

Difference and Differential Equations

African Institute of Mathematics Texts

Editorial Board

Editor 1

Editor 2

Editor 3

Editor 4

Modelling with Difference and
Differential Equations

Jacek Banasiak
University of KwaZulu-Natal



copyright information here

Contents

Preface *page xi*

1	Basic models leading to difference equations	1
1.1	Basic difference equations of finance mathematics	1
1.1.1	Compound interest and loan repayments	1
1.1.2	Some money related models	4
1.2	Difference equations of population theory	8
1.2.1	Single equations for unstructured population models	8
1.2.2	Structured populations and linear systems of difference equations	19
1.2.3	General structured population models	25
1.2.4	Markov chains	28

1.2.5	Interacting populations and nonlinear systems of difference equations	30
2	Basic differential equations models	38
2.1	Equations related to financial mathematics	39
2.1.1	Continuously compounded interest and loan repayment	39
2.1.2	Continuous models in economical applications	41
2.2	Other models leading to exponential growth formula	44
2.3	Continuous population models: first order equations	45
2.3.1	Exponential growth	46
2.3.2	Logistic differential equation	48
2.3.3	Other population models with restricted growth	50
2.4	Equations of motion: second order equations	51
2.4.1	A waste disposal problem	52
2.4.2	Motion in a changing gravitational field	53
2.5	Equations coming from geometrical modelling	54
2.5.1	Satellite dishes	54
2.5.2	The pursuit curve	56
2.6	Modelling interacting quantities – systems of differential equations	59
2.6.1	Two compartment mixing – a system of linear equations	59
2.6.2	Continuous population models	61

2.6.3	Continuous model of epidemics – a system of nonlinear differential equations	65
2.6.4	Predator–prey model – a system of nonlinear equations	67
3	Solutions and applications of discrete mod- els	70
3.1	Inverse problems – estimates of the growth rate	70
3.2	Drug release	73
3.3	Mortgage repayment	74
3.4	Conditions for the Walras equilibrium	76
3.5	Some explicitly solvable nonlinear models	78
4	Basic differential equation models – solu- tions	82
4.1	What are differential equations?	82
4.2	Cauchy problem for first order equations	89
4.3	Miscellaneous applications	100
4.3.1	Exponential growth	100
4.3.2	Continuous loan repayment	102
4.3.3	The Neo-classical model of Economic Growth	104
4.3.4	Logistic equation	105
4.3.5	The waste disposal problem	107
4.3.6	The satellite dish	113
4.3.7	Pursuit equation	117
4.3.8	Escape velocity	120
4.4	Exercises	124
5	Qualitative theory for a single equation	126

5.1	Direct application of difference and differential equations	126
5.1.1	Sustainable harevesting	126
5.1.2	Maximal concentration of a drug in an organ	127
5.1.3	A nonlinear pendulum	128
5.2	Equilibria of first order equations	129
5.2.1	Equilibria for differential equations	130
5.2.2	Crystal growth—a case study	137
5.2.3	Equilibrium points of difference equations	144
6	From discrete to continuous models and back	170
6.1	Discretizing differential equations	171
6.1.1	The Euler method	171
6.1.2	The time-one map	172
6.1.3	Discrete and continuous exponential growth	173
6.1.4	Logistic growth in discrete and continuous time	174
6.1.5	Discrete models of seasonally changing population	178
6.2	A comparison of stability results for differential and difference equations	182
7	Simultaneous systems of equations and higher order equations	189
7.1	Systems of equations	189
7.1.1	Why systems?	189
7.1.2	Linear systems	190
7.1.3	Algebraic properties of systems	193

7.1.4	The eigenvalue-eigenvector method of finding solutions	199
7.2	Second order linear equations	223
7.2.1	Homogeneous equations	225
7.2.2	Nonhomogeneous equations	228
7.2.3	Applications	236
8	Qualitative theory of differential and difference equations	245
8.1	Introduction	245
8.2	The phase-plane and orbits	248
8.3	Qualitative properties of orbits	252
8.4	An application to predator-prey models	257
8.4.1	Lotka-Volterra model	257
8.4.2	Modified Lotka-Volterra model	262
8.5	Stability of linear systems	267
8.5.1	Planar linear systems	267
8.6	Stability of equilibrium solutions	276
8.6.1	Linear systems	277
8.6.2	Nonlinear systems	279
Appendix 1 Methods of solving first order difference equations		286
Appendix 2 Basic solution techniques in differential equations		288
<i>References</i>		315
<i>Index</i>		316

Preface

Engineers, natural scientists and, increasingly, researchers and practitioners working in economical and social sciences, use mathematical models of the systems they are investigating. Models give simplified descriptions of real-life problems so that they can be expressed in terms of mathematical equations which can be, hopefully, solved in one way or another. Mathematical modelling is a subject difficult to teach but it is what applied mathematics is about. The difficulty is that there are no set rules, and the understanding of the 'right' way to model can be only reached by familiarity with a number of examples. This, together with basic techniques for solving the resulting equations, is the main content of this course.

Despite these difficulties, applied mathematicians have a procedure that should be applied when building models. First of all, there must be a phenomenon of interest that one wants to describe or, more importantly, to explain and make predictions about. Observation of this phenomenon allows to make hypotheses about which quantities are most

relevant to the problem and what are the relations between them so that one can devise a hypothetical mechanism that can explain the phenomenon. The purpose of a model is then to formulate a description of this mechanism in quantitative terms, that is, as mathematical equations, and the analysis of the resulting equations. It is then important to interpret the solutions or other information extracted from the equations as the statements about the original problem so that they can be tested against the observations. Ideally, the model also leads to predictions which, if verified, lend authenticity to the model. It is important to realize that modelling is usually an iterative procedure as it is very difficult to achieve a balance between simplicity and meaningfulness of the model: often the model turns out to be too complicated to yield itself to an analysis, and often it is over-simplified so that there is insufficient agreement between the actual experimental results and the results predicted from the model. In both these cases we have to return to the first step of modelling and try to remedy the ills.

The first step in modelling is the most creative but also the most difficult, involving often a concerted effort of specialists in many diverse fields. Hence, though we describe a number of models in detail, starting from first principles, the main emphasis of the course is on the later stages of the modelling process, that is: introducing mathematical symbols and writing assumptions as equations, analysing and/or solving these equations and interpreting their solutions in the language of the original problem and reflecting on whether the answers seem reasonable.

In most cases discussed here a model is a representation

of a process, that is, it describes a change of the states of some system in time. There are two ways of describing what happens to a system: discrete and continuous. Discrete models correspond to the situation in which we observe a system in regular finite time intervals, say, every second or every year and relate the observed state of the system to the states at the previous instants. Such a system is modelled through difference equations. In the continuous cases we treat time as a continuum allowing observation of the system at any instant. In such a case the model expresses relations between the rates of change of various quantities rather than between the states at various times and, as rates of change are given by derivatives, the model is represented by differential equations.

In the next two sections of this chapter we shall present some simple discrete and continuous models. These models are presented here as an illustration of the above discussion. Their analysis, and a discussion of more advanced models, will appear later in the course.

1

Basic models leading to difference equations

1.1 Basic difference equations of finance mathematics

1.1.1 Compound interest and loan repayments

Compound interest is relevant to loans or deposits made over longer periods. The interest is added to the initial sum at regular intervals, called conversion periods, and the new amount, rather than the initial one, is used for calculating the interest for the next conversion period. The fraction of a year occupied by the conversion period is denoted by α so that the conversion period of 1 month is given by $\alpha = 1/12$. Instead of saying that the conversion period is 1 month we say that the interest is compounded monthly.

For an annual interest rate of $p\%$ and conversion period equal to α , the interest earned for the period is equal to $ap\%$ of the amount on deposit at the start of the period,

that is

$$\begin{pmatrix} \text{amount on} \\ \text{deposit} \\ \text{after } k + 1 \\ \text{conversion} \\ \text{periods} \end{pmatrix} = \begin{pmatrix} \text{amount on} \\ \text{deposit} \\ \text{after } k \\ \text{conversion} \\ \text{periods} \end{pmatrix} + \frac{\alpha p}{100} \begin{pmatrix} \text{amount on} \\ \text{deposit} \\ \text{after } k \\ \text{conversion} \\ \text{periods} \end{pmatrix}$$

To express this as a difference equation, for each k let $S(k)$ denote the amount on deposit after k conversion periods. Thus

$$S(k + 1) = S(k) + \frac{\alpha p}{100} S(k) = S(k) \left(1 + \frac{\alpha p}{100} \right)$$

which is a simple first-order (that is, expressing the relation only between the consecutive values of the unknown sequence) difference equation. Here, S_k follows the geometric progression so that

$$S(k) = \left(1 + \frac{\alpha p}{100} \right)^k S_0 \quad (1.1)$$

gives the so-called compound interest formula. If we want to measure time in years, then $k = t/\alpha$ where t is time in years. Then (1.1) takes the form

$$S(t) = \left(1 + \frac{\alpha p}{100} \right)^{t/\alpha} S_0 \quad (1.2)$$

It is worthwhile to introduce here the concept of effective interest rate. First we note that in (1.2) with $S_0 = 1$

$$S(1) = \left(1 + \frac{\alpha p}{100} \right)^{1/\alpha} = 1 + \frac{p}{100} + \dots > 1 + \frac{p}{100}$$

so if the interest is compounded several times per year the increase in savings is bigger than if it was compounded

1.1 Basic difference equations of finance mathematics 3

annually. This is the basis of defining the effective interest rate r_{eff} (relative to the conversion period), namely

$$1 + r_{eff} = \left(1 + \frac{\alpha p}{100}\right)^{1/\alpha}; \quad (1.3)$$

that is, r_{eff} is the interest rate which, compounded annually, would give the same return as the interest p compounded with conversion period α .

A slight modification of the above argument can be used to find the equation governing a loan repayment. The scheme described here is usually used for the repayment of house or car loans. Repayments are made at regular intervals and usually in equal amounts to reduce the loan and to pay the interest on the amount still owing.

It is supposed that the compound interest at $p\%$ is charged on the outstanding debt with the conversion period equal to the same fraction α of the year as the period between the repayments. Between payments, the debt increases because of the interest charged on the debt still outstanding after the last repayment. Hence

$$\left\{ \begin{array}{l} \text{debt after} \\ k + 1 \text{ payments} \end{array} \right\} = \left\{ \begin{array}{l} \text{debt after} \\ k \text{ payments} \end{array} \right\} + \left\{ \begin{array}{l} \text{interest} \\ \text{on this debt} \end{array} \right\} - \{\text{payment}\}$$

To write this as a difference equation, let D_0 be the initial debt to be repaid, for each k let the outstanding debt after the k th repayment be D_k , and let the payment made after each conversion period be R . Thus

$$D_{k+1} = D_k + \frac{\alpha p}{100} D_k - R = D_k \left(1 + \frac{\alpha p}{100}\right) - R. \quad (1.4)$$

We note that if the instalment was paid at the beginning

of the conversion period, the equation would take a slightly different form

$$D_{k+1} = D_k - R + \frac{\alpha p}{100}(D_k - R) = (D_k - R) \left(1 + \frac{\alpha p}{100}\right). \quad (1.5)$$

The reason for the change is that the interest is calculated from the debt D_k reduced by the payment R done at the beginning of the conversion period. These equations are more difficult to solve. We shall discuss general methods of solving first order difference equations in Section 1 of Chapter 3.

The modelling process in these two examples was very simple and involved only translation of given rules into mathematical symbols. This was due to the fact that there was no need to discover these rules as they are explicitly stated in bank's regulations. In the next sections we shall attempt to model behaviour of more complicated systems and then modelling will involve making some hypotheses about the rules governing them.

1.1.2 Some money related models

Walras equilibrium

We study pricing of a certain commodity. According to Walras the market price is the equilibrium price; that is, it is the price at which demand D for this commodity equals supply S of it. Let $p(n)$ denotes the price in period n . The assumptions are that $D(n) = f(p(n))$ where f is a decreasing function and $S(n) = g(p(n-1))$, where g is an increasing function. The fact that S depends on $p(n-1)$ and D depends on $p(n)$ is due to the fact that producers need some time to react to changing prices whereas consumers

react almost immediately. Thus, following the equilibrium hypothesis

$$f(p(n)) = g(p(n-1)),$$

which is highly nonlinear first order equation. If f is strictly decreasing and the ranges of f and g are the same, this equation can be solved for $p(n)$ giving

$$p(n) = f^{-1}(g(p(n-1))).$$

Still, to proceed we have to make some further assumptions on the functions f and g . The simplest functions satisfying the assumptions are

$$f(p(n)) = -m_d p(n) + b_d, \quad g(p(n-1)) = m_s p(n-1) + b_s$$

where $m_d, m_s, b_d > 0, b_s \geq 0$ are constants. Coefficients m_d and m_s are called, respectively, consumers' and suppliers' sensitivity to price. For these assumptions we obtain the following linear equation for price:

$$p(n) = -\frac{m_s}{m_d} p(n-1) + \frac{b_d - b_s}{m_d}. \quad (1.6)$$

The Keynesian National Income Model

In market economy, the national income $Y(n)$ of a country in a given period n can be written as

$$Y(n) = C(n) + I(n) + G(n), \quad (1.7)$$

where

- $C(n)$ is the consumer expenditure for purchase of consumer goods;
- $I(n)$ is the private investment for buying capital equipment;

- $G(n)$ is government expenditure.

There are various model for the above functions. We use widely accepted assumptions introduced by Samuelson [1]. The consumption satisfies

$$C(n) = \alpha Y(n - 1); \quad (1.8)$$

that is the consumer expenditure is proportional to the income in the preceding year. It is natural that $0 < \alpha < 1$. The investment satisfies

$$I(n) = \beta(C(n) - C(n - 1)), \quad (1.9)$$

so that the private investment is induced by the increase in consumption rather than by the consumption itself. The constant β is positive that is acceleration of consumption results in an increased investment while deceleration causes its decrease. Finally, it is assumed that the government expenditure remains constant over the years and we rescale the variables to have

$$G(n) = 1. \quad (1.10)$$

Inserting (1.8) into (1.9) results in

$$I(n) = \alpha\beta(Y(n - 1) - Y(n - 2))$$

so that (1.7) can be written as the second order linear difference equation:

$$Y(n + 2) - \alpha(1 + \beta)Y(n + 1) + \alpha\beta Y(n) = 1. \quad (1.11)$$

Gambler's ruin This problem involves a different type of modelling with roots in the probability theory. Problems of this type are common in the theory of Markov chains, see [?].

1.1 Basic difference equations of finance mathematics 7

A gambler plays a sequence of games against an adversary. The probability that the gambler wins R 1 in any given game is q and the probability of him losing R 1 is $1 - q$. He quits the game if he either wins a prescribed amount of N rands, or loses all his money; in the latter case we say that he has been ruined. Let $p(n)$ denotes the probability that the gambler will be ruined if he starts gambling with n rands. We build the difference equation satisfied by $p(n)$ using the following argument. Firstly, note that we can start observation at any moment, that is, the probability of him being ruined with n rands at the start is the same as the probability of him being ruined if he acquires n rands at any moment during the game. If at some moment during the game he has n rands, he can be ruined in two ways: by winning the next game and ruined with $n + 1$ rand, or by losing and then being ruined with $n - 1$ rands. Thus

$$p(n) = qp(n + 1) + (1 - q)p(n - 1). \quad (1.12)$$

Replacing n by $n + 1$ and dividing by q , we obtain

$$p(n + 2) - \frac{1}{q}p(n + 1) + \frac{1 - q}{q}p(n) = 0, \quad (1.13)$$

with $n = 0, 1, \dots, N$. This is a second order linear difference equation which requires two side conditions. While in the previous cases the side (initial) conditions were natural and we have not ponder on them, here the situation is slightly untypical. Namely, we know that the probability of ruin starting with 0 rands is 1, hence $p(0) = 1$. Further, if the player has N rands, then he quits and cannot be ruined so that $p(N) = 0$. These are not initial conditions but an example of two-point conditions; that is, conditions pre-

scribed at two arbitrary points. Such problems not always have a solution.

1.2 Difference equations of population theory

1.2.1 Single equations for unstructured population models

In many fields of human endeavour it is important to know how populations grow and what factors influence their growth. Knowledge of this kind is important in studies of bacterial growth, wildlife management, ecology and harvesting.

Many animals tend to breed only during a short, well-defined, breeding season. It is then natural to think of the population changing from season to season and therefore time is measured discretely with positive integers denoting breeding seasons. Hence the obvious approach for describing the growth of such a population is to write down a suitable difference equation. Later we shall also look at populations that breed continuously (e.g. human populations).

We start with population models that are very simple and discuss some of their more realistic variants.

1.2.1.1 Exponential growth – linear first order difference equations

Let us start with insect-type (so-called semelparous) populations. Insects often have well-defined annual non-overlapping generations - adults lay eggs in spring/summer and then die. The eggs hatch into larvae which eat and grow and then overwinter in a pupal stage. The adults emerge from the pupae in spring. We take the census of adults in the

breeding seasons. It is then natural to describe the population as the sequence of numbers

$$N_0, N_1, \dots, N_k$$

where N_k is the number of adults in the k -th breeding season.

The simplest assumption to make is that there is a functional dependence between subsequent generations

$$N_{n+1} = f(N_n), \quad n = 0, 1, \dots \quad (1.14)$$

Let us introduce the number R_0 , which is the average number of eggs laid by an adult. R_0 is called the *basic reproductive ratio* or the *intrinsic growth rate*. The simplest functional dependence in (1.14) is

$$N_{n+1} = R_0 N_n, \quad n = 0, 1, \dots \quad (1.15)$$

which describes the situation that the size of the population is determined only by its fertility.

The exponential (or Malthusian) equation (1.15) has a much larger range of applications. In general, in population theory the generations can overlap. Looking at large populations in which individuals give birth to new offspring but also die after some time, we can treat population as a whole and assume that the population growth is governed by the average behaviour of its individual members. Thus, we make the following assumptions:

- Each member of the population produces in average the same number of offspring.
- Each member has an equal chance of dying (or surviving) before the next breeding season.

- The ratio of females to males remains the same in each breeding season

We also assume

- Age differences between members of the population can be ignored.
- The population is isolated - there is no immigration or emigration.

Suppose that on average each member of the population gives birth to the same number of offspring, β , each season. The constant β is called per-capita birth rate. We also define μ as the probability that an individual will die before the next breeding season and call it the per-capita death rate. Thus:

- (a) the number of individuals born in a particular breeding season is directly proportional to the population at the start of the breeding season, and
- (b) the number of individuals who have died during the interval between the end of consecutive breeding seasons is directly proportional to the population at the start of the breeding season.

Denoting by N_k the number of individuals of the population at the start of the k th breeding season, we obtain

$$N_{k+1} = N_k - \mu N_k + \beta N_k,$$

that is

$$N_{k+1} = (1 + \beta - \mu)N_k. \quad (1.16)$$

This equation reduces to (1.15) by putting $\mu = 1$ (so that the whole adult population dies) and $\beta = R_0$.

Equation (1.15) is easily solvable yielding

$$N_k = R_0^k N_0, \quad k = 0, 1, 2, \dots \quad (1.17)$$

We see that the behaviour of the model depends on R_0 . If $R_0 < 1$, then the population decreases towards extinction, but with $R_0 > 1$ it grows indefinitely. Such a behaviour over long periods of time is not observed in any population so that we see that the model is over-simplified and requires corrections.

1.2.1.2 Models leading to nonlinear difference equations

In a real population, some of the R_0 offspring produced by each adult will not survive to be counted as adults in the next census. If we denote by $S(N)$ the *survival rate*; that is, fraction that survives, then the Malthusian equation is replaced by

$$N_{k+1} = R_0 S(N_k) N_k, \quad k = 0, 1, \dots \quad (1.18)$$

which may be alternatively written as

$$N_{k+1} = F(N_k) N_k = f(N_k), \quad k = 0, 1, \dots \quad (1.19)$$

where $F(N)$ is per capita production of a population of size N . Such models, with density dependent growth rate, lead to nonlinear equations.

We introduce most typical nonlinear models.

Beverton-Holt type models.

Let us look at the model (1.19)

$$N_{k+1} = F(N_k) N_k, \quad k = 0, 1, \dots,$$

where $F(N_k) = R_0 S(N_k)$. We would like the model to display a *compensatory* behaviour; that is, mortality should

balance the increase in numbers. For this we should have $NS(N) \approx \text{const.}$ Also, for small N , $S(N)$ should be approximately 1 as we expect very small intra-species competition and thus the growth should be exponential with the growth rate R_0 . A simple function of this form is

$$S(N) = \frac{1}{1 + aN}$$

leading to

$$N_{k+1} = \frac{R_0 N_k}{1 + aN_k}.$$

If we introduce the concept of carrying capacity of the environment K and assume that the population having reached K , will stay there; that is, if $N_k = K$ for some k , then $N_{k+m} = K$ for all $m \geq 0$, then

$$K(1 + aK) = R_0 K$$

leading to $a = (R_0 - 1)/K$ and the resulting model, called the *Beverton-Holt model*, takes the form

$$N_{k+1} = \frac{R_0 N_k}{1 + \frac{R_0 - 1}{K} N_k}. \quad (1.20)$$

As we said earlier, this model is compensatory.

A generalization of this model is called the *Hassell* or again *Beverton-Holt* model, and reads

$$N_{k+1} = \frac{R_0 N_k}{(1 + aN_k)^b}. \quad (1.21)$$

Substitution $x_k = aN_k$ reduces the number of parameters giving

$$x_{k+1} = \frac{R_0 x_k}{(1 + x_k)^b} \quad (1.22)$$

which will be analysed later.

The logistic equation.

The Beverton-Holt models are best applied to semelparous insect populations but was also used in the context of fisheries. For populations surviving to the next cycle it is more informative to write the difference equation in the form

$$N_{k+1} = N_k + R(N_k)N_k, \quad (1.23)$$

so that the increase in the population is given by $R(N) = R_0 S(N)N$. Here we assume that no adults die (death can be incorporated by introducing factor $d < 1$ in front of the first N_k or modifying $S(N)$ which would lead to the equation of the same form).

As before, the function R can have different forms but must satisfy the requirements:

- (a) Due to overcrowding, $R(N)$ must decrease as N increases until N equals the carrying capacity K ; then $R(K) = 0$ and, as above, $N = K$ stops changing.
- (b) Since for N much smaller than K there is small intra-species competition, we should observe an exponential growth of the population so that $R(N) \approx R_0$ as $N \rightarrow 0$; here R_0 is called the unrestricted growth rate of the population.

Constants R_0 and K are usually determined experimentally.

In the spirit of mathematical modelling we start with the simplest function satisfying these requirements. The simplest function is a linear function which, to satisfy (a) and (b), must be chosen as

$$R(N) = -\frac{R_0}{K}N + R_0.$$

Substituting this formula into (1.23) yields the so-called

discrete logistic equation

$$N_{k+1} = N_k + R_0 N_k \left(1 - \frac{N_k}{K}\right), \quad (1.24)$$

which is still one of the most often used discrete equations of population dynamics.

While the above arguments may seem to be of *bunny-out-of-the-hat* type, it could be justified by generalizing (1.16). Indeed, assume that the mortality μ is not constant but equals

$$\mu = \mu_0 + \mu_1 N,$$

where μ_0 corresponds to death of natural caused and μ_1 could be attributed to cannibalism where one adult eats/kills on average μ_1 portion of the population. Then (1.16) can be written as

$$N_{k+1} = N_k + (\beta - \mu_0) N_k \left(1 - \frac{N_k}{\frac{\beta - \mu_0}{\mu_1}}\right) \quad (1.25)$$

which is (1.24) with $R_0 = \beta - \mu_0$ and $K = (\beta - \mu_0)/\mu_1$.

In the context of insect population, where there are no survivors from the previous generation, the above equation reduces to

$$N_{k+1} = R_0 N_k \left(1 - \frac{N_k}{K}\right). \quad (1.26)$$

By substitution

$$x_n = \frac{1}{1 + R_0} \frac{N_k}{K}, \quad \mu = 1 + R_0$$

we can reduce (1.24) to a simpler form

$$x_{n+1} = \mu x_n (1 - x_n) \quad (1.27)$$

We observe that the logistic equation, especially with S given by (1.28) is an extreme example of the scramble competition.

The Ricker equation

The problem with the discrete logistic equation is that large (close to K) populations can become negative in the next step. Although we could interpret a negative populations as extinct, this may not be the behaviour that would actually happen. Indeed, the model was constructed so as to have $N = K$ as a stationary population. Thus, if we happen to hit exactly K , then the population survives but if we even marginally overshoot, the population becomes extinct.

One way to avoid such problems with negative population is to replace the density dependent survival rate by

$$S(N_k) = \left(1 - \frac{N_k}{K}\right)_+ . \quad (1.28)$$

to take into account that S cannot be negative. However, this model also leads to extinction of the population if it exceeds K which is not always realistic.

