of $m^*(x)$ gives

$$m^*\left(\frac{x}{y}\right) = 100\frac{\frac{m^*(x)}{100}}{\frac{m^*(y)}{100}} = 100\frac{m^*(x)}{m^*(y)}.$$

It is apparent that the basic functional equation is not scale invariant. We have to introduce a generalization.

7.2. Formal Justice: A Generalized Approach

In order to correct the model so that it is scale invariant, we have to revise our definition of formal justice. We will say that a wage scale m satisfies the criterion for formal justice relative to ratio scales if there exists a (scale-dependent) constant $c > 0$ such that

$$m\left(\frac{x}{y}\right) = c\frac{m(x)}{m(y)}, \qquad x, y \in \mathbb{R}^+. \tag{7.3}$$

The addition of a scaling factor c addresses the lack of scale invariance in the basic functional equation. In order to solve (7.3), we first rescale the function $m(x)$ by setting $\overline{m}(x) = c^{-1}m(x)$. Using this definition and relation (7.3) to evaluate $\overline{m}(x/y)$ gives

$$\overline{m}\left(\frac{x}{y}\right) = \frac{1}{c}m\left(\frac{x}{y}\right) = \frac{cm(x)}{cm(y)} = \frac{\frac{1}{c}m(x)}{\frac{1}{c}m(y)},$$

so

$$\overline{m}\left(\frac{x}{y}\right) = \frac{\overline{m}(x)}{\overline{m}(y)}, \qquad x, y \in \mathbb{R}^+.$$

We have, therefore, shown that the rescaled wage function satisfies the basic functional equation, and, therefore, by Theorem 7.1 $\overline{m}(x)$ will be of the type $\overline{m}(x) = x^p$. Equation (7.3) will therefore have solutions

$$m(x) = c\overline{m}(x) = cx^p, \qquad x \in \mathbb{R}^+,$$

where $p \in \mathbb{R}$.

7.2.1. Testing Wage Scales. Our results may be used to determine whether a given wage scale is fair. For example, suppose that three employees A, B, and C work for a company which determines their salaries based on the number of years they have been employed, i.e., based on their seniority. Employee A has been working for twenty-five years and earns $\$2,000,000$ per year. Employee B, after sixteen years, earns $\$60,000$ per year and Employee C makes $\$40,000$, having only worked for three years.

To determine whether these wages satisfy formal justice, we examine how well they fit on the graph of one of our solution curves $m(x) = cx^p$. If there were only two data points (two employees), the parameters c and p could be chosen for a perfect fit. For three or more data points, we determine c and p by a log-log least-squares fit. After that, it depends upon accepted margins who among the employees is considered underpaid or overpaid, respectively. Recalling that $h(t) = \ln m(e^t)$, we have

$$h(t) = \ln m(e^t) = \ln\left[c(e^t)^p\right] = \ln c + pt,$$

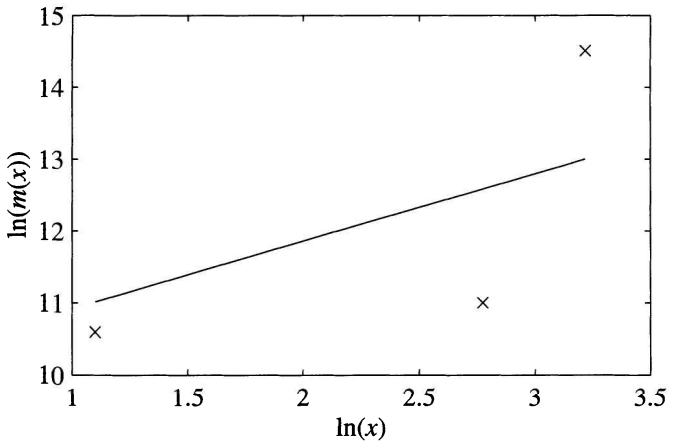


Figure 3. Transformed wage scale described in the text. Here, x is the qualification (in this case seniority) and $m(x)$ is the compensation or salary. Clearly, formal justice is violated.

and, therefore, for a fair wage scale the graph of $h(t)$ is a straight line. A wage distribution will be fair if the converted data are a good fit of the linear least-squares fit found by plotting the logarithms of the salaries versus the logarithms of the (measured) qualifications.

Figure 3 displays the linear least-squares fit curve for employees A, B, and C, graphed on a log-log scale. The figure suggests that employees A and B are significantly overpaid and underpaid, respectively. You might have guessed that.

7.3. Multiple Qualifications

So far we have only modelled the situation where compensation is dependent upon one qualification. In reality salaries and wages are based on multiple qualifications; i.e., $m = m(x_1, \ldots, x_N)$ is a function of N measured compensable properties, and x_i denotes the measurement of the ith qualification.

If we consider the case $N = 2$, then we will say that $m = m(x, y)$ satisfies the criterion of formal justice if our previous formal justice criterion applies with respect to each variable while the other one is

kept fixed. Therefore, we hold one variable constant and look at the function only as a function of one (the other) variable. Define

$$m_{y_o}(x) = m(x,y)|_{y=y_o}, \tag{7.4}$$

$$m_{x_o}(y) = m(x,y)|_{x=x_o}. \tag{7.5}$$

Then formal justice is defined by the validity of the two coupled functional equations

$$m_y\left(\frac{x_1}{x_2}\right) = c_1(y)\frac{m_y(x_1)}{m_y(x_2)}, \tag{7.6}$$

$$m_x\left(\frac{y_1}{y_2}\right) = c_2(x)\frac{m_x(y_1)}{m_x(y_2)}, \tag{7.7}$$

where x and y denote the measured qualifications. The c_i will remain fixed as long as the respective variable is fixed, and thus formal justice will hold, but c_1 and c_2 can really be dependent on y and x, respectively. Under mild continuity assumptions on m (see the solution of the basic functional equations presented earlier) equations (7.6) and (7.7) will have solutions $m_y(x) = c_1(y)x^{p(y)}$ and $m_x(y) = c_2(x)y^{q(x)}$, where $c_1(y)$, $c_2(x)$, $p(y)$, and $q(x)$ are parametric functions to be determined. This is the content of our last theorem in this chapter.

Theorem 7.2. *If $m(x,y)$ is continuous such that $m : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ and satisfies equations* (7.6) *and* (7.7)*, then m is of the form*

$$m(x,y) = cx^p y^q e^{\alpha \ln x \ln y}, \qquad p,q,\alpha \in \mathbb{R},$$

where c is some positive constant. All functions of this type satisfy equations (7.6) *and* (7.7)*.*

Remark 7.3. The continuity assumption can be greatly weakened without weakening the assertion. For example, it is sufficient that m be continuous at one single point in the quadrant of nonnegative values x, y. Even this condition is not necessary. We refer to advanced texts on functional equations (for example, [**A**]) for weaker conditions. However, it must be mentioned that the assertion of the theorem does not remain true if no conditions other than the functional equations are imposed—in that case, the functional equations have (very esoteric) solutions which are not of the given type. These solutions are not *measurable*, unbounded on any interval of positive

length, and altogether ungraphable. The proof of their existence is based on the concept of a *Hamel basis* of the real numbers over the rational numbers.

Proof. We first check that the function $m(x,y) = cx^p y^q e^{\alpha \ln x \ln y}$ satisfies equations (7.6) and (7.7). Fixing x, write

$$m_x(y) = cx^p y^q e^{\alpha \ln x \ln y} = cx^p y^{q+\alpha \ln x}.$$

As x is fixed, this is of the form $m_x(y) = c_2(x)y^{q(x)}$, as required. Similarly, fixing y gives

$$m_y(x) = cx^p y^q e^{\alpha \ln x \ln y} = cy^q x^{p+\alpha \ln y} = c_1(y)x^{p(y)}.$$

Here we show that all continuous solutions of (7.6) and (7.7) are of the stated type. To this end observe that $m_y(x) = m(x,y) = m_x(y)$, so

$$m_y(x) = c_1(y)x^{p(y)} = c_2(x)y^{q(x)} = m_x(y).$$

Taking the logarithm of both sides yields

$$\ln c_1(y) + p(y)\ln x = \ln c_2(x) + q(x)\ln y, \qquad x, y > 0, \tag{7.8}$$

when $x = y = 1$, this gives $c_1(1) = c_2(1) =: c$. Setting just $y = 1$, equation (7.8) becomes

$$\ln c + p(1)\ln x = \ln c_2(x), \tag{7.9}$$

which identifies $\ln c_2(x)$. Similarly, letting $x = 1$ in equation (7.8) gives

$$\ln c_1(y) = \ln c + q(1)\ln y, \tag{7.10}$$

and this identifies $\ln c_1(y)$. After substituting (7.9) and (7.10) into relation (7.8) and simplifying, we have

$$q(1)\ln y + p(y)\ln x = p(1)\ln x + q(x)\ln y.$$

In order to determine $p(y)$ and $q(x)$ we group like terms

$$[q(1) - q(x)]\ln y = [p(1) - p(y)]\ln x,$$

and separation of variables yields

$$\frac{\ln y}{p(1) - p(y)} = \frac{\ln x}{q(1) - q(x)}. \tag{7.11}$$

Let us now assume that for all $x \neq 1$ we have $q(x) \neq q(1)$, and that for all $y \neq 1$, $p(y) \neq p(1)$. (We leave it to the reader to decide what happens if, say, there is an $x \neq 1$ such that $q(x) = q(1)$.) Under these assumptions, both sides of equation (7.11) will be equal to some constant denoted as C. Therefore,

$$\ln y = C\,[p(1) - p(y)]\,,$$

and solving for $p(y)$ we have

$$p(y) = p(1) - \frac{1}{C} \ln y. \tag{7.12}$$

Likewise, $q(x)$ is found to be

$$q(x) = q(1) - \frac{1}{C} \ln x.$$

Solving equation (7.10) for $c_1(y)$ now gives $c_1(y) = ce^{q(1)\ln y}$; hence

$$m(x,y) = m_y(x) = c_1(y)x^{p(y)} = ce^{q(1)\ln y}x^{p(y)}.$$

After substituting $p(y)$ from (7.12) this becomes

$$m(x,y) = cy^{q(1)}x^{p(1)-\frac{1}{C}\ln y} = cy^{q(1)}x^{p(1)}x^{-\frac{1}{C}\ln y},$$

and since $x = e^{\ln x}$, this reduces to

$$m(x,y) = cy^{q(1)}x^{p(1)}e^{-\frac{1}{C}\ln x \ln y}.$$

Setting $\alpha = -1/C$ gives the general solution of $m(x,y)$ as

$$m(x,y) = cy^{q(1)}x^{p(1)}e^{\alpha \ln x \ln y}$$

as claimed. □

In conclusion, if $N = 2$, all continuous m satisfying our criterion of formal justice are of the form

$$m(x,y) = cy^q x^p e^{\alpha \ln x \ln y}. \tag{7.13}$$

When dealing with N qualifications, m can be proved to be of the form

$$m(x_1,\ldots,x_N) = \prod_{i=0}^{2^N-1} \left[\exp\left(\alpha_i \prod_{j\in M_i} \ln x_j\right)\right].$$

Here, $M_i \subset \{1, \ldots, N\}$. There are 2^N such subsets. For example, when $N = 2$, there will be 2^2 subsets: $\emptyset$, $\{1\}$, $\{2\}$, and $\{1, 2\}$ corresponding to $i = 0, 1, 2, 3$, respectively. So,

$$m(x_1, x_2) = e^{\alpha_0} e^{\alpha_1 \ln x_1} e^{\alpha_2 \ln x_2} e^{\alpha_3 \ln x_1 \ln x_2},$$

which simplifies to

$$m(x_1, x_2) = c x_1^a x_2^b e^{\alpha \ln x_1 \ln x_2},$$

where α, a, b, and c are constants. This is consistent with equation (7.13). The cases where $N > 2$ may be proved with an induction argument.

7.4. Exploration: Exotic Solutions of Cauchy's Functional Equation

Among others things, we learned in this chapter that the Cauchy functional equation

$$f(x + y) = f(x) + f(y) \tag{7.14}$$

possesses only solutions of the type $f(x) = cx$, provided that f is continuous at at least one point. So if there are other solutions, they must be functions which have no continuity points at all. Although such functions are unlikely to be of any practical significance, the question arises whether there are such solutions of (7.14) at all. Surprisingly, the answer is *yes*.

The existence of such exotic solutions of (7.14) was first established in the early part of the 20th century and rests on the concept of a *Hamel basis* of the real numbers over the rational numbers. Hamel bases are a well-known concept in vector space theory; for our present purposes it is sufficient to consider only the real numbers.

Definition 7.4. A *Hamel basis* H of the real numbers over the rational numbers is a set $H \subset \mathbb{R}$ such that every $x \in \mathbb{R}$ can be uniquely (up to zero terms) represented as a finite sum

$$x = \sum_{k=1}^{n} q_k h_k, \qquad q_k \in \mathbb{Q}, \quad h_k \in H.$$

Here, the rational coefficient q_k and the integer n will depend on x. The representation is unique except for the possible addition of zeros (thought of as $0 \cdot h$ with $h \in H$).

A Hamel basis of this type is a rather elusive set. It must be an uncountable set because otherwise we could prove that the real numbers are a countable set (which we know is not true). It must be a linearly independent set in the sense that the identity

$$\sum_{k=1}^{n} r_k h_k = 0, \quad r_k \in \mathbb{Q},$$

will always imply that all $r_k = 0$. Hence, if (for example) $\sqrt{2} \in H$, no rational multiple of $\sqrt{2}$ can be in H.

The existence of a Hamel basis of $\mathbb{R}$ over $\mathbb{Q}$ follows from Zorn's lemma and the fact that $\mathbb{R}$ is a linear vector space over $\mathbb{Q}$. This is beyond the scope of our text, but we state the result.

Theorem 7.5. *The real numbers possess Hamel bases over the rational numbers.*

Note that we say nothing about uniqueness. There are many Hamel bases.

Hamel bases are the tool of choice to produce the most general solutions of (7.14). Let H be any Hamel basis of $\mathbb{R}$ over $\mathbb{Q}$. For $h \in H$, choose $f(h)$ arbitrary (!). For $x \in \mathbb{R}$, write $x = \sum_k^n q_k h_k$, as is possible by the definition of the Hamel basis, and set $f(x) := \sum_k^n q_k f(h_k)$, which defines f for all real x.

Theorem 7.6. *The function f defined in this way is a solution of* (7.14)*, and every solution of* (7.14) *satisfies*

$$f\left(\sum_{k}^{n} q_k h_k\right) = \sum_{k}^{n} q_k f(h_k).$$

Proof. The second part of the assertion was in fact proved in the text, where we first analysed the Cauchy equation. As for the first part, let x and y be in $\mathbb{R}$. Then x and y have unique representations

(except for zero coefficients)

$$x = \sum_{k=1}^{m} p_k h_k, \qquad y = \sum_{k=1}^{m} q_k h_k$$

(because we can always add terms $0 \cdot h_k$, it is no restriction of generality to assume the same number of terms in both sums). $x + y$ has the (unique) representation

$$x + y = \sum_{k=1}^{m} (p_k + q_k) h_k,$$

and therefore, by definition of f,

$$\begin{aligned} f(x+y) &= \sum_{k=1}^{m} (p_k + q_k) f(h_k) \\ &= \sum_{k=1}^{m} p_k f(h_k) + \sum_{k=1}^{m} q_k f(h_k) = f(x) + f(y). \quad \square \end{aligned}$$

So now we know how to *find* general solutions of (7.14). All we have to do is take a Hamel basis and choose arbitrary values of f on that set. Easy? Wrong. There is unfortunately no way of *finding* a Hamel basis. The theorem quoted earlier which states existence does just that—it tells us that there is such as thing as a Hamel basis, but does not tell us how to find it. The proof of existence (based on Zorn's lemma) is not constructive.