Another approach is to try to find a model in which large values of N_k produce very small, but still positive, values of N_{k+1} . Thus, a population well over the carrying capacity crashes to very low levels but survives. Let us find a way in which this can be modelled. Consider the per capita population change

$$\frac{\Delta N}{N} = f(N).$$

First we note that it is impossible for f to be less than -1 – this would mean that an individual could die more than once. We also need a decreasing f which is non-zero

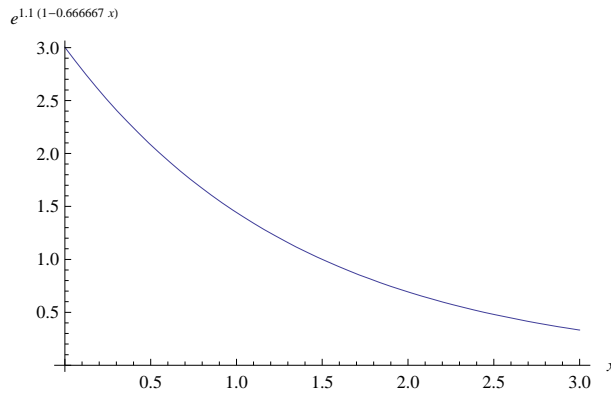


Fig. 1.1. The function $f(x) = e^{r(1-x/K)}$

(= R_0) at 0. One such function can be recovered from the Beverton-Holt model, another simple choice is an exponential shifted down by 1:

$$\frac{\Delta N}{N} = ae^{-bN} - 1,$$

which leads to

$$N_{k+1} = aN_k e^{-bN_k}.$$

If, as before, we introduce the carrying capacity K and require it give stationary population, we obtain

$$b = \frac{\ln a}{K}$$

and, letting for simplicity $r = \ln a$, we obtain the so-called *Ricker equation*

$$N_{k+1} = N_k e^{r(1-\frac{N_k}{K})}. \quad (1.29)$$

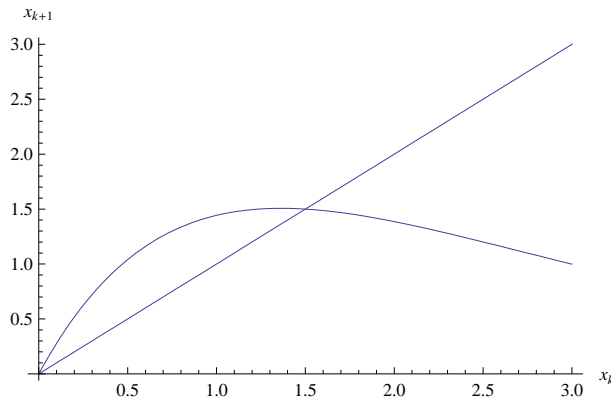


Fig. 1.2. The relation $x_{n+1} = x_n e^{r(1-x_n/K)}$

We note that if $N_k > K$, then $N_{k+1} < N_k$ and if $N_k < K$, then $N_{k+1} > N_k$. The intrinsic growth rate R_0 is given by $R_0 = e^r - 1$ but, using the Maclaurin formula, for small r we have $R_0 \approx r$.

Allee type equations

In all previous models with density dependent growth rates the bigger the population (or the higher the density), the slower the growth. However, in 1931 Warder Clyde Allee noticed that in small, or dispersed, populations individual chances of survival decrease which can lead to extinction of the populations. This could be due to the difficulties of finding a mating partner or more difficult cooperation in e.g., organizing defence against predators. Models having this property can also be built within the considered framework by introducing two thresholds: the carrying capacity K and a parameter $0 < L < K$ at which the behaviour of

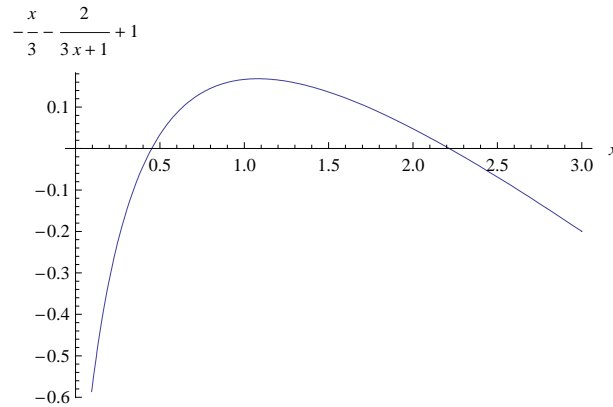


Fig. 1.3. The function $1 - \frac{N_k}{K} - \frac{A}{1+BN_k}$

the population changes so that $\Delta N/N < 0$ for $0 < N < L$ and $N > K$ and $\Delta N/N > 0$ for $L < N < K$. If

$$\Delta N/N = f(N),$$

then the resulting difference equation is

$$N_{k+1} = N_k + N_k f(N_k)$$

and the required properties can be obtained by taking $f(N) \leq 0$ for $0 < N < L$ and $N > K$ and $f(N) \geq 0$ for $L < N < K$. A simple model like that is offered by choosing $f(N) = (L - N)(N - K)$ so that

$$N_{k+1} = N_k(1 + (L - N_k)(N_k - K)). \quad (1.30)$$

Another model of this type, see [?], which can be justified by modelling looking of a mating partner or introducing a generalized predator (that is, preying also on other species),

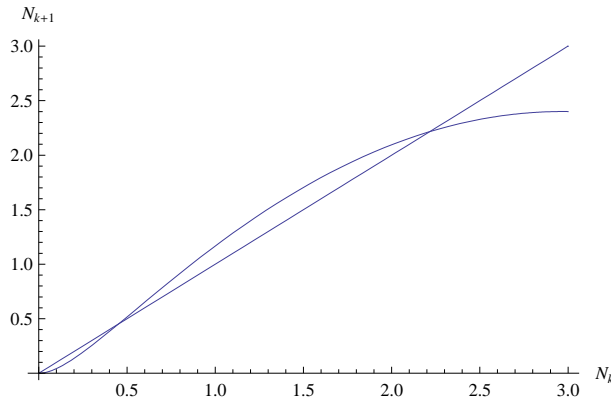


Fig. 1.4. The relation $N_{k+1} = N_k + N_k f(N_k)$ for an Allee model

has the form

$$N_{k+1} = N_k \left(1 + \lambda \left(1 - \frac{N_k}{K} - \frac{A}{1 + BN_k} \right) \right) \quad (1.31)$$

where $\lambda > 0$ and

$$1 < A < \frac{(BK + 1)^2}{4KB}, \quad BK > 1. \quad (1.32)$$

However, since $x \rightarrow (x + 1)^2/4x$ is an increasing function for $x > 1$ and equals 1 for $x = 1$, the second condition is redundant.

1.2.2 Structured populations and linear systems of difference equations

There are two main problems with models introduced above. One is that all individuals in the population have the same age, the other is that the described population does not

interact with others. Trying to remedy these deficiencies leads to systems of equations.

1.2.2.1 *Fibonacci rabbits and related models*

We start with what possibly is the first formulated problem related to populations with age structure. Leonardo of Pisa, called Fibonacci, in his famous book *Liber abaci*, published in 1202, formulated the following problem:

A certain man put a pair of rabbits in a place surrounded on all sides by a wall. How many rabbits can be produced from that pair in a year if it is supposed that every month each pair begets a new pair which from the second month on becomes productive?

To formulate the mathematical model we also assume that no deaths occur in the period of observation. Also the monthly census of the population is taken just before births for this month take place; that is, we count the end of the given month, when the births are taking place, to this month. Then, for the end of month $k + 1$ we can write

$$\begin{aligned} \left\{ \begin{array}{l} \text{number present} \\ \text{in month } k + 1 \end{array} \right\} &= \left\{ \begin{array}{l} \text{number present} \\ \text{in month } k \end{array} \right\} \\ &+ \left\{ \begin{array}{l} \text{number born} \\ \text{in month } k \end{array} \right\} \end{aligned}$$

Since rabbits become productive only two months after birth and produce only one pair per month, we can write

$$\left\{ \begin{array}{l} \text{number born} \\ \text{in month } k \end{array} \right\} = \left\{ \begin{array}{l} \text{number present} \\ \text{in month } k - 1 \end{array} \right\}.$$

To justify the last statement, we recall that the census is taken just before the end of the month (when the births

occur). Thus, pairs present by the end of month $k - 1$ were born a month earlier, in month $k - 2$, and thus two months later, at the end of month k , will give births to pairs observed in the census taken in month N_{k+1} .

Denoting by N_k the number of pairs at the end of month k and combining the two equations above, we obtain the so-called Fibonacci equation

$$N_{k+1} = N_k + N_{k-1}, \quad k = 1, 2, \dots \quad (1.33)$$

This is a linear difference equation of second order since it gives the value of N_k at time k in terms of its values at two times immediately preceding k .

It is clear that (1.33) as a model describing a population of rabbits is oversimplified. Later we introduce more adequate population models. Here we note that there are biological phenomena for which (1.33) provides an exact fit. One of them is family tree of honeybees. Honeybees live in colonies and one of the unusual features of them is that not every bee has two parents. To be more precise, let us describe a colony in more detail. First, in any colony there is one special female called the queen. Further, there are worker bees who are female but they produce no eggs. Finally, there are drones, who are male and do no work, but fertilize the queen's eggs. Drones are borne from the queen's unfertilised eggs and thus they have a mother but no father. On the other hand, the females are born when the queen has mated with a male and so have two parents. In Fig. 1.5 we present a family tree of a drone. It is clear that the number of ancestors k th generations earlier exactly satisfies (1.33).

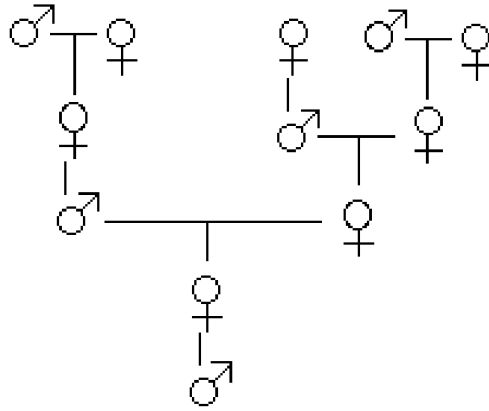


Fig. 1.5. The family tree of a drone

1.2.2.2 Models with age structure

Writing the Fibonacci model in the form of a single equation makes it neat and compact some information, however, is lost. In particular, it is impossible to find long time ratio between adults and juveniles. Such a question is of paramount importance in population theory where the determination of stable age structure of the population is vital for designing e.g. pension funds and health care systems.

It is possible to re-write the Fibonacci model to make such a detailed analysis feasible. We note that each month the population is represented by two classes of rabbits, adults $v_1(k)$ and juveniles $v_0(k)$ and thus the state of the

population is described by the vector

$$\mathbf{v}(k) = \begin{pmatrix} v_0(k) \\ v_1(k) \end{pmatrix}$$

Since the number of juvenile (one-month old) pairs in month $k+1$ is equal to the number of adults in month n (remember, we take the census before birth cycle in a given month, so these are newborns from a month before) and the number of adults is the number of adults from the month before and the number of juveniles from the month ago who became adults. In other words

$$\begin{aligned} v_0(k+1) &= v_1(k) \\ v_1(k+1) &= v_0(k) + v_1(k) \end{aligned} \quad (1.34)$$

or, in a more compact form

$$\mathbf{v}(k+1) = \mathcal{L}\mathbf{v}(k) := \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \mathbf{v}(k). \quad (1.35)$$

We note formal similarity with the the exponential growth equation (1.15) which suggest that the solution can be written in the same form as (1.17):

$$\mathbf{v}(k) = \mathcal{L}^k \mathbf{v}(0). \quad (1.36)$$

However, at this moment we do not have efficient tools for calculation powers of matrices—these will be developed in Chapter ??.

The idea leading to (1.35) immediately lends itself to generalization. Assume that instead of pairs of individuals, we are tracking only females and that the census is taken immediately before the reproductive period. Further, assume that there is an oldest age class n and if no individual can

stay in an age class for more than one time period (which means that all females who are of age n at the beginning of the time period die during this period). Note that it is **not** the case for Fibonacci rabbits. We introduce the survival rate s_i and the age dependent maternity function m_i ; that is, s_i is probability of survival from age $i - 1$ to age i (or conditional probability of survival of a female to age i provided she survived till $i - 1$), and each female of age i produces m_i offspring in average. Hence, $s_1 m_i$ is the average number of female offspring produced by each female of the age i who survived to the census. In this case, the evolution of the population can be described by the system of difference equations

$$\mathbf{v}(n+1) = \mathcal{L}\mathbf{v}(n)$$

where \mathcal{L} is the $n \times n$ matrix

$$\mathcal{L} := \begin{pmatrix} s_1 m_1 & s_1 m_2 & \cdots & s_1 m_{n-1} & s_1 m_n \\ s_2 & 0 & \cdots & 0 & 0 \\ 0 & s_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & s_n & 0 \end{pmatrix}, \quad (1.37)$$

The matrix of the form (1.37) is referred to as a *Leslie matrix*. To shorten notation we often denote $f_i = s_1 m_i$ and are referred to as the (effective) age specific *fertility*.

A generalization of the Leslie matrix can be obtained by assuming that a fraction τ_i of i -th population stays in the

same population. This gives the matrix

$$\mathcal{L} := \begin{pmatrix} f_1 + \tau_1 & f_2 & \cdots & f_{n-1} & f_n \\ s_2 & \tau_2 & \cdots & 0 & 0 \\ 0 & s_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & s_n & \tau_n \end{pmatrix}, \quad (1.38)$$

Such matrices are called *Usher matrices*.

In most cases $f_i \neq 0$ only if $\alpha \leq i \leq \beta$ where $[\alpha, \beta]$ is the fertile period. For example, for a typical mammal population we have three stages: immature (pre-breeding), breeding and post-breeding. If we perform census every year, then naturally a fraction of each class remains in the same class. Thus, the transition matrix in this case is given by

$$\mathcal{L} := \begin{pmatrix} \tau_1 & f_2 & 0 \\ s_2 & \tau_2 & 0 \\ 0 & s_3 & \tau_3 \end{pmatrix}, \quad (1.39)$$

On the other hand, in many insect populations, reproduction occurs only in the final stage of life and in such a case $f_i = 0$ unless $i = n$.

1.2.3 General structured population models

Leslie matrices fit into a more general mathematical structure describing evolution of populations divided in states, or subpopulations, not necessarily related to age. For example, we can consider clusters of cells divided into classes with respect to their size, cancer cells divided into classes on the basis of the number of copies of a particular gene

responsible for its drug resistance, or a population divided into subpopulations depending on the geographical patch they occupy in a particular moment of time. Let us suppose we have n states. Each individual in a given state j contributes on average, say, a_{ij} individuals in state i . Typically, this is due to the state j individual:

- migrating to i -th subpopulation with probability p_{ij} ;
- contributing to a birth of an individual in i -th subpopulation with probability b_{ij} ;
- surviving with probability $1 - d_j$ (thus d_j is probability of dying),

other choices and interpretations are, however, also possible.

If we assume that the evolution follows the above rules, then we can write the balance of individuals in population i at time $k + 1$:

$$v_i(k+1) = (1 - d_i)v_i(k) + \sum_{\substack{j=1 \\ j \neq i}}^n p_{ij}v_j(k) + \sum_{j=1}^n b_{ij}v_j(k), \quad (1.40)$$

where, as before, $v_i(k)$ is the number of individuals at time k in state i . Hence, a_{ij} are non-negative but otherwise arbitrary numbers. Denoting with $\mathbf{v}(k) = (v_1(k), \dots, v_n(k))$ we can write (1.40) in the matrix form

$$\mathbf{v}_{k+1} = \mathcal{A}\mathbf{v}_k, \quad (1.41)$$

where

$$\mathcal{A} := \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1\ n-1} & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2\ n-1} & a_{2n} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{n\ n-1} & a_{nn} \end{pmatrix}. \quad (1.42)$$

Thus

$$\mathbf{v}_k = \mathcal{A}^k \mathbf{v}_0,$$

where \mathbf{v}_0 is the initial distribution of the population between the subpopulations.

Example 1.1 [?] *Any chromosome ends with a telomer which protects it against damage during the DNA replication process. Recurring divisions of cells can shorten the length of telomers and this process is considered to be responsible for cell's aging. If telomer is too short, the cell cannot divide which explains why many cell types can undergo only a finite number of divisions. Let us consider a simplified model of telomer shortening. The length of a telomer is a natural number from 0 to n , so cells with telomer of length i are in subpopulation i . A cell from subpopulation i can die with probability μ_i and divide (into 2 daughters). Any daughter can have a telomer of length i with probability a_i and of length $i - 1$ with probability $1 - a_i$. Cells of 0 length telomer cannot divide and thus will die some time later. To find coefficients of the transition matrix, we see that the average production of offspring with telomer of length i by a parent of the same class is*

$$2a_i^2 + 2a_i(1 - a_i) = 2a_i,$$

(2 daughters with telomer of length i produced with probability a_i^2 and 1 daughter with telomer of length $i - 1$ produced with probability $2a_i(1 - a_i)$). Similarly, average production of an daughters with length $i - 1$ telomer is $2(1 - a_i)$. However, to have offspring, the cell must survived from one census to another which happens with probability $1 - \mu_i$. Hence, defining $r_i = 2a_i(1 - \mu_i)$ and $d_i = 2(1 - a_i)(1 - \mu_i)$, we have

$$\mathcal{A} := \begin{pmatrix} 0 & d_1 & 0 & \cdots & 0 \\ 0 & r_1 & d_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & r_n \end{pmatrix}. \quad (1.43)$$

The model can be modified to make it closer to reality by allowing, for instance, shortening of telomers by different lengths or consider models with more telomers in a cell and with probabilities depending on the length of all of them.

1.2.4 Markov chains

A particular version of (1.42) is obtained when we assume that the total population has constant size so that no individual dies and no new individual can appear, so that the only changes occur due to migration between states. In other words, $b_{ij} = d_j = 0$ for any $1 \leq i, j \leq n$ and thus $a_{ij} = p_{ij}$ is the fraction of j -th subpopulation which, on average, moves to the i -th subpopulation or, using a probabilistic language, probabilities of such a migration. Then, in addition to the constraint $p_{ij} \geq 0$ we must have $p_{ij} \leq 1$ and, since the total number of individuals contributed by the state j to all other states must equal to the number of

individuals in this state, we must have

$$v_j = \sum_{1 \leq i \leq n} p_{ij} v_i$$

we obtain

$$\sum_{1 \leq i \leq n} p_{ij} = 1,$$

or

$$p_{ii} = 1 - \sum_{\substack{j=1 \\ j \neq i}}^n p_{ij}, \quad i = 1, \dots, n, \quad (1.44)$$

In words, the sum of entries in each column must be equal to 1. This expresses the fact that each individual must be in one of the n states at any time.

Matrices of this form are called *Markov matrices*.

We can check that, indeed, this condition ensures that the size of the population is constant. Indeed, the size of the population at time k is $N(k) = v_1(k) + \dots + v_n(k)$ so that

$$\begin{aligned} N(k+1) &= \sum_{1 \leq i \leq n} v_i(k+1) = \sum_{1 \leq i \leq n} \left(\sum_{1 \leq j \leq n} p_{ij} v_j(k) \right) \\ &= \sum_{1 \leq j \leq n} v_j(k) \left(\sum_{1 \leq i \leq n} p_{ij} \right) = \sum_{1 \leq j \leq n} v_j(k) \\ &= N(k). \end{aligned} \quad (1.45)$$

1.2.5 Interacting populations and nonlinear systems of difference equations

1.2.5.1 Models of an epidemic: system of difference equations

SI model

In nature, various population interact with each other co-existing in the same environment. This leads to systems of difference equations. As an illustration we consider a model for spreading of measles epidemic.

Measles is a highly contagious disease, caused by a virus and spread by effective contact between individuals. It affects mainly children. Epidemic of measles have been observed in Britain and the US roughly every two or three years.

Let us look at the development of measles in a single child. A child who has not yet been exposed to measles is called a susceptible. Immediately after the child first catches the disease, there is a latent period where the child is not contagious and does not exhibit any symptoms of the disease. The latent period lasts, on average, 5 to 7 days. After this the child enters the contagious period. The child is now called infective since it is possible for another child who comes in contact with the infective to catch the disease. This period last approximately one week. After this the child recovers, becomes immune to the disease and cannot be reinfected.

For simplicity we assume that both latent and contagious periods last one week. Suppose also that most interactions between children occur on weekend so that the numbers of susceptibles and infectives remains constant over the rest of the week. Since the typical time in the model is one week,

we shall model the spread of the disease using one week as the unit of time.

To write down the equations we denote

$$I(k) = \left\{ \begin{array}{c} \text{number of} \\ \text{infectives in week } k \end{array} \right\}$$

and

$$S(k) = \left\{ \begin{array}{c} \text{number of} \\ \text{susceptibles during week } k \end{array} \right\}$$

To develop an equation for the number of infectives we consider the number of infectives in week $k+1$. Since the recuperation period is one week after which an infective stops to be infective, no infectives from week k will be present in week $k+1$. Thus we have

$$\begin{aligned} I(k+1) &= \left\{ \begin{array}{c} \text{number of} \\ \text{infectives in week } k+1 \end{array} \right\} \\ &= \left\{ \begin{array}{c} \text{number of susceptibles} \\ \text{who caught measles in week } k \end{array} \right\} \end{aligned}$$

It is generally thought that the number of new births is an important factor in measles epidemic. Thus

$$\begin{aligned} S(k+1) &= \left\{ \begin{array}{c} \text{number of} \\ \text{births in week } k \end{array} \right\} + \left\{ \begin{array}{c} \text{number of} \\ \text{susceptibles in week } k \end{array} \right\} \\ &\quad - \left\{ \begin{array}{c} \text{number of susceptibles} \\ \text{who caught measles in week } k \end{array} \right\} \end{aligned}$$

We assume further that the number of births each week is a constant B . Finally, to find the number of susceptibles infected in a week it is assumed that a single infective infects a constant fraction f of the total number of susceptibles.

Thus, if fS_k is the number of susceptibles infected by a single infective so, with a total of I_k infectives, then

$$\left\{ \begin{array}{l} \text{number of susceptibles} \\ \text{who caught measles in week } k \end{array} \right\} = fS(k)I(k).$$

This kind of assumption presupposing random encounters is known as the *mass action law* and is one of the simplest and therefore most used nonlinear models of interactions. Combining the obtained equations we obtain the system

$$\begin{aligned} I(k+1) &= fS(k)I(k), \\ S(k+1) &= S(k) - fS(k)I(k) + B, \end{aligned} \quad (1.46)$$

where B and f are constant parameters of the model.

SIR model

The model (1.46) can be generalized in various ways to cater for different scenarios. One of typical generalization is introducing explicitly one more class R (from removed/recovered) describing individuals who had contracted the disease but either died or recovered. Note that this class is implicitly present in the SI model but since we assumed that individuals became immune after recovery, this class did not have any effect on classes S and I and can be discarded from the model. The SIR model discussed below allows for various scenarios after infection.

Let us consider the population divided into three classes: susceptibles S , infectives I and removed (recovered or dead) R . We assume that the model is closed; that is we do not consider any births in the process. Thus the total population $N = S + I + R$ is constant. This allows a better characterization of the ‘infectiveness’ coefficient f in (1.46). Namely, if the probability of an individual meeting some-

one within one cycle from time k to time $k + 1$ is α' , then meeting a susceptible is $\alpha'S/N$. Further, assume that a fraction α'' of these encounters results in an infection. We denote $\alpha = \alpha'\alpha''$. Thus the number of encounters resulting in infection within one cycle is $\alpha SI/N$. Moreover, we assume that a fraction β of individuals (except from class S) can become susceptible (could be reinfected) and a fraction γ of infectives move to R . Reasoning as in the previous paragraph we obtain the system

$$\begin{aligned} S(k+1) &= S(k) - \frac{\alpha}{N}I(k)S(k) + \beta(I(k) + R(k)) \\ I(k+1) &= I(k) + \frac{\alpha}{N}I(k)S(k) - \gamma I(k) - \beta I(k) \\ R(k+1) &= R(k) - \beta R(k) + \gamma I(k) \end{aligned} \quad (1.47)$$

We observe that

$$S(k+1) + I(k+1) + R(k+1) = S(k) + I(k) + R(k) = \text{const} = N$$

in accordance with the assumption.

This can be used to reduce (1.47) to a two dimensional (SI -type) system

$$\begin{aligned} S(k+1) &= S(k) - \frac{\alpha}{N}I(k)S(k) + \beta(N - S(k)) \\ I(k+1) &= I(k)(1 - \gamma - \beta) + \frac{\alpha}{N}I(k)S(k). \end{aligned} \quad (1.48)$$

The modelling indicates that we need to assume $0 < \gamma + \beta < 1$ and $0 < \alpha < 1$.

1.2.5.2 Host-parasitoid system

Discrete difference equation models apply most readily to groups such as insect population where there is rather natural division of time into discrete generations. A model

which has received a considerable attention from experimental and theoretical biologists is the *host-parasitoid* system. Let us begin by introducing definition of a parasitoid. Predators kill their prey, typically for food. Parasites live in or on a host and draw food, shelter, or other requirements from that host, often without killing it. Female parasitoids, in turn, typically search and kill, but do not consume, their hosts. Rather, they *oviposit* (deposit eggs) on, in, or near the host and use it as a source of food and shelter for the developing young. There are around 50000 species of wasp-like parasitoids, 15000 of fly-type parasitoids and 3000 species of other orders.

Typical of insect species, both host and parasitoid have a number of life-stages that include eggs, larvae, pupae and adults. In most cases eggs are attached to the outer surface of the host during its larval or pupal stage, or injected into the host's flesh. The larval parasitoids develop and grow at the expense of their host, consuming it and eventually killing it before they pupate.

A simple model for this system has the following set of assumptions:

- (i) Hosts that have been parasitized will give rise to the next generation of parasitoids.
- (ii) Hosts that have not been parasitized will give rise to their own progeny.
- (iii) The fraction of hosts that are parasitized depends on the rate of *encounter* of the two species; in general, this fraction may depend on the densities of one or both species.

It is instructive to consider this minimal set of interactions first and examine their consequences. We define:

- $N(k)$ – density (number) of host species in generation k ,
- $P(k)$ – density (number) of parasitoid in generation k ,
- $f = f(N(k), P(k))$ – fraction of hosts not parasitized,
- λ – host reproductive rate,
- c – average number of viable eggs laid by parasitoid on a single host.