Nonetheless, we can gather a lot of information on solutions of (7.14) which are not of the form $f(x) = cx$. Recall that the graph of a function f is defined to be the set

$$G = \{(x, y) \in \mathbb{R}^2 : y = f(x)\}.$$

Theorem 7.7. *Suppose that f is a solution of* (7.14) *which is not of the type $f(x) = cx$. Then the graph of f is dense in $\mathbb{R}^2$; in other words, for any point P in the Euclidean plane $\mathbb{R}^2$ and any number $\epsilon > 0$ there is a point Q on the graph of f such that $|P - Q| < \epsilon$.*

Proof. If f is not of the type $f(x) = cx$, there must be real numbers $x_1 \neq 0$ and $x_2 \neq 0$ such that $f(x_1)/x_1 \neq f(x_2)/x_2$. In other words,

the determinant

$$\begin{vmatrix} x_1 & f(x_1) \\ x_2 & f(x_2) \end{vmatrix} \neq 0.$$

This means that the vectors $\vec{p}_1 = (x_1, f(x_1))$ and $\vec{p}_2 = (x_2, f(x_2))$ are linearly independent, so any $\vec{p} \in \mathbb{R}^2$ can be represented uniquely as

$$\vec{p} = \rho_1 \vec{p}_1 + \rho_2 \vec{p}_2.$$

Moreover, the set of all vectors $\vec{p}$ which can be represented in this way with rational coefficients ρ_1, ρ_2 is dense in $\mathbb{R}^2$. If we now let $(r_1, r_2) \in \mathbb{Q}^2$, then

$$\begin{aligned} r_1\vec{p}_1 + r_2\vec{p}_2 &= (r_1x_1 + r_2x_2, r_1f(x_1) + r_2f(x_2)) \\ &= (r_1x_1 + r_2x_2, f(r_1x_2 + r_2x_2)), \end{aligned}$$

where the fact that f satisfies (7.14) is used in the last identity. The assertion of the theorem follows. □

There is an immediate yet impressive corollary to this theorem.

Corollary 7.8. *If f satisfies* (7.14) *and is not of the form $f(x) = cx$, then the image of any nonempty interval (a, b) with $a < b$ is dense in $\mathbb{R}$.*

Further, our results show that the results from this chapter generalize as follows.

Corollary 7.9. *If f solves* (7.14) *and is either*

- *continuous at at least one point, or*
- *monotone on an interval of positive length, or*
- *bounded from one side (above or below) on an interval of positive length,*

then there is a $c \in \mathbb{R}$ such that for all x, $f(x) = cx$.

Well, we handled the first case in the chapter; in the second or third case it is clear that the graph of f cannot be dense in the plane, and so the last theorem gave us the answer. Much more material on functional equations may be found in the book by Aczél and Dhombres [**A**].

Exercises

(1) Let $h : \mathbb{R} \to \mathbb{R}$ be such that for all $s, t \in \mathbb{R}$, $h(t+s) = h(t)+h(s)$. Show that
   (a) $h(-t) = -h(t)$,
   (b) $h(t-s) = h(t) - h(s)$.

(2) Let m be the function $m(xy) = m(x) + m(y)$ defined on $\mathbb{R}^+$ for all $x, y > 0$. Assuming m is continuous, show that there is a constant c such that $m(x) = c \ln x$.
   Hint: You may use the fact that all continuous solutions of the functional equation from Exercise 1 are of the type $h(t) = ct$.

(3) (a) Suppose that $f : \mathbb{R} \to \mathbb{R}$ is continuous and satisfies the equation
$$f(x+y) = 2f(x) + f(y)$$
   for all $x, y \in \mathbb{R}$. Find f.
   (b) Determine all continuous $f : \mathbb{R} \to \mathbb{R}$ such that for all $x, y \in \mathbb{R}$,
$$f(x+y) = 2f(x) + f(y) + 1.$$

(4) Three employees of a company with otherwise equal qualifications have worked for the number of years indicated and earn the annual salary stated below. Use the method of least-square fits to construct a *fair* salary curve, and determine who is under or overpaid, respectively. (You may use a pocket calculator or a PC. You may also use MAPLE and its `with stats` package.)

Employee	Years	Annual Salary
1	9	\$48,000
2	14	\$57,000
3	20	\$65,000

(5) Find all continuous functions $f(x, y)$ of two real variables with the property that
$$\begin{aligned} f_y(x) &:= f(x,y) = C(y)e^{p(y)x}, \\ f_x(y) &:= f(x,y) = D(x)\ln[q(x)\,y] \end{aligned}$$

(i.e., for each fixed y, f behaves like an exponential in x and for each fixed x, f behaves like an logarithm in y.)

(6) In 2003 the base salaries of players in the NFL were as follows: Rookies, \$225,000; 2nd year, \$300,000; 3rd year, \$375,000; 4th–6th year, \$450,000; 7th–9th year, \$655,000; 10th year, \$755,000. Would you rate this salary scale as fair?

This page intentionally left blank

Chapter 8

Traffic Dynamics: A Microscopic Model

Concepts and Tools: Differential-delay equations, solution procedures, numerical methodology

This and the next chapter focuses on the dynamics of traffic flow and some models which are applicable to this problem. There are essentially three types of models which examine this situation: microscopic models, which focus on the individual cars and investigate deterministic or stochastic interactions; kinetic models predicting the statistical distribution of cars with respect to their location and velocity; and macroscopic models, which are partial differential equations of conservation type relating density and flux. We begin by setting up a microscopic model and will use it to motivate conservation laws discussed in the next chapter.

8.1. The Braking Force

Suppose there are N vehicles in a traffic lane, each of equal length L and mass m, and that the front of the ith car is at position $x_i(t)$ at time t, where $i = 1, \ldots, N$ as shown by Figure 1. Because there is only a single lane of traffic, the model is only valid while the order

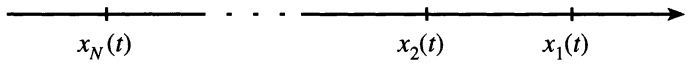


Figure 1. As shown, the front of each vehicle has position $x_i(t)$ at time t for $1 \leq i \leq N$. The position of the front of the lead car is denoted by x_1.

of the cars, $x_1 > x_2 > \cdots > x_N$, is preserved. If two vehicles with labels i and $i-1$ collide at some time T, then $x_i(T) + L = x_{i-1}(T)$, and the model loses its meaning for all cars after the $(i-1)$st and all times $t \geq T$. We will suppose that each driver has a reaction time τ, where, for simplicity, we assume that this reaction time is identical for all drivers. The velocity of the ith vehicle is $v_i(t) = x_i'(t)$, where x' indicates differentiation with respect to time.

Thus the braking force $F_{bi}(t)$ of a given vehicle is related to its deceleration by Newton's second law,

$$F_{bi}(t) = (\text{mass})(\text{acceleration}) = m\, x_i''(t+\tau),$$

where m is the mass of the vehicle and $x_i''(t+\tau)$ is the deceleration of the ith vehicle at the delayed time $t+\tau$. The magnitude of this braking force will depend on the relative velocity and the relative distance to the car in the x_{i-1} position, i.e., the car ahead of the vehicle under consideration. If, for example, $v_{i-1} \ll v_i$ (the car in front is travelling at a much slower pace than the car following), then the braking force should be large; however, if $v_{i-1} \approx v_i$, very little braking should be required.

The relative distance between the two vehicles, given by $|x_i(t) - x_{i-1}(t)|$, will also have an effect on the braking force. If the two cars are travelling at approximately the same speed and are close when the leading car slows down (i.e., the case where $v_{i-1} < v_i$ and $x_{i-1} \approx x_i$), then the braking force should be large as little space is available for deceleration. Assumptions consistent with these observations are that the braking force be directly proportional to the relative speed, and inversely proportional to the relative distance. The braking force is then given by

$$mx_i''(t+\tau) = F_{bi}(t) = A\frac{x_i'(t) - x_{i-1}'(t)}{|x_i(t) - x_{i-1}(t)|},$$

where $|x_i(t) - x_{i-1}(t)|$ is the relative distance between the ith and $(i-1)$st cars, and A is an a priori unknown positive proportionality constant.

If we set $\lambda = A/m$, then, because by definition $x_{i-1} > x_i$, the acceleration becomes

$$x_i''(t+\tau) = v_i'(t+\tau) = \lambda \frac{d}{dt} \ln |x_i(t) - x_{i-1}(t)| .$$

Integrating this to determine the velocity $v_i(t+\tau)$ gives

$$v_i(t+\tau) = x_i'(t+\tau) = \lambda \ln |x_i(t) - x_{i-1}(t)| + \alpha_i, \tag{8.1}$$

which holds for $i = 2, 3, \ldots, N$. Notice that this identity does not hold for $i = 1$ since the leading car is not affected by any other vehicle.

8.2. Density and Flux at Equilibrium

We now focus on equilibria situations and macroscopic variables such as density and flux, which are yet to be defined. For example, it is commonly observed that when cars are travelling through a tunnel, the velocity of each vehicle will decrease as the density of the cars increases. To give a reasonable definition of this density, consider an interval of length 2ϵ for some suitably chosen $\epsilon > 0$, large relative to L (the size of each car) but small relative to the macroscopic scale of the road. The density at some point x_o, denoted as $\rho(x_o, t)$, is defined by counting the number of vehicles in the ϵ-interval about x_o and dividing by the length of the interval. In other words, the density at x_o is given by

$$\rho(x_o, t) = \frac{\text{number of cars in } (x_o - \epsilon, x_o + \epsilon) \text{ at time } t}{2\epsilon}, \tag{8.2}$$

where x_o is some position in the road. Notice that this definition depends on ϵ; however, it can be proved that for sufficiently small ϵ (but still large compared to L) and car distances that vary slowly as we scan the road, the dependence on ϵ becomes weak, and disappears altogether in equilibrium scenarios. An equilibrium is a situation where all cars are at equal distance from the leading car, and all cars are moving at the same speed. In this case the above definition gives

$$\rho = \frac{1}{|x_i - x_{i-1}|}.$$

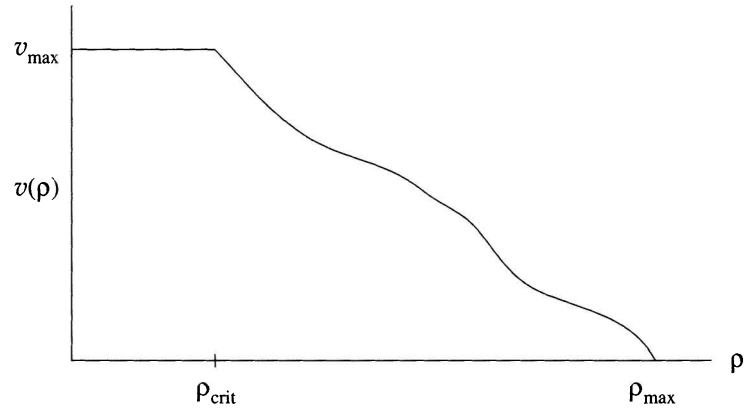


Figure 2. A typical fundamental diagram for the traffic flow is depicted. Vehicles travel at a velocity $v_{\max}$ until a critical density ρ_{crit} is reached at which point the traffic will slow and eventually halt.

Notice that the minimal value for the denominator is L. Hence the maximal possible density, associated with *bumper-to-bumper* traffic, is $\rho_{\max} = 1/L$.

We will assume that the speed v observed at a location will depend only on the density ρ at that location;[1] i.e., that $v(x,t) = v(\rho(x,t))$. Observations suggest that there is a critical (observable) density ρ_{crit} such that for densities in the range $0 \leq \rho \leq \rho_{\text{crit}}$, $v(\rho)$ will be the maximal velocity, denoted as $v_{\max}$, and equal to the speed limit. See Figure 2. At the critical density ρ_{crit}, traffic will begin to slow down and halt altogether at the maximum density $\rho_{\max}$, so that $v(\rho_{\max}) = 0$.

It is also observed that $v'(\rho) \leq 0$. Our next objective is to determine $v(\rho)$ for $\rho \geq \rho_{\text{crit}}$. This dependency is known as a *fundamental diagram.* The maximum density, $\rho_{\max}$, will be reached when the cars in the tunnel are bumper-to-bumper and no longer moving. The maximum number of cars in an interval $(z_o - \epsilon, z_o + \epsilon)$ will then be $2\epsilon/L$.

[1]This assumption is controversial, but reasonable for equilibrium situations where all vehicles travel at the same speed.

Therefore, using equation (8.2), $\rho_{\max}$ is

$$\rho_{\max} = \left(\frac{2\epsilon}{L}\right)\left(\frac{1}{2\epsilon}\right) = \frac{1}{L}, \tag{8.3}$$

as stated earlier.

8.2.1. At Equilibrium. Consider a situation in which all cars move at the same speed v, a distance $d > 0$ from one another. If each of these cars has length L, then the density ρ is

$$\rho = \frac{1}{d+L}, \qquad d, L > 0. \tag{8.4}$$

In the equilibrium situation the speed will be the same for all vehicles and will therefore not depend on i. From expression (8.1)

$$v = \lambda \ln(d+L) + \alpha_i = \lambda \ln(d+L) + \alpha,$$

where $d+L$ is the relative distance between the fronts of two successive vehicles, and α_i must also be independent of i and has therefore been replaced by α. Using (8.4), this becomes

$$v = \lambda \ln\left(\frac{1}{\rho}\right) + \alpha, \tag{8.5}$$

with parameters λ and α. We determine α from the observation that $v(\rho_{\max}) = 0$. Setting $\rho = \rho_{\max}$ in (8.5) gives

$$v(\rho_{\max}) = \lambda \ln\left(\frac{1}{\rho_{\max}}\right) + \alpha = 0,$$

and solving for α yields $\alpha = \lambda \ln(\rho_{\max})$; thus, (8.5) becomes

$$v(\rho) = \lambda \ln\left(\frac{1}{\rho}\right) + \lambda \ln(\rho_{\max}) = -\lambda \ln\left(\frac{\rho}{\rho_{\max}}\right) \tag{8.6}$$

for $\rho \geq \rho_{\text{crit}}$. Moreover, $v(\rho)$ must be continuous at $\rho = \rho_{\text{crit}}$. Setting $\rho = \rho_{\text{crit}}$ gives the maximum velocity or speed limit, denoted as $v_{\max}$, as

$$v_{\max} = v(\rho_{\text{crit}}) = -\lambda \ln\left(\frac{\rho_{\text{crit}}}{\rho_{\max}}\right),$$

which yields

$$\lambda = \frac{v_{\max}}{\ln\left(\dfrac{\rho_{\max}}{\rho_{\text{crit}}}\right)}.$$

Substituting this into equation (8.6) we have

$$v(\rho) = v_{\max} \ln\left(\frac{\rho_{\max}}{\rho}\right)\left[\ln\left(\frac{\rho_{\max}}{\rho_{\text{crit}}}\right)\right]^{-1},$$

which ultimately yields the result

$$v(\rho) = \begin{cases} v_{\max} & \text{for } \rho \le \rho_{\text{crit}}, \\ v_{\max} \ln\left(\frac{\rho_{\max}}{\rho}\right)\left[\ln\left(\frac{\rho_{\max}}{\rho_{\text{crit}}}\right)\right]^{-1} & \text{for } \rho > \rho_{\text{crit}}. \end{cases} \tag{8.7}$$

Notice that the velocity is constant until the critical density ρ_{crit} is reached, then it decays logarithmically.