Then our assumptions 1)–3) lead to:

$$\begin{aligned} N(k+1) &= \lambda N(k) f(N(k), P(k)), \\ P(k+1) &= c N(k) (1 - f(N(k), P(k))). \end{aligned} \quad (1.49)$$

To proceed we have to specify the rate of encounter f . One of the earliest models is the Nicholson-Bailey model. *The Nicholson-Bailey model*

Nicholson and Bailey added two assumptions to the list (i)–(iii).

- (iv) Encounters occur randomly. The number of encounters N_e of the host with the parasitoid is therefore proportional to the product of their densities (numbers):

$$N_e = \alpha NP,$$

where α is a constant, which represents the searching efficiency of the parasitoids (law of mass action).

- (v) Only the first encounter between a host and parasitoid is significant (once the host has been parasitized it gives rise exactly c parasitoid progeny; a

second encounter with an egg laying parasitoid will not increase or decrease this number.

Based on the latter assumption, we have to distinguish only between those hosts that have had no encounters and those that had n encounters, $n \geq 1$. Because the encounters are random, one can represent the probability of r encounters by some distribution based on the average number of encounters that take place per unit time.

Interlude-the Poisson distribution. One of the simplest distributions used in such a context is the Poisson distribution. It is a limiting case of the binomial distribution: if the probability of an event occurring in a single trial is p and we perform n trials, then the probability of exactly r events is

$$b(n, p; r) = \binom{n}{r} p^r (1-p)^{n-r}.$$

Average number of events in $\mu = np$. If we assume that the number of trials n grows to infinity in such a way that the average number of events μ stays constant (so p goes to zero), then the probability of exactly r events is given by

$$\begin{aligned} p(r) &= \lim_{n \rightarrow \infty} b(n, \mu/n; r) = \lim_{n \rightarrow \infty} \frac{n!}{r!(n-r)!} \frac{\mu^r}{n^r} \left(1 - \frac{\mu}{n}\right)^{n-r} \\ &= \frac{e^{-\mu} \mu^r}{r!}, \end{aligned}$$

which is the Poisson distribution. In the case of host-parasitoid interaction, the average number of encounters per host per unit time is

$$\mu = \frac{N_e}{N},$$

that is, by 4.,

$$\mu = aP.$$

Hence, the probability of a host not having any encounter with parasitoid in a period of time from k to $k + 1$ is

$$p(0) = e^{-aP(k)}.$$

Assuming that the parasitoids search independently and their searching efficiency is constant a , leads to the Nicholson-Bailey system

$$\begin{aligned} N(k+1) &= \lambda N(k)e^{-aP(k)}, \\ P(k+1) &= cN(k)(1 - e^{-aP(k)}) \end{aligned} \quad (1.50)$$

2

Basic differential equations models

As we observed in the previous section, the difference equation can be used to model quite a diverse phenomena but their applicability is limited by the fact that the system should not change between subsequent time steps. These steps can vary from fraction of a second to years or centuries but they must stay fixed in the model. There are however numerous situations when the changes can occur instantaneously. These include growth of populations in which breeding is not restricted to specific seasons, motion of objects where the velocity and acceleration changes every instant, spread of epidemic with no restriction on infection times, and many others. In such cases it is not feasible to model the process by relating the state of the system at a particular instant to the earlier states (though this part remains as an intermediate stage of the modelling process) but we have to find relations between the rates of change of quantities relevant to the process. Rates of change are typically expressed as derivatives and thus continuous time modelling leads to differential equations that express re-

lations between the derivatives rather than to difference equations that express relations between the states of the system in subsequent moments of time.

In what follows we shall derive basic differential equations trying to provide continuous counterparts of some discrete systems described above.

2.1 Equations related to financial mathematics

2.1.1 Continuously compounded interest and loan repayment

Many banks now advertise continuous compounding of interest which means that the conversion period α of Subsection 1.1 tends to zero so that the interest is added to the account on the continual basis. If we measure now time in years, that is, Δt becomes the conversion period, and p is the annual interest rate, then the increase in the deposit between time instants t and $t + \Delta t$ will be

$$S(t + \Delta t) = S(t) + \Delta t \frac{p}{100} S(t). \quad (2.1)$$

which, dividing by Δt and passing with Δt to zero, as suggested by the definition of continuously compounded interest, yields the differential equation

$$\frac{dS}{dt} = \bar{p}S, \quad (2.2)$$

where $\bar{p} = p/100$. This is a first order (only the first order derivative of the unknown function occurs) linear (the unknown function appears only by itself, not as an argument of any function) equation. It is easy to check that it has the solution

$$S(t) = S_0 e^{\bar{p}t} \quad (2.3)$$

where S_0 is the initial deposit made at time $t = 0$.

To compare this formula with the discrete one (1.1) we note that in t years we have $k = t/\alpha$ conversion periods

$$S(t) = N_k = (1 + \bar{p}\alpha)^k S_0 = (1 + \bar{p}\alpha)^{t/\alpha} S_0 = \left((1 + \bar{p}\alpha)^{1/\bar{p}\alpha} \right)^{\bar{p}t}.$$

From calculus we know that

$$\lim_{x \rightarrow 0^+} (1 + x)^{1/x} = e,$$

and that the sequence is monotonically increasing. Thus, if the interest is compounded very often (almost continuously), then practically

$$S(t) \approx S_0 e^{\bar{p}t},$$

which is exactly (2.3). It is clear that after 1 year the initial investment will increase by the factor $e^{\bar{p}}$ and, recalling (1.3), we have the identity

$$1 + r_{eff} = e^{\bar{p}}, \quad (2.4)$$

which can serve as the definition of the effective interest rate when the interest is compounded continuously. This relation can be of course obtained by passing with α to 0 in (1.3). Typically, the exponential can be calculated even on a simple calculator, contrary to (1.1). Due to monotonic property of the limit, the continuously compounded interest rate is the best one can get. However, the differences in return are negligible. A short calculation reveals that if one invests R10000 at $p = 15\%$ in banks with conversion periods 1 year, 1 day and with continuously compounded interest, then the return will be, respectively, R11500, R11618 and R11618.3. That is why the continuous formula (2.3) can be used as a good approximation for the real return.

Similar argument can be used for the loan repayment. Assume that the loan is being paid off continuously, a rate $\rho > 0$ per annum. Then, after short period of time Δt the change in the debt D can be written, similarly to (1.4), as

$$D(t + \delta t) = D(t) + \Delta t \bar{p} D(t) - \rho \Delta t$$

where $\alpha = 1/\Delta t$ is the conversion period (time unit is 1 year so that Δt is a fraction of 1 year) and $\bar{p} = p/100$ with p being the annual interest rate (in percents). As before, we divide by Δt and, taking $\Delta t \rightarrow 0$ we obtain

$$\frac{dD}{dt} - \bar{p}D = -\rho \quad (2.5)$$

with the initial condition $D(0) = D_0$ corresponding to the initial debt. Eq. (4.10) is an example of nonhomogeneous linear differential equation which are discussed in Appendix A2.2.2. Discussion of this equation will be carried on in Chapter ??.

2.1.2 Continuous models in economical applications

A Keynesian model

A continuous variant of the discrete Keynesian model discussed earlier is offered by the following argument. We define the aggregate demand D as $D = C + I + G$ where, as before, C is the individual consumption, I is private investment and G is the government expenditure; Y is the national income. If $Y = D$, then the economy is in equilibrium, but if $Y \neq D$, then an adjustment of the income is

required. It is assumed that

$$\frac{dY}{dt} = k(D - Y)$$

for some constant $k > 0$; that is, the national economy responds positively for an excess demand. For a simple closed economy with investment $I(t) = \bar{I}$ and government spending $G(t) = \bar{G}$ constant, we can write $D(t) = C(t) + \bar{I} + \bar{G}$. In general, C is an increasing function of Y : $C(Y) = f(Y)$ with $C'(Y) > 0$. Therefore the equation can be written in the form

$$Y' = k(C(Y) - Y + \bar{I} + \bar{G}) = kfY \quad (2.6)$$

where $f(Y) = C(Y) - Y + \bar{I} + \bar{G}$. Often the affine function C is used: $C(Y) = c_0 + cY$ with $c, c_0 > 0$. In this case we obtain

$$Y' = k(c - 1)Y + k(c_0\bar{I} + \bar{G}). \quad (2.7)$$

As in the loan repayment, this is a nonhomogeneous first order linear differential equation.

The Neo-classical Model of Economic Growth

More sophisticated models of economic growth involve a production function Y which is a function of capital input K and labour input L :

$$Y = Y(K, L).$$

P is the total production; that is, the monetary value of all goods produced in a year. A widely used form of P is the Cobb-Douglas production function

$$Y(K, L) = AK^\alpha L^\beta,$$

for some constants A, α, β . The model we consider here is based on [2, 3]. It is assumed that

- the labour input grows at a constant rate:

$$L' = nL$$

for some constant n ;

- All savings S (which are a fraction of the total production Y : $S = sY$) are invested into the capital K formation I . Investment is assumed to follow the equation $K' = I - \delta K$. Thus

$$sY = K' + \delta K$$

with $s, \delta > 0$;

- Production takes place under so-called constant return condition; that is, if both K and L increase by a factor a , then Y will increase by a .

let us first consider the last point. Writing it in a mathematical language we obtain

$$Y(aK, aL) = aY(K, L)$$

and taking $a = 1/L$ we have

$$Y(K, L) = LY \left(\frac{K}{L}, 1 \right) = Lf \left(\frac{K}{L} \right)$$

for some function of one variable f . Note that for the Cobb-Douglas production function we must have $a^\alpha a^\beta = a$ which yields $\alpha + \beta = 1$ or $\beta = 1 - \alpha$; that is

$$Y(K, L) = K^\alpha L^{1-\alpha} = L \left(\frac{K}{L} \right)^\alpha.$$

Denote $k = K/L$. Using this information, we differentiate k to get

$$k' = \frac{K'}{L} - \frac{KL'}{L^2} = \frac{sY - \delta K}{L} - kn = s \frac{Y}{L} - \delta = sf(k) - (\delta + n)k.$$

Using the Cobb-Douglas function, we arrive at

$$k' = sk^\alpha - \lambda k \quad (2.8)$$

where $\lambda = \delta + n$. The final equation is a nonlinear first order equation which is called the Bernoulli equation. Such equations are considered in Appendix A2.2.3 and also considered later in the book.

2.2 Other models leading to exponential growth formula

Exponential growth models appear in numerous applications where the rate of change of some quantity is proportional to the amount present. We briefly describe two often used models of this type.

Radioactive decay

Radioactive substances undergo a spontaneous decay due to the emission of α particles. The mass of α particles is small in comparison with the sample of the radioactive material so it is reasonable to assume that the decrease of mass happens continuously. Experiments indicate that rate of decrease is proportional to mass of the sample still present

This principle immediately leads to the equation

$$N' = -kN, \quad (2.9)$$

where N is the number of radioactive particles present in the sample and k is the proportionality constant.

Absorption of drugs

Another important process which also leads to an exponential decay model is the absorption of drugs from the bloodstream into the body tissues. The significant quantity to

2.3 Continuous population models: first order equation 45

monitor is the concentration of the drug in the bloodstream, which is defined as the amount of drug per unit volume of blood. Observations show that the rate of absorption of the drug, which is equal to the rate of decrease of the concentration of the drug in the bloodstream, is proportional to the concentration of the drug in the bloodstream. Thus

{rate of decrease of concentration} is proportional to {concentration}

As before, the concentration c of the drug in the bloodstream satisfies

$$c' = -\gamma c, \quad (2.10)$$

with γ being the proportionality constant.

2.3 Continuous population models: first order equations

In this subsection we will study first order differential equations which appear in the population growth theory. At first glance it appears that it is impossible to model the growth of species by differential equations since the population of any species always change by integer amounts. Hence the population of any species can never be a differentiable function of time. However, if the population is large and it increases by one, then the change is very small compared to a given population. Thus we make the approximation that large populations changes continuously (and even differentiable) in time and, if the final answer is not an integer, we shall round it to the nearest integer. A similar justification applies to our use of t as a real variable: in absence of specific breeding seasons, reproduction can occur

at any time and for sufficiently large population it is then natural to think of reproduction as occurring continuously.

Let $N(t)$ denote the size of a population of a given isolated species at time t and let Δt be a small time interval. Then the population at time $t + \Delta t$ can be expressed as

$$\begin{aligned} N(t + \Delta t) - N(t) &= \text{number of births in } \Delta t \\ &\quad - \text{number of deaths in } \Delta t. \end{aligned}$$

It is reasonable to assume that the number of births and deaths in a short time interval is proportional to the population at the beginning of this interval and proportional to the length of this interval, thus

$$N(t + \Delta t) - N(t) = \beta(t, N(t))N(t)\Delta t - \mu(t, N(t))N(t)\Delta t. \quad (2.11)$$

Taking $r(t, N)$ to be the difference between the birth and death rate coefficients at time t for the population of size N we obtain

$$N(t + \Delta t) - N(t) = r(t, N(t))\Delta t N(t).$$

Dividing by Δt and passing with $\Delta t \rightarrow 0$ gives the equation

$$\frac{dN}{dt} = r(t, N)N. \quad (2.12)$$

Because of the unknown coefficient $r(t, N)$, depending on the unknown function N , this equation is impossible to solve. The form of r has to be deduced by other means.

2.3.1 Exponential growth

The simplest possible $r(t, N)$ is a constant and in fact such a model is used in a short-term population forecasting. So

2.3 Continuous population models: first order equation 47

let us assume that $r(t, N(t)) = r$ so that

$$\frac{dN}{dt} = rN. \quad (2.13)$$

It is exactly the same equation as (2.2). A little more general solution to it is given by

$$N(t) = N(t_0)e^{r(t-t_0)}, \quad (2.14)$$

where $N(t_0)$ is the size of the population at some fixed initial time t_0 .

To be able to give some numerical illustration to this equation we need the coefficient r and the population at some time t_0 . We use the data of the U.S. Department of Commerce: it was estimated that the Earth population in 1965 was 3.34 billion and that the population was increasing at an average rate of 2% per year during the decade 1960-1970. Thus $N(t_0) = N(1965) = 3.34 \times 10^9$ with $r = 0.02$, and (2.14) takes the form

$$N(t) = 3.34 \times 10^9 e^{0.02(t-1965)}. \quad (2.15)$$

To test the accuracy of this formula let us calculate when the population of the Earth is expected to double. To do this we solve the equation

$$N(T + t_0) = 2N(t_0) = N(t_0)e^{0.02T},$$

thus

$$2 = e^{0.02T}$$

and

$$T = 50 \ln 2 \approx 34.6 \text{ years.}$$

This is in an excellent agreement with the present observed

value of the Earth population and also gives a good agreement with the observed data if we don't go too far into the past. On the other hand, if we try to extrapolate this model into a distant future, then we see that, say, in the year 2515, the population will reach $199980 \approx 200000$ billion. To realize what it means, let us recall that the Earth total surface area 167400 billion square meters, 80% of which is covered by water, thus we have only 3380 billion square meters to our disposal and there will be only $0.16m^2$ ($40cm \times 40cm$) per person. Therefore we can only hope that this model is not valid for all times.

2.3.2 Logistic differential equation

Indeed, as for discrete models, it is observed that the linear model for the population growth is satisfactory as long as the population is not too large. When the population gets very large (with regard to its habitat), these models cannot be very accurate, since they don't reflect the fact that the individual members have to compete with each other for the limited living space, resources and food available. It is reasonable that a given habitat can sustain only a finite number K of individuals, and the closer the population is to this number, the slower is its growth. Again, the simplest way to take this into account is to take $r(t, N) = r(K - N)$ and then we obtain the so-called *continuous logistic model*

$$\frac{dN}{dt} = rN \left(1 - \frac{N}{K} \right), \quad (2.16)$$

which proved to be one of the most successful models for describing a single species population. This equation is still

Fig. 2.1. Comparison of actual population figures (points) with those obtained from equation (2.15)

first order equation but a non-linear one (the unknown function appears as an argument of the non-linear (quadratic) function $rx(1 - x/K)$). Eq. (2.16) is an example of a separable equation methods of solution of which are introduced in Appendix A2.2.1.

2.3.2.1 Interlude: spread of information

The logistic equation found numerous application in various contexts. We describe one such application in modelling

the spread of information in a fixed size community. Let us suppose that we have a community of constant size C and N members of this community have some important information. How fast this information is spreading? To find an equation governing this process we adopt the following assumptions:

- the information is passed when a person knowing it meets a person that does not know it;
- the rate at which one person meets other people is a constant f

Hence in a time interval Δt a person knowing the news meets $f\Delta t$ people and, in average, $\Delta t f(C - N)/C$ people who do not know it. If N people had the information at time t , then the increase in the time Δt will be

$$N(t + \Delta t) - N(t) = fN(t) \left(1 - \frac{N(t)}{C}\right) \Delta t$$

so that, as before,

$$\frac{dN}{dt} = fN \left(1 - \frac{N}{C}\right).$$

As in the discrete case, Eq. (2.16) can be generalized in many ways.

2.3.3 Other population models with restricted growth

Alternatively, as in the discrete case, we can obtain (2.16) by taking in (2.11) constant birth rate β but introduce density dependent mortality rate

$$\mu(N) = \mu_0 + \mu_1 N.$$

Then the increase in the population over a time interval Δt is given by

$$N(t + \Delta t) - N(t) = \beta N(t)\Delta t - \mu_0 N(t)\Delta t - \mu_1 N^2(t)\Delta t$$

which, upon dividing by Δt and passing with it to the limit, gives

$$\frac{dN}{dt} = (\beta - \mu_0)N - \mu_1 N^2$$

which is another form of (2.16).

In general, we can obtain various models suitable for particular applications by appropriately choosing β and μ in (2.11). In particular, by again taking β constant and

$$\mu(N) = \mu_0 + \mu_1 N^\theta$$

for some positive constant θ . The same argument as above results in the Bernoulli equation

$$\frac{dN}{dt} = (\beta - \mu_0)N - \mu_1 N^{\theta+1}. \quad (2.17)$$

We have already encountered this equation in the economic model (2.8).

2.4 Equations of motion: second order equations

Second order differential equations appear often as equations of motion. This is due to the Newton's law of motion that relates the acceleration of the body, that is, the second derivative of the position y with respect to time t , to the (constant) body's mass m and the forces F acting on it:

$$\frac{d^2 y}{dt^2} = \frac{F}{m}. \quad (2.18)$$

We confined ourselves here to a scalar, one dimensional case with time independent mass. The modelling in such cases concern the form of the force acting on the body. We shall consider two such cases in detail.

2.4.1 A waste disposal problem

In many countries toxic or radioactive waste is disposed by placing it in tightly sealed drums that are then dumped at sea. The problem is that these drums could crack from the impact of hitting the sea floor. Experiments confirmed that the drums can indeed crack if the velocity exceeds $12m/s$ at the moment of impact. The question now is to find out the velocity of a drum when it hits the sea floor. Since typically the waste disposal takes place at deep sea, direct measurement is rather expensive but the problem can be solved by mathematical modelling.

As a drum descends through the water, it is acted upon by three forces W, B, D . The force W is the weight of the drum pulling it down and is given by mg , where g is the acceleration of gravity and m is the mass of the drum. The buoyancy force B is the force of displaced water acting on the drum and its magnitude is equal to the weight of the displaced water, that is, $B = g\rho V$, where ρ is the density of the sea water and V is the volume of the drum. If the density of the drum (together with its content) is smaller than the density of the water, then of course the drum will be floating. It is thus reasonable to assume that the drum is heavier than the displaced water and therefore it will start drowning with constant acceleration. Experiments (and also common sense) tell us that any object

moving through a medium like water, air, etc. experiences some resistance, called the drag. Clearly, the drag force acts always in the opposite direction to the motion and its magnitude increases with the increasing velocity. Experiments show that in a medium like water for small velocities the drag force is proportional to the velocity, thus $D = cV$. If we now set $y = 0$ at the sea level and let the direction of increasing y be downwards, then from (2.18)

$$\frac{d^2y}{dt^2} = \frac{1}{m} \left(W - B - c \frac{dy}{dt} \right). \quad (2.19)$$

This is a second order (the highest derivative of the unknown function is of second order) and linear differential equation.

2.4.2 Motion in a changing gravitational field

According to Newton's law of gravitation, two objects of masses m and M attract each other with force of magnitude

$$F = G \frac{mM}{d^2}$$

where G is the gravitational constant and d is the distance between the objects' centres. Since at the Earth's surface the force is equal (by definition) to $F = mg$, the gravitational force exerted on a body of mass m at a distance y above the surface is given by

$$F = - \frac{mgR^2}{(y + R)^2},$$

where the minus sign indicates that the force acts towards Earth's centre. Thus the equation of motion of an object

of mass m projected upward from the surface is

$$m \frac{d^2 y}{dt^2} = -\frac{mgR^2}{(y+R)^2} - c \left(\frac{dy}{dt} \right)^2$$

where the last term represents the air resistance which, in this case, is taken to be proportional to the square of the velocity of the object. This is a second order nonlinear differential equation.

2.5 Equations coming from geometrical modelling

2.5.1 Satellite dishes

In many applications, like radar or TV/radio transmission it is important to find the shape of a surface that reflects parallel incoming rays into a single point, called the focus. Conversely, constructing a spot-light one needs a surface reflecting light rays coming from a point source to create a beam of parallel rays. To find an equation for a surface satisfying this requirement we set the coordinate system so that the rays are parallel to the x -axis and the focus is at the origin. The sought surface must have axial symmetry, that is, it must be a surface of revolution obtained by rotating some curve C about the x -axis. We have to find the equation $y = y(x)$ of C . Using the notation of the figure, we let $M(x, y)$ be an arbitrary point on the curve and denote by T the point at which the tangent to the curve at M intersects the x -axis. It follows that the triangle TOM is isosceles and

$$\tan \sphericalangle OTM = \tan \sphericalangle TMO = \frac{dy}{dx}$$

Fig. 2.2. Geometry of a reflecting surface

where the derivative is evaluated at M . By symmetry, we can assume that $y > 0$. Thus we can write

$$\tan \sphericalangle OTM = \frac{|MP|}{|TP|},$$

but $|MP| = y$ and, since the triangle is isosceles, $|TP| = |OT| \pm |OP| = |OM| \pm |OP| = \sqrt{x^2 + y^2} + x$, irrespectively of the sign of x . Thus, the differential equation of the curve

C is

$$\frac{dy}{dx} = \frac{y}{\sqrt{x^2 + y^2} + x}. \quad (2.20)$$

This is a nonlinear, so-called homogeneous, first order differential equation. As we shall see later, it is not difficult to solve, if one knows appropriate techniques, yielding a parabola, as expected from the Calculus course.

2.5.2 The pursuit curve

What is the path of a dog chasing a rabbit or the trajectory of self-guided missile trying to intercept an enemy plane? To answer this question we must first realize the principle used in controlling the chase. This principle is that at any instant the direction of motion (that is the velocity vector) is directed towards the chased object.

To avoid technicalities, we assume that the target moves with a constant speed v along a straight line so that the pursuit takes place on a plane. We introduce the coordinate system in such a way that the target moves along the y -axis in the positive direction, starting from the origin at the time $t = 0$, and the pursuer starts from a point at the negative half of the x -axis, see Fig. 2.3. We also assume that the pursuer moves with a constant speed u . Let $M = M(x(t), y(t))$ be a point at the curve C , having the equation $y = y(x)$, corresponding to the time t of the pursuit at which $x = x(t)$. At this moment the position of the target is $(0, vt)$. Denoting $y' = \frac{dy}{dx}$, from the principle of the pursuit we obtain

$$y' = -\frac{vt - y}{x}, \quad (2.21)$$

Fig. 2.3. The pursuit curve

where we have taken into account that $x < 0$. In this equation we have too many variables and we shall eliminate t as we are looking for the equation of the trajectory in x, y variables. Solving (2.21) with respect to x we obtain

$$x = -\frac{vt - y}{y'}, \quad (2.22)$$

whereupon, using the assumption that v is a constant and remembering that $x = x(t), y = y(x(t))$ and $y' = y'(x(t))$,

differentiating with respect to t we get

$$\frac{dx}{dt} = \frac{(-v + y' \frac{dx}{dt}) y' + (vt - y) y'' \frac{dx}{dt}}{(y')^2}.$$

Multiplying out and simplifying we get

$$0 = -vy' + (vt - y)y'' \frac{dx}{dt}$$

whereupon, using (2.22) and solving for $\frac{dx}{dt}$, we obtain

$$\frac{dx}{dt} = -\frac{v}{xy''}. \quad (2.23)$$

On the other hand, since we know that the speed of an object moving according to parametric equation $(x(t), y(t))$ is given by

$$u = \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2} = \sqrt{1 + (y')^2} \left| \frac{dx}{dt} \right|, \quad (2.24)$$

where we used the formula for parametric curves

$$\frac{dy}{dx} = \frac{\frac{dy}{dt}}{\frac{dx}{dt}},$$

whenever $dx/dt \neq 0$. From the formulation of the problem it follows that $dx/dt > 0$ (it would be unreasonable for the dog to start running away from the rabbit) hence we can drop the absolute value bars in (2.23). Thus, combining (2.23) and (2.24) we obtain the equation of the pursuit curve

$$xy'' = -\frac{v}{u} \sqrt{1 + (y')^2}. \quad (2.25)$$

This is a nonlinear second order equation having, however, a nice property of being reducible to a first order equation

and thus yielding a closed form solutions. We shall analyse this equation in Chapter ?? deal with such equations later on.

2.6 Modelling interacting quantities – systems of differential equations

In many situations we have to model evolutions of two (or more) quantities that are coupled in the sense that the state of one of them influences the other and conversely. We have seen this type of interactions in the discrete case when we modelled spread of a measles epidemic. It resulted then in a system of difference equations. Similarly, in the continuous case the evolution of interacting populations will lead to a system of differential equations. In this subsection we shall discuss modelling of such systems that results in both linear and non-linear systems.