We emphasize that this fundamental diagram for the speed as a function of density only applies for the equilibrium situation described. However, in spite of our rather simplistic approach (the assumptions regarding the braking forces are just first guesses) the function given by (8.7) provides a qualitatively reasonable fundamental diagram.

8.2.2. The Maximal Traffic Flux at Equilibrium. The traffic flux, denoted as $j = j(\rho)$, is defined as the number of cars passing through a given point per unit of time, and is given by

$$j(\rho) = \left(\frac{\text{cars}}{\text{distance}}\right)\left(\frac{\text{distance}}{\text{time}}\right) = \rho v(\rho).$$

For the velocity-density relationship established for the equilibrium situation represented by (8.7), one has

$$j(\rho) = \begin{cases} \rho v_{\max} & \text{for } \rho \le \rho_{\text{crit}}, \\ \rho v_{\max} \ln\left(\frac{\rho_{\max}}{\rho}\right)\left[\ln\left(\frac{\rho_{\max}}{\rho_{\text{crit}}}\right)\right]^{-1} & \text{for } \rho > \rho_{\text{crit}}. \end{cases} \tag{8.8}$$

To determine the maximum value of the flux, we differentiate equation (8.8). For $\rho \ge \rho_{\text{crit}}$ this gives

$$j'(\rho) = v_{\max} \ln\left(\frac{\rho_{\max}}{\rho}\right)\left[\ln\left(\frac{\rho_{\max}}{\rho_{\text{crit}}}\right)\right]^{-1} - v_{\max}\left[\ln\left(\frac{\rho_{\max}}{\rho_{\text{crit}}}\right)\right]^{-1},$$

then grouping like terms and equating this to zero yields the condition that

$$v_{\max}\left[\ln\left(\frac{\rho_{\max}}{\rho_{\text{crit}}}\right)\right]^{-1}\left[\ln\left(\frac{\rho_{\max}}{\rho}\right) - 1\right] = 0.$$

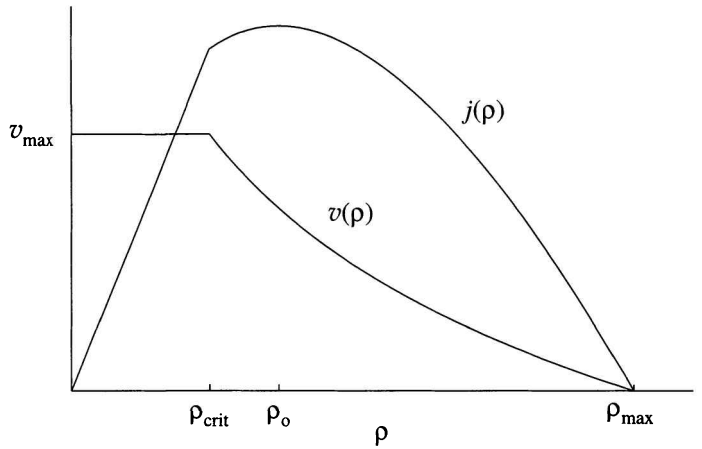


Figure 3. The fundamental diagram of speed as a function of density and the corresponding flux as a function of density. The particular choice of the maximum flux occurring at $\rho_o = \rho_{max}/e > \rho_{crit}$ is justified in the text.

This holds if and only if $\ln \rho_{max} - \ln \rho = 1$ which, solving for ρ, occurs at the density $\rho_o = \rho = \rho_{max}/e$.

If $\rho_o > \rho_{crit}$, then the maximal flux occurs at ρ_o. If $\rho_o \leq \rho_{crit}$, then the maximum flux will occur at ρ_{crit}. This is because for any $0 \leq \rho \leq \rho_{crit}$, $j'(\rho) = v_{max} > 0$. So j is increasing on this interval and reaches its maximum at the right-hand end point. Figure 3 shows a fundamental diagram given by expression (8.7) and the graph of the corresponding $j(\rho)$ in the case where $\rho_o = \rho_{max}/e > \rho_{crit}$.

The critical density ρ_{crit} is, of course, found by observation (it may well depend on the temperament and driving experience of the local drivers and, therefore, differ markedly between, say, Los Angeles and Paris), but we suggest that it should in any case be expected that $\rho_{crit} < \rho_{max}/e$. Recall that $e < 3$, and at $1/3$ maximal density there are just two car lengths between any two cars; at this point nobody would want to drive the speed limit, unless the speed limit is drastically reduced due to road construction, congestion, or other incidental reasons.

8.3. A Case Study: Propagation of a Perturbation

The example we discuss here is presented in Mesterton-Gibbons [**K**]. Our discussion adds detail and generality.

Assume that for times $t < 0$, a platoon of vehicles has been moving in equilibrium configuration at the optimal density ρ_o computed above, and with the corresponding speed. Suppose that the lead car crosses the point $x_o = 0$ on the road at time $t = 0$ and goes through a braking maneuver.

In practice, it is easier and more informative to compute the *difference* of the ensuing reactions of the following cars from their positions had the first car not changed its speed. This difference is called the *perturbation displacement*, and we denote it by $z_i(t)$ (for the ith car). To set up the correct differential-delay equations for the z_i, we proceed as follows.

First, while equilibrium persists, the displacement associated with the equilibrium speed v of the ith car at time t is

$$y_i(t) = vt - (i-1)(d+L). \tag{8.9}$$

Here, as before, d is the distance between two successive vehicles and L is the length of each car. Clearly $y_i' = v$. From equation (8.6), we know that the velocity for a vehicle in the equilibrium situation is given as

$$v = \lambda\left(\ln \rho_{\max} - \ln \rho\right),$$

which, since by our assumption the cars are spaced at the optimal density $\rho_o = \rho_{\max}/e$, yields

$$v = \lambda\left(\ln \rho_{\max} - \ln \rho_o\right) = \lambda\left(\ln \rho_{\max} - \ln \rho_{\max} + 1\right) = \lambda. \tag{8.10}$$

We also saw in Section 8.2.1 that, at equilibrium, $\alpha_i = \alpha = \lambda \ln \rho_{\max}$. Therefore, the identity (8.10) implies that $\alpha_i = v \ln \rho_{\max}$. Thus, equation (8.1) becomes

$$v_i(t+\tau) = v \ln |x_i(t) - x_{i-1}(t)| + v \ln \rho_{\max},$$

and this becomes

$$v_i(t+\tau) = v \ln\left(\rho_{\max}\left[x_{i-1}(t) - x_i(t)\right]\right) \tag{8.11}$$

(as $x_{i-1} > x_i$, it is not necessary to take the absolute value of the distance between two vehicles).

Second, assume that starting at time $t = 0$, the lead driver brakes for a short time (say for one second) and then accelerates again to the old speed v. To be specific we set

$$v_1(t) = \begin{cases} v & \text{for } t \le 0, \\ v(1 - b(t)) & \text{for } t > 0. \end{cases}$$

Here, $b = b(t)$ is a function which is zero for $t < 0$ and zero for $t > t_1$, and smooth and nonnegative for $t \in [0, t_1]$. This function models the braking and acceleration process. We will denote by $B(t)$ the function

$$B(t) = \int_0^t b(s)\, ds.$$

Clearly $B(t) = 0$ for $t < 0$, and $B(t) = \int_0^{t_1} b(s)\, ds$ for $t > t_1$. A specific example for a function b of the described type is

$$b(t) = \begin{cases} 0 & \text{for } t \le 0, \\ kte^{(t_o - t)/t_o} & \text{for } t > 0, \end{cases}$$

where for this example $t_1 = \infty$, t_o is the time at which the deceleration stops, and $k > 0$ is a parameter that relates to the braking force. Figure 4 illustrates the speed of the first car as a result of this particular braking profile.

Integrating the velocity function of the lead car with respect to t and using $x_1(0) = 0$ gives the position of the lead car as a function of time:

$$x_1(t) = \begin{cases} vt & \text{for } t \le 0, \\ v[t - B(t)] & \text{for } t > 0. \end{cases} \tag{8.12}$$

The perturbation displacement, $z_1(t)$, is given by the difference of this true position $x_1(t)$ and the *would-have-been*, or unperturbed, position $y_1(t) = vt$ of the lead car. As a result,

$$z_1(t) = \begin{cases} 0 & \text{for } t \le 0, \\ -vB(t) & \text{for } t > 0. \end{cases} \tag{8.13}$$

Figure 5 shows the perturbation displacement for the lead car for the function given in the above example, where, for $t > 0$, $B(t) = kt_o[t_o - (t + t_o)e^{-t/t_o}]e$.

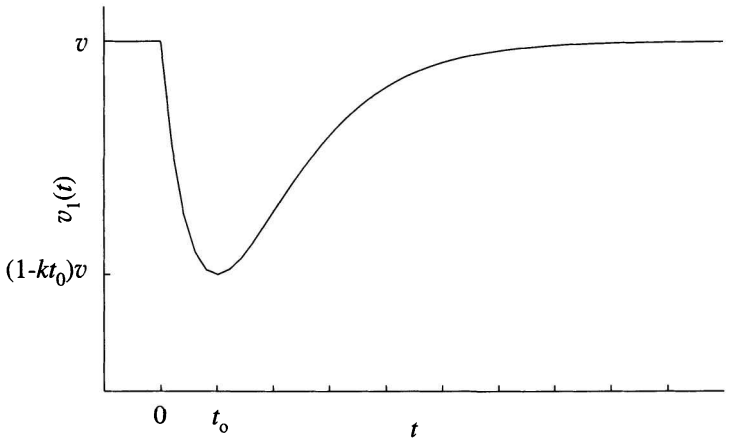


Figure 4. The lead car travels at an equilibrium velocity v until time $t = 0$. At this point, it will begin braking for $t = t_o$ seconds, and then resume its previous velocity. The resulting velocity perturbation for $b(t) = kte^{(t_o-t)/t_o}$ for $t > 0$ is shown.

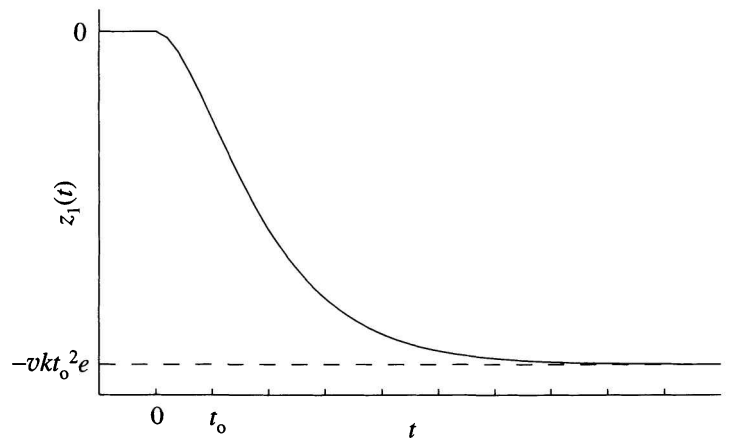


Figure 5. Shown here is the perturbation displacement of the lead car given by (8.13) when the braking profile is given by $b(t) = kte^{(t_o-t)/t_o}$ for $t > 0$.

Third, to avoid a collision, the remaining cars are forced to brake as well. For $i \geq 2$, the perturbation displacement of the ith car,

defined by $z_i(t) := x_i(t) - y_i(t)$, is given by

$$z_i(t) = \begin{cases} 0 & \text{for } t \le 0, \\ x_i(t) - vt + (i-1)(d+L) & \text{for } t > 0. \end{cases}$$

The variables z_i are defined such that they would all remain identically zero if equilibrium were to persist. As we are now investigating what happens if a perturbation occurs, we also must explore how the constraints on the $\{x_i\}$, namely, $x_{i-1} - x_i > L$, translate into constraints on the $\{z_i\}$. This is easy: by subtracting z_i from z_{i-1} we find

$$L < x_{i-1}(t) - x_i(t) = z_{i-1}(t) - z_i(t) + d + L,$$

which implies the constraint (specific for the scenario under consideration)

$$z_{i-1}(t) + d > z_i(t).$$

It must be kept in mind that violation of this constraint for some pair $(i-1, i)$ means that a collision has occurred and that the model has lost its validity.

We complete the setup of a system of differential-delay equations for the current perturbation problem. Recalling that for the equilibrium situation $d + L = \rho^{-1} = e/\rho_{\max}$ gives

$$z_i(t) = \begin{cases} 0 & \text{for } t \le 0, \\ x_i(t) - vt + \frac{(i-1)e}{\rho_{\max}} & \text{for } t > 0, \end{cases}$$

and substituting $x_i = y_i + z_i$ into equation (8.11) yields

$$\frac{d}{dt}\left[y_i(t+\tau) + z_i(t+\tau)\right] = v \ln\left\{\rho_{\max}\left[y_{i-1}(t) + z_{i-1}(t) - y_i(t) - z_i(t)\right]\right\}.$$

Simplifying, we obtain the system of equations

$$v + \frac{d}{dt} z_i(t+\tau) = v \ln\left\{\rho_{\max}\left[\frac{e}{\rho_{\max}} + z_{i-1}(t) - z_i(t)\right]\right\} \tag{8.14}$$

for $2 \le i \le N$ with initial conditions $z_i(t) = 0$ for $t < 0$, $1 \le i \le N$, and $z_1(t)$ given by expression (8.13).

8.3.1. The Solution Procedure. In principle, the following procedure allows an explicit solution of (8.14). To do this, substitute $t \to t - \tau$ into equation (8.14) which produces

$$\frac{d}{dt} z_i(t) = v(-\ln e) + v \ln \{e + \rho_{\max} [z_{i-1}(t-\tau) - z_i(t-\tau)]\}.$$

After simplifying, we obtain for $2 \le i \le N$

$$\frac{d}{dt} z_i(t) = v \ln \left\{1 + \frac{\rho_{\max}}{e} [z_{i-1}(t-\tau) - z_i(t-\tau)]\right\} \tag{8.15}$$

with initial conditions $z_i(t-\tau) = 0$ for $t < \tau$. We already know $z_1(t)$ with (8.13); to determine $z_2(t)$ consider first the interval $[0, \tau)$. Since $t - \tau < 0$ on this interval, $z_2(t-\tau) = 0$; therefore, for the case $i = 2$ and for all $t \in [0, \tau)$, we have

$$\frac{d}{dt} z_2(t) = v \ln \left[1 + \frac{\rho_{\max}}{e} z_1(t-\tau)\right], \tag{8.16}$$

where z_1 is known everywhere. In fact, it is identical to zero on the domain in (8.16). Thus by integrating (8.16) we can compute $z_2(t)$ on $[0, \tau)$. This is now used to find z_2 on the next interval $[\tau, 2\tau)$, on which expression (8.15) becomes

$$\frac{d}{dt} z_2(t) = v \ln \left\{1 + \frac{\rho_{\max}}{e} [z_1(t-\tau) - z_2(t-\tau)]\right\},$$

where z_1 is explicitly known and $z_2(t-\tau)$ on $[\tau, 2\tau)$ is known from the previous step from (8.16). Continuing with this process, $z_2(t)$ can thus be found on any interval by using the values of $z_1(t)$ and $z_2(t)$ from the previous interval by integration. Repeating this procedure recursively will produce the function $z_2(t)$ for all $t \ge 0$. Knowledge of z_2 on $[0, (i+1)\tau]$ will similarly permit us to compute z_3 explicitly on $[0, i\tau]$, and so on. The procedure is graphically depicted in Figure 6.

The method described above is conceptually very simple but not terribly practical. The reason is that the integrations, while completely explicit, tend to become more and more complex for the later time intervals and the higher indices i. We present a simple example that demonstrates the reasons behind this.

Example 8.1. Let $\phi_1(t) = \phi_2(t) = 0$ for $t < 0$, $\phi_1(t) = t + t^2/6$ for all $t \ge 0$, and suppose that for all $t > 0$,

$$\frac{d}{dt} \phi_2(t) = \phi_1(t-1) - 2\phi_2(t-1).$$

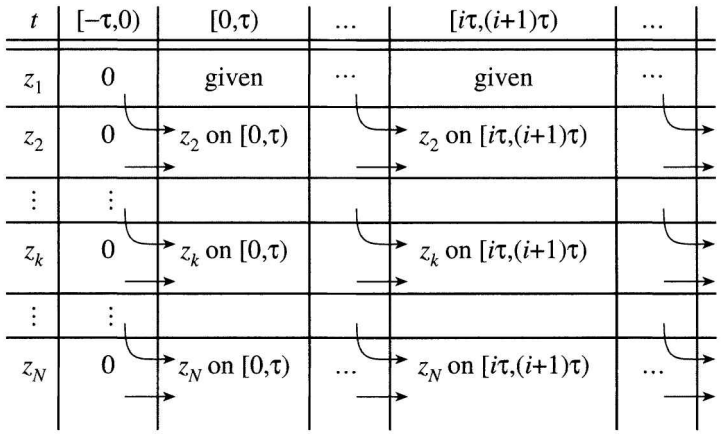


Figure 6. In this example, N is the number of cars, and time is broken into disjoint intervals of length τ. The arrows depict the dependencies between the various time intervals. In general, $z_i(t)$ is determined from z_i and z_{i-1} on the previous interval: $z_i(t-\tau)$ and $z_{i-1}(t-\tau)$. For example, $z_3(t)$ on the interval $[\tau, 2\tau)$ is determined by the values of z_2 and z_3 on the interval $[0, \tau)$. Note that $z_i(t) = 0 \ \forall i$ whenever $t < 0$.

We set the task of computing $\phi_2(t)$ on $[1, 2)$. This is done as follows. First, determine $\phi_2(t)$ on all previous intervals and for the interval $[0, 1)$ there is only one, namely $[0, 1)$. So for $t \in [0, 1)$,

$$\frac{d}{dt}\phi_2(t) = \phi_1(t-1) - 2\phi_2(t-1) = 0$$

since $t - 1 < 0$ on $[0, 1)$. Thus $\phi_2(t) = 0$ on $[0, 1)$. Now, using the value of $\phi_2(t)$ and $\phi_1(t)$ on $[0, 1)$, it is possible to determine $\phi_2(t)$ on the next interval $[1, 2)$. For this interval

$$\frac{d}{dt}\phi_2(t) = \phi_1(t-1) - 2\phi_2(t-1) = (t-1) + \frac{1}{6}(t-1)^2,$$

which, upon integrating, gives $\phi_2(t)$ on $[1, 2)$ as

$$\phi_2(t) = \frac{(t-1)^2}{2} + \frac{(t-1)^3}{18}.$$

It is transparent from this result that the calculation of ϕ_2 on the next interval, $t \in [2, 3)$, will involve integrations of a third-order polynomial, producing a fourth-order polynomial, and so on.

Systems of differential-delay equations like (8.15) can in principle be solved explicitly by the elementary recursive method demonstrated in the previous example. However, the rapidly growing complexity of the explicit integrations suggests that numerical approximations may be a more practical alternative. We describe how Euler's method could be used in this context and present the results of some numerical calculations.

8.3.2. Using Euler's Method to Approximate Solutions of the Differential-Delay Equations. The first step in this rough numerical procedure is to replace the left-hand side of (8.15) by the finite difference

$$\frac{d}{dt}z_i(t) \approx \frac{z_i(t+h) - z_i(t)}{h}, \qquad 0 < h \ll 1. \tag{8.17}$$

Let $h = 1/K$ where $K \in \mathbb{N}$ and $1/K$ is considered to be one time step in our simulation. It is natural to take the reaction time τ as an integer multiple of h, i.e., $\tau = M/K$ where $K > 0$. The general time t will be counted in units of h, that is $t = n/K$ with $n \in \mathbb{N}$.

Let the perturbation displacement of a driver after n such units of time be denoted as $Z_i(n)$, such that $Z_i(n) := z_i(n/K)$. $Z_i(n+1)$ is then

$$Z_i(n+1) = z_i\left(\frac{n}{K} + \frac{1}{K}\right) = z_i\left(\frac{n}{K} + h\right). \tag{8.18}$$

Using this and the definition of h gives (8.17) as

$$\frac{d}{dt}z_i(t) \approx \frac{z_i(t+h) - z_i(t)}{h} = K\left[Z_i(n+1) - Z_i(n)\right]. \tag{8.19}$$

Substituting (8.19) and (8.18) into the formula for the perturbation displacement given by equation (8.15) and using the definition of τ yields the algebraic system

$$K[Z_i(n+1) - Z_i(n)] = v \ln\left\{1 + \frac{\rho_{\max}}{e}\left[Z_{i-1}(n-M) - Z_i(n-M)\right]\right\} \tag{8.20}$$

for $i = 2, 3, \ldots, N$, $Z_i(n-M) = 0$ for $n \leq M$, and $Z_1(n) = z_1(n/K)$ is given by (8.13).

System (8.20) is easily solved recursively by a computer. For the purposes of the simulation we will assume that the cars are travelling at an equilibrium speed of 100 km/hr with a distance of about 3 car lengths between successive vehicles. If we assume that an average vehicle has a length of $L = 6$ m, then this situation is realized with a choice of $\rho_{\max} = 40$ cars/km so that $d = 19$ m. Furthermore, we will assume that there are a total of $N = 5$ vehicles in the platoon and that the lead vehicle momentarily decelerates from 100 km/hr to $100(1 - k)$ km/hr over an interval of $t_o = 1$ second according to $b(t) = kte^{(t_o - t)/t_o}$. Notice that in this case the asymptotic value of the perturbation displacement is $vkt_o^2 e$. For $k = 0.2$, $vkt_o^2 e \approx 15.1$ m or nearly 2.5 car lengths.

Results for three different values of the reaction time $\tau = 0.5$, 1.5, and 2 seconds where k takes on the values $k = 0.1$ and $k = 0.2$ are displayed in Figure 7. Rather than plotting the perturbation displacement $z_i(t)$ of each vehicle, we instead plot the quantities $z_i(t) - (i-1)d$. In this way the condition for no collision ($z_i < z_{i-1} + d$) is violated when the curves cross. It is obvious from these calculations that for a given k and reaction time τ the deceleration of the lead vehicle causes instabilities in the perturbation displacements which may lead to collisions down the platoon of cars. In fact, for $k = 0.2$, $\tau = 2$, the model is no longer valid for cars 4 and 5 beyond time $t = 12.9$ seconds as a collision has occurred.

Remark 8.2. Approximation (8.19) is an Euler-type approximation of a derivative and thus significant truncation errors are to be expected. Hence our discrete approximation is quite poor. Remember, however, that our model is based on a very rough qualitative argument, so the validity of our differential-delay equations themselves is very questionable. Given this shortcoming, is it wise to work hard towards numerical procedures with higher accuracy?

8.4. Exploration: Peano's Existence Theorem

Among other things, we learned in this chapter that systems of differential-delay equations are easily and uniquely solvable (even explicitly), and we described the solution process in detail. This fact, and the ensuing properties of the computed solutions can be used for a

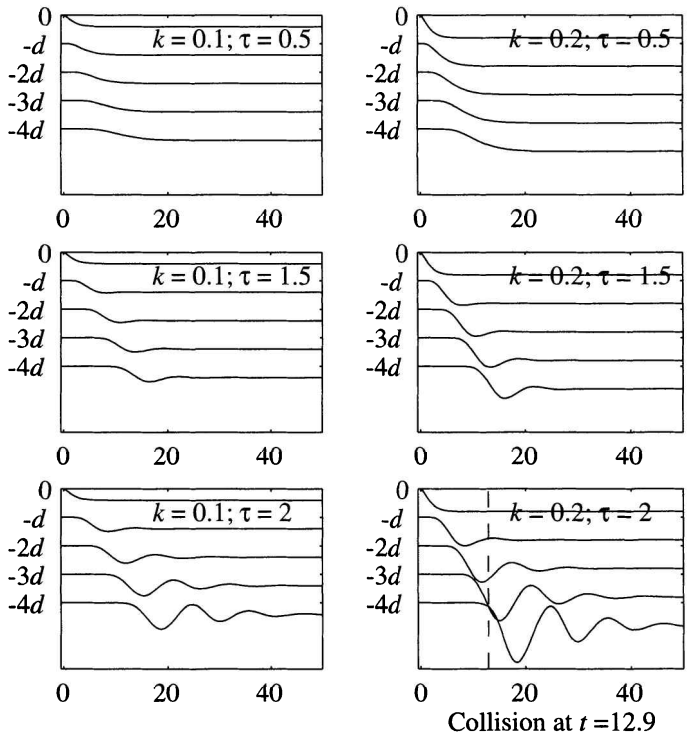


Figure 7. Shown is the displacement of each car in a platoon of five when the lead car slows temporarily. The abscissa is time in seconds and the ordinate is the shifted perturbation displacements $z_i(t) - (i-1)d$, $i = 1, \ldots, N$ in meters. Note that the condition for no collision ($z_i < z_{i-1} + d$) is violated when the curves cross. The left column of plots corresponds to $k = 0.1$ and the right column corresponds to $k = 0.2$. Reaction times for each driver take on three values: $\tau = 0.5$, $\tau = 1.5$ and $\tau = 2$ seconds.

proof of one of the more famous existence results for initial value problems for ordinary differential equations, first proved by the Italian mathematician G. Peano in 1890. What Peano proved is that the initial value problem

$$y' = f(x, y), \qquad y(x_0) = y_0,$$

possesses a solution passing through the point (x_0, y_0) if the function f is continuous in a neighbourhood of this point. Notice that this condition is weaker than the Lipschitz continuity of f required for the standard existence and uniqueness theorem (compare with the discussion in Chapters 3 and 5). However, the assertion is also weaker. There is no statement about uniqueness! As the examples from Chapter 3 show, mere continuity of f does not guarantee uniqueness. Recall that the initial value problem

$$y' = f(y), \qquad y(0) = 0,$$

with $f(y) = |y|^{1/2}$ possesses infinitely many solutions.

We prove a slightly weaker version of the Peano existence result.

Theorem 8.3. *Suppose that the function $f(x, y)$ is continuous and bounded on the set $[x_0, x_0 + a] \times \mathbb{R}$, where $a > 0$. Then there exists at least one function $y(x)$ defined on $[x_0, x_0 + a]$ so that in this interval*

$$y' = f(x, y), \qquad y(x_0) = y_0. \tag{8.21}$$

In particular, y is continuously differentiable.

We will prove this by using approximating differential-delay equations, the concept of equicontinuity, and a theorem from real analysis known as the Arzelà–Ascoli theorem.

Definition 8.4. A set S of continuous functions $S = \{f\}$ defined on an interval $[a, b]$ is called *equicontinuous* if for every $\epsilon > 0$ and every $x \in [a, b]$ there is a $\delta > 0$ such that for all $f \in S$

$$|x - y| < \delta \implies |f(x) - f(y)| < \epsilon,$$

for all $y \in [a, b]$.

For comparison, the statement that all of the $f \in S$ are continuous is, "For every $\epsilon > 0$, every $f \in S$, and every $x \in [a, b]$ there is a $\delta > 0$ such that

$$|x - y| < \delta \implies |f(x) - f(y)| < \epsilon,$$

for all $y \in [a, b]$", so that δ can depend on ϵ, x and the particular $f \in S$. What characterizes the concept of equicontinuity is the fact that δ must be independent of $f \in S$.

Example 8.5. Let $C > 0$ and let

$$L_C := \{f : \forall x_1, x_2 \in [a,b],\ |f(x_1) - f(x_2)| \leq C|x_1 - x_2|\}.$$

Verify that these Lipschitz-continuous functions form an equicontinuous set.

The Arzelà–Ascoli theorem is a powerful and famous result about bounded equicontinuous sets of functions.

Theorem 8.6 (Arzelà–Ascoli). *Every bounded and equicontinuous sequence of functions $\{g_n\}$ on $[a,b]$ contains a subsequence which converges uniformly to a continuous limit function g on $[a,b]$.*

This theorem is proved in texts on real analysis and requires a fair amount of preparation beyond the scope of our text. Uniform convergence means that the subsequence, which we shall simply denote again by $\{g_n\}$, converges in the sense that

$$\lim_{n\to\infty} \sup_{x\in[a,b]} |g_n(x) - g(x)| = 0.$$

Notice that this is stronger than just pointwise convergence. Can you think of a sequence of continuous functions on $[0,1]$ that will converge to 0 pointwise but not uniformly? Remember that such a sequence cannot be bounded and equicontinuous (because then it would satisfy the conditions of the Arzelà–Ascoli theorem). After these preparations we are finally ready to sketch the proof of Theorem 8.3.

Proof (Theorem 8.3). The initial value problem (8.21) is equivalent to the integral equation

$$y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))\, dt$$

in that solutions of one are solutions of the other, and so we will show that the integral equation is solvable. Now let $\alpha > 0$ be a parameter and consider the approximating functions y_α defined by

$$y_\alpha(x) = \begin{cases} x_0 & \text{for } x \leq x_0, \\ x_0 + \displaystyle\int_{x_0}^{x} f(t, y_\alpha(t-\alpha))\, dt & \text{for } x_0 < x \leq x_0 + a. \end{cases} \tag{8.22}$$

We now study the y_α successively on the intervals $[x_0 + (k-1)\alpha, x_0 + k\alpha]$ for $k = 1, 2, \ldots, N$ together with $[x_0 + N\alpha, x_0 + a]$, where

$N = \lfloor a/\alpha \rfloor$. One can verify, as in the solution procedure for differential-delay equations presented in this chapter, that all the functions y_α are uniquely defined on the entire interval $[x_0, x_0 + a]$. Moreover, they form a bounded and equicontinuous family of functions. We can easily see by using our boundedness assumption on f ($|f(x,y)| \leq C$ for some $C > 0$) and the representation given by (8.22) that for all $\alpha > 0$

$$|y_\alpha(x)| \leq |x_0| + aC \qquad \text{and} \qquad |y'_\alpha(x)| \leq C.$$

Now let $\{\alpha_n\}$ be a sequence which converges to zero, and denote (for simplicity) $y_n := y_{\alpha_n}$. By the Arzelà–Ascoli theorem this sequence contains a uniformly convergent subsequence (which we denote again by y_n) whose limit we call $y(t)$. Now using the estimate

$$\begin{aligned} |y_n(t-\alpha_n) - y(t)| &\leq |y_n(t-\alpha_n) - y_n(t)| + |y_n(t) - y(t)| \\ &\leq C\alpha_n + |y_n(t) - y(t)|, \end{aligned}$$

we see that $y_n(t-\alpha_n)$ also converges uniformly to $y(t)$. The last step is to pass to the limit $n \to \infty$ in (8.22) and use the uniform continuity of f on bounded sets (here we have finally used the continuity of f) to conclude that y satisfies the integral equation. □

The Peano result (and this proof) generalizes readily to systems of ordinary differential equations, and the methodology shown here is a standard tool in approximation procedures for partial differential equations.