2.6.1 Two compartment mixing – a system of linear equations

Let us consider a system consisting of two vats of equal capacity containing a diluted dye: the concentration at some time t of the dye in the first vat is $c_1(t)$ and in the second is $c_2(t)$. Suppose that the pure dye is flowing into the first vat at a constant rate r_1 and water is flowing into the second vat at a constant rate r_2 (in, say, litres per minute). Assume further that two pumps exchange the contents of both vats at constant rates: p_1 from vat 1 to vat 2 and conversely at p_2 . Moreover, the diluted mixture is drawn off vat 2 at a rate R_2 . The flow rates are chosen so that the

volumes of mixture in each vat remain constant, equal to V , that is $r_1 + p_2 - p_1 = r_2 - p_2 - R_2 = 0$. We have to find how the dye concentration in each vat changes in time.

Let $x_1(t)$ and $x_2(t)$ be the volumes of dye in each tank at $t \geq 0$. Thus, the concentrations c_1 and c_2 are defined by $c_1 = x_1/V$ and $c_2 = x_2/V$. We shall consider what happens to the volume of the dye in each vat during a small time interval from t to Δt . In vat 1

$$\begin{aligned} x_1(t + \Delta t) - x_1(t) &= \left\{ \begin{array}{l} \text{volume of} \\ \text{of pure dye} \\ \text{flowing into vat 1} \end{array} \right\} \\ &+ \left\{ \begin{array}{l} \text{volume of} \\ \text{dye in mixture 2} \\ \text{flowing into vat 1} \end{array} \right\} - \left\{ \begin{array}{l} \text{volume of} \\ \text{dye in mixture 1} \\ \text{flowing out of vat 1} \end{array} \right\} \\ &= r_1 \Delta t + p_2 \frac{x_2(t)}{V} \Delta t - p_1 \frac{x_1(t)}{V} \Delta t, \end{aligned}$$

and in vat 2, similarly,

$$\begin{aligned} x_2(t + \Delta t) - x_2(t) &= \left\{ \begin{array}{l} \text{volume of} \\ \text{dye in mixture 1} \\ \text{flowing into vat 2} \end{array} \right\} \\ &- \left\{ \begin{array}{l} \text{volume of} \\ \text{dye in mixture 2} \\ \text{flowing out of vat 2} \end{array} \right\} - \left\{ \begin{array}{l} \text{volume of dye in} \\ \text{mixture 2 flowing} \\ \text{from vat 2 vat 1} \end{array} \right\} \\ &= p_1 \frac{x_1(t)}{V} \Delta t - R_2 \frac{x_2(t)}{V} \Delta t - p_2 \frac{x_2(t)}{V} \Delta t. \end{aligned}$$

As before, we dividing by Δt and passing with it to zero we obtain the following simultaneous system of linear dif-

2.6 Modelling interacting quantities – systems of differential equations

$$\begin{aligned}\frac{dx_1}{dt} &= r_1 + p_2 \frac{x_2}{V} - p_1 \frac{x_1}{V} \\ \frac{dx_2}{dt} &= p_1 \frac{x_1}{V} - (R_2 + p_2) \frac{x_2}{V}.\end{aligned}\quad (2.26)$$

2.6.2 Continuous population models

Let us consider a model with population divided into n subpopulations, as in Subsection 1.2.3, but with transitions between occurring very quickly. This warrants describing the process in continuous time. Note that this in natural way excludes age structured populations discussed earlier as those models were constructed assuming discrete time. Continuous time age structure population models require a different approach leading to partial differential equation and thus are beyond the scope of this lecture notes.

Let $v_i(t)$ denotes the number of individuals in subpopulation i at time t and consider the change of the size of this population in a small time interval Δt . Over this interval, an individual from a j -th subpopulation can undergo the same processes as in the discrete case; that is,

- move to i -th subpopulation with (approximate) probability $p_{ij}\Delta t$;
- contribute to the birth of an individual in i -th subpopulation with probability $b_{ij}\Delta t$;
- die with probability $d_j\Delta t$.

Thus, the number of individuals in class i at time $t + \Delta t$ is:
the number of individuals in class i at time t - the number of deaths in class i + the number of births in class i do to interactions with individuals in all other classes + the number of

individuals who migrated to class i from all other classes - the number of individuals who migrated from class i to all other classes,

or, mathematically,

$$\begin{aligned} v_i(t + \Delta t) &= v_i(t) - d_i \Delta t v_i(t) + \sum_{j=1}^n b_{ij} \Delta t v_j(t) \\ &= \sum_{\substack{j=1 \\ j \neq i}}^n (p_{ij} \Delta t v_j(t) - p_{ji} \Delta t v_i(t)), \quad i = 1, \dots, n, \end{aligned} \quad (2.27)$$

To make the notation more compact, we denote $q_{ij} = b_{ij} + p_{ij}$ for $i \neq j$ and

$$q_{ii} = b_{ii} - d_i - \sum_{\substack{j=1 \\ j \neq i}}^n p_{ji}.$$

Using this notation in (2.27), dividing by Δt and passing to the limit with $\Delta t \rightarrow 0$ we obtain

$$v'_i(t) = \sum_{j=1}^n q_{ij} v_j(t), \quad i = 1, \dots, n, \quad (2.28)$$

or

$$\mathbf{v}' = \mathcal{Q}\mathbf{v}, \quad (2.29)$$

where $\mathcal{Q} = \{q_{ij}\}_{1 \leq i, j \leq n}$.

Let us reflect for a moment on similarities and differences between continuous and discrete time models. To simplify the discussion we shall focus on processes with no births or deaths events: $b_{ij} = d_j = 0$ for $1 \leq i, j \leq n$. As in the discrete time model, the total size of the population at any

given time t is given by $N(t) = v_1(t) + \dots + v_n(t)$. Then, the rate of change of N is given by

$$\begin{aligned}
 \frac{dN}{dt} &= \sum_{1 \leq i \leq n} \frac{dv_i(t)}{dt} = \sum_{i=1}^n \left(\sum_{j=1}^n q_{ij} v_j(t) \right) & (2.30) \\
 &= \sum_{i=1}^n q_{ii} v_i(t) + \sum_{i=1}^n \left(\sum_{\substack{j=1 \\ j \neq i}}^n q_{ij} v_j(t) \right) \\
 &= - \sum_{i=1}^n v_i(t) \left(\sum_{\substack{j=1 \\ j \neq i}}^n p_{ji} \right) + \sum_{i=1}^n \left(\sum_{\substack{j=1 \\ j \neq i}}^n p_{ij} v_j(t) \right) \\
 &= - \sum_{i=1}^n v_i(t) \left(\sum_{\substack{j=1 \\ j \neq i}}^n p_{ji} \right) + \sum_{j=1}^n v_j(t) \left(\sum_{\substack{i=1 \\ i \neq j}}^n p_{ij} \right) \\
 &= - \sum_{i=1}^n v_i(t) \left(\sum_{\substack{j=1 \\ j \neq i}}^n p_{ji} \right) + \sum_{i=1}^n v_i(t) \left(\sum_{\substack{j=1 \\ j \neq i}}^n p_{ji} \right) = 0,
 \end{aligned}$$

where we used the fact that i, j are dummy variables.

Remark 2.1 *The change of order of summation can be*

justified as follows

$$\begin{aligned}
\sum_{i=1}^n \left(\sum_{\substack{j=1 \\ j \neq i}}^n p_{ij} v_j \right) &= \sum_{i=1}^n \left(\sum_{j=1}^n p_{ij} v_j \right) - \sum_{i=1}^n p_{ii} v_i \\
&= \sum_{j=1}^n \left(\sum_{i=1}^n p_{ij} v_j \right) - \sum_{j=1}^n p_{jj} v_j = \sum_{j=1}^n v_j \left(\sum_{i=1}^n p_{ij} - p_{jj} \right) \\
&= \sum_{j=1}^n v_j \left(\sum_{\substack{i=1 \\ i \neq j}}^n p_{ij} \right).
\end{aligned}$$

Hence, $N(t) = N(0)$ for all time and the process is conservative.

The continuous process to certain extent can be compared to analysis of the period increments in the discrete time process:

$$\begin{aligned}
\mathbf{v}(k+1) - \mathbf{v}(k) &= (-I + \mathcal{P})\mathbf{v}(k) \tag{2.31} \\
&= \begin{pmatrix} -1 + p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & -1 + p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ p_{n1} & p_{n2} & \cdots & -1 + p_{nn} \end{pmatrix} \mathbf{v}(k),
\end{aligned}$$

The 'increment' matrix has the property that each row adds up to zero due to (1.44). However, it is important to remember that the coefficients p_{ij} in the continuous case are not probabilities and thus they do not add up to zero. In fact, they can be arbitrary numbers and represent probability rates with $p_{ij}\Delta t$ being approximate interstate transition probabilities.

2.6.3 Continuous model of epidemics – a system of nonlinear differential equations

The measles epidemic discussed earlier was modelled as a system of non-linear difference equations. The reason for the applicability of difference equations was the significant latent period between catching the disease and becoming contagious. If this period is very small (ideally zero) it is more reasonable to construct a model involving coupled differential equations. For the purpose of formulating the model we divide the population into three groups: susceptibles (who are not immune to the disease), infectives (who are capable of infecting susceptibles) and removed (who have previously had the disease and may not be reinfected because they are immune, have been quarantined or have died from the disease). The symbols S, I, R will be used to denote the number of susceptibles, infectives and removed, respectively, in the population at time t . We shall make the following assumptions on the character of the disease:

- (a) The disease is transmitted by close proximity or contact between an infective and susceptible.
- (b) A susceptible becomes an infective immediately after transmission.
- (c) Infectives eventually become removed.
- (d) The population of susceptibles is not altered by emigration, immigration, births and deaths.
- (e) Each infective infects a constant fraction β of the susceptible population per unit time (mass action law).
- (f) The number of infectives removed is proportional to the number of infectives present.

As mentioned earlier, it is assumption (b) that makes a differential rather than difference equation formulation more reasonable. Diseases for which this assumption is applicable include diphtheria, scarlet fever and herpes. Assumption (e) is the same that used in difference equation formulation. It is valid provided the number of infectives is small in comparison to the number of susceptibles.

To set up the differential equations, we shall follow the standard approach writing first difference equations over arbitrary time interval and then pass with the length of this interval to zero. Thus, by assumptions (a), (c) and (d), for any time t

$$S(t + \Delta t) = S(t) - \left\{ \begin{array}{l} \text{number of susceptibles} \\ \text{infected in time } \Delta t \end{array} \right\},$$

by assumptions (a), (b) and (c)

$$I(t + \Delta t) = I(t) + \left\{ \begin{array}{l} \text{number of susceptibles} \\ \text{infected in time } \Delta t \end{array} \right\} \\ - \left\{ \begin{array}{l} \text{number of infectives} \\ \text{removed in time } \Delta t \end{array} \right\},$$

and by assumptions (a), (c) and (d)

$$R(t + \Delta t) = R(t) + \left\{ \begin{array}{l} \text{number of infectives} \\ \text{removed in time } \Delta t \end{array} \right\}.$$

However, from assumptions (c) and (f)

$$\left\{ \begin{array}{l} \text{number of susceptibles} \\ \text{infected in time } \Delta t \end{array} \right\} = \beta SI \Delta t \\ \left\{ \begin{array}{l} \text{number of infectives} \\ \text{removed in time } \Delta t \end{array} \right\} = \gamma I \Delta t.$$

Combining all these equations and dividing by Δt and passing with it to 0 we obtain the coupled system of nonlinear differential equations

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI, \\ \frac{dI}{dt} &= \beta SI - \gamma I, \\ \frac{dR}{dt} &= \gamma I,\end{aligned}\tag{2.32}$$

where α, β are proportionality constants. Note that R does not appear in the first two equations so that we can consider separately and then find R by direct integration. The first two equations are then a continuous analogue of the system (1.46) with $B = 0$. Note that a simpler form of the equation for I in the discrete case follows from the fact that due to precisely one week recovering time the number of removed each week is equal to the number of infectives the previous week so that these two cancel each other in the equation for I .

2.6.4 Predator–prey model – a system of nonlinear equations

Systems of coupled nonlinear differential equations similar to (2.32) appear in numerous applications. One of the most famous is the Lotka–Volterra, or predator–prey model, created to explain why in a period of reduced fishing during the World War I, the number of sharks and other predators substantially increased. We shall describe it on the original example of small fish – sharks interaction.

To describe the model, we consider two populations: of

smaller fish and sharks, with the following influencing factors taken into account.

- (i) Populations of fish and sharks display an exponential growth when considered in isolation. However, the growth rate of sharks in absence of fish is negative due to the lack of food.
- (ii) Fish is preyed upon by sharks resulting in the decline in fish population. It is assumed that each shark eats a constant fraction of the fish population.
- (iii) The population of sharks increases if there is more fish. The additional number of sharks is proportional to the number of available fish.
- (iv) Fish and sharks are being fished indiscriminately, that is, the number of sharks and fish caught by fishermen is directly proportional to the present populations of fish and sharks, respectively, with the same proportionality constant.

If we denote by x and y the sizes of fish and shark populations, then an argument, similar to that leading to (2.32), gives the following system

$$\begin{aligned}\frac{dx}{dt} &= (r - f)x - \alpha xy, \\ \frac{dy}{dt} &= -(s + f)y + \beta xy\end{aligned}\quad (2.33)$$

where α, β, r, s, f are positive constants.

We note that a more general form of (2.33):

$$\begin{aligned}\frac{dx}{dt} &= x(a + bx + cy), \\ \frac{dy}{dt} &= y(d + ey + fx)\end{aligned}\quad (2.34)$$

with constants a, b, c, d, e, f of arbitrary sign can describe a wide range of interactions between two populations. For instance, if $b < 0$, then we have a logistic growth of the population x in absence of y indicating an intraspecies competition. If $c > 0$ then, contrary to the previous case, the presence of the species y benefits x so that if both c and f are positive we have cooperating species. On the other hand, the situation when both f and c are negative arises when one models populations competing for the same resource.

3

Solutions and applications of discrete models

In this chapter we shall go through several difference equations introduced in Chapter 1 which admit closed form solutions and describe some further applications.

3.1 Inverse problems – estimates of the growth rate

Most population models contain parameters which are not given and must be determined by fitting the model to the observable data. We shall discuss two simple examples of this type.

Growth rate in an exponential model A total of 435 bass fish were introduced in 1979 and 1981 into a bay. In 1989, the commercial net catch alone was 4 000 000 kg. Since the growth of this population was so fast, it is reasonable to assume that it obeyed the Malthusian law $N_{k+1} = R_0 N_k$. Assuming that the average weight of a bass fish is 1.5 kg, and that in 1999 only 10% of the bass population was caught, we find lower and upper bounds for r . Recalling the formula

(1.17) we have

$$N(k) = N_0 R_0^k$$

where we measure k in years and $R_0 > 1$ (as we have growth). Let us denote by N_1 and N_2 the amounts of fish introduced in 1979 and 1981, respectively, so that $N_1 + N_2 = 435$. Thus, we can write the equation

$$N(1989) = N_1 R_0^{1989-1979} + N_2 R_0^{1989-1981} = N_1 R_0^{10} + N_2 R_0^8.$$

Since we do not know N_1 nor N_2 we observe that $R_0^2 > 1$ and thus

$$N(1989) \leq N_1 R_0^{10} + N_2 R_0^{10} = 435 R_0^{10}.$$

Similarly

$$N(1989) \geq N_1 R_0^8 + N_2 R_0^{10} = 435 R_0^8.$$

Hence

$$\sqrt[10]{\left(\frac{N(1989)}{435}\right)} \leq R_0 \leq \sqrt[8]{\left(\frac{N(1989)}{435}\right)}.$$

Now, the data of the problem give as $N(1989) = 10 \times 4000000/1.5 \approx 26666666$ and so

$$2.39 \leq R_0 \leq 2.97.$$

Growth rate and capacity of the environment in the logistic model

Suppose that a population grows according to the logistic equation

$$N(k+1) = N(k) + R_0 N(k) \left(1 - \frac{N(k)}{K}\right), \quad (3.1)$$

with r and K unknown. We easily see that if $N(k+1) = N(k)$ implies $N(k) = 0$ or $N(k) = K$ so that the population is strictly growing provided it is not extinct or the largest possible in the given environment.

It follows that to determine R_0 and K it is enough to know three subsequent measurements of the population size. Since the model is autonomous, we can call them N_0, N_1 and N_2 . Thus

$$\begin{aligned} N(1) &= N(0) + R_0 N(0) \left(1 - \frac{N(0)}{K}\right), \\ N(2) &= N(1) + R_0 N(1) \left(1 - \frac{N(1)}{K}\right) \end{aligned}$$

hence

$$\frac{N(1) - N(0)}{N(2) - N(1)} = \frac{N(0)(K - N(0))}{N(1)(K - N(1))}.$$

Denoting

$$Q = \frac{N(1)(N(1) - N(0))}{N(0)(N(2) - N(1))}$$

we obtain $K - N(0) = Q(K - N(1))$ so

$$K = \frac{N(0) - QN(1)}{1 - Q},$$

which is possible as $Q \neq 1$. Indeed, $Q = 1$ implies $N(1)^2 = N(0)N(2)$ or $N(1)/N(0) = N(2)/N(1)$. But then (3.1) implies $N(0) = N(1)$, contrary to the result at the beginning of the paragraph. Having determined K , we get

$$R_0 = \frac{K(N(1) - N(0))}{N(0)(K - N(0))}.$$

3.2 Drug release

Assume that a dose D_0 of a drug, that increases its concentration in the patient's body by c_0 , is administered at regular time intervals $t = 0, T, 2T, 3T \dots$. Between the injections the concentration c of the drug decreases according to the differential equation $c' = -\gamma c$, where γ is a positive constant. It is convenient here to change slightly the notational convention and denote by c_n the concentration of the drug just after the n th injection, that is, c_0 is the concentration just after the initial (zeroth) injection, c_1 is the concentration just after the first injection, that is, at the time T , etc. We are to find formula for c_n and determine whether the concentration of the drug eventually stabilizes.

In this example we have a combination of two processes: continuous between the injections and discrete in injection times. Firstly, we observe that the process is discontinuous at injection times so we have two different values for $c(nT)$: just before the injection and just after the injection (assuming that the injection is done instantaneously). To avoid ambiguities, we denote by $c(nT^-)$ the concentration just before the n th injection and by c_n the concentration just after, in accordance with the notation introduced above. Thus, between the n th and $n + 1$ st injection the concentration changes according to the exponential law

$$c((n + 1)T^-) = c_n e^{-\gamma T}$$

so that over each time interval between injection the concentration decreases by a constant fraction $a = e^{-\gamma T} < 1$. Thus, we are able to write down the difference equation for concentrations just after $n + 1$ st injection as

$$c_{n+1} = ac_n + c_0. \quad (3.2)$$

We can write down the solution using (1.7) as

$$c_n = c_0 a^n + c_0 \frac{a^n - 1}{a - 1} = -\frac{c_0}{1 - a} a^{n+1} + \frac{c_0}{1 - a}.$$

Since $a < 1$, we immediately obtain that $\bar{c} = \lim_{n \rightarrow \infty} c_n = \frac{c_0}{1 - a} = \frac{c_0}{1 - e^{-\gamma T}}$.

Similarly, the concentration just before n th injection is

$$\begin{aligned} c(nT) &= c_{n-1} e^{-\gamma T} = e^{-\gamma T} \left(\frac{c_0}{e^{-\gamma T} - 1} e^{-\gamma T n} + \frac{c_0}{1 - e^{-\gamma T}} \right) \\ &= \frac{c_0}{1 - e^{\gamma T}} e^{-\gamma T n} + \frac{c_0}{e^{\gamma T} - 1} \end{aligned}$$

and for the long run $\underline{c} = \lim_{n \rightarrow \infty} c(nT) = \frac{c_0}{e^{\gamma T} - 1}$.

For example, using $c_0 = 14$ mg/l, $\gamma = 1/6$ and $T = 6$ hours we obtain that after a long series of injections the maximal concentration, attained immediately after injections, will stabilize at around 22 mg/l. The minimal concentration, just before injection, will stabilize at around $\underline{c} = 14/e - 1 \approx 8.14$ mg/l. This effect is illustrated at Fig. 3.1.

3.3 Mortgage repayment

In Subsection 1.1 we discussed the difference equation governing long-term loan repayment:

$$D(k+1) = D(k) + \frac{\alpha p}{100} D(k) - R = D(k) \left(1 + \frac{\alpha p}{100} \right) - R, \quad (3.3)$$

where D_0 is the initial debt to be repaid, for each k , $D(k)$ is the outstanding debt after the k th repayment, the payment made after each conversion period is R , $p\%$ is the annual interest rate and α is the conversion period, that is, the

Fig. 3.1. Long time behaviour of the concentration $c(t)$.

number of payments in one year. To simplify notation we denote $r = \alpha p/100$

Using again (1.7) we obtain the solution

$$\begin{aligned} D(k) &= (1+r)^k D_0 - R \sum_{i=0}^{k-1} (1+r)^{k-i-1} \\ &= (1+r)^k D_0 - \left((1+r)^k - 1 \right) \frac{R}{r} \end{aligned}$$

This equation gives answers to a number of questions relevant in taking a loan. For example, if we want to know

what will be the monthly instalment on a loan of D_0 to be repaid in n payments, we observe that the loan is repaid in n instalments if $D(n) = 0$, thus we must solve

$$0 = (1+r)^n D_0 - ((1+r)^n - 1) \frac{R}{r}$$

in R , which gives

$$R = \frac{rD_0}{1 - (1+r)^{-n}}. \quad (3.4)$$

For example, taking a mortgage of R200000 to be repaid over 20 years in monthly instalments at the annual interest rate of 13% we obtain $\alpha = 1/12$, hence $r = 0.0108$, and $n = 20 \times 12 = 240$. Therefore

$$R = \frac{0.0108 \cdot 200000}{1 - 1.0108^{-240}} \approx R2343.15.$$

3.4 Conditions for the Walras equilibrium

Let us recall the model describing evolution of the price of a commodity (1.6):

$$p(n) = -\frac{m_s}{m_d} p(n-1) + \frac{b_d - b_s}{m_d}. \quad (3.5)$$

The equilibrium would be the price such that $p(n+1) = p(n)$. To find conditions for existence of such a price, let us solve this equation. Using (1.7) we see that

$$\begin{aligned} p(n) &= \left(-\frac{m_s}{m_d}\right)^n - \frac{b_d - b_s}{m_d + m_s} \left(\left(-\frac{m_s}{m_d}\right)^n - 1\right) \\ &= \left(-\frac{m_s}{m_d}\right)^n \left(1 - \frac{b_d - b_s}{m_d + m_s}\right) + \frac{b_d - b_s}{m_d + m_s}. \end{aligned}$$

Since $m_d, m_s > 0$, we have three cases to consider:

- (i) If $m_s/m_d < 1$; that is the suppliers are less sensitive to price than consumers, then

$$\lim_{n \rightarrow \infty} p(n) = \frac{b_d - b_s}{m_d + m_s} =: p_\infty.$$

Substituting $p(n-1) = p_\infty$ in (3.5) we obtain

$$\begin{aligned} p(n+1) &= -\frac{m_s}{m_d} \frac{b_d - b_s}{m_d + m_s} + \frac{b_d - b_s}{m_d} \\ &= \frac{b_d - b_s}{m_d} \left(\frac{-m_s}{m_d + m_s} + 1 \right) \\ &= \frac{b_d - b_s}{m_d + m_s} = p_\infty \end{aligned} \quad (3.6)$$

so that p_∞ is the equilibrium price. If we start from any other price p_0 , then the price of the commodity will oscillate around p_∞ tending to as $n \rightarrow \infty$. Such an equilibrium point is called asymptotically stable

- (ii) If $m_s = m_d$, then $p(n)$ will take on only two values: p_0 and

$$p(1) = -p_0 + \frac{b_d - b_s}{m_d}.$$

Indeed, we have

$$p(2) = -\left(-p_0 + \frac{b_d - b_s}{m_d}\right) + \frac{b_d - b_s}{m_d} = p_0$$

and the cycle repeats itself. Thus the price oscillates around $p_\infty = (b_d - b_s)/2$ (which is the midpoint of the possible values of the solution) but will not get closer to it with time. We call such an equilibrium stable but not asymptotically stable (or not attracting). In economic applications the price p_∞ in both cases is called stable.

- (iii) If $m_d/m_s > 1$; that is suppliers are more sensitive to price than consumers, then $p(n)$ will oscillate with $|p(n)|$ increasing to infinity. Thus, though p_∞ is an equilibrium price by (3.6), it is not a stable price.

3.5 Some explicitly solvable nonlinear models

We recall that the Beverton-Holt-Hassel equation equation (1.21) can be simplified to

$$x(n+1) = \frac{R_0 x(n)}{(1+x(n))^b}. \quad (3.7)$$

While for general b this equation can display a very rich dynamics, which will be looked at later on, for $b = 1$ it can be solved explicitly. So, let us consider:

$$x(n+1) = \frac{R_0 x(n)}{1+x(n)} \quad (3.8)$$

Writing (3.8) as

$$x(n+1) = \frac{R_0}{1 + \frac{1}{x(n)}}$$

we see that the substitution $y(n) = 1/x(n)$ converts it to

$$y(n+1) = \frac{1}{R_0} + \frac{1}{R_0} y(n)$$

Using (1.7) we find

$$y(n) = \frac{1}{R_0} \frac{R_0^{-n} - 1}{R_0^{-1} - 1} + R_0^{-n} y(0) = \frac{1 - R_0^n}{R_0^n (1 - R_0)} + R_0^{-n} y(0)$$

if $R_0 \neq 1$ and

$$y(n) = n + y(0)$$

for $R_0 = 1$. From these equations we see that $x(n) \rightarrow R_0 - 1$ if $R_0 > 1$ and $x(n) \rightarrow 0$ if $R_0 \leq 1$ as $n \rightarrow \infty$. It is maybe unexpected that the population faces extinction if $R_0 = 1$ (which corresponds to every individual giving on average birth to one offspring). However, the density depending factor causes some individuals to die between reproductive seasons which means the the population with $R_0 = 1$ in fact decreases with every cycle.

The logistic equation

In general the discrete logistic equation does not admit closed form solution. However, some special cases can be solved by an appropriate substitution. We shall look at two such cases. Consider

$$x(n+1) = 2x(n)(1-x(n))$$

and use substitution $x(n) = 1/2 - y(n)$. Then

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

so that

$$y(n+1) = 2y^2(n)$$

Then $y(n) > 0$ for $n > 0$ provided $y(0) > 0$ and we can take the logarithm of both sides getting

$$\ln y(n+1) = 2 \ln y(n) + 2$$

which, upon substitution $z(n) = \ln y(n)$ becomes the inhomogeneous linear equation

$$z(n+1) = 2z(n) + 2$$

which, according (1.7) has the form

$$z(n) = 2^n z(0) + 2(2^n - 1).$$

Hence

$$\begin{aligned} x(n) &= \frac{1}{2} - \exp(2(2^n - 1)) \exp\left(2^n \ln\left(\frac{1}{2} - x_0\right)\right) \\ &= \frac{1}{2} - \exp(2(2^n - 1)) \exp\left(2^{n-1} \ln\left(\frac{1}{2} - x_0\right)^2\right) \end{aligned}$$

where the second formula should be used if $x_0 > 1/2$ so that $y(0) < 0$. We note that for $x_0 = 1/2$ we have

$$x_1 = 2 \frac{1}{2} \frac{1}{2} = \frac{1}{2} = x_0$$

so that we obtain a constant solution ($x = 1/2$ is an equilibrium point).

Another particular logistic equation which can be solved by substitution is

$$x(n+1) = 4x(n)(1-x(n)). \quad (3.9)$$

First we note that since the function $f(x) = 4x(1-x) \leq 1$ for $0 \leq x \leq 1$, $0 \leq x_{n+1} \leq 1$ if $x(n)$ has this property. Thus, assuming $0 \leq x_0 \leq 1$, we can use the substitution

$$x(n) = \sin^2 y(n) \quad (3.10)$$

so that

$$\begin{aligned} x(n+1) &= \sin^2 y(n+1) = 4 \sin^2 y(n)(1 - \sin^2 y(n)) \\ &= 4 \sin^2 y(n) \cos^2 y(n) = \sin^2 2y(n). \end{aligned}$$

This gives a family of equations

$$y(n+1) = \pm 2y(n) + k\pi, \quad k \in \mathbb{Z}.$$

However, bearing in mind that our aim is to find $x(n)$ given

by (3.10) we can discard $k\pi$ as well as the minus sign and focus on

$$y(n+1) = 2y(n).$$

This is the geometric progression and we get

$$y(n) = C2^n,$$

where C is an arbitrary constant, as the general solution. Hence

$$x(n) = \sin^2 C2^n$$

where C is to be determined from $x_0 = \sin^2 C$. What is remarkable in this example is that, as we see later, the dynamics generated by (3.9) is chaotic despite the fact that there is an explicit formula for the solution.

4

Basic differential equation models – solutions

In the previous chapter we have introduced several objects which we called differential equations. Here we shall make this concept more precise, we discuss various approaches to solving a differential equation and provide solutions to the models of the previous chapters. Since these notes mainly are about modelling, we refer the reader to dedicated texts to learn more about the theory of differential equations. However, to make the presentation self-consistent, we provide basic facts and methods in the appendix.

4.1 What are differential equations?

What precisely do we mean by a differential equation? The more familiar notion of an algebraic equation, like for example the quadratic equation $x^2 - 4x - 5 = 0$, states something about a number x . It is sometimes called an open statement since the number x is left unspecified, and the statement's truth depends on the value of x . Solving the

equation then amounts to finding values of x for which the open statement turns into a true statement.

Algebraic equations arise in modelling processes where the unknown quantity is a number (or a collection of numbers) and all the other relevant quantities are constant. As we observed in the first chapter, if the data appearing in the problem are variable and we describe a changing phenomenon, then the unknown will be rather a function (or a sequence). If the changes occur over very short interval, then the modelling usually will have to balance small increments of this function and the data of the problem and will result typically in an equation involving the derivatives of the unknown function. Such an equation is called a differential equation.

Differential equations are divided into several classes. The main two classes are ordinary differential equations (ODEs) and partial differential equations (PDEs). As suggested by the name, ODEs are equations where the unknown function is a function of one variable and the derivatives involved in the equation are ordinary derivatives of this function. A partial differential equation involves functions of several variables and thus expresses relations between partial derivatives of the unknown function.

In this course we shall be concerned solely with ODEs and systems of ODEs. Symbolically, the general form of ODE is

$$F(y^{(n)}, y^{(n-1)}, \dots, y', y, t) = 0, \quad (4.1)$$

where F is a given function of $n + 2$ variables. For example, the equation of exponential growth can be written as $F(y', y, t) = y' - ry$ so that the function F is a function of

two variables (constant with respect to t) and acting into r . Systems of differential equations can be also written in the form (4.1) if we accept that both F and y (and all the derivatives of y) can be vectors. For example, in the case of the epidemic spread (2.32) we have a system of ODEs which can be written as

$$\mathbf{F}(\mathbf{y}, t) = 0,$$

with three-dimensional vector $\mathbf{y} = (S, I, R)$ and the vector $\mathbf{F} = (F_1, F_2, F_3)$ with $F_1(S, I, R, t) = -\beta SI$, $F_2(S, I, R, t) = \beta SI - \gamma I$ and $F_3(S, I, R, t) = \gamma I$.

What does it mean to solve a differential equation? For algebraic equations, like the one discussed at the beginning, we can apply the techniques learned in the high school finding the discriminant of the equation $\Delta = (-4)^2 - 4 \cdot 1 \cdot (-5) = 36$ so that $x_{1,2} = 0.5(4 \pm 6) = 5, -1$. Now, is this the solution to our equation? How can we check it? The answer is given above – the solution is a number (or a collection of numbers) that turns the equation into a true statement. In our case, $5^2 - 20 - 5 = 0$ and $(-1)^2 - 4(-1) - 5 = 0$, so both numbers are solutions to the equation.

Though presented in a simple context, this is a very important point.

To solve a problem is to find a quantity that satisfies all the conditions of this problem.

This simple truth is very often forgotten as students tend to apply mechanically steps they learned under say, "techniques for solving quadratic equations" or "techniques of integration" labels and look for answers or model solutions "out there" though the correctness of the solution in most cases can be checked directly.

The same principle applies to differential equations. That is, to solve the ODE (4.1) means to find an n -times continuously differentiable function $y(t)$ such that for any t (from some interval)

$$F(y^{(n)}(t), y^{(n-1)}(t), \dots, y'(t), y(t), t) \equiv 0.$$

Once again, there are many techniques for solving differential equations. Some of them give only possible candidates for solutions and only checking that these suspects really turn the equation into the identity can tell us whether we have obtained the correct solution or not.

Example 4.1 *As an example, let us consider which of these functions $y_1(t) = 30e^{2t}$, $y_2(t) = 30e^{3t}$ and $y_3(t) = 40e^{2t}$ solves the equation $y' = 2y$. In the first case, LHS is equal to $60e^{2t}$ and RHS is $2 \cdot 30e^{2t}$ so that $LHS = RHS$ and we have a solution. In the second case we obtain $LHS = 90e^{3t} \neq 2 \cdot 30e^{3t} = RHS$ so that y_2 is not a solution. In the same way we find that y_3 satisfies the equation.*

Certainly, being able to check whether a given function is a solution is not the same as actually finding the solution. Thus, this example rises the following three questions.

- (i) Can we be sure that a given equation possesses a solution at all?
- (ii) If we know that there is a solution, are there systematic methods to find it?
- (iii) Having found a solution, can we be sure that there are no other solutions?

Question 1 is usually referred to as the **existence problem** for differential equations, and Question 3 as the **unique-**

ness problem. Unless we deal with very simple situations these should be addressed before attempting to find a solution. After all, what is the point of trying to solve equation if we do not know whether the solution exists, and whether the solution we found is the one we are actually looking for, that is, the solution of the real life problem the model of which is the differential equation.

Let us discuss briefly Question 1 first. Roughly speaking, we can come across the following situations.

- (i) No function exists which satisfies the equation.
- (ii) The equation has a solution but no one knows what it looks like.
- (iii) The equation can be solved in a closed form, either in elementary functions,
or in quadratures.

Case 1 is not very common in mathematics and it should never happen in mathematical modelling. In fact, if a given equation was an exact reflection of a real life phenomenon, then the fact that this phenomenon exists would ensure that the solution to this equation exists also. For example, if we have an equation describing a flow of water, then the very fact that water flows would be sufficient to claim that the equation must have a solution. However, in general, models are imperfect reflections of real life and therefore it may happen that in the modelling process we missed a crucial fact, rendering thus the final equation unsolvable. Thus, checking that a given equation is solvable serves as an important first step in validation of the model. Unfortunately, these problems are usually very difficult and require quite advanced mathematics that is beyond the scope of

this course. On the other hand, all the equations we will be dealing with are classical and the fundamental problems of existence and uniqueness for them have been positively settled at the beginning of the 20th century.

Case 2 may look somewhat enigmatic but, as we said above, there are advanced theorems allowing to ascertain the existence of solution without actually displaying them. This should be not surprising: after all, we know that the Riemann integral of any continuous function exists though in many cases we cannot evaluate it explicitly.

Even if we do not know a formula for the solution, the situation is not hopeless. Knowing that the solution exists, we have an array of approximate, numerical methods at our disposal. Using them we are usually able to find the numerical values of the solution with arbitrary accuracy. Also, very often we can find important features of the solution without knowing it. These feature include e.g. the long time behaviour of the solution, that is, whether it settles at a certain equilibrium value or rather oscillates, whether it is monotonic etc. These questions will be studied by in the final part of our course.

Coming now to Case 3 and to an explanation of the meaning of the terms used in the subitems, we note that clearly an ideal situation is if we are able to find the solution as an algebraic combination of elementary functions

$$y(t) = \text{combination of elementary functions like :} \\ \sin t, \cos t, \ln t, \text{ exponentials, polynomials...}$$

Unfortunately, this is very rare for differential equation. Even the simplest cases of differential equations involving only elementary functions may fail to have such solutions.

Example 4.2 For example, consider is the equation

$$y' = e^{-t^2}.$$

Integrating, we find that the solution must be

$$y(t) = \int e^{-t^2} dt$$

but, on the other hand, it is known that this integral cannot be expressed as a combination of elementary functions.

This brings us to the definition of *quadratures*. We say that an equation is *solvable in quadratures* if a solution to this equation can be written in terms of integrals of elementary functions (as above). Since we know that every continuous function has an antiderivative (though often we cannot find this antiderivative explicitly), it is almost as good as finding the explicit solution to the equation.

Having dealt with Questions 1 and 2 above, that is, with existence of solutions and solvability of differential equations, we shall move to the problem of uniqueness. We have observed in Example 4.1 that the differential equation by itself defines a family of solutions rather than a single function. In this particular case this class depend on an arbitrary parameter. Another simple example of a second order differential equation $y'' = t$, solution of which can be obtained by a direct integration as $y = \frac{1}{6}t^3 + C_1t + C_2$, shows that in equations of the second order we expect the class of solutions to depend on 2 arbitrary parameters. It can be then expected that the class of solutions for an n th order equation will contain n arbitrary parameters. Such a full class is called the *general solution* of the differential equation. By imposing the appropriate number of *side con-*

ditions we can specify the constants obtaining thus a *special solution* - ideally one member of the class.

A side condition may take all sorts of forms, like "at $t = 15$, y must have the value of 0.4" or "the area under the curve between $t = 0$ and $t = 24$ must be 100". Very often, however, it specifies the initial value of $y(0)$ of the solution and the derivatives $y^k(0)$ for $k = 1, \dots, n - 1$. In this case the side conditions are called the *initial conditions*.

After these preliminaries we shall narrow our consideration to a particular class of problems for ODEs.

4.2 Cauchy problem for first order equations

In this section we shall be concerned with *first order* ordinary differential equations which are solved with respect to the derivative of the unknown function, that is, with equations which can be written as

$$\frac{dy}{dt} = f(t, y), \quad (4.2)$$

where f is a given function of two variables.

In accordance with the discussion of the previous session, we shall be looking for solutions to the following Cauchy problem

$$\begin{aligned} y' &= f(t, y), \\ y(t_0) &= y_0 \end{aligned} \quad (4.3)$$

where we abbreviated $\frac{dy}{dt} = y'$, and t_0 and y_0 are some given numbers.

Several comments are in place here. Firstly, even though in such a simplified form, the question of solvability of the problem (4.3) is almost as difficult as that of (4.1). Before

we embark on studying this problem, we again emphasize that to solve (4.3) is to find a function $y(t)$ that is continuously differentiable at least in some interval (t_1, t_2) containing t_0 , that satisfies

$$\begin{aligned}y'(t) &\equiv f(t, y(t)) \quad \text{for all } t \in (t_1, t_2) \\y(t_0) &= y_0.\end{aligned}$$

Let consider the following example.

Example 4.3 *Check that the function $y(t) = \sin t$ is a solution to the problem*

$$\begin{aligned}y' &= \sqrt{1 - y^2}, \quad t \in (0, \pi/2), \\y(\pi/2) &= 1\end{aligned}$$

Solution. *LHS: $y'(t) = \cos t$, RHS: $\sqrt{1 - y^2} = \sqrt{1 - \sin^2 t} = |\cos t| = \cos t$ as $t \in (0, \pi/2)$. Thus the equation is satisfied. Also $\sin \pi/2 = 1$ so the "initial" condition is satisfied.*

Note that the function $y(t) = \sin t$ is not a solution to this equation on a larger interval $(0, a)$ with $a > \pi/2$ as for $\pi/2 < t < 3\pi/2$ we have LHS: $y'(t) = \cos t$ but RHS: $\sqrt{1 - y^2} = |\cos t| = -\cos t$, since $\cos t < 0$.

How do we know that a given equation has a solution? For an equation in the (4.2) form the answer can be given in relatively straightforward terms, though it is still not easy to prove.

Theorem 4.1 [Peano] *If the function f in (4.3) is continuous in some neighbourhood of the point (t_0, y_0) , then the problem (4.3) has at least one solution in some interval (t_1, t_2) containing t_0 .*

Thus, we can safely talk about solutions to a large class of ODEs of the form (4.2) even without knowing their explicit formulae.

As far as uniqueness is concerned, we know that the equation itself determines a class of solutions; for first order ODE this class is a family of functions depending on one arbitrary parameter. Thus, in principle, imposing one additional condition, as e.g. in (4.3), we should be able to determine this constant so that the Cauchy problem (4.3) should have only one solution. Unfortunately, in general this is not so as demonstrated in the following example.

Example 4.4 *The Cauchy problem*

$$\begin{aligned} y' &= \sqrt{y}, & t > 0 \\ y(0) &= 0, \end{aligned}$$

has at least two solutions: $y \equiv 0$ and $y = \frac{1}{4}t^2$.

Fortunately, there is a large class of functions f for which (4.3) has exactly one solution. This result is known as the Picard Theorem which we state below.

Theorem 4.2 [Picard] *Let the function f in (4.3) be continuous in the rectangle $\mathcal{R} : |t - t_0| \leq a, |y - y_0| \leq b$ for some $a, b > 0$ and satisfy there the Lipschitz condition with respect to y :*

$$\exists_{0 \leq L < +\infty} \forall_{(t, y_1), (t, y_2) \in \mathcal{R}} |f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2| \quad (4.4)$$

Compute

$$M = \max_{(t, y) \in \mathcal{R}} |f(t, y)|$$

and define $\alpha = \min\{a, b/M\}$. Then the initial value problem (4.3) has exactly one solution at least on the interval $t_0 - \alpha \leq t \leq t_0 + \alpha$.

Remark 4.1 *If f is such that f_y is bounded in \mathcal{R} , then (4.4) is satisfied.*

Picard's theorem gives local uniqueness; that is, for any point (t_0, y_0) around which the assumptions are satisfied, there is an interval over which there is only one solution of the given Cauchy problem. However, taking a more global view, it is possible that we have two solutions $y_1(t)$ and $y_2(t)$ which coincide over the interval of uniqueness mentioned above but branching for larger times. If we assume that any point of the plane is the uniqueness point, such a scenario is impossible. In fact, if $y_1(t) = y_2(t)$ over some interval I , then by continuity of solutions, there is the largest t , say t_1 , having this property. Thus, $y_1(t_1) = y_2(t_1)$ with $y_1(t) \neq y_2(t)$ for some $t > t_1$. Thus, the point $(t_1, y_1(t_1))$ would be the point violating Picard's theorem, contrary to the assumption.

An important consequence of the above is that we can glue solutions together to obtain solution defined on a possibly larger interval. If $y(t)$ is a solution to (4.3) defined on an interval $[t_0 - \alpha, t_0 + \alpha]$ and $(t_0 + \alpha, y(t_0 + \alpha))$ is a point around which the assumption of Picard's theorem is satisfied, then there is a solution passing through this point defined on some interval $[t_0 + \alpha - \alpha', t_0 + \alpha + \alpha']$, $\alpha' > 0$. These two solutions coincide on $[t_0 + \alpha - \alpha', t_0 + \alpha]$ and therefore, by the first part, they must coincide over the whole interval of their common existence and therefore constitute

a solution of the original Cauchy problem defined at least on $[t_0 - \alpha, t_0 + \alpha + \alpha']$.

In many applications it is important to determine whether the solution exists for all times. To provide a relevant result first we introduce the *maximal interval of existence* of a solution of the differential equation: $[t_0, t_0 + \alpha^*)$ is said to be the (forward) maximal interval of existence for a solution $y(t)$ to (4.3) if there is no solution $y_1(t)$ on an interval $[t_0, t_0 + \alpha^+)$, where $\alpha^+ > \alpha^*$, satisfying $y(t) = y_1(t)$ for $t \in [t_0, t_0 + \alpha^*)$. We can also consider backward intervals trying to extend the solution for $t < t_0$. We note that the forward (backward) interval of existence is open from the right (left).

Theorem 4.3 *If we assume that f in (4.3) satisfies the assumptions of Picard's theorem on any rectangle $\mathcal{R} \subset \mathbb{R}^2$, then $[t_0, t_0 + \alpha^*)$ is a finite forward maximal interval of existence of $y(t)$ if and only if*

$$\lim_{t \rightarrow t_0 + \alpha^*} |y(t)| = \infty. \quad (4.5)$$

Proof. See e.g. [?].

The above theorems are of utmost importance, both theoretical and practical. We illustrate some applications below.

Example 4.5 *We consider one of the simplest differential equations, introduced in (2.2) and (2.13)*

$$\frac{dy}{dt} = ay. \quad (4.6)$$

This is a separable equation discussed in more details in

(A2.2.1) and thus can be solved easily. Assuming that $y(t) \neq 0$ for any t , we have

$$\frac{1}{y(t)} \frac{dy}{dt} = \frac{d}{dt} \ln |y(t)|,$$

so that

$$\frac{d}{dt} \ln |y(t)| = a,$$

and, by direct integration,

$$\ln |y(t)| = at + c_1$$

where c_1 is an arbitrary constant of integration. Taking exponentials of both sides yields

$$|y(t)| = \exp(at + c_1) = c_2 \exp(at)$$

where c_2 is an arbitrary positive constant: $c_2 = \exp c_1 > 0$. We have to get rid of the absolute value bars at $y(t)$. To do this observe that in the derivation we required that $y(t) \neq 0$ for any t , thus y , being a continuous function, must be of a constant sign. Hence,

$$y(t) = \pm c_2 \exp(at) = c_3 \exp(at) \quad (4.7)$$

where this time c_3 can be either positive or negative.

Are these all the possible solutions to (4.6)? Solution (4.7) was derived under provision that $y \neq 0$. We clearly see that $y \equiv 0$ is a solution to (4.6) but, fortunately, this solution can be incorporated into (4.7) by allowing c_3 to be zero.

However, we still have not ruled out the possibility that the solution can cross the x -axis at one or more points. To prove that this is impossible, we must resort to the Picard

theorem. First of all we note that the function $f(t, y)$ is here given by

$$f(t, y) = ay$$

and $|f(t, y_1) - f(t, y_2)| = a|y_1 - y_2|$ so that, if f satisfies assumptions of the Picard theorem on any closed rectangle $\mathcal{R} \subset \mathbb{R}^2$ with Lipschitz constant $L = a$. If there were a solution satisfying $y(t_0) = 0$ for some t_0 , then from the uniqueness part of Picard's theorem, this solution should be identically zero, as $y(t) \equiv 0$ is a solution to this problem. In other words, if a solution to (4.6) is zero at some point, then it is identically zero.

Example 4.6 Another example of a nonuniqueness than in Example 4.4 is offered by

$$\begin{aligned} y' &= (\sin 2t)y^{1/3}, \quad t \geq 0 \\ y(0) &= 0, \end{aligned} \tag{4.8}$$

Direct substitution shows that we have at least 3 different solutions to this problem: $y_1 \equiv 0$, $y_2 = \sqrt{8/27} \sin^3 t$ and $y_3 = -\sqrt{8/27} \sin^3 t$. These are shown in the figure 4.6.

To illustrate applications of Theorem 4.3 let us consider the following two examples.

Example 4.7 The solution of the initial value problem

$$\begin{aligned} y' &= 1 + y^2, \\ y(0) &= 0, \end{aligned}$$

is given by $y(t) = \tan t$. This solution is defined only for $-\pi/2 < t < \pi/2$. Let us check this equation against the Picard Theorem. We have $f(t, y) = 1 + y^2$ and $f_y(t, y) = 2y$

Fig 2.1 Multiple solutions of the problem (4.8).

and both functions are continuous on the whole plane. Let R be the rectangle $|t| \leq a$, $|y| \leq b$, then

$$M = \max_{(t,y) \in R} |f(t,y)| = 1 + b^2,$$

and the solution exists for

$$|t| \leq \alpha = \min\left\{a, \frac{b}{1 + b^2}\right\}.$$

Since a can be arbitrary, the maximal interval of existence predicted by the Picard Theorem is the maximum of $b/(1 + b^2)$ which is equal to $1/2$.

This shows that it may happen that the Picard theorem does not give the best possible answer - that is why it is sometimes called "the local existence theorem". On the other hand, the right hand side of the equation satisfies the

assumptions of the Picard theorem everywhere and thus the solution ends its existence at finite $t = +\pi/2$ with "a bang" in accordance with Theorem 4.3.

Example 4.8 Find the solution to the following initial value problem

$$y' = -y^{-1}(1 + y^2) \sin t, \quad y(0) = 1.$$

In a standard way we obtain

$$\int_1^y \frac{r dr}{1 + r^2} = - \int_0^t \sin s ds,$$

which gives

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

Solving this equation for $y(t)$ gives

$$y(t) = \pm (2e^{-4 \sin^2 t/2} - 1)^{1/2}.$$

To determine which sign we should take we note that $y(0) = 1 > 0$, thus the solution is given by

$$y(t) = (2e^{-4 \sin^2 t/2} - 1)^{1/2}.$$

Clearly, this solution is only defined when

$$2e^{-4 \sin^2 t/2} - 1 \geq 0,$$

that is

$$e^{4 \sin^2 t/2} \leq 2.$$

Since the natural logarithm is increasing we may take logarithms of both sides preserving the direction of inequality.

Fig 2.2 The graph of the solution in Example 4.8.

We get this way

$$4 \sin^2 t/2 \leq \ln 2$$

and consequently

$$\left| \frac{t}{2} \right| \leq \sin^{-1} \frac{\sqrt{\ln 2}}{2}.$$

Therefore, the solution $y(t)$ exists only on the open interval $(-2 \sin^{-1} \frac{\sqrt{\ln 2}}{2}, 2 \sin^{-1} \frac{\sqrt{\ln 2}}{2})$. However, contrary to the previous example, the solution does not blow up at the endpoints, but simply vanishes. We note that this does not violate Theorem 4.3 as at the point the solution vanishes ($y(t) \rightarrow 0$ as $2 \sin^{-1} \frac{\sqrt{\ln 2}}{2}$) the right hand side is not Lipschitz continuous.

It is equally important to have easy to use criteria ensuring

that solutions of (4.3) are defined for all $t \in \mathbb{R}$. Theorem 4.3 combined with the Gronwall lemma (see Lemma 2.1) allows to give quite a general result of this type.

Example 4.9 We say that f appearing in (4.3) is globally Lipschitz if the constant L in (4.4) can be chosen for all $(t, y_1), (t, y_2) \in \mathbb{R}^2$. Hence, assume that f satisfies assumptions of the Picard theorem in any rectangle $\mathcal{R} \subset \mathbb{R}^2$ and is globally Lipschitz. Let $[t_0, t_{max})$ is the forward maximal interval of existence for a solution to (4.3). Then for $t \leq t_{max}$,

$$\begin{aligned}
 |y(t)| &\leq |y_0| + \int_{t_0}^t |f(s, y(s))| ds \\
 &\leq |y_0| + \int_{t_0}^t |f(s, y(s)) - f(s, y_0)| ds + \int_{t_0}^t |f(s, y_0)| ds \\
 &\leq |y_0| + \int_{t_0}^t |f(s, y_0)| ds + L \int_{t_0}^t |y(s) - y_0| ds \\
 &\leq |y_0| + \int_{t_0}^t |f(s, y_0)| ds + L \int_{t_0}^t |y_0| ds + L \int_{t_0}^t |y(s)| ds \\
 &\leq |y_0| + \int_{t_0}^t |f(s, y_0)| ds + L(t - t_0)|y_0| + L \int_{t_0}^t |y(s)| ds
 \end{aligned}$$

If $y(t)$ were not defined for all t , then by Proposition 4.3, $|y(t)|$ would become unbounded as $t \rightarrow t_{max}$ for some t_{max} .