Exercises

(1) Consider the system of differential-delay equations:

$$\begin{aligned} \frac{d}{dt}\phi_1(t+1) &= 1 - \frac{1}{2}t, \\ \frac{d}{dt}\phi_2(t+1) &= 2\phi_1(t) - [\phi_2(t)]^2, \end{aligned}$$

with the conditions $\phi_1(t) = 0$ for $t < 0$, and $\phi_2(t) = 0$ for $t < 0$. Find $\phi_2(2.5)$.

(2) We introduced the traffic dynamics model

$$z_j''(t+\tau) = \lambda \frac{z_j'(t) - z_{j-1}'(t)}{|z_j(t) - z_{j-1}(t)|}$$

with the idea that the braking force is directly proportional to the (negative) relative velocity and inversely proportional to the distance between the $(j-1)$st and jth car. Now assume that $z_{j-1}'(t) > z_j'(t)$, i.e., the $(j-1)$st car speeds away from the jth car. Does this model still apply for this scenario or should it be changed? What modification might be suggested?

(3) Reproduce the results of the numerical simulation in Section 8.3.2.

(4) Suppose we change the model so that the acceleration is proportional to $x_{i-1}' - x_i'$ and inversely proportional to the density. Using the forward Euler approximation, derive the resulting system of differential-delay equations and perform a numerical simulation. How do your results compare to the situation when the acceleration is proportional to the density as modelled in the text?

Chapter 9

Traffic Dynamics: Macroscopic Modelling

Concepts and Tools: Scalar conservation laws, partial differential equations

This chapter further examines the dynamics of traffic flow, but we will now ignore individual cars and instead focus on *macroscopic* traffic variables such as density

$$\rho = \rho(x,t),$$

flux

$$j = j(x,t),$$

and average speed

$$u = u(x,t).$$

The three are related by the simple identity

$$j = u\rho.$$

Conservation laws are partial differential equations of first-order linking density and flux. We will discuss the case of one scalar equation, where flux is given in terms of density by a fundamental diagram.

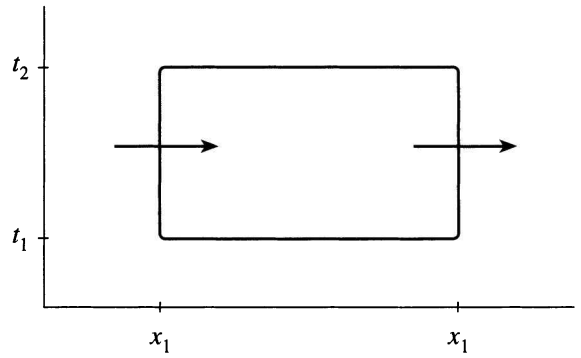


Figure 1. A part of the road in space and time is represented by the rectangle $[x_1, x_2] \times [t_1, t_2]$.

The solution method for such equations is the method of characteristics, but, as we will see, there are solution forming discontinuities (shock waves) which are physically realistic and require a weakening of the solution concept.

9.1. Scalar Conservation Laws

Consider a stretch of the road in space and time as shown by Figure 1. Let $\rho(x,t)$ be the traffic density (cars per unit length) and $j(x,t)$ the traffic flux (cars per unit time) such that

$$j(x,t) := u(x,t)\rho(x,t),$$

where u is the observed speed at location x at time t. We assume that both ρ and j are nonnegative functions. As well, we will make the (overly simplistic) assumption that the speed u is a function of density alone, i.e., $u = u(\rho)$. This means that we assume validity of a fundamental diagram as described in Chapter 8. The expected type of dependence of u on ρ will be discussed later.

The number of cars entering $[x_1, x_2]$ through the point x_1 during the time interval $[t_1, t_2]$ is given by $\int_{t_1}^{t_2} j(x_1, t)\, dt$, and the number of cars leaving $[x_1, x_2]$ through x_2 is $\int_{t_1}^{t_2} j(x_2, t)\, dt$. The number of cars in the space interval $[x_1, x_2]$ at time t_1 is given by $\int_{x_1}^{x_2} \rho(x, t_1)\, dx$, and the number at time t_2 is similarly given by $\int_{x_1}^{x_2} \rho(x, t_2)\, dx$.

Therefore, the conservation of cars on $[x_1, x_2]$ during the time interval $[t_1, t_2]$ requires

(9.1)
$$\int_{x_1}^{x_2} \rho(x, t_2)\, dx - \int_{x_1}^{x_2} \rho(x, t_1)\, dx = \int_{t_1}^{t_2} j(x_1, t)\, dt - \int_{t_1}^{t_2} j(x_2, t)\, dt.$$

Suppose now that ρ and j are continuously differentiable with respect to both x and t. We can then express the left-hand side of equation (9.1) as

$$\int_{x_1}^{x_2} [\rho(x, t_2) - \rho(x, t_1)]\, dx = \int_{x_1}^{x_2} \int_{t_1}^{t_2} \frac{\partial}{\partial t} \rho(x, t)\, dt\, dx. \tag{9.2}$$

The right-hand side can be rewritten similarly; therefore, (9.1) becomes

$$\int_{x_1}^{x_2} \int_{t_1}^{t_2} \frac{\partial}{\partial t} \rho(x, t)\, dt\, dx = -\int_{t_1}^{t_2} \int_{x_1}^{x_2} \frac{\partial}{\partial x} j(x, t)\, dx\, dt$$

or more simply,

$$\int_{t_1}^{t_2} \int_{x_1}^{x_2} \left[\frac{\partial}{\partial t} \rho(x, t) + \frac{\partial}{\partial x} j(x, t) \right] dx\, dt = 0. \tag{9.3}$$

This calculation holds for any choice of the rectangle $[x_1, x_2] \times [t_1, t_2]$. Hence, by the Fundamental Theorem of the Calculus of Variations (for a simple version see Lemma 9.1 below) this implies that

$$\frac{\partial}{\partial t} \rho(x, t) + \frac{\partial}{\partial x} j(x, t) = 0,$$

which is a first-order conservation law.

Taking $j = j(\rho)$, where we assume the validity of a fundamental diagram, and using subscripts to denote partial derivatives in the above equation yields

$$\rho_t(x, t) + [j(\rho)]_x(x, t) = 0.$$

As we will see later on, this conservation law predicts, among other things, the formation and propagation of rarefaction and shock waves.

Lemma 9.1. *If $f(x, t)$ is a continuous function defined on $\mathbb{R}^2$ such that*

$$\iint_R f(x, y)\, dx\, dy = 0$$

for each rectangle $R \subseteq \mathbb{R}^2$, then $f(x, y) \equiv 0$ for all (x, y).

Proof. Suppose that there exists a pair of coordinates (x_o, y_o) such that $f(x_o, y_o) \neq 0$. Without loss of generality assume that $f(x_o, y_o) > 0$. Since f is continuous, there is a $\delta > 0$ such that $f(x, y) > f(x_o, y_o)/2$ whenever $|x - x_o| < \delta$ and $|y - y_o| < \delta$. Therefore, if we let

$$R_\delta = \{(x, y) : |x - x_o| < \delta \text{ and } |y - y_o| < \delta\},$$

then

$$\iint_{R_\delta} f(x, y)\, dx\, dy \geq \frac{1}{2} \iint_{R_\delta} f(x_o, y_o)\, dx\, dy. \tag{9.4}$$

By assumption, the left-hand side is zero; consequently, (9.4) implies that

$$0 \geq 2\delta^2 f(x_o, y_o),$$

a contradiction. Thus $f(x, y) \equiv 0$. □

9.1.1. Simplification of the Conservation Law. The conservation law that we have established is

$$\frac{\partial}{\partial t}\rho(x, t) + \frac{\partial}{\partial x} j(x, t) = 0,$$

a first-order partial differential equation with two unknowns. To reduce this to one equation with one unknown, which we need to obtain solvable initial value problems, we have to add a state equation (or fundamental diagram) which relates the unknowns ρ and j. As an example of such a state equation, recall from the microscopic theory in the previous chapter that in a steady equilibrium, $u(\rho)$ was found to be

$$u(\rho) = \begin{cases} v_{\max} & \text{for } \rho < \rho_{\text{crit}}, \\ v_{\max} \ln\left(\frac{\rho_{\max}}{\rho}\right) \left[\ln\left(\frac{\rho_{\max}}{\rho_{\text{crit}}}\right)\right]^{-1} & \text{for } \rho \geq \rho_{\text{crit}}. \end{cases}$$

Thus $j(\rho)$ is given by

$$j(\rho) = \begin{cases} \rho v_{\max} & \text{for } \rho < \rho_{\text{crit}}, \\ \rho v_{\max} \ln\left(\frac{\rho_{\max}}{\rho}\right) \left[\ln\left(\frac{\rho_{\max}}{\rho_{\text{crit}}}\right)\right]^{-1} & \text{for } \rho \geq \rho_{\text{crit}}. \end{cases}$$

This fundamental diagram has the right behaviour structurally: linear increase of j for small ρ, then levelling off until a maximum is reached, then rapid decrease until j becomes zero at bumper-to-bumper traffic. However, the diagram was derived under equilibrium

assumptions which will in general not be satisfied; moreover, its derivative has a jump discontinuity at $\rho = \rho_{\text{crit}}$.

This jump is probably unrealistic, as is the whole specific formula of this state equation. From now on we will simply assume a general $j = j(\rho)$, differentiable for the admissible ρ, and of the basic form suggested by the equilibrium calculation. The simplest examples are given by $j(\rho) = a\rho(b-\rho)$ with a parameter $a > 0$ and $b = \rho_{\max}$. If the derivative of $j(\rho)$ exists and the solution ρ itself is smooth enough, then the equation $\rho_t + j_x = 0$ takes the form

$$\rho_t + j'(\rho)\rho_x = 0. \tag{9.5}$$

This is a first-order partial differential equation (PDE), and the method of choice to solve such equations is the method of characteristics.

9.2. Solving Initial Value Problems for First-Order PDEs

Suppose next that an initial density $\rho(x,0) = \rho_o(x)$ is given. To develop the tools necessary for the solution of initial value problems of this type, we first examine the much simpler model where

$$j(\rho) = c_1 - c_2\rho, \qquad c_1, c_2 > 0,$$

so that $j'(\rho) = -c_2$. Substituting j' into equation (9.5) produces the linear, homogeneous first-order PDE

$$\rho_t - c_2\rho_x = 0, \qquad \rho(x,0) = \rho_o(x),$$

which we can solve by using the following observation.

Consider a smooth function $f(x(t),t)$, where $(x(t),t)$ is a given curve. Using the chain rule, the total derivative of this function with respect to time t is

$$\frac{d}{dt}f(x(t),t) = \frac{\partial}{\partial t}f(x(t),t) + \frac{\partial}{\partial x}f(x(t),t)x'(t).$$

Therefore, if for the PDE under consideration we consider curves such that $x'(t) = -c_2$, then

$$\frac{d}{dt}\rho(x(t),t) = \rho_t - c_2\rho_x = 0$$

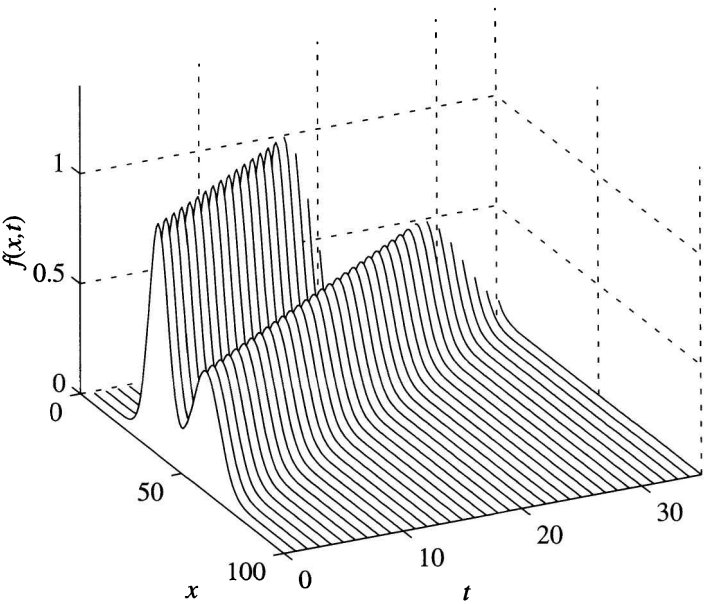


Figure 2. The initial traffic profile at time $t = 0$ is preserved and simply propagates towards smaller values of x at a velocity c_2.

(i.e., along such curves, ρ will be constant). To find these so-called *characteristic base curves* we solve $x'(t) = -c_2$ with $x(0) = x_o$. This yields $x(t) = x_o - c_2 t$. Thus, in order for ρ to be a solution to the conservation law $\rho_t - c_2 \rho_x = 0$ with initial conditions $\rho(x, 0) = \rho_o(x)$, it is necessary that

$$\rho(x(t), t) = \rho(x_o - c_2 t, t) = \text{const.} \tag{9.6}$$

Setting $t = 0$ gives

$$\rho(x_o, 0) = \text{const} = \rho_o(x_o) \tag{9.7}$$

and therefore, combining (9.6) and (9.7) gives

$$\rho(x(t), t) = \rho_o(x_o) = \rho_o(x(t) + c_2 t) \tag{9.8}$$

for any value of x. The initial traffic profile persists and moves backward at velocity c_2, as illustrated in Figure 2.

This method of solution, known as the method of characteristics, generalizes to any first-order linear conservation laws of the type

$$\rho_t + f(x,t)\rho_x = 0, \qquad \rho(x,0) = \rho_o(x), \tag{9.9}$$

where $f(x,t)$ is given.

9.2.1. Solving Initial Value Problems with the Method of Characteristics. A characteristic base curve for equation (9.9) is defined as the solution to

$$x'(t) = f(x,t), \qquad x(0) = x_o.$$

Then a solution $\rho(x(t),t)$ satisfies

$$\frac{d}{dt}\rho(x(t),t) = \rho_t + x'(t)\rho_x = \rho_t + f(x,t)\rho_x = 0, \tag{9.10}$$

and this implies, of course, that $\rho(x(t),t) = \rho_o(x_o)$. Each value of x_o determines a unique characteristic base curve if f is such that the initial value problems for the differential equation $x' = f(x,t)$ are uniquely solvable (we will assume that f is smooth enough for this). In particular, the characteristic base curves cannot cross, as such crossings would be in violation of unique solvability of the initial value problems. To clarify these ideas, we solve a particular initial value problem.