Denoting

$$c = |y_0| + \int_{t_0}^{t_{max}} |f(s, y_0)| ds + L(t_{max} - t_0)|y_0|$$

which is finite as f is continuous for all t , we can write the above inequality as

$$|y(t)| \leq c + L \int_{t_0}^t |y(s)| ds$$

for any $t_0 \leq t \leq t_{max}$. Using now Gronwall's lemma, we obtain

$$|y(t)| \leq c \exp L(t - t_{max}) \leq c \exp L(t_{max} - t_0)$$

contradicting thus the definition of t_{max} .

4.3 Miscellaneous applications

4.3.1 Exponential growth

One of the best known applications of the exponential growth/decay equation is the so-called radiocarbon dating used for dating of samples which contain once living matter, like fossils etc. The so-called radiocarbon dating is based on the observation that the element carbon appears in nature as a mixture of stable Carbon-12 (C^{12}) and radioactive Carbon-14 (C^{14}) and the ratio between them remains constant throughout history. Thus, while an animal or a plant is living, the ratio C^{12}/C^{14} in its tissue is a known constant. When it dies, however, there is no new carbon absorbed by tissues and, since the radioactive C^{14} decays, the ratio C^{14}/C^{12} decreases as the amount of C^{14} decreases.

To be able to use the equation (2.9):

$$N' = -kN,$$

and its solution $N(t) = N(t_0)e^{-k(t-t_0)}$, where t_0 is the initial time of measurement, we must find a way to determine the value of k for C^{14} . What can be directly measured is the amount of particles which remain in a sample after some time (through the mass of the sample). The parameter which is most often used when dealing with radioactive materials is the so-called *half-life* defined as the time T after which only half of the original amount of particles remains. This is a constant depending only on the material and not on the original amount of particles or the moment in time we start to observe the sample. Indeed, by definition, the relation between k and T is found from the equation

$$N(T + t_0) = 0.5N(t_0) = N(t_0)e^{-kT}$$

that is $kT = \ln 2$ so that the solution is given by

$$N(t) = N(t_0)e^{-(t-t_0) \ln 2/T}. \quad (4.9)$$

It is clear that after time T the amount of radioactive particles in a sample halves, irrespectively of the initial amount of particles and initial time t_0 ; so that indeed the amount of particles halves after every period of length T .

To demonstrate how this formula is applied in concrete case, let us estimate the age of a of charcoal found in 2003 in a prehistoric cave in which the ratio C^{12}/C^{14} was determined to be 14.5% of its original value. The half-life of C^{14} is 5730 years.

The crucial step here is to translate the absolute numbers of C^{14} appearing in (4.9) into ratio C^{14}/C^{12} which is the

only information available from the experiment. Imagine a reference sample containing certain amount $N_{14}(t_0)$ of C^{14} and $N_{12}(t_0)$ of C^{12} at some time t_0 . Then, at time t we will have $N_{14}(t)$ of C^{14} but for C^{12} the amount does not change: $N_{12}(t) = N_{12}(t_0)$. Thus, dividing

$$N_{14}(2003) = N(t_0)e^{-(2003-t_0) \ln 2/5730}$$

by constant N_{12} we will get the equation governing the evolution of the ratio C^{14}/C^{12}

$$0.145 \frac{N_{14}(t_0)}{N_{12}(t_0)} = \frac{N_{14}(2003)}{N_{12}(2003)} = \frac{N_{14}(2003)}{N_{12}(t_0)} = \frac{N_{14}(t_0)}{N_{12}(t_0)} e^{-(2003-t_0) \ln 2/5730}$$

Thus

$$0.145 = e^{-(2003-t_0) \ln 2/5730}$$

or

$$t_0 = 2003 + \frac{5730 \ln 0.145}{\ln 2}.$$

Numerically, we obtain $t_0 = -13960$; that is, 13 960 BC.

4.3.2 *Continuous loan repayment*

Let us begin with the equation for the continuous loan repayment

$$\frac{dD}{dt} - \bar{p}D = -\rho \quad (4.10)$$

with the initial condition $D(0) = D_0$ which represents the original amount borrowed from a bank. Following the approach of Subsection A2.2.2, we can write the equation as

$$\frac{d}{dt}(De^{-\bar{p}t}) = -\rho e^{-\bar{p}t}$$

which, upon integration from 0 to t and using the initial condition, gives

$$D(t)e^{-\bar{p}t} - D_0 = \frac{\rho}{\bar{p}}(e^{-\bar{p}t} - 1)$$

hence

$$D(t) = D_0e^{\bar{p}t} + \frac{\rho}{\bar{p}}(1 - e^{\bar{p}t}).$$

As in Section 3.3, we are interested in determining the instalments for a given loan D_0 , annual interest rate p and the repayment period T . As before, we must have $D(T) = 0$; that is

$$\rho = \frac{\bar{p}D_0e^{\bar{p}T}}{e^{\bar{p}T} - 1}. \quad (4.11)$$

Let us test this formula against the numerical data used in Section 3.3. We have $D_0 = R200000$, $\bar{p} = 0.013$ and $T = 20$ (remember we use years as a unit of time and not the conversion period as in the discrete case). We get

$$\rho = \frac{0.13 \times 200000 \exp(0.13 \times 20)}{\exp(0.13 \times 20) - 1} = 28086.1.$$

We must remember that this is the annual rate of payment, thus the amount paid per month is $R = 2340.5$. As in the case of continuously compounded interest, it is slightly better than if instalments were paid on the monthly basis but clearly the much simpler formula (4.11) can be used as a good approximation of the exact one.

This numerical observation can be confirmed mathematically by noting that (4.11) is the limit of (3.4) as $\alpha \rightarrow 0$. Indeed, taking into account that the annual payment $\rho_\alpha = R/\alpha$ and $n = T/\alpha$ where T is time in years and $1/\alpha$ is the

number of payments in a year, we get

$$\lim_{\alpha \rightarrow 0} \rho_\alpha = \lim_{\alpha \rightarrow 0} \frac{\bar{p}D_0}{(1 - (1 + \alpha\bar{p})^{1/\alpha\bar{p}})^{-T\bar{p}}} = \frac{\bar{p}D_0 e^{\bar{p}T}}{1 - e^{-\bar{p}T}},$$

which, after simple algebra, becomes exactly (4.11). Moreover, the limit is monotonic, so that indeed $\rho_\alpha < \rho$ for any $\alpha > 0$.

4.3.3 *The Neo-classical model of Economic Growth*

Let us return to the Bernoulli equation (2.8) describing the evolution of the ratio of the capital to the labour $k = K/L$

$$k' = sk^\alpha - \lambda k \quad (4.12)$$

where α and λ are parameters of the models. Following the procedure outlined in Subsection A2.2.3 we define $x = k^{1-\alpha}$ which gives and substituting gives

$$x' = -(1-\alpha)\lambda x + (1-\alpha)s.$$

which is a linear equation of the form (2.7). The solution is

$$x(t) = \left(x_0 - \frac{s}{\lambda}\right) e^{-(1-\alpha)\lambda t} + \frac{s}{\lambda}$$

or, in the original variable k ,

$$k^{1-\alpha} = \left(k_0^{1-\alpha} - \frac{s}{\lambda}\right) e^{-(1-\alpha)\lambda t} + \frac{s}{\lambda}.$$

Since $\lambda > 0$ and $0 < \alpha < 1$, we see that

$$\lim_{t \rightarrow \infty} k(t) = \left(\frac{s}{\lambda}\right)^{\frac{1}{1-\alpha}}$$

so that the model is asymptotically stable.

4.3.4 Logistic equation

Let us consider the Cauchy problem for the logistic equation:

$$\begin{aligned}\frac{dN}{dt} &= R_0 N \left(1 - \frac{N}{K}\right), \\ N(t_0) &= N_0,\end{aligned}\tag{4.13}$$

where R_0 denotes the unrestricted growth rate and K the carrying capacity of the environment. Since the right-hand side of the equation does not contain t , we immediately recognize it as a separable equation. Hence, separating variables and integrating we obtain

$$\frac{K}{R_0} \int_{N_0}^N \frac{ds}{(K-s)s} = t - t_0.$$

To integrate the left-hand side we use partial fractions

$$\frac{1}{(K-s)s} = \frac{1}{K} \left(\frac{1}{s} + \frac{1}{K-s} \right)$$

which gives

$$\begin{aligned}\frac{K}{R_0} \int_{N_0}^N \frac{ds}{(K-s)s} &= \frac{1}{R_0} \int_{t_0}^t \left(\frac{1}{s} + \frac{1}{K-s} \right) ds \\ &= \frac{1}{R_0} \ln \frac{N}{N_0} \left| \frac{K-N_0}{K-N} \right|.\end{aligned}$$

Since $N = K$ and $N = 0$ are solution of the logistic equation, the Picard theorem ensures that $N(t)$ cannot reach K in any finite time, so if $N_0 < K$, then $N(t) < K$ for any t , and if $N_0 > K$, then $N(t) > K$ for all $t > 0$. Therefore

$(K - N_0)/(K - N(t))$ is always positive and

$$R_0(t - t_0) = \ln \frac{N}{N_0} \frac{K - N_0}{K - N}.$$

Exponentiating, we get

$$e^{R_0(t-t_0)} = \frac{N(t)}{N_0} \frac{K - N_0}{K - N(t)}$$

or

$$N_0(K - N(t))e^{R_0(t-t_0)} = N(t)(K - N_0).$$

Bringing all the terms involving N to the left-hand side and multiplying by -1 we get

$$N(t) \left(N_0 e^{R_0(t-t_0)} + K - N_0 \right) = N_0 K e^{R_0(t-t_0)},$$

thus finally

$$N(t) = \frac{N_0 K}{N_0 + (K - N_0) e^{-R_0(t-t_0)}}. \quad (4.14)$$

Let us examine (4.14) to see what kind of population behaviour it predicts. First observe that we have

$$\lim_{t \rightarrow \infty} N(t) = K,$$

hence our model correctly reflects the initial assumption that K is the maximal capacity of the habitat. Next, we obtain

$$\frac{dN}{dt} = \frac{R_0 N_0 K (K - N_0) e^{-R_0(t-t_0)}}{(N_0 + (K - N_0) e^{-R_0(t-t_0)})^2}$$

thus, if $N_0 < K$, the population monotonically increases, whereas if we start with the population which is larger than

the capacity of the habitat, then such a population will decrease until it reaches K . Also

$$\frac{d^2N}{dt^2} = R_0 \frac{d}{dt}(N(K-N)) = N'(K-2N) = N(K-N)(K-2N)$$

from which it follows that, if we start from $N_0 < K$, then the population curve is convex down for $N < K/2$ and convex up for $N > K/2$. Thus, as long as the population is small (less than half of the capacity), then the rate of growth increases, whereas for larger population the rate of growth decreases. This results in the famous *logistic* or *S-shaped* curve which is presented below for particular values of parameters $R_0 = 0.02, K = 10$ and $t_0 = 0$ resulting in the following function:

$$N(t) = \frac{10N_0}{N_0 + (10 - N_0)e^{-0.2t}}.$$

To show how this curve compare with the real data and with the exponential growth we take the experimental coefficients $K = 10.76$ billion and $R_0 = 0.029$. Then the logistic equation for the growth of the Earth population will read

$$N(t) = \frac{N_0(10.76 \times 10^9)}{N_0 + ((10.76 \times 10^9) - N_0)e^{-0.029(t-t_0)}}.$$

We use this function with the value $N_0 = 3.34 \times 10^9$ at $t_0 = 1965$. The comparison is shown on Fig. 4.2.

4.3.5 The waste disposal problem

Let us recall that the motion of a drum of waste dumped into the sea is governed by the equation (2.19)

$$\frac{d^2y}{dt^2} = \frac{1}{m} \left(W - B - c \frac{dy}{dt} \right). \quad (4.15)$$

Fig. 4.1. Logistic curves with $N_0 < K$ (dashed line) and $N_0 > K$ (solid line) for $K = 10$ and $R_0 = 0.02$

The drums are dropped into the 100m deep sea. Experiments show that the drum could brake if its velocity exceeds 12m/s at the moment of impact. Thus, our aim is to determine the velocity of the drum at the sea bed level. To obtain numerical results, the mass of the drum is taken to be 239 kg, while its volume is 0.208 m³. The density of the sea water is 1021 kg/m³ and the drag coefficient is experimentally found to be $c = 1.18\text{kg/s}$. Thus, the mass of water displaced by the drum is 212.4 kg.

Equation (4.15) can be re-written as the first order equation for the velocity $V = dy/dt$.

$$V' + \frac{c}{m}V = g - \frac{B}{m}. \quad (4.16)$$

Since the drum is simply dumped into the sea, its initial

Fig. 4.2. Human population on Earth. Comparison of observational data (points), exponential growth (solid line) and logistic growth (dashed line)

velocity $V(0) = 0$. Since (4.16) is a linear equation, we find the integration factor $\mu(t) = e^{tc/m}$ and the general solution of the full equation is obtained as

$$V(t) = e^{-tc/m} \left(g - \frac{B}{m} \right) \int e^{tc/m} dt = \frac{mg - B}{c} (1 + Ce^{-tc/m})$$

for some constant C . Using the initial condition $V(0) = 0$, we find $C = -1$ so that

$$V(t) = \frac{mg - B}{c} (1 - e^{-tc/m}). \quad (4.17)$$

Integrating once again, we find

$$y(t) = \frac{mg - B}{c} \left(t + \frac{m}{c} e^{-tc/m} \right) + C_1.$$

To determine C_1 we recall that the coordinate system was set up in such a way that $y = 0$ was at the sea surface so we can take the initial condition to be $y(0) = 0$. Thus we obtain the equation

$$0 = y(0) = \frac{mg - B}{c} \frac{m}{c} + C_1,$$

so that

$$y(t) = \frac{mg - B}{c} \left(t + \frac{m}{c} e^{-tc/m} \right) - \frac{m(mg - B)}{c^2}. \quad (4.18)$$

Equation (4.17) expresses the velocity of the drum as a function of time t . To determine the impact velocity, we must compute the velocity at time t at which the drum hits the ocean floor, that is we have to solve for t the equation (4.18) with $y(t) = 100\text{m}$. Explicit solution of this equation is obviously impossible so let us try some other method.

As a first attempt, we notice from (4.17) that $V(t)$ is an increasing function of time and that it tends to a finite limit as $t \rightarrow \infty$. This limit is called the terminal velocity and is given by

$$V_T = \frac{mg - B}{c}. \quad (4.19)$$

Thus, for any time t the velocity is smaller than V_T and if $V_T < 12\text{m/s}$, we can be sure that the velocity of the drum when it hits the sea floor is also smaller than 12 m/s and it will not crack upon the impact. Substituting the data to (4.19) we obtain

$$V_T = \frac{(239 - 212.4)9.81}{1.18} \approx 221\text{m/s},$$

which is clearly way too large.

However, the approximation that gave the above figure is

far too crude - this is the velocity the drum would eventually reach if it was allowed to descend indefinitely. As this is clearly not the case, we have to find the way to express the velocity as a function of the position y (see Subsection A2.2.5). This velocity, denoted by $v(y)$, is very different from $V(t)$ but they are related through

$$V(t) = v(y(t)).$$

By the chain rule of differentiation

$$\frac{dV}{dt} = \frac{dv}{dy} \frac{dy}{dt} = V \frac{dv}{dy} = v \frac{dv}{dy}.$$

Substituting this into (4.16) we obtain

$$mv \frac{dv}{dy} = (mg - B - cv). \quad (4.20)$$

We have to supplement this equation with an appropriate initial condition. For this we have

$$v(0) = v(y(0)) = V(0) = 0.$$

This is a separable equation which we can solve explicitly. Firstly, we note that since $v < V_T = (mg - B)/c$, $mg - B - cv > 0$ all the time. Thus, we can divide both sides of (4.20) by $mg - B - cv$ and integrate, getting

$$\int_0^v \frac{r dr}{mg - B - cr} = \frac{1}{m} \int_0^y ds = \frac{y}{m}.$$

To find the left-hand side integral, we note that the degree of the numerator is the same as the degree of the denomi-

nator so that we have to decompose

$$\begin{aligned} \frac{r}{mg - B - cr} &= -\frac{1}{c} \frac{-cr}{mg - B - cr} \\ &= -\frac{1}{c} \frac{-mg + B + mg - Bcr}{mg - B - cr} = -\frac{1}{c} \left(\frac{-mg + B}{mg - B - cr} + 1 \right). \end{aligned}$$

Thus

$$\begin{aligned} \int_0^v \frac{r dr}{mg - B - cr} &= -\frac{1}{c} \int_0^v dr + \frac{mg - B}{c} \int_0^v \frac{dr}{mg - B - cr} \\ &= -\frac{v}{c} - \frac{mg - B}{c^2} \ln \frac{mg - B - cv}{mg - B}, \end{aligned}$$

and we obtain the solution

$$\frac{y}{m} = -\frac{v}{c} - \frac{mg - B}{c^2} \ln \frac{mg - B - cv}{mg - B}. \quad (4.21)$$

It seems that the situation here is as hopeless as before as we have $y = y(v)$ and we cannot find $v(y)$ explicitly. However, at least we have a direct relation between the quantities of interest, and not through intermediate parameter t that is irrelevant for the problem, as before. Thus, we can easily graph y as a function of v and estimate $v(100)$ from the graph shown at the Fig. 4.3. We can also answer the question whether the velocity at $y = 100\text{m}$ is higher than the critical velocity $v = 12\text{m/s}$. To do this, we note that from (4.20) and the fact that $v < V_T$ we can infer that v is an increasing function of y . Let us find what y corresponds to $v = 12\text{m/s}$. Using the numerical data, we obtain

$$y(12) \approx 68.4\text{m},$$

that is, the drum will reach the velocity of 12m/s already at the depth of 68.4m . Since v is a strictly increasing function

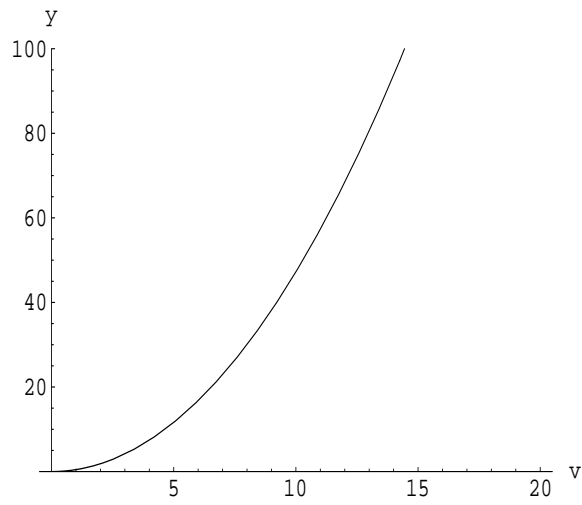


Fig. 4.3. The depth as a function of velocity of the drum

of y , the velocity at 100m will be much higher and therefore the drum could crack on impact.

4.3.6 The satellite dish

In Subsection 2.5.1 we obtained the equation (2.20) for a reflecting surface:

$$\frac{dy}{dx} = \frac{y}{\sqrt{x^2 + y^2} + x}. \quad (4.22)$$

Now we shall solve this equation. We observe that the right-hand side can be written as

$$\frac{\frac{y}{x}}{\sqrt{1 + \left(\frac{y}{x}\right)^2 + 1}},$$

for $x > 0$. This suggests the substitution used for homogeneous equations $z = y/x$. Since $y' = z'x + z$, we obtain

$$z'x\sqrt{1+z^2} + z'x + z\sqrt{1+z^2} + z = z,$$

which, after a simplification, can be written as

$$z' \left(\frac{1}{z} + \frac{1}{z\sqrt{z^2+1}} \right) = -\frac{1}{x}.$$

Integrating and using $z, x > 0$ we obtain

$$\ln z + \int \frac{dz}{z\sqrt{1+z^2}} = -\ln x + C'. \quad (4.23)$$

There are several ways to integrate the second term. We use the hyperbolic substitution but first we simplify it:

$$\int \frac{dz}{z\sqrt{1+z^2}} = \int \frac{dz}{z^2\sqrt{1+z^{-2}}} = -\int \frac{du}{\sqrt{1+u^2}}$$

where $u = z^{-1}$. Then, taking $u = \sinh \xi$ gives $du = \cosh \xi d\xi$ and $\sqrt{1+z^2} = \sqrt{1+\sinh^2 \xi} = \sqrt{\cosh^2 \xi} = \cosh \xi$, as $\cosh \xi$ is always positive. Thus we obtain

$$\int \frac{du}{\sqrt{1+u^2}} = \int d\xi = \xi,$$

where we skipped the constant of integration as it already appears in (4.23). Then

$$u = \frac{e^\xi - e^{-\xi}}{2}$$

and, denoting $t = e^\xi$ we transform the above into a quadratic equation:

$$t^2 - 2tu - 1 = 0$$

with solutions

$$t_{1,2} = u \pm \sqrt{u^2 + 1}$$

Since $e^\xi > 0$, we must take the positive solution which gives

$$e^\xi = u + \sqrt{u^2 + 1} = \frac{1 + \sqrt{z^2 + 1}}{z}$$

and

$$\xi = -\ln z + \ln(1 + \sqrt{z^2 + 1});$$

that is,

$$\int \frac{dz}{z\sqrt{1+z^2}} = \ln z - \ln(1 + \sqrt{z^2 + 1})$$

up to an additive constant. Returning to (4.23) we obtain

$$\ln \frac{z^2}{1 + \sqrt{z^2 + 1}} = -\ln x/C$$

for some constant $C > 0$. Thus

$$\frac{z^2}{1 + \sqrt{z^2 + 1}} = \frac{C}{x},$$

and, returning to the original unknown function $z = y/x$,

$$\frac{y^2}{x + \sqrt{y^2 + x^2}} = C,$$

which, after some algebra, gives

$$y^2 - 2Cx = C^2. \quad (4.24)$$

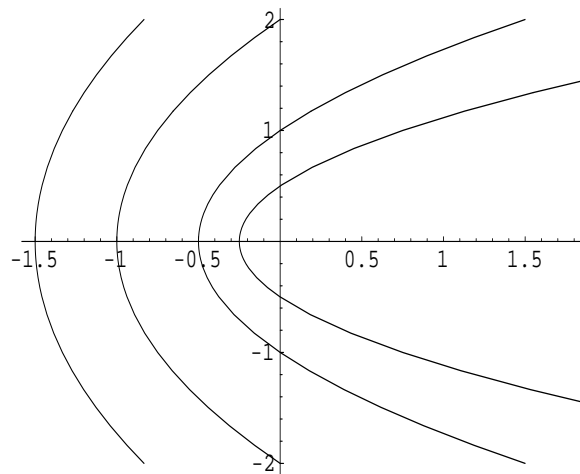


Fig. 4.4. Different shapes of parabolic curves corresponding to various values of the constant C . In each case the focus is at the origin.

This is an equation of the parabola with the vertex at $x = -C/2$ and with focus at the origin.

We note that this equation was obtained under the assumption that $x > 0$ so, in fact, we do not have the full parabola at this moment. The assumption $x > 0$ was, however, purely technical, and all calculations above, with only minor changes, can be repeated for $x < 0$. Another way of showing that (4.24) is valid for $-C/2 < x < 0$ (and also for $y < 0$) is by direct substitution. In fact, $y = \pm\sqrt{2Cx + C^2}$ so that

$$LHS = \frac{dy}{dx} = \pm \frac{C}{\sqrt{2Cx + C^2}},$$

and

$$\begin{aligned} RHS &= \frac{y}{\sqrt{y^2 + x^2 + x}} = \frac{\pm\sqrt{2Cx + C^2}}{\sqrt{x^2 + 2Cx + C^2 + x}} \\ &= \frac{\pm\sqrt{2Cx + C^2}}{\sqrt{(x + C)^2 + x}} = \pm \frac{C}{\sqrt{2Cx + C^2}}, \end{aligned}$$

where we used the fact that $x \geq -C/2$ so that $x + C > 0$. Thus LHS = RHS for any $x \geq -C/2$ and (4.24) gives the solution to the equation in the whole range of independent variables.