Example 9.2. Consider the initial value problem

$$\rho_t + (x\sin t)\rho_x = 0, \qquad \rho_o(x) = 1 + \frac{1}{1+x^2}. \tag{9.11}$$

The equation is a linear conservation equation (9.9) with $f(x,t) = x\sin t$. The characteristic base curves for this problem are solutions of

$$\frac{dx}{dt} = f(x,t) = x\sin t, \qquad x(0) = x_o. \tag{9.12}$$

Separating the variables and integrating equation (9.12), we obtain

$$\int \frac{dx}{x} = \int \sin t\, dt.$$

Hence,

$$\ln x = -\cos t + c, \tag{9.13}$$

where c is an integration constant. Solving for $x(t)$ in (9.13) and using the initial condition $x(0) = x_o$ gives

$$x(t) = x_o e^{1-\cos t}.$$

The function ρ is conserved along the characteristic base curves,

$$\rho(x(t), t) = \rho(x_o), \qquad x_o = x(t)e^{-1+\cos t},$$

and from

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

we find that

$$\rho\,(x, t) = 1 + \frac{1}{1 + x^2 e^{-2+2\cos t}}.$$

Now that we have developed the method, it is clear that it also applies to nonhomogeneous initial value problems of the type

$$\rho_t + f(x,t)\rho_x = g(x,t), \qquad \rho(x,0) = \rho_o(x).$$

Observe that along characteristic base curves defined by $x'(t) = f(x,t)$ with $x(0) = x_o$,

$$\frac{d}{dt}\rho(x(t),t) = g(x(t),t). \tag{9.14}$$

The curves $t \to (t, x(t), \rho(x(t),t))$ are called characteristic curves. Their projections into the (x,t)-plane are the characteristic base curves. Consequently, we compute the solution as

$$\rho(x(t),t) = \rho_o(x_o) + \int_0^t g(x(\tau),\tau)\,d\tau. \tag{9.15}$$

Notice that this is not a conservation law since in general $\rho'(x(t),t) \neq 0$.

The process extends similarly to initial value problems of the form

$$\rho_t + f(x,t)\rho_x = g(\rho, x, t).$$

As before, the characteristic base curves are the solutions to $x'(t) = f(x,t)$, and

$$\frac{d}{dt}\rho(x(t),t) = g(\rho(x(t),t), x(t), t)$$

is an ordinary differential equation for $\rho(x(t),t)$.

Example 9.3. Consider the initial value problem

$$\rho_t + e^t \rho_x = 2\rho, \qquad \rho_o(x) = 1 + \sin^2 x.$$

The characteristic base curves satisfy $x' = e^t$, $x(0) = x_o$. Solving for $x(t)$ yields $x(t) = x_o + e^t - 1$, and along these curves

$$\frac{d}{dt}\rho(x(t), t) = 2\rho(x(t), t).$$

Hence

$$\rho(x(t), t) = e^{2t}\rho(x(0), 0) = \rho(x_o)e^{2t}.$$

Substituting the initial value for $\rho_o(x)$, we have

$$\rho(x(t), t) = \rho(x_o)e^{2t} = \left(1 + \sin^2 x_o\right) e^{2t},$$

and after replacing x_o with $x - e^t + 1$,

$$\rho\,(x, t) = \left[1 + \sin^2(x + 1 - e^t)\right] e^{2t}.$$

9.2.2. Return to Nonlinear Scalar Conservation Laws. The previous sections developed the method of characteristics to solve various linear or semilinear initial value problems. We now investigate how this method can be used for nonlinear conservation equations like our traffic model.

Consider the nonlinear scalar conservation law given by

$$\rho_t + j'(\rho)\rho_x = 0 \tag{9.16}$$

with characteristic base curves that satisfy

$$x'(t) = j'(\rho(x, t)), \qquad x(0) = x_o.$$

If we assume that there exists a solution $x(t)$ to this equation (this assumes implicitly that we have a sufficiently smooth ρ), then the conservation law may be written as

$$\frac{d}{dt}\rho(x(t), t) = 0.$$

Thus, as before, $\rho(x(t), t) = \rho(x_o)$; i.e., ρ is constant along the characteristics. The characteristic equation therefore reduces to

$$x'(t) = j'(\rho(x(t), t)) = j'(\rho(x_o)),$$

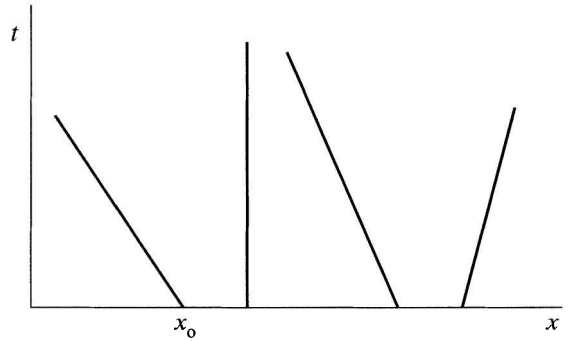


Figure 3. The characteristic base curves of the nonlinear conservation law (9.16) are straight lines as shown. The slope of each line is the reciprocal of the speed of the propagation along the line.

which, since $\rho_o(x_o)$ is a constant, is easily integrated: $x(t)$ has the form

$$x(t) = x_0 + j'(\rho(x_o))t \tag{9.17}$$

so the nonlinear scalar conservation (9.16) has characteristic base curves that are straight lines and computable explicitly. Figure 3 shows these characteristics wherein each characteristic has slope $[j'(\rho(x_o))]^{-1}$ corresponding to a propagation speed $j'(\rho(x_o))$.

9.3. The Green Light Problem

Having developed a method for solving initial value problems we now consider the Green Light problem, simply defined as the initial value problem where at time zero, $\rho_o(x) = \rho_{\max}$ for $x \leq 0$ and $\rho_o(x) = 0$ for $x > 0$.

The idea is that there is a (red) traffic light at $x = 0$, where traffic is standing bumper to bumper behind the traffic light ($x \leq 0$), the road ahead is empty, and the traffic light turns green at time 0. For simplicity we take

$$j(\rho) = \begin{cases} \rho(1-\rho) & \text{for } \rho \in [0,1], \\ 0 & \text{for } \rho > 1. \end{cases} \tag{9.18}$$

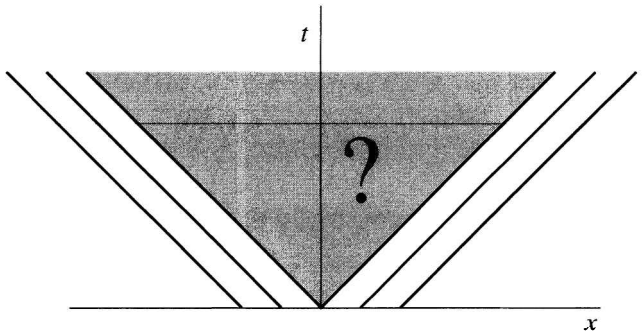


Figure 4. The characteristics associated with the initial density for the Green Light problem are shown. As is evident from the figure, it is not possible to obtain information for the shaded region using this method as no characteristics extend into the area associated with the discontinuity in the density function.

This simple traffic flux function assumes the normalization $\rho_{\max} = 1$. While the light remains red, traffic is, of course, not moving; fluxes on either side of the light are $j = 0$. For $0 < \rho < 1$, we obtain $j'(\rho) = 1 - 2\rho$. As stated,

$$\rho_o(x) = \begin{cases} 1 \text{ for } x \leq 0, \\ 0 \text{ for } x > 0. \end{cases}$$

The characteristic base lines associated with this initial density satisfy

$$x'(t) = j'(\rho_o) = \begin{cases} -1 \text{ for } x \leq 0, \\ 1 \text{ for } x > 0, \end{cases}$$

as depicted by Figure 4. The figure reveals an inadequacy in our approach: because there is a discontinuity in the initial density, the shaded area in Figure 4 is not reached by any characteristic base lines.

Thus, the method as it stands does not allow us to obtain information for this region. There is a trivial solution where all the cars simply stay put and nobody moves (the first driver at the light is asleep). The characteristic base lines in the shaded region will then emerge from the density jump at the traffic light. This solution is what is known as an unphysical shock wave. More about this later.

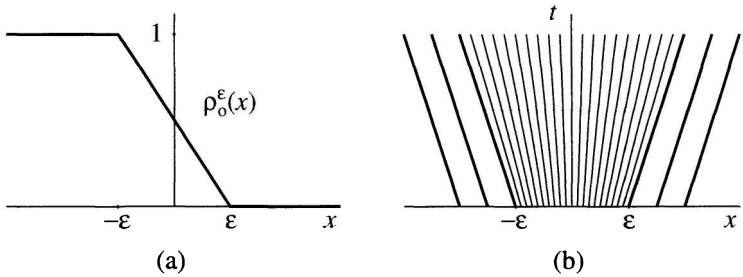


Figure 5. (a) The new initial density ρ_o^ϵ converges to ρ_o as $\epsilon \to 0$ and yields a smoother approximation. (b) Displayed are the characteristics which result from the redefined density. This solution is known as a rarefaction fan.

This solution is not the one of practical interest. We will now see that there is a realistic solution, known as a *rarefaction wave.* We resolve the problem by considering a smooth approximation of ρ_o. Define a smoothed initial density ρ_o^ϵ which is continuous and converges (pointwise everywhere except at $x = 0$) to ρ_o as $\epsilon \to 0$. For example, as shown by Figure 5(a), if we take ρ_o to be

$$\rho_o^\epsilon(x) = \begin{cases} 1 & \text{for } x \le -\epsilon, \\ \frac{1}{2} - \frac{x}{2\epsilon} & \text{for } -\epsilon < x \le \epsilon, \\ 0 & \text{for } x > \epsilon, \end{cases} \tag{9.19}$$

where $\epsilon > 0$ (i.e., if we assume that the lead drivers venture beyond the traffic light even before the light changes), then the characteristics must satisfy

$$x'(t) = j'(\rho_o(x_o)) = \begin{cases} -1 \text{ for } x_o \le -\epsilon, \\ \frac{x_o}{\epsilon} \text{ for } -\epsilon < x_o \le \epsilon, \\ 1 \text{ for } x_o > \epsilon. \end{cases}$$

The set of these characteristic base lines is known as the rarefaction fan and is shown by Figure 5(b). Unlike in the previous scenario, the x, t plane is now covered completely by characteristic base lines. As x_o goes from $-\epsilon$ to ϵ, the velocity changes linearly. In the next section we consider what happens as we let $\epsilon \to 0$.

9.3.1. The Rarefaction Wave. We denote the solution of the initial value problem with the initial density ρ_o^ϵ as $\rho^\epsilon(x,t)$. Since

$$x'(t) = j'(\rho_o^\epsilon(x_o)) = \frac{x_o}{\epsilon}$$

for $x_o \in [-\epsilon, \epsilon]$, we have from (9.17) that

$$\rho^\epsilon(x(t),t) = \rho\left(x_o + \frac{x_o}{\epsilon}t, t\right) = \text{const.} \tag{9.20}$$

As in Section 9.2, we set $t = 0$ in (9.19) to see that

$$\rho^\epsilon(x_o,0) = \text{const} = \rho_o^\epsilon(x_o) = \frac{1}{2} - \frac{x_o}{2\epsilon}. \tag{9.21}$$

Therefore, setting $x = x_o(1 + t/\epsilon)$ in (9.20), solving for x_o, and then using (9.21) yields the explicit solution

$$\rho^\epsilon(x,t) = \frac{1}{2} - \frac{x}{2\epsilon + 2t} = \frac{1}{2}\left(1 - \frac{x}{\epsilon + t}\right). \tag{9.22}$$

Taking the limit of (9.22) as $\epsilon \to 0$ removes the regularization and determines the behaviour of the ρ in this previously unknown region to be

$$\lim_{\epsilon \to 0} \rho^\epsilon(x,t) = \rho(x,t) = \frac{1}{2}\left(1 - \frac{x}{t}\right), \qquad t > 0, \tag{9.23}$$

where t is held fixed as shown in Figure 6(a).

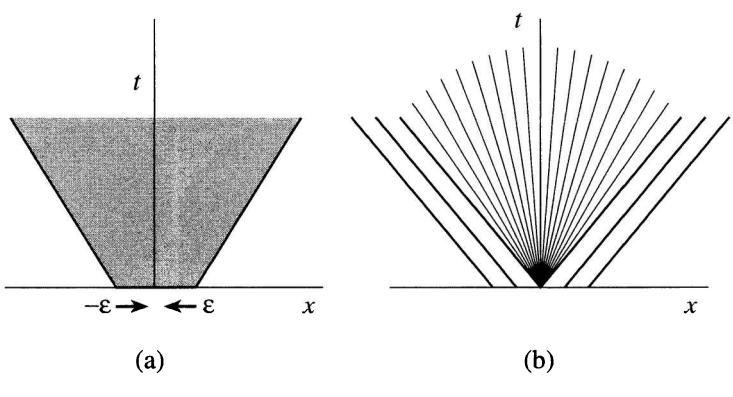


Figure 6. (a) The region for which the solution of $\rho^\epsilon(x,t)$ is valid when $\epsilon \to 0$. (b) The resulting rarefaction fan is a physically acceptable solution to the Green Light problem.

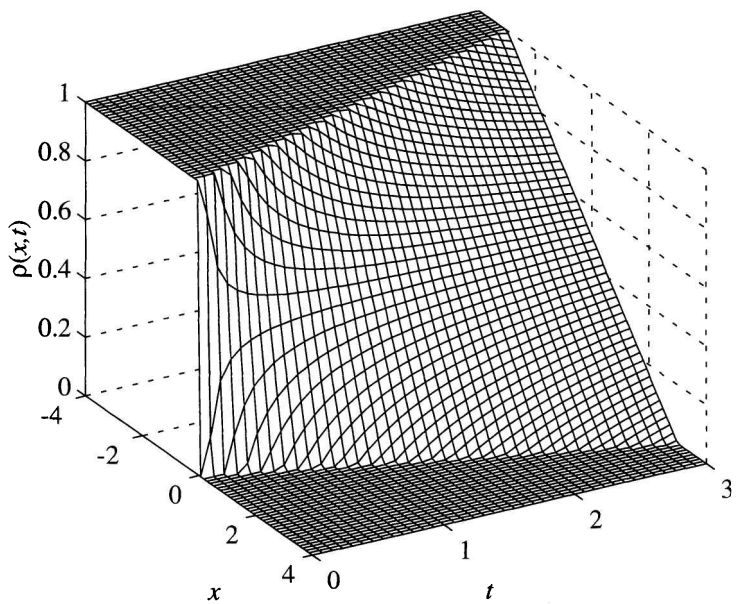


Figure 7. The time dependent density profile for the Green Light problem is illustrated. Notice that the region of density $\rho = 1$ propagates backward (negative x) with increasing t, whereas the $\rho = 0$ region propagates forward (positive x). Joining these two regions is an expanding rarefaction fan.

As is apparent from the figure, the limit function is linear in x over the rarefaction fan. If we are in the region $-1 < x/t < 1$ where $\rho \in (0, 1)$, then (9.23) gives the flux in this region as

$$j(\rho(x)) = \rho(1-\rho) = \frac{1}{4}\left(1 - \frac{x^2}{t^2}\right), \qquad t > 0.$$

Figure 6(b) shows the diverging characteristics for this region as a rarefaction fan. The associated rarefaction wave is a physically acceptable solution to the Green Light problem. For any $t > 0$

$$\rho(x,t) = \begin{cases} 0 & \text{for } x > t, \\ \frac{1}{2}\left(1 - \frac{x}{t}\right) & \text{for } -t < x \leq t, \\ 1 & \text{for } x \leq -t, \end{cases}$$

depicted in Figure 7.

9.4. Smooth Initial Data, and General Scalar Conservation Laws

The procedure described in the previous sections indicates how to solve initial value problems for smooth initial data. As there is no reason to focus only on specific traffic models, we consider general scalar conservation laws

$$\rho_t + j(\rho)_x = 0, \qquad \rho(x,0) = \rho_0, \tag{9.24}$$

and assume that the functions $j(\rho)$, $j'(\rho)$ and the initial value ρ_0 are all continuously differentiable functions. To find a smooth solution of (9.24), we use the method of characteristics as described in the previous sections. A characteristic base line emerging from x_o will satisfy the equation

$$\dot{x} = j'(\rho(x(t),t)) = j'(\rho_0(x_o)),$$

and after integration we find

$$x(t) = j'(\rho_0(x_o))t + x_o. \tag{9.25}$$

To find the characteristic base line (yes, it is a straight line) passing through the point (x,t), we therefore have to find the starting point x_o, which satisfies

$$x = x_o + j'(\rho_0(x_o))t.$$

The implicit function theorem (a classical and very powerful theorem proved in real analysis courses) guarantees that the latter equation has a unique solution provided that t is small enough. Under suitable assumptions on j and ρ_0, there may always be a unique solution, and then the solution to (9.24) is determined for all $t \geq 0$. In general, however, the unique solvability will be lost after some time t_0 depending on ρ_0; for example, the characteristic base lines may cross (many solutions).

9.4.1. General Rarefaction Waves. The analysis producing the rarefaction wave in our traffic example generalizes nicely to the general equation. Suppose that we have $\rho_0(x) = r_0$ for $x \leq x_o$, $\rho_0(x) = r_1$ for $x > x_o$, and that $j'(r_0) < j'(r_1)$. The cone between the characteristic base lines $x_l(t) := x_o + j'(r_0)t$ and $x_r(t) := x_o + j'(r_1)t$ will

not be reached by any characteristic base lines because of the jump in the initial value ρ_0. However, we can fill the gap with a rarefaction wave type solution. If we make the ansatz

$$\rho(x,t) = f\left(\frac{x - x_o}{t}\right)$$

inside this cone, then a short calculation shows that the conservation equation will be satisfied if f satisfies the identity

$$-y + j'(f(y)) = 0$$

for $y \in [j'(r_0), j'(r_1)]$, i.e., if f is a right inverse of j' on this interval. So the necessary requirement for this construction is that j' possesses an inverse on the interval; in general, it suffices that j' is continuous and monotone. It is a simple exercise (see Exercise 4) that the so-defined f takes the values r_0 and r_1 on the two edges of the cone.

In our previous traffic example, j was quadratic in ρ, hence j' was linear in ρ. But linear functions (unless they are constant) are strictly monotone and have strictly monotone inverses, which are again linear functions. This explains the linear rarefaction profile we obtained in the previous section. More examples of this will be encountered shortly, when we discuss general *Riemann problems.*

The mathematics of nonlinear scalar conservation laws that we outlined here shows that initial values with jumps are needed in this theory for two reasons. First, they are physically completely realistic (remember the Green Light problem). Second, even if we start with smooth data, the phenomenon of intersecting characteristic baselines will produce jumps in the solutions after a finite time (see Exercise 5), and then we have to deal with jumps anyway.

We have learned how to handle rarefactions. Now we will deal with intersecting characteristics, which lead to the formation and propagation of *shock waves.*

9.5. Intersecting Characteristics

We return now to the context of traffic models, although the generalization of our analysis to general conservation laws is rather straightforward.

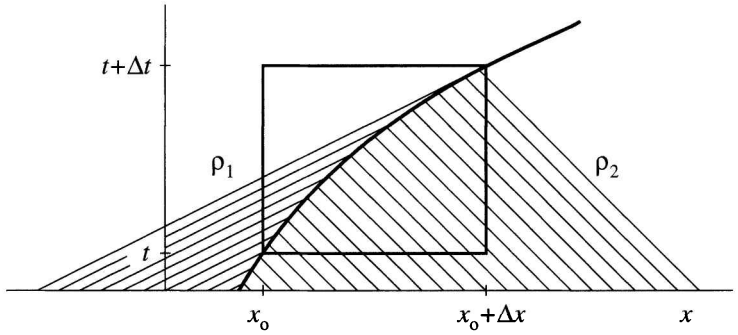


Figure 8. The curve which results from the intersection of converging characteristics is illustrated. The curve is determined by the Rankine–Hugoniot shock conditions.

In the section on the Green Light problem the initial density profile resulted in characteristic base lines that were seen to diverge and never intersect. Suppose now that we face instead an initial density ρ_o given by

$$\rho_o(x) = \begin{cases} 0 \text{ for } x \leq 0, \\ 1 \text{ for } x > 0. \end{cases}$$

Notice that there is maximal density and no flux for $x > 0$ and zero density and no flux for $x < 0$. The characteristic baselines now intersect on some curve, as sketched in Figure 8. On the curve of intersection the *solution* formed by the transport along characteristic base lines will have a discontinuity, and the partial differential equation loses its meaning on this curve. We will return to the integral form of the conservation law to determine the curve of intersection.

The characteristic baselines in our example satisfy

$$x'(t) = j'(\rho) = \begin{cases} 1 \text{ for } x \leq 0, \\ -1 \text{ for } x > 0. \end{cases}$$

We observe that the nature of the discontinuity in ρ, and thus the discontinuity in j, forces the characteristic baselines to intersect. As stated the differential equation loses its meaning at these locations. We will assume that there exists a curve on which the characteristics intersect and compute this curve.

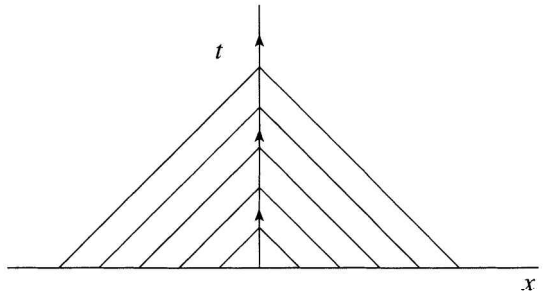


Figure 9. The curve of intersecting characteristics, called a shock, is the straight line with slope $\Delta x/\Delta t = (j_2 - j_1)/(\rho_2 - \rho_1)$.

Figure 8 displays such a curve. We use the integral formulation (9.1) to establish

$$\int_x^{x+\Delta x} \rho(y, t+\Delta t)\, dy - \int_x^{x+\Delta x} \rho(y,t)\, dy = \int_t^{t+\Delta t} j(x,\tau)\, d\tau - \int_t^{t+\Delta t} j(x+\Delta x,\tau)\, d\tau,$$

where Δt and Δx denote (small) changes in t and x so that (x,t) and $(x+\Delta x, t+\Delta t)$ are both on the curve. As both ρ and j are (in the example) constant on either side of the curve, we can write this as

$$\rho_1 \Delta x - \rho_2 \Delta x = j_1 \Delta t - j_2 \Delta t, \tag{9.26}$$

where ρ_i and $j_i, i = 1,2$, are the constant values of ρ and j to the left and right of the curve, respectively. Grouping like terms and rearranging (9.26), we obtain

$$\frac{\Delta x}{\Delta t} = \frac{j_2 - j_1}{\rho_2 - \rho_1}. \tag{9.27}$$

The left-hand side here is constant if ρ_1 and ρ_2 are constant. Therefore, since $\Delta x/\Delta t$ is constant, the intersection curve of the converging characteristics is a straight line with slope $s = \Delta x/\Delta t$. In our example $\Delta x/\Delta t = 0$ is shown by Figure 9.

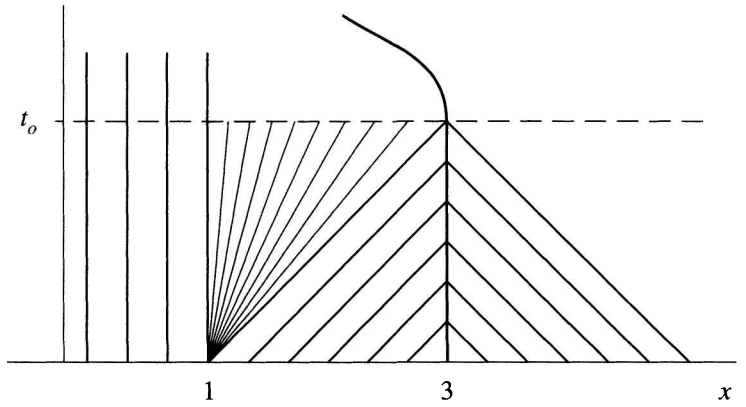


Figure 10. For the density function of Example 9.4 the resulting characteristics diverge at one discontinuity of $\rho_o(x)$ and converge at the other.

The condition given by expression (9.27) is known as the Rankine–Hugoniot condition on the shock speed, and the discontinuity propagating along this line of interaction is called a shock wave. In our example the shock wave is stationary (which is perfectly reasonable, as traffic is standing and no new cars arrive to add to the jam). Next we apply the new concepts to a more complicated example.

Example 9.4. We consider the conservation law $\rho_t + j_x = 0$ with the traffic flux given by $j(\rho) = 4\rho(2-\rho)$ and the initial value $\rho_o(x)$ is

$$\rho_o(x) = \begin{cases} 1 \text{ for } x \leq 1, \\ \frac{1}{2} \text{ for } 1 < x \leq 3, \\ \frac{3}{2} \text{ for } x > 3. \end{cases}$$

Wherever ρ is sufficiently smooth, the conservation law takes the form

$$\rho_t + j'(\rho)\rho_x = \rho_t + 8(1-\rho)\rho_x = 0$$

with

$$j'(\rho) = \begin{cases} 0 \text{ for } x \leq 1, \\ 4 \text{ for } 1 < x \leq 3, \\ -4 \text{ for } x > 3. \end{cases}$$

Figure 10 shows the associated characteristics.

For the converging characteristics, one obtains a shock propagating at speed s given by the Rankine–Hugoniot conditions (9.27)

$$s = \frac{j(3/2) - j(1/2)}{3/2 - 1/2} = 0.$$

The shock is stationary and remains at $x = 3$ until it interacts with the rarefaction wave. It is easy to determine how the shock will behave after it begins to interact with the rarefaction wave emerging from $x = 1$. Let $\epsilon > 0$ be sufficiently small and replace $\rho = 1/2$ by $\rho = 1/2 + \epsilon$ in the Rankine–Hugoniot condition. Computing the corresponding shock speed s, one finds that

$$s(\epsilon) = \frac{j\left(\frac{3}{2}\right) - j\left(\frac{1}{2} + \epsilon\right)}{\frac{3}{2} - \left(\frac{1}{2} + \epsilon\right)} = \frac{3 - \left(3 + 4\epsilon - 4\epsilon^2\right)}{1 - \epsilon} = -4\epsilon,$$

and so the shock will bend to the left. We can do much better. To compute exactly what happens after the shock and rarefaction waves begin to interact we will use the explicit density distribution of the rarefaction fan. The interaction begins at time $t_o = 1/2$ and position $\sigma_o = 3$ (t_o is determined from the equation $1 + 4t_o = 3$). After the interaction has started, the shock wave will no longer propagate along a straight line. Rather, the location of the shock wave is described in parameterized form by $\sigma = \sigma(t)$, with $\sigma(1/2) = 3$. The shock speed after $t_o = 1/2$ is described by (9.27) as

$$\frac{d\sigma}{dt} = \frac{j_2 - j_1}{\rho_2 - \rho_1} = \frac{j\left(\frac{3}{2}\right) - j\left(1 - \frac{\sigma(t)-1}{8t}\right)}{\frac{3}{2} - \left(1 - \frac{\sigma(t)}{8t}\right)} = -2\left(1 - \frac{\sigma - 1}{4t}\right), \tag{9.28}$$

where we have inserted the explicit form for the rarefaction fan density

$$\rho(x,t) = 1 - \frac{x-1}{8t}.$$

Standard solution procedures for ordinary differential equations show that the solution to (9.28) is

$$\sigma(t) = 1 + 4\left(\sqrt{2t} - t\right), \qquad \frac{1}{2} \le t \le 2.$$

The fan closes off at $t = 2$ since $\sigma(2) = 1$, the location of the $\rho = 1$ characteristics. At $t = 2$ a new shock emerges at $x = 1$, with left density $\rho_1 = 1$ and right density $\rho_2 = 3/2$. From (9.27) we find a

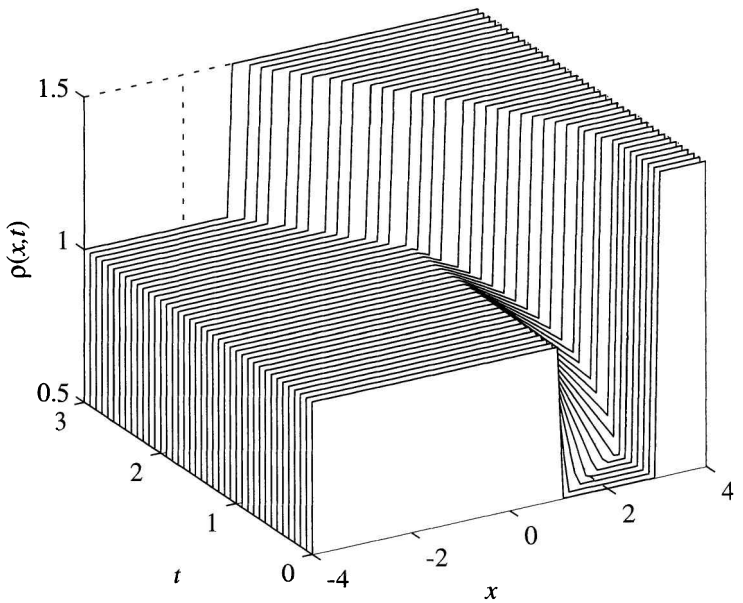


Figure 11. The solution has three separate domains. For $t \in [0, 1/2]$, a rarefaction fan spreads from $x = 1$ to $x = 3$. When it comes into contact with the characteristic at $x = 3$, it bends back to the left, eventually closing off the fan at $t = 2$. For $t > 2$, the jump in the density propagates with a speed of $s = -2$.

shock speed $s = -2$. Collecting all these results produces the full explicit solution

$$0 \le t \le 1/2: \qquad \rho(x,t) = \begin{cases} 1 & \text{for } x \le 1, \\ 1 - \frac{x-1}{8t} & \text{for } 1 < x \le 1 + 4t, \\ 1/2 & \text{for } 1 + 4t < x \le 3, \\ 3/2 & \text{for } x > 3; \end{cases}$$

$$1/2 < t \le 2: \qquad \rho(x,t) = \begin{cases} 1 & \text{for } x \le 1, \\ 1 - \frac{x-1}{8t} & \text{for } 1 < x \le 1 + 4(\sqrt{2t} - t), \\ 3/2 & \text{for } x > 1 + 4(\sqrt{2t} - t); \end{cases}$$

$$t > 2: \qquad \rho(x,t) = \begin{cases} 1 & \text{for } x \le 5 - 2t, \\ 3/2 & \text{for } x > 5 - 2t, \end{cases}$$

which is depicted in Figure 11.

9.5.1. Riemann Problems. Initial value problems for conservation laws with piecewise constant initial data, such as in Example 9.4, are known as *Riemann problems.* We present a second example in detail.