4.3.7 Pursuit equation

In this paragraph we shall provide the solution to the pursuit equation

$$xy'' = -\frac{v}{u}\sqrt{1 + (y')^2}. \quad (4.25)$$

Firstly, we observe that this a second order equation that does not contain the unknown function but only its higher derivatives. Thus, following the approach of Subsection A2.2.5 we introduce the new unknown $z = y'$ reducing (4.25) to a first order equation:

$$xz' = -k\sqrt{1 + z^2}$$

where we denoted $k = v/u$. This is a separable equation with non-vanishing right-hand side, so that we do not have stationary solutions. Separating variables and integrating, we obtain

$$\int \frac{dz}{\sqrt{1 + z^2}} = -k \ln(-C'x)$$

for some constant $C' > 0$, where we used the fact that in the model $x < 0$. Integration (for example as in the previous paragraph) gives

$$\ln(z + \sqrt{z^2 + 1}) = \ln C(-x)^{-k},$$

with $C = (C')^{-k}$, hence

$$z + \sqrt{z^2 + 1} = C(-x)^{-k},$$

from where, after some algebra,

$$z = \frac{1}{2} \left(C(-x)^{-k} - \frac{1}{C}(-x)^k \right). \quad (4.26)$$

Returning to the original unknown function y , where $y' = z$, and integrating the above equation, we find

$$y(x) = \frac{1}{2} \left(\frac{1}{C(k+1)}(-x)^{k+1} - \frac{C}{(1-k)}(-x)^{-k+1} \right) + C_1.$$

Let us express the constants C_1 and C through initial conditions. We assume that the pursuer started from the position $(x_0, 0)$, $x_0 < 0$ and that at the initial moment the target was at the origin $(0, 0)$. Using the principle of the pursuit, we see that the initial direction was along the x -axis, that is, we obtain the initial conditions in the form

$$y(x_0) = 0, \quad y'(x_0) = 0.$$

Since $y' = z$, substituting $z = 0$ and $x = x_0$ in (4.26), we obtain

$$0 = y'(x_0) = z(x_0) = C(-x_0)^{-k} - \frac{1}{C}(-x_0)^k$$

which gives

$$C = (-x_0)^k,$$

so that

$$y(x) = -\frac{x_0}{2} \left(\frac{1}{k+1} \left(\frac{x}{x_0} \right)^{k+1} - \frac{1}{1-k} \left(\frac{x}{x_0} \right)^{-k+1} \right) + C_1.$$

To determine C_1 we substitute $x = x_0$ and $y(x_0) = 0$ above getting

$$0 = -\frac{x_0}{2} \left(\frac{1}{k+1} + \frac{1}{k-1} \right) + C_1$$

thus

$$C_1 = \frac{kx_0}{k^2 - 1}.$$

Finally,

$$y(x) = -\frac{x_0}{2} \left(\frac{1}{k+1} \left(\frac{x}{x_0} \right)^{k+1} - \frac{1}{1-k} \left(\frac{x}{x_0} \right)^{-k+1} \right) + \frac{kx_0}{k^2 - 1}.$$

This formula can be used to obtain two important pieces of information: the time and the point of interception. The interception occurs when $x = 0$. Thus

$$y(0) = \frac{kx_0}{k^2 - 1} = \frac{vux_0}{v^2 - u^2}.$$

Since $x_0 < 0$ and the point of interception must be on the upper semi-axis, we see that for the interception to occur, the speed of the target v must be smaller than the speed of the pursuer u . This is of course clear from the model, as the pursuer moves along a curve and has a longer distance to cover.

The duration of the pursuit can be calculated by noting that the target moves with a constant speed v along the y axis from the origin to the interception point $(0, y(0))$ so

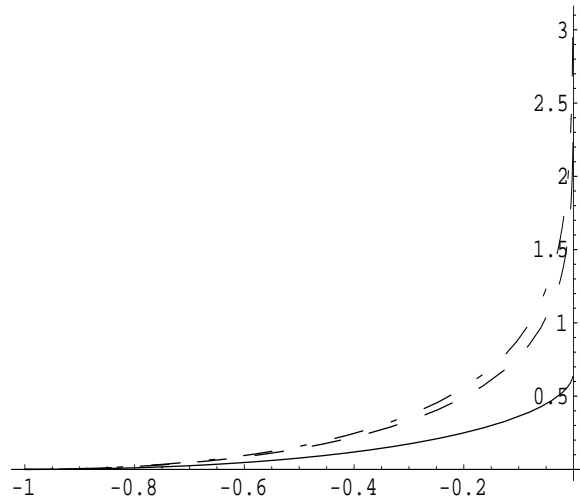


Fig. 4.5. Pursuit curve for different values of k . $k = 0.5$ (solid line), $k = 0.9$ (dashed line), $k = 0.99$ (dot-dashed line).

that

$$T = \frac{y(0)}{v} = \frac{ux_0}{v^2 - u^2}.$$

4.3.8 *Escape velocity*

The equation of motion of an object of mass m projected upward from the surface of a planet was derived at the end of Subsection 2.4. The related Cauchy problem reads

$$\begin{aligned} m \frac{d^2 y}{dt^2} &= -\frac{mgR^2}{(y+R)^2} - c(y) \left(\frac{dy}{dt} \right)^2 \\ y(0) &= R, \quad y'(0) = v_0, \end{aligned}$$

where the initial conditions tell us that the missile was shot from the surface with the initial velocity v_0 and we allow the air resistance coefficient to change with height. Rather than solve the full Cauchy problem, we shall address the question of the existence of the *escape velocity*; that is, whether there exists an initial velocity which would allow the object to escape from planet's gravitational field.

The equation is of the form (2.20); that is, it does not contain explicitly the independent variable. To simplify calculations, first we shall change the unknown function according to $z = y + R$ (so that z is the distance from the centre of the planet) and next introduce $F(z) = z'$ so that $z'' = F_z F$, see (2.21). Then the equation of motion will take the form

$$F_z F + C(z)F^2 = -\frac{gR^2}{z^2}, \quad (4.27)$$

where $C(z) = c(z - R)/m$. Noting that

$$F_z F = \frac{1}{2} \frac{d}{dz} F^2$$

and denoting $F^2 = G$ we reduce (4.27) to the linear differential equation

$$G_z + 2C(z)G = -\frac{2gR^2}{z^2}. \quad (4.28)$$

We shall consider three forms for C .

Case 1. $C(z) \equiv 0$ (airless moon).

In this case (4.28) becomes

$$G_z = -\frac{2gR^2}{z^2}.$$

which can be immediately integrated from R to z giving

$$G(z) - G(R) = 2gR^2 \left(\frac{1}{z} - \frac{1}{R} \right).$$

Returning to the old variables $G(z) = F^2(z) = v^2(z)$, where v is the velocity of the missile at the distance z from the centre of the moon, we can write

$$v^2(z) - v^2(R) = 2gR^2 \left(\frac{1}{z} - \frac{1}{R} \right).$$

The missile will escape from the moon if its speed remains positive for all times – if it stops at any finite z , then the gravity pull will bring it back to the moon. Since $v(z)$ is decreasing, its minimum value will be the limit at infinity so that, passing with $z \rightarrow \infty$, we must have

$$v^2(R) \geq 2gR$$

and the escape velocity is

$$v(R) = \sqrt{2gR}.$$

Case 2. Constant air resistance.

If we are back on Earth, it is not reasonable to assume that there is no air resistance during motion. Let us investigate the next simple case with $c = \text{constant}$. Then we have

$$G_z + 2CG = -\frac{2gR^2}{z^2}, \quad (4.29)$$

where $C = c/m$. The integrating factor equals e^{2cz} so that we obtain

$$\frac{d}{dz} (e^{2cz} G(z)) = -2gR^2 \frac{e^{2Cz}}{z^2},$$

and, upon integration,

$$e^{2Cz}v^2(z) - e^{2CR}v_0^2 = -2gR^2 \int_R^z e^{2Cs} s^{-2} ds,$$

or

$$v^2(z) = e^{-2Cz} \left(e^{2CR}v_0^2 - 2gR^2 \int_R^z e^{2Cs} s^{-2} ds \right). \quad (4.30)$$

Consider the integral

$$I(z) = \int_R^z e^{2Cs} s^{-2} ds.$$

Since $\lim_{s \rightarrow \infty} e^{2Cs} s^{-2} = \infty$, we have also

$$\lim_{s \rightarrow \infty} \int_R^z e^{2Cs} s^{-2} ds = \infty.$$

Since $\int_R^R e^{2Cs} s^{-2} ds = 0$ and because $e^{2CR}v^2(R)$ is independent of z , from the Darboux theorem we see that, no matter what the value of v_0 is, for some $z_0 \in [R, \infty)$ the right-hand side of (4.30) becomes 0 and thus $v^2(z_0) = 0$. Thus, there is no initial velocity v_0 for which the missile will escape the planet.

Case 3. Variable air resistance.

By passing from no air resistance at all ($c = 0$) to a constant air resistance we definitely overshoot since the air becomes thinner with height and thus its resistance decreases. Let

us consider one more case with $C(z) = k/z$ where k is a proportionality constant. Then we obtain

$$G_z + \frac{2k}{z}G = -\frac{2gR^2}{z^2}. \quad (4.31)$$

The integrating factor equals z^{2k} so that we obtain

$$\frac{d}{dz} (z^{2k}G(z)) = -2gR^2 z^{2k-2},$$

and, upon integration,

$$z^{2k}v^2(z) - R^{2k}v_0^2 = -2gR^2 \int_R^z s^{2k-2} ds.$$

Using the same argument, we see that the escape velocity will exist if and only if

$$\lim_{z \rightarrow \infty} \int_R^z s^{2k-2} ds < +\infty$$

and from the properties of improper integral we infer that we must have $2k - 2 < -1$ or

$$k < \frac{1}{2}.$$

Of course, from physics $k \geq 0$. Thus, the escape velocity is given by

$$v_0 = \sqrt{\frac{2gR}{1-2k}}.$$

4.4 Exercises

- (i) Show that if the function f in (4.3) satisfies assumptions of the Picard theorem on any rectangle $\mathcal{R} \subset \mathbb{R}^2$

additionally satisfies $|f(t, y)| \leq K$ on \mathbb{R}^2 then any solution to (4.3) exists for all $t \in \mathbb{R}$.

5

Qualitative theory for a single equation

In most cases it is impossible to find an explicit solution to a given differential equation. However, one can often deduce properties of solution and answer some relevant questions by analyzing the right-hand side of the equation. Let us first illustrate this idea on few examples.

5.1 Direct application of difference and differential equations

5.1.1 Sustainable harevesting

Let us consider a population of fish living in a pond, which grows according to the logistic equation (1.24)

$$N(k+1) = N(k) + R_0 N(k) \left(1 - \frac{N(k)}{K}\right).$$

We know that this equation only can be solved in some particular cases. However, even without solving it we can draw a conclusion of some importance for fisheries.

The basic idea of a sustainable economy is to find an

optimal level of fishing: too much harvesting would deplete the fish population beyond recovery and too little would provide insufficient return from the industry. To maintain the population at a constant level, only the increase in the population should be harvested during any one season. In other words, the harvest should be $H(k) = N(k+1) - N(k)$. Using the equation, we find

$$H(k) = R_0 N(k) \left(1 - \frac{N(k)}{K} \right).$$

Hence, to maximize the harvest at each k the population should be kept at the size $N(k) = N$ for which the right hand side is the absolute maximum. We note that the right hand side is a quadratic function:

$$f(N) = R_0 N \left(1 - \frac{N}{K} \right)$$

and it is easy to find that the maximum is attained at $N = K/2$, that is, the population should be kept at around half of the carrying capacity of the environment. Thus, the maximum sustainable yield is $H = R_0 K/4$.

5.1.2 Maximal concentration of a drug in an organ

In Section 3.2 we introduced a model describing the concentration of a drug in the bloodstream when the drug is injected at discrete time intervals. If the drug is fed externally intravenously at a constant rate a then a usual modelling procedure leads to the following equation for the concentration c

$$u' = -\gamma u + a, \quad (5.1)$$

where, as before, γ is the rate of the drug's absorption. Let the initial concentration is $c(0) = c_0$. Though it is easy to solve this equation, we present its qualitative analysis providing quick answers to several natural questions. The employed techniques can be used for more complicated problems and the answers can be tested against the known solution.

Let us find the maximum concentration of the drug if the original concentration satisfies $c_0 < a/\gamma$. First, note that $u = a/\gamma$ is a solution to (5.1) (an equilibrium solution). Thus, by the Picard theorem, any solution either stays below or above the line $u = a/\gamma$. But if $u(t) < a/\gamma$, then $u'(t) > 0$ (and $u(t) > a/\gamma$, then $u'(t) < 0$). Summarizing, if $c_0 < a/\gamma$, then $u(t)$ is increasing and $u(t) < a/\gamma$ for all $t \geq 0$. Thus, there is a limit $\lim_{t \rightarrow \infty} u(t) = u_\infty \leq a/\gamma$. Using the Mean Value Theorem for any t and some fixed h we can write

$$u(t+h) - u(t) = u'(t + \theta_t h)h = -\gamma u(t + \theta_t h)h + ah$$

for some $0 \leq \theta_t \leq 1$. Passing to the limit as $t \rightarrow \infty$, we get $u_\infty = \gamma/a$ so that the (asymptotic) maximum concentration of the drug is γ/a .

We note that the argument used above is a particular case of reasoning employed in the proof of Theorem 6.1.

5.1.3 A nonlinear pendulum

Consider the equation

$$u'' + \sin u = 0 \tag{5.2}$$

which is the 'non-linearized' version of the well-known harmonic oscillator equation $u'' + u = 0$. Suppose $0 < u(0) < \pi$

and $u'(0) < 0$. We shall prove that u steadily decreases until it reaches 0. First, since $u'(0) < 0$ and, as a solution of a second order equation, $u'(t)$ is continuous, we have $u'(t) < 0$ on $[0, T)$ for some $T > 0$. If $u(t)$ stops decreasing without reaching 0, then it must attain a local minimum at some t_0 satisfying $0 < u(t_0) < \pi$ and $u'(t_0) = 0$. But at this point $-\sin u(t_0) > 0$ so $u''(t_0) > 0$ and hence u has a local maximum at $t = t_0$. This contradiction proves the thesis.

In the remaining part of the chapter we shall formalize and generalize the ideas used in the previous three examples.

5.2 Equilibria of first order equations

Most equations cannot be solved in closed form. However, as we have seen above, a number of pertinent questions can be answered by a careful analysis of the equation itself, without solving it. One of the typical problems is to determine whether the system is stable; that is, whether if we allow it to run for sufficiently long time (which, in the case of difference equations, means many iterations) it eventually will settle at some state, which clearly should be an equilibrium.

To explain, in both cases of difference and differential equation, by an equilibrium point, or an equilibrium solution, we understand a solution which does not change; that is, which is constant with respect to the independent variable. Since, however, in the differential equation

$$x' = f(x) \tag{5.3}$$

the right hand side describes the rate of change of a given quantity whereas in the difference equation

$$x(n + 1) = f(x(n)) \quad (5.4)$$

the right hand side gives the amount of the quantity in the state $n + 1$ in relation to the amount present in the previous state, the theories are different and will be treated in separate subsections.

We also note that, since finding equilibria is considerably easier than solving the equation, knowing that the system will converge to a particular equilibrium allows to regard the latter as an approximate solution.

We start with presenting a stability theory for differential equations as it is much simpler than in the case of difference equations.

5.2.1 *Equilibria for differential equations*

In this section we shall focus on autonomous differential equations (5.3). We recall that the word autonomous refers to the fact that f does not explicitly depend on time. To fix attention we shall assume that f is an everywhere defined function satisfying assumptions of the Picard theorem on the whole real line.

By $x(t, t_0, x_0)$ we denote the flow of (5.3) which is the solution of the Cauchy problem

$$\frac{d}{dt}x(t, t_0, x_0) = f(x(t, t_0, x_0)), \quad x(t_0, t_0, x_0) = x_0.$$

In most cases we take $t_0 = 0$ and then we shall write $x(t, 0, x_0) = x(t, x_0)$. $y(t, y_0)$.

Let us specify the notion the *equilibrium point* or the

stationary solution in the present context. We note that if (5.3) has a solution that is constant in time, $x(t) \equiv x_0$, called a stationary solution, then such a solution satisfies $x'(t) = 0$ and consequently

$$f(x_0) = 0. \quad (5.5)$$

Conversely, if the equation $f(x) = 0$ has a solution, which we call an equilibrium point, then, since f is independent of time, such a solution is a number, say x_0 . If we now consider a function defined by $x(t) = x_0$, then $x'(t) \equiv 0$ and consequently

$$0 = x'(t) = x'_0 = f(x_0)$$

and such a solution is a stationary solution. Summarizing, equilibrium points are solutions to the algebraic equation (5.5) and, treated as constant functions, they are (the only) stationary solutions to (5.3).

Equilibrium points play another important rôle for differential equations – they are the only limiting points of bounded solutions as $t \rightarrow \infty$. Let us first note the following lemma.

Lemma 5.1 *If x_0 is not an equilibrium point of (5.3), then $y(t, x_0)$ is never equal to equilibrium point. In other words,*

$$f(x(t, x_0)) \neq 0$$

for any t for which the solution exists.

Proof. An equilibrium point x^* generates a time independent solution given by $x(t) = x^*$. Thus, if $x(t_1, x_0) = x^*$ for some t_1 , then (t_1, x_0) belongs to two different solutions which contradicts the uniqueness which is ensured by the

assumption that at each point of the plane the assumptions of the Picard theorem are satisfied. ■

From the above lemma it follows that if f has several equilibrium points, then the stationary solutions corresponding to these points divide the (t, x) plane into horizontal strips such that any solution remains always confined to one of them. We shall formulate and prove a theorem that strengthens this observation.

Theorem 5.1 *Let $x(t, x_0)$ be a non-stationary solution of (5.3) with $x_0 \in \mathbb{R}$ and $I_{max} = (t_-, t_+)$ be its maximal interval of existence. Then $x(t, x_0)$ is either a strictly decreasing or a strictly increasing function of t . Moreover $x(t, x_0)$ either diverges to $+\infty$ or $-\infty$ or converges to an equilibrium point, as $t \rightarrow t_{\pm}$. In the latter case $t_{\pm} = \pm\infty$.*

Proof. Assume that for some $t_* \in I_{max}$ the solution $x(t) := x(t, x_0)$ has a local maximum or minimum $x_* = x(t_*)$. Since $x(t)$ is differentiable, we must have $x'(t_*) = 0$ but then $f(x_*) = 0$ which makes x_* the equilibrium point of f . This means that a non-stationary solution $x(t)$ reaches an equilibrium in finite time which contradicts Lemma 5.1. Thus, if $x(t)$ is not a stationary solution, then it cannot attain local maxima or minima and thus must be either strictly increasing or strictly decreasing.

Since the solution is monotonic it either diverges to $\pm\infty$ (depending on whether it decreases or increases) or converges to finite limits as $t \rightarrow t_{\pm}$. Let us focus on the right end point t_+ of I_{max} . If $x(t)$ converges as $t \rightarrow t_+$ then $t_+ = \infty$ by Theorem 4.3. Thus

$$\lim_{t \rightarrow \infty} x(t, x_0) = \bar{x}.$$

Without compromising generality, we assume further that x is an increasing function. If \bar{x} is not an equilibrium point, then from continuity (the Darboux property) the values of $x(t)$ must fill the interval $[x_0, \bar{x})$ and this interval cannot contain any equilibrium point as the existence of such would violate the Picard theorem. Thus, for any $\tilde{x} \leq \bar{x}$, $f(\tilde{x})$ is strictly positive and, integrating the equation, we obtain

$$t(x) - t(x_0) = \int_{x_0}^x \frac{ds}{f(s)}. \quad (5.6)$$

Passing with t to infinity (since $t(\bar{x}) = \infty$), we see that the left-hand side becomes infinite and so

$$\int_{x_0}^{\bar{x}} \frac{ds}{f(s)} = \infty.$$

By assumption, the interval of integration is finite so that the only way the integral could become infinite is if $1/f(\bar{s}) = \infty$ or $f(\bar{s}) = 0$ for some $s \in [x_0, \bar{x}]$. The only point which can have this property is $s = \bar{x}$, thus \bar{x} is an equilibrium point. ■

Remark 5.1 *The proof that the only finite limit point of a solution is an equilibrium can be carried out also as in Subsection 5.1.2. However, Eq. (5.6) is of independent value as it gives a formula for the ‘blow-up’ time of the solution $x(t, x_0)$. To wit, let the interval $[x_0, \infty)$ be free of equilibria and that $x(t, x_0)$ is increasing for $t > 0$. Then*

$\lim_{t \rightarrow t_+} x(t, x_0) = \infty$ so that, by (5.6)

$$t_+ - t(x_0) = \int_{x_0}^{\infty} \frac{ds}{f(s)}$$

and, in particular, we see that if $1/f$ is integrable at $+\infty$, then the maximal existence time is finite and we have the so-called blow-up of the solution in finite time. On the other hand, if $1/f$ is not integrable, then $t_{max} = +\infty$. We note that the latter occurs if $f(s)$ grows not faster than z as $s \rightarrow \infty$ giving thus another proof of the result of Example 4.9. If $f(s)$ behaves, say, as s^2 as $s \rightarrow \infty$, then the integral of the right hand side is finite and thus $t_{max} < \infty$ – we have seen this in Example 4.7.

Remark 5.2 *It is important to emphasize that the assumption that f satisfies assumptions of the Picard theorem everywhere on \mathbb{R}^2 is crucial. If there are non-Lipschitzian points, then the behaviour of solutions close to such points is not covered by Theorem 6.1 as we have seen in Example 4.8.*

Let us summarize the possible scenarios for an autonomous equation (5.3). Assume that y_* is a single equilibrium point of f with $f(y) < 0$ for $y < y_*$ and $f(y) > 0$ for $y > y_*$. If the initial condition satisfies $y_0 < y_*$, then the solution $y(t, y_0)$ decreases so it diverges either to $-\infty$ or to an equilibrium point. Since there is no equilibrium point smaller than y_0 , the solution must diverge to $-\infty$. Similarly, for $y_0 > y_*$ we see that $y(t, y_0)$ must diverge to infinity. Conversely, assuming that y_* is a single stationary point of f with $f(y) > 0$ for $y < y_*$ and $f(y) < 0$ for $y > y_*$, we

see that if $y_0 < y_*$, then the solution $y(t, y_0)$ increases so it converges to y_* . Similarly, for $y_0 > y_*$, we see that $y(t, y_0)$ must decrease converging again to y_* . If there are more than one equilibrium point, then the behaviour of the solution is a combination of the above scenarios. Assume, for example, that f has two equilibrium points $y_1 < y_2$ and is positive for $y < y_1$, negative for $y_1 < y < y_2$ and again positive for $y > y_2$. Thus, for $y_0 < y_1$, $y(t, y_0)$ increases converging to y_1 , for $y_1 < y_0 < y_2$ we have $y(t, y_0)$ decreasing and converging to y_1 and, finally, for $y_0 > y_2$, $y(t, y_0)$ increases to infinity.

Example 5.1 *Let us consider the Cauchy problem for the logistic equation*

$$y' = y(1 - y), \quad u(0) = t_0. \quad (5.7)$$

We have solved this problem in Section 4.3. Let us now get as many information as possible about the solutions to this problem without actually solving it. Firstly, we observe that the right-hand side is given by $f(y) = y(1 - y)$ which is a polynomial and therefore at each point of \mathbb{R}^2 the assumptions of Picard's theorem are satisfied, that is, through each point (t_0, y_0) there passes only one solution of (5.7). However, f is not a globally Lipschitz function so that this solutions may be defined only locally, on small time intervals.

The second step is to determine equilibrium points and stationary solutions. From

$$y(1 - y) = 0.$$

we see that $y \equiv 0$ and $y \equiv 1$ are the only equilibrium solutions. Moreover, $f(y) < 0$ for $y < 0$ and $y > 1$ and

$f(y) > 0$ for $0 < y < 1$. From Picard's theorem (uniqueness) it follows then that solutions starting from $y_0 < 0$ will stay strictly negative, starting from $0 < y_0 < 1$ will stay in this interval and, finally those with $y_0 > 1$ will be larger than 1, for all times of their respective existence, as they cannot cross equilibrium solutions. Then, from Theorem 6.1, we see that the solutions with negative initial condition are decreasing and therefore tend to $-\infty$ for increasing times (in fact, they blow-up (become infinite) for finite times) as integrating the equation, we obtain

$$t(y) = \int_{y_0}^y \frac{d\eta}{\eta(1-\eta)}$$

and we see that passing with y to $-\infty$ on the right-hand side we obtain a finite number (the improper integral exists) giving the time of blow-up.

Next, solutions with $0 < y_0 < 1$ are bounded and thus defined for all times by Proposition 4.3. They are increasing and thus must converge to the larger equilibrium point, that is

$$\lim_{t \rightarrow \infty} y(t, y_0) = 1.$$

Finally, if we start with $y_0 > 1$, then the solution $y(t, y_0)$ will be decreasing and thus bounded, satisfying again

$$\lim_{t \rightarrow \infty} y(t, y_0) = 1.$$

We can learn even more about the shape of the solution curves. Differentiating the equation with respect to time and using the product rule, we obtain

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

Since for each solution (apart from the stationary ones), y' has fixed sign, we see that the inflection points can exist only on solutions starting at $y_0 \in (0, 1)$ and occur precisely at $y = 1/2$ - for this value of y the solution changes from being convex downward to being convex upward. In the two other cases, the second derivative is of constant sign, giving the solution convex upward for negative solutions and convex downward for solutions larger than 1.

We see that we got essentially the same picture as by solving the equation with much less work.

5.2.2 Crystal growth—a case study

In many applications, such as photographic film production, it is important to be able to manufacture crystals of a given size. The method is based on a process called ‘Ostwald ripening’. The process begins by adding a mixture of small crystals of various sizes to a certain solvent and kept mixed. Ostwald ripening is based on the following observation: if one allows the process to continue for a long time, either all crystal grains will dissolve into solution, or all the grains will become the same size. Hence, technologically, one has to arrange the conditions, such as the concentration, so that the second possibility occurs.

To start our modelling process, we will make some simplifying assumptions. First we assume that all crystals are of the same shape and differ only in size which can be described by a single real parameter, i.g, they may be boxes with edges (La, Lb, Lc) where a, b, c are fixed, reference, positive numbers and L is a real positive variable. If, in-

stead of to crystals, we apply the model to aerosols, we can think about balls with radius L .

Consider a volume of fluid containing an amount of dissolved matter (solute) with (uniform) concentration $c(t)$ at time t . There is a saturation concentration c^* , which is the maximum solute per unit volume that the fluid can hold. If $C(t) > c^*$, then the excess solute precipitates out in solid form; that is, in crystal form. Actually, for the precipitation $c(t)$ must be bigger than a certain quantity $c_L > c^*$ which depends on the size of precipitating grains. The threshold constant c_L is given by the Gibbs-Thomson relation

$$c_L = c^* e^{\Gamma/L}$$

where Γ is a physical quantity that depends on the shape of the crystals, its material properties and temperature (which here is assumed fixed). Hence, if $c(t) > c_L$, then material will come out of the solution and deposit onto the crystals, characterized by the size L , and if $c(t) < c_L$, then the material will dissolve back from the crystals. Using the Gibbs-Thomson relation, we define

$$L^*(t) = \frac{\Gamma}{\log \frac{c(t)}{c^*}}, \quad (5.8)$$

Note that function $L^*(t) < L$ if and only if $c(t) > c_L$ so that $L^*(t) < L$, then the crystals will grow. A semi-empirical law taking into account this observation is

$$\frac{dL}{dt} = G(L, c(t)), \quad (5.9)$$

where

$$G(L, c(t)) = \begin{cases} k_g \left(c(t) - c^* e^{\frac{\Gamma}{L}} \right)^g & \text{if } L > L^*(t), \\ -k_d \left(c^* e^{\frac{\Gamma}{L}} - c(t) \right)^d & \text{if } L < L^*(t), \end{cases} \quad (5.10)$$

and where k_g, k_d, g, d are positive constants with

$$1 \leq g, d \leq 2. \quad (5.11)$$

As expected, for $c(t) > c^*$ we have $dL/dt > 0$; that is, the crystal grows. Conversely, $c(t) < c^*$ we have $dL/dt < 0$ and the crystal shrinks.

We assume that initially there are N different crystal sizes characterized by sizes $L_j = x_j^*$ with μ_j^* crystals of size x_j^* per unit volume, $j = 1, \dots, N$, ordered as

$$0 < x_1^* < \dots < x_N^*.$$

We assume that the crystals do no coalesce or split. Moreover, according to (5.10), crystals which of the same size grow at the same rate. Hence, at any time t we will have N classes of crystals with sizes $x_1(t), \dots, x_N(t)$ (some $x_j(t)$ may be, however, zero if the crystals of this size completely dissolve. Thus we obtain the system of N equations

$$\frac{dx_j}{dt} = G(x_j, c(t)), \quad j = 1, \dots, N, \quad (5.12)$$

which is coupled through the unknown concentration $c(t)$. The formula for $c(t)$ can be obtained as the sum of the initial concentration c_0 and the amount which was dissolved from the crystals initially present (in a unit volume):

$$c(t) = c_0 + \rho k_v \sum_{j=1}^N \mu_j^* (x_j^*)^3 - \rho k_v \sum_{j=1}^N \mu_j^* (x_j(t))^3, \quad (5.13)$$

where k_v is a geometric parameter relating L^3 to the crystal volume ($k_v = abc$ in the case of a box discussed earlier or $k_v = 4\pi/3$ in the case of a sphere), and ρ is the mass density of the solid phase of the material. With this, we can write (5.12) in a more explicit form:

$$\frac{dx_j}{dt} = G_j(x_1, \dots, x_N), \quad j = 1, \dots, N \quad (5.14)$$

For further use we introduce

$$\mu_j = \rho k_v \mu_j^*, \quad c_1 = c_0 + \sum_{j=1}^N \mu_j (x_j^*)^3. \quad (5.15)$$

Note that c_1 is the total amount of the material per unit volume in either crystal or solution form.

5.2.2.1 The case of one crystal size

In the case when $N = 1$ we have

$$\frac{dx}{dt} = G(x), \quad x(0) = x^*, \quad (5.16)$$

where

$$G(x) = \begin{cases} k_g (c_1 - \mu x^3 - c^* e^{\frac{r}{x}})^g & \text{if } c_1 - \mu x^3 > c^* e^{\frac{r}{x}}, \\ -k_d (c^* e^{\frac{r}{x}} - (c_1 - \mu x^3))^d & \text{if } c_1 - \mu x^3 < c^* e^{\frac{r}{x}}, \end{cases} \quad (5.17)$$

We observe that since $g, d \geq 1$, G is continuously differentiable on each set $\{x > 0; c_1 - \mu x^3 - c^* e^{\frac{r}{x}} \leq 0\}$. Thus, it is Lipschitz continuous for $x > 0$. The first question is to determine points at which G changes sign. Denote $f(x) = c_1 - \mu x^3 - c^* e^{\frac{r}{x}}$ so that (5.16) can be written as

$$\frac{dx}{dt} = \begin{cases} k_g (f(x))^g & \text{if } f(x) > 0, \\ -k_d (-f(x))^d & \text{if } f(x) < 0, \end{cases} \quad (5.18)$$

Lemma 5.2 *There exist at most two positive solutions of the equation*

$$f(x) = c_1 - \mu x^3 - c^* e^{\frac{\Gamma}{x}} = 0. \quad (5.19)$$

Proof. We have $\lim_{x \rightarrow 0^+} f(x) = \lim_{x \rightarrow \infty} f(x) = -\infty$. Further

$$f'(x) = -3\mu x^2 + \frac{\Gamma c^*}{x^2} e^{\frac{\Gamma}{x}}$$

and

$$f''(x) = -6\mu x - \frac{\Gamma c^*}{x^3} e^{\frac{\Gamma}{x}} - \frac{\Gamma^2 c^*}{x^4} e^{\frac{\Gamma}{x}}$$

so that $f''(x) < 0$ for all $x > 0$. Therefore f' has at most one zero and thus, by the Rolle theorem, f can have at most two solutions. \square

In what follows we focus on the case when we have exactly two solutions denote $0 < \xi_1 < x_2$. Note that in practice this can be always achieved by taking the initial concentration c_0 large enough, so that $f(x_0) > 0$ for some chosen x_0 . Then $\xi_1 < x_0 < \xi_2$.

We can apply Theorem ?? to describe the evolution of the crystal's size depending on the initial condition.

Proposition 5.1

- (i) $x(t) = \xi_1$ and $x(t) = \xi_2$ are stationary solutions of (5.18);
- (ii) If $x^* > \xi_2$, then $x(t)$ is decreasing with $\lim_{t \rightarrow \infty} x(t) = \xi_2$;
- (iii) If $\xi_1 < x^* < \xi_2$, then $x(t)$ is increasing with $\lim_{t \rightarrow \infty} x(t) = \xi_2$;
- (iv) If $x^* < \xi_1$, then $x(t)$ is decreasing and there is finite time t_0 such that $x(t_0) = 0$.

Proof. Items (i)-(iii) follow directly from Theorem ?? (note that the solutions exists for all positive times as they are bounded). We have to reflect on (iv) as Theorem ?? is not directly applicable here (G does not satisfy assumptions of the Picard theorem on the whole real line, in fact, $x = 0$ clearly is not the point of Lipschitz continuity). However, the idea of the proof still is applicable. Indeed, clearly $x(t)$ decreases, that is $x(t) \leq x^* < \xi_1$ for all $t \in [0, t_{max})$ and, as f increases for $x < \xi_1$ we have $dx/dt = G(x(t)) \leq G(x_*) = c < 0$. Thus, $x(t) \leq ct + x^*$ and $x(t_0) = 0$ for $t \leq -x^*/c$. Alternatively, we have

$$t = -\frac{1}{k_d} \int_{x^*}^{x(t)} \frac{ds}{(c^* e^{\Gamma/s} - c_1 + \mu s^3)^d}$$

and the time t_0 at which $x(t_0) = 0$ is given by

$$t_0 = \frac{1}{k_d} \int_0^{x^*} \frac{ds}{(c^* e^{\Gamma/s} - c_1 + \mu s^3)^d}$$

Since

$$\lim_{s \rightarrow 0^+} \frac{c^* e^{d\Gamma/s}}{(c^* e^{\Gamma/s} - c_1 + \mu s^3)^d} = 1$$

and $\int_0^{x^*} e^{-d\Gamma/s} ds < +\infty$, the improper integral above converges giving $t_0 < +\infty$. \square

5.2.2.2 *The case of multiple crystal sizes*

Now let us consider the general case of crystals with N sizes. We start with the following observation:

Lemma 5.3 *If*

$$c_0 > c^*, \quad (5.20)$$

then $c(t) > c^*$ *for all* $t > 0$. *Moreover*

$$x_j(t) \leq \left(\frac{c_1}{\mu_j} \right)^{\frac{1}{3}}. \quad (5.21)$$

Proof. Since $c(t)$ is a continuous function, the set $\{t > 0; c(t) = c_0\}$ is closed and bounded from below and thus it contains the smallest element t_0 ; that is, t_0 is the first time at which $c(t_0) = 0$. Hence $c(t) > 0$ for $t < t_0$ and, since c is differentiable, $\frac{dc}{dt}|_{t=t_0} \leq 0$. On the other hand, by (5.12),

$$\frac{dc}{dt}|_{t=t_0} = -3 \sum_{i=1}^N \mu_i x_i(t_0)^2 \frac{dx_i}{dt}|_{t=t_0} = -3 \sum_{i=1}^N \mu_i x_i^2(t_0) G(x_i(t_0), c^*) > 0$$

as by (5.10), $G(x_j, c^*) < 0$ unless $x_j(t_0) = 0$ for all $j = 1, \dots, N$. But in the latter case we would have, by (5.13),

$$c^* = c(t_0) = c_0 + \sum_{j=1}^N \mu_j (x_j^*)^3 \geq c_0,$$

which contradicts (5.20).

To prove (5.21) we note that, again by (5.13),

$$0 < c(t) < c_1 - \sum_{j=1}^N \mu_j (x_j(t))^3,$$

so that

$$\sum_{j=1}^N \mu_j (x_j(t))^3 < c_1$$

which yields (5.21) since all summands are nonnegative. This completes the proof of the lemma. \square

In the next lemma we show that the difference between sizes of crystals increases in time.

Lemma 5.4 *If $x_{j+1}^* > x_j^*$, then $x_{j+1}(t) > x_j(t)$ for all t as long as $x_j(t) > 0$.*

Proof. First we observe that $G(x, c)$ is an increasing function of x , hence if $x_{j+1}(t) > x_j(t)$ at some time t , then

$$\frac{d}{dt}(x_{j+1}(t) - x_j(t)) = G(x_{j+1}(t), c(t)) - G(x_j(t), c(t)) > 0.$$

To shorten notation, let $f(t) = x_{j+1}(t) - x_j(t)$ and $g(t) = G(x_{j+1}(t), c(t)) - G(x_j(t), c(t))$ and fix t for which $f(t) > 0$. We have the situation that $f'(t) = g(t) > 0$ and $g(t) > 0$ as long as $f(t) > 0$. Since f' is continuous, we can argue as in the previous lemma that if $f'(\tau) = 0$ for some $\tau > t$, then there must be the first time $t_0 > t$ for which this happens. Then

$$0 = g(t_0) = G(x_{j+1}(t_0), c(t_0)) - G(x_j(t_0), c(t_0)).$$

Since G is monotonic in x , this means that $x_{j+1}(t_0) = x_j(t_0)$.

Since $x_{j+1}(t) > x_j(t)$, with $f'(t) > 0$, then there must be a point $t < \bar{t} < t_0$ at which $f'(\bar{t}) = 0$ which contradicts the assumption that t_0 is the first such time (precisely, by the Mean Value Theorem, there is point $0 \leq t' \leq t_0$ such that $f'(t') < 0$ but by continuity of f' and the Darboux property $f'(t'') = 0$ for some $t'' < t'$).

Let us recall that the curve $L^*(t)$ determines whether a crystal grows or shrinks: if $x_j(t) > L_j^*(t)$, then $x_j(t)$ is growing and if $x_j(t) < L_j^*(t)$, then $x_j(t)$ shrinks.

As we noted earlier, some crystals can completely dissolve in finite time. Let us denote by k the number of all crystal sizes that have disappeared in finite time. Since the number of crystal classes is finite, we can say that after a certain time t_0 only the crystals of sizes

$$x_{k+1}(t), x_{k+2}(t), \dots, x_N(t), \quad t > t_0,$$

are present in the system. Clearly, it is possible that $k = 0$.

We present the main theorem of this section.

Theorem 5.2 *All crystals which do not belong to the largest cohort x_N will dissolve in finite time.*

Proof. We begin by noting that during their lifetime a crystal can grow and shrink in various periods of time. The change occurs when $x_j(t)$ crosses the line $L^*(t)$. It follows that, while $x_j(t)$, $k + 1 \leq j < N$, can cross $L^*(t)$ several times, $x_N(t)$ can do at most once. Indeed, at the point of intersection we have $dx_N/dt = G(x_N, c(t^*)) = 0$ but

$$\frac{dL^*}{dt} = -\frac{\Gamma}{\log \frac{c(t)}{c^*}} \frac{1}{c(t)} \frac{dc}{dt} < 0$$

since at $t = t^*$ we have

$$\frac{dc}{dt} = -3 \sum_{j=k+1}^N \mu_j x_j(t^*) G(x_j(t^*), c(t^*)) > 0$$

as $G(x_j(t^*), c(t^*)) < 0$ on account of $x_j(t^*) < x_N(t^*) = L^*(t^*)$.

5.2.3 Equilibrium points of difference equations

Definition 5.1 A point x^* in the domain of f is said to be an equilibrium point of (5.4) if it is a fixed point of f ; that is, if $f(x^*) = x^*$.

In other words, x^* is a constant solution of (5.4).

Graphically, an equilibrium point is the the x -coordinate of the point where the graph of f intersects the diagonal $y = x$. This is the basis of the so-called cobweb method of finding equilibria and analyse their stability, which is described later.

Definition 5.2

- (i) The equilibrium x^* is stable if for given $\epsilon > 0$ there is $\delta > 0$ such that for any x and for any $n > 0$, $|x - x^*| < \delta$ implies $|f^n(x) - x^*| < \epsilon$ for all $n > 0$. If x^* is not stable, then it is called unstable (that is, x^* is unstable if there is $\epsilon > 0$ such that for any $\delta > 0$ there are x and n such that $|x - x^*| < \delta$ and $|f^n(x) - x^*| \geq \epsilon$.)
- (ii) A point x^* is called attracting if there is $\eta > 0$ such that

$$|x(0) - x^*| < \eta \text{ implies } \lim_{n \rightarrow \infty} x(n) = x^*.$$

If $\eta = \infty$, x^* is called a global attractor or globally attracting.

- (iii) The point x^* is called an asymptotically stable equilibrium if it is stable and attracting. If $\eta = \infty$, then x^* is said to be globally asymptotically stable equilibrium.

Example 5.2 Consider the logistic equation

$$x(n+1) = 3x(n)(1-x(n)). \quad (5.22)$$

The equation for the equilibrium points reads

$$x = 3x(1-x)$$

which gives $x_0 = 0$ and $x_1 = 2/3$.

5.2.3.1 The Cobweb Diagrams

We start with an important graphical method for analysing the stability of equilibrium (and periodic) points of (5.4). Since $x(n+1) = f(x(n))$, we may draw a graph of f in the $(x(n), x(n+1))$ system of coordinates. Then, given $x(0)$, we pinpoint the value $x(1)$ by drawing the vertical line through $x(0)$ so that it also intersects the graph of f at $(x(0), x(1))$. Next, draw a horizontal line from $(x(0), x(1))$ to meet the diagonal line $y = x$ at the point $(x(1), x(1))$. A vertical line drawn from the point $(x(1), x(1))$ will meet the graph of f at the point $(x(1), x(2))$. In this way we may find $x(n)$. This is illustrated in Fig. 5.1 where we presented several steps of drawing the cobweb diagram for the logistic equation (5.22) with $x_0 = 0.2$. On the basis of the diagram we can conjecture that $x_1 = 2/3$ is an asymptotically stable equilibrium as the solution converges to it as n becomes large. However, to be sure, we need to develop analytical tools.

5.2.3.2 Analytic criterion for stability

Theorem 5.3 Let x^* be an equilibrium point of the difference equation

$$x(n+1) = f(x(n)) \quad (5.23)$$

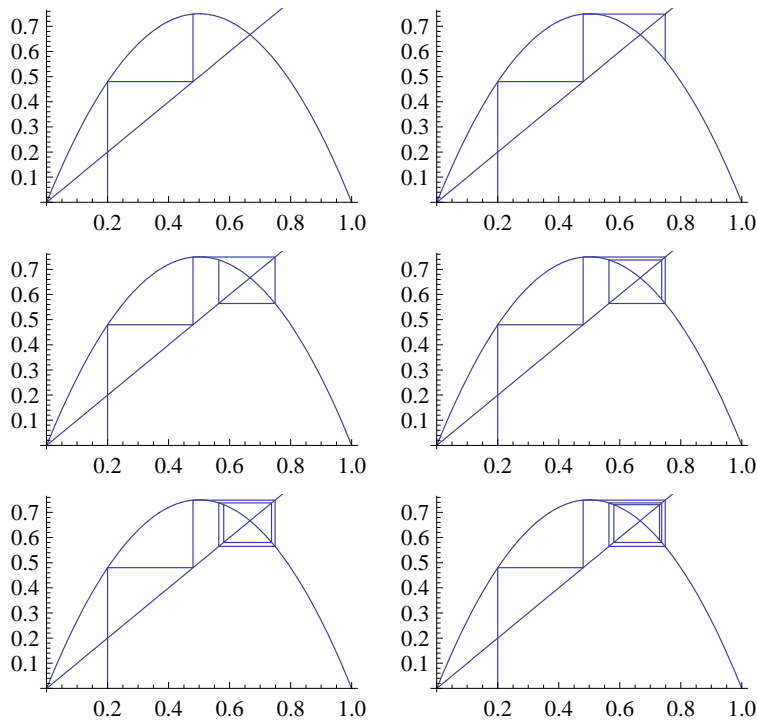


Fig. 5.1. Cobweb diagram of a logistic difference equation

where f is continuously differentiable at x^* . Then:

- (i) If $|f'(x^*)| < 1$, then x^* is asymptotically stable;
- (ii) If $|f'(x^*)| > 1$, then x^* is unstable.

Proof. Suppose $|f'(x^*)| < M < 1$. Then $|f'(x)| \leq M < 1$ over some interval $J = (x^* - \gamma, x^* + \gamma)$ by the property of local preservation of sign for continuous function. Now, we

have

$$|x(1) - x^*| = |f(x(0)) - f(x^*)|.$$

By the Mean Value Theorem, there is $\xi \in [x(0), x^*]$ such that

$$|f(x(0)) - f(x^*)| = |f'(\xi)||x(0) - x^*|.$$

Hence

$$|f(x(0)) - f(x^*)| \leq M|x(0) - x^*|,$$

and therefore

$$|x(1) - x^*| \leq M|x(0) - x^*|.$$

Since $M < 1$, the inequality above shows that $x(1)$ is closer to x^* than $x(0)$ and consequently $x(1) \in J$. By induction,

$$|x(n) - x^*| \leq M^n|x(0) - x^*|.$$

For given ϵ , define $\delta = \epsilon/2M$. Then $|x(n) - x^*| < \epsilon$ for $n > 0$ provided $|x(0) - x^*| < \delta$ (since $M < 1$). Furthermore $x(n) \rightarrow x^*$ and $n \rightarrow \infty$ so that x^* is asymptotically stable.

To prove the second part of the theorem, we observe that, as in the first part, there is $\epsilon > 0$ such that on $J = (x^* - \epsilon, x^* + \epsilon)$ on which $|f'(x)| \geq M > 1$. Take arbitrary $\delta > 0$ smaller than ϵ and x satisfying $|x - x^*| < \delta$. Using again the Mean Value Theorem

$$|f(x) - x^*| = |f'(\xi)||x - x^*|$$

for some ξ between x^* and x so that

$$|f(x) - x^*| \geq M|x - x^*|.$$

If $f(x)$ is outside J , then we are done. If not, we can repeat the argument getting $|f^2(x) - x^*| \geq M^2|x - x^*|$,

that is, $f^2(x)$ which is further away from x^* than $f(x)$. If it is in J we can continue the procedure till $|f^n(x) - x^*| \geq M^n|x - x^*| > \epsilon$ for some n . \square

Equilibrium points with $|f'(x^*)| \neq 1$ are called *hyperbolic*.

What happens if the equilibrium point is non-hyperbolic? Before we give the answer to this question, let us reflect on the geometry of the preceding theorem. In this discussion we assume that $f'(x^*) > 0$. The equilibrium x^* is stable if the graph of $y = f(x)$ is less steep than the graph of $y = x$; that is, the graph of f crosses the line $y = x$ from above to below as x increases. This ensures that the cobweb iterations from the left are increasing and from the right are decreasing converging to x^* . On the contrary, x^* is unstable if the graph of f crosses $y = x$ from below—then the cobweb iterations will move away from x^* . If $f'(x^*) = 1$, then the graph of f is tangent to the line $y = x$ at $x = x^*$ but the stability properties follow from the geometry. If $f''(x^*) \neq 0$, then the graph of f will be (locally) either entirely above or entirely below the line $y = x$ and then the picture will be the same as in the unstable case either to the left, or to the right, of x^* . Hence x^* is unstable in this case (remember that for instability it is sufficient to display, for any neighbourhood of x^* , only one diverging sequence of iterations emanating from this neighbourhood). On the other hand, if $f''(x^*) = 0$, then x^* is an inflection point and the graph of f crosses the line $y = x$. This case is essentially the same as when $|f'(x^*)| \neq 1$ —the equilibrium is stable if the graph of f crosses $y = x$ from above and unstable if it does it from below. A quick reflection ascertains that the former occurs when $f'''(x^*) < 0$ and the latter if $f'''(x^*) > 0$.

Summarizing, the following theorem holds.

Theorem 5.4 *Let x^* be an isolated equilibrium with $f'(x^*) =$*

1. *Then*

- (i) *If $f''(x^*) \neq 0$, then x^* is unstable.*
- (ii) *If $f''(x^*) = 0$ and $f'''(x^*) > 0$, then x^* is unstable.*
- (iii) *If $f''(x^*) = 0$ and $f'''(x^*) < 0$, then x^* is asymptotically stable.*

The case of $f'(x^*) = -1$ is more difficult. First we note, that if $f(x) = -x + 2x^*$; that is f is the linear function passing producing equilibrium at $x = x^*$ with $f'(x^*) = -1$, then iterations starting from $x_0 \neq x^*$ produce solution taking only two values (compare item (ii) of Section 3.4) oscillating around x^* . Thus, if $-1 < f'(x^*) < 0$, then f passes from below of $y = -x + 2x^*$ to above as x increases and so the stability follows from the fact that subsequent iterations oscillate around x^* getting closer to x^* with each iteration. On the contrary, if $f'(x^*) < -1$ the oscillating iterations move away from x^* . If $f'(x^*) = -1$, then the graph of f crosses the line $y = x$ at the right angle. Hence, the stability depends on fine details of the shape of f close to x^* . Unfortunately, using an argument similar to the case with $f'(x^*) = 1$ and considering the relation of the graph of f with the graph of $y = -x + 2x^*$ produces only partial result: x^* will be stable if $f''(x^*) = 0$ and $f'''(x^*) > 0$ (as then the graph of f will have the same shape as in the stable case, crossing the line $y = -x + 2x^*$ from below. However, the stability of x^* can be achieved in a more general situation. First we note that x^* is an equilibrium of $g(x) := f(f(x))$ as well and it is a stable

equilibrium of f if and only if it is stable for g . This statement follows from continuity of f : if x^* is stable for g , then $|g^n(x_0) - x^*| = |f^{2n}(x_0) - x^*|$ is small for x_0 sufficiently close to x^* . But then $|f^{2n+1}(x_0) - x^*| = |f(f^{2n}(x_0)) - f(x^*)|$ is also small by continuity of f . The reverse is obvious. We have

$$g(x)' = f'(f(x))f'(x)$$

so $g'(x^*) = 1$ and we can apply Theorem 5.3 to g . Hence

$$g''(x) = f''(f(x))[f'(x)]^2 + f'(f(x))f''(x)$$

and, since $f(x^*) = x^*$ and $f'(x^*) = -1$,

$$g''(x^*) = 0.$$

Using the chain rule once again, we find

$$g'''(x^*) = -2f'''(x^*) - 3[f''(x^*)]^2.$$

Hence, we can note

Theorem 5.5 *Suppose that at an equilibrium point x^* we have $f'(x^*) = -1$. Define*

$$S(x^*) = -f'''(x^*) - \frac{3}{2}(f''(x^*))^2. \quad (5.24)$$

Then x^ is asymptotically stable if $S(x^*) < 0$ and unstable if $S(x^*) > 0$.*

Example 5.3 *Consider the equation*

$$x(n+1) = x^2(n) + 3x(n).$$

Solving $f(x) = x^2 + 3x = x$, we find that $x = 0$ and $x = -2$ are the equilibrium points. Since $f'(0) = 3 > 1$, we

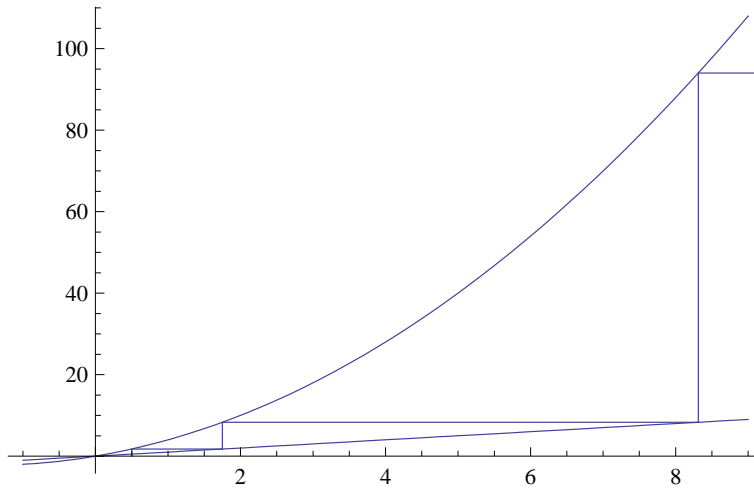


Fig. 5.2. Unstable character of the equilibrium