Suppose that the traffic flux is given by the simple function

$$j(\rho) = 3\rho(2-\rho) = 6\rho - 3\rho^2,$$

and $\rho_o(x)$ is

$$\rho_o(x) = \begin{cases} \frac{1}{2} \text{ for } x \leq 1, \\ \frac{3}{2} \text{ for } 1 < x < 3, \\ 1 \text{ for } x \geq 3. \end{cases}$$

Here the traffic flux is zero at $\rho = 2$, which must be interpreted as the maximal density. The characteristics associated with this equation are

$$x'(t) = j'(\rho(x,t)) = 6(1-\rho),$$

so that

$$x'(t) = \begin{cases} 3 \text{ for } x \leq 1, \\ -3 \text{ for } 1 < x < 3, \\ 0 \text{ for } x \geq 3. \end{cases}$$

There must be a shock where the characteristics converge. From the Rankine–Hugoniot condition (9.27), we determine the shock speed s using

$$s = \frac{j_2 - j_1}{\rho_2 - \rho_1}$$

and evaluating this over the region $-\infty < x < 3$, where the characteristic base lines converge yields

$$s = \frac{j(3/2) - j(1/2)}{3/2 - 1/2} = \frac{9/4 - 9/4}{3/2 - 1/2} = 0.$$

The result is the stationary shock displayed in Figure 12.

Figure 12 further illustrates the interaction of the various pieces of the solution. At the point $x = 3$, a linear profile rarefaction fan will emerge by interpolation between the characteristics associated with $\rho = 3/2$ and $\rho = 1$. After time $t_o = 2/3$, this rarefaction fan will interact with the shock, causing it to bend to the right, as shown in Figure 12. We leave it as an exercise to the student to determine

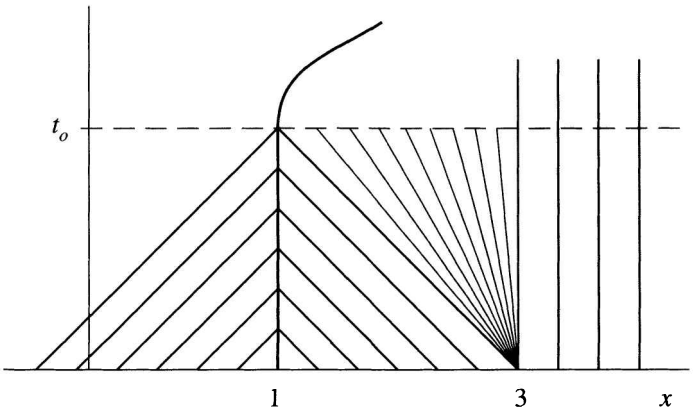


Figure 12. A stationary shock at $x = 1$ (the case where $s = 0$) is displayed. A rarefaction fan emerges at $x = 3$ and eventually interacts with the shock causing it to bend to the right as shown.

the curve along which the shock propagates after the beginning of the interaction. The main step is the determination of the density in the rarefaction fan. Note that it must be of the form

$$\rho(x,t) = a + b\left(\frac{x-3}{t}\right)$$

for some a and b. The student is left to fill in the details.

9.5.2. Unphysical Shocks. Entropy Conditions. A rarefaction fan was found to be an acceptable solution to the Green Light problem considered in Section 9.3. However, we saw there that this initial value problem also possesses a shock wave solution. Recall that for this example $j(\rho) = \rho(1-\rho)$ and $\rho_o(x)$ was given by

$$\rho_o(x) = \begin{cases} 1 \text{ for } x \leq 0, \\ 0 \text{ for } x > 0. \end{cases}$$

The shock speed will be

$$s = \frac{j_2 - j_1}{\rho_2 - \rho_1} = \frac{0-0}{0-1} = 0,$$

and the result is the stationary shock shown by Figure 13(a).

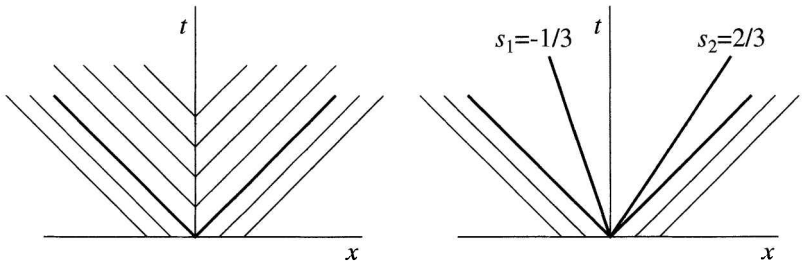


Figure 13. (a) The associated characteristics emanate from the stationary shock. (b) An intermediate value of $\rho_o = 1/3$ generates two shocks which emanate from the origin.

In this particular case, the associated characteristics emanate from the shock. Since the origin of the values emerging from the shock is unclear, these *unphysical shocks* are in violation of the principle of causality and are therefore not considered an acceptable solution to the initial value problem at hand. Shocks are considered to be valid solutions only if characteristics terminate at a shock and not vice versa.

Another way of solving this initial value problem is to choose an intermediate value so that $\rho_o(x)$ becomes

$$\rho_o(x) = \begin{cases} 1 \text{ for } x < 0, \\ \frac{1}{3} \text{ for } x = 0, \\ 0 \text{ for } x > 0. \end{cases}$$

This results in two separate shocks emanating from $\rho = 0$: one with speed

$$s_1 = \frac{j(1/3) - j(1)}{1/3 - 1} = -\frac{1}{3}$$

and the other with speed

$$s_2 = \frac{j(1/3) - j(0)}{1/3 - 0} = \frac{2}{3}.$$

As is apparent from Figure 13(b), the resulting shocks are again in violation of causality. It is clear that more and more unphysical shocks can be placed into the gap which ought to be filled by the rarefaction wave.

9.5.3. Riemann Problems Are Numerical Tools. Realistic initial values can be approximated (in a mathematically precise sense) by piecewise constant functions; the solutions of the corresponding Riemann problems are then approximating the solutions of the original initial value problems. As Riemann problems are explicitly solvable, they can, therefore, be used effectively for numerical purposes. See [**J**] for a good graduate text where this idea is discussed in detail.

Exercises

(1) Use the method of characteristics to solve the following initial value problems:
   (a) $\rho_t + 2\rho_x = 0, \quad \rho_o(x) = e^{-x^2}$.
   (b) $\rho_t + 2t\rho_x = 0, \quad \rho_o(x) = e^{-x^2}$.
   (c) $\rho_t + j_x = 0$, $j = j(\rho) = 2\rho - \rho^2$, $\rho_o(x) = \begin{cases} 1 & \text{for } x \leq 0, \\ 0 & \text{for } x > 0. \end{cases}$

(2) Repeat problem 1(c) with the initial condition,

$$\rho_o(x) = \begin{cases} \frac{1}{2} & \text{for } x \leq 0, \\ \frac{5}{4} & \text{for } x > 0. \end{cases}$$

Show that this leads to a shock wave propagating in the traffic flow, and determine the shock speed.

(3) Consider the scenario presented in Section 9.5.1. Compute the curve along which the shock wave propagates after it begins to interact with the rarefaction wave.

(4) Consider the conservation law $\rho_t + j'(\rho)\rho_x = 0$ with the piecewise constant initial condition

$$\rho_o(x) = \begin{cases} r_o \text{ for } x \leq x_o, \\ r_1 \text{ for } x > x_o. \end{cases}$$

In addition, suppose that $j'(r_o) < j'(r_1)$.
   (a) Show that if

$$\rho(x,t) = f\left(\frac{x - x_o}{t}\right)$$

satisfies the conservation law, then f must be the right inverse of j'. That is, $j'(f(y)) = y$ for all $y \in [j'(r_o), j'(r_1)]$.

(b) Prove that if a real valued function $f : A \to B$ is continuous and monotone on its domain, then it has a unique inverse $g : B \to A$ with $g \circ f = 1_A$, $f \circ g = 1_B$.

(c) Suppose that j' is continuous and monotone on its domain so that the f defined in part (b) exists. Show that f takes the values r_0 and r_1 on the two edges of the cone between the characteristic base curves $x_l(t) := x_o + j'(r_0)t$ and $x_r(t) := x_o + j'(r_1)t$.

(5) In this problem we investigate the so-called Burger equation

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}\left(\frac{1}{2}\rho^2\right) = 0$$

so that $j(\rho) = \rho^2/2$.

(a) Show that the initial value problem for $\rho_0(x) = e^x$ is uniquely and smoothly solvable (it cannot be done explicitly because it involves the solution of a transcendental equation).

(b) Show that the initial value problem for the Burger equation with initial value

$$\rho_0(x) = \begin{cases} 1 & \text{for } x < 0, \\ 1 - x & \text{for } 0 \le x \le 1, \\ 0 & \text{for } x > 1, \end{cases}$$

possesses a continuous solution while $0 \le t < 1$. Show that characteristic baselines cross at $t = 1$. Where do they cross?

Bibliography

[A] Aczél, J. & Dhombres, J. (1989). *Functional Equations in Several Variables.* New York: Cambridge University Press.

[B] Boots, B., Okabe, A., & Sugihara, K. (1992). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams.* New York: John Wiley & Sons, Inc.

[C] Clark, C. (1976). *Mathematical Bioeconomics: The Optimal Management of Renewable Resources.* John Wiley & Sons, Inc.

[D] Fox, R.W. & McDonald, A.T. (1985). *Introduction to Fluid Mechanics.* New York: John Wiley & Sons, Inc.

[E] Giordano, F.R. & Weir, M.D. (1985). *A First Course in Mathematical Modeling.* Monterey: Brooks–Cole.

[F] Gough, T.E. & Illner, R. (1999). *Modeling Crystallization Dynamics When the Avrami Model Fails.* VLSI Design, **9** (4), 377-383.

[G] Gough, T.E., Rowat, T.E. & Illner, R. (1998). *Modelling the solid state reaction* $CO_2 \cdot C_2H_2 \rightarrow CO_2 + C_2H_2$. Chemical Physics Letters, **298** (1), 196-200.

[H] Illner, R. (1994). *Formal Justice and Functional Equations.* Mathematics Magazine, **67** (3), 214-219.

[I] Jordan, D.W. & Smith, P. (1987). *Nonlinear Ordinary Differential Equations.* Oxford: Claredon Press.

[J] Le Veque, R.J. (1992). *Numerical Methods for Conservation Laws.* Basel: Birkhäuser.

[K] Mesterton–Gibbons, M. (1989). *A Concrete Approach to Mathematical Modelling.* California: Addison–Wesley, 34-35, 57-58, 76-83, 115-123.

[L] Quine, M.P. & Watson, D.F. (1984). *Radial Generation of n-Dimensional Poisson Processes.* Journal of Applied Probability, **21**, 548-557.

[M] Rowat, T.E. (1997). *Stoichiometry and Stability of Binary Phase Crystals Formed between Acetylene and Nitrous Oxide/Carbon Dioxide.* University of Victoria. Ph.D. Thesis.

[N] Salinger, G.L. & Sears, F.W. (1975). *Thermodynamics, Kinetic Theory and Statistical Thermodynamics.* Massachusetts: Addison–Wesley.

[O] Sołtan, K.E. (1987). *The Causal Theory of Justice.* California: University of California Press.

[P] Sparrow, C. (1982). *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors.* New York: Springer–Verlag.

[Q] Stelson, H.E. (1957). *Mathematics of Finance.* Princeton: Van Nostrand.

[R] Streeter, V.L. (1961). *Handbook of Fluid Dynamics.* New York: McGraw–Hill, 3-12.

[S] Symon, K.R. (1971). *Mechanics.* Massachusetts: Addison–Wesley.

[T] Taylor, Sir Geoffrey, F.R.S. (1950). *The Formation of a Blast Wave by a Very Intense Explosion. II: The Atomic Explosion of 1945.* Proceedings of the Royal Society, Series A, **201**, 175-186.

[U] Wan, F.Y.M. (1989). *Mathematical Models and Their Analysis.* New York: Harper & Row.

[V] Wunderlich, B. (1976). *Macromolecular Physics, vol. 2: Crystal Nucleation, Growth, Annealing.* San Diego: Academic Press Inc. 132-147.

[W] Zima, P. & Brown, R.L. (1993). *Mathematics of Finance.* Toronto: McGraw–Hill Ryerson Ltd.

Titles in This Series

27 **Reinhard Illner, C. Sean Bohun, Samantha McCollum, and Thea van Roode,** Mathematical modelling: A case studies approach, 2005

26 **Robert M. Hardt, Editor,** Six themes on variation, 2004

25 **S. V. Duzhin and B. D. Chebotarevsky,** Transformation groups for beginners, 2004

24 **Bruce M. Landman and Aaron Robertson,** Ramsey theory on the integers, 2004

23 **S. K. Lando,** Lectures on generating functions, 2003

22 **Andreas Arvanitoyeorgos,** An introduction to Lie groups and the geometry of homogeneous spaces, 2003

21 **W. J. Kaczor and M. T. Nowak,** Problems in mathematical analysis III: Integration, 2003

20 **Klaus Hulek,** Elementary algebraic geometry, 2003

19 **A. Shen and N. K. Vereshchagin,** Computable functions, 2003

18 **V. V. Yaschenko, Editor,** Cryptography: An introduction, 2002

17 **A. Shen and N. K. Vereshchagin,** Basic set theory, 2002

16 **Wolfgang Kühnel,** Differential geometry: curves - surfaces - manifolds, 2002

15 **Gerd Fischer,** Plane algebraic curves, 2001

14 **V. A. Vassiliev,** Introduction to topology, 2001

13 **Frederick J. Almgren, Jr.,** Plateau's problem: An invitation to varifold geometry, 2001

12 **W. J. Kaczor and M. T. Nowak,** Problems in mathematical analysis II: Continuity and differentiation, 2001

11 **Michael Mesterton-Gibbons,** An introduction to game-theoretic modelling, 2000

10 **John Oprea,** The mathematics of soap films: Explorations with Maple®, 2000

9 **David E. Blair,** Inversion theory and conformal mapping, 2000

8 **Edward B. Burger,** Exploring the number jungle: A journey into diophantine analysis, 2000

7 **Judy L. Walker,** Codes and curves, 2000

6 **Gérald Tenenbaum and Michel Mendès France,** The prime numbers and their distribution, 2000

5 **Alexander Mehlmann,** The game's afoot! Game theory in myth and paradox, 2000

4 **W. J. Kaczor and M. T. Nowak,** Problems in mathematical analysis I: Real numbers, sequences and series, 2000

3 **Roger Knobel,** An introduction to the mathematical theory of waves, 2000

TITLES IN THIS SERIES

2 **Gregory F. Lawler and Lester N. Coyle,** Lectures on contemporary probability, 1999

1 **Charles Radin,** Miles of tiles, 1999

Mathematical modelling is a subject without boundaries. It is the means by which mathematics becomes useful to virtually any subject. Moreover, modelling has been and continues to be a driving force for the development of mathematics itself. This book explains the process of modelling real situations to obtain mathematical problems that can be analyzed, thus solving the original problem.

The presentation is in the form of case studies, which are developed much as they would be in true applications. In many cases, an initial model is created, then modified along the way. Some cases are familiar, such as the evaluation of an annuity. Others are unique, such as the fascinating situation in which an engineer, armed only with a slide rule, had 24 hours to compute whether a valve would hold when a temporary rock plug was removed from a water tunnel.

Each chapter ends with a set of exercises and some suggestions for class projects. Some projects are extensive, as with the explorations of the predator-prey model; others are more modest.

The text was designed to be suitable for a one-term course for advanced undergraduates. The selection of topics and the style of exposition reflect this choice. The authors have also succeeded in demonstrating just how enjoyable the subject can be.

This is an ideal text for classes on modelling. It can also be used in seminars or as preparation for mathematical modelling competitions.

ISBN 978-0-8218-3650-7

9 780821 836507

STML/27

For additional information and updates on this book, visit

www.ams.org/bookpages/stml-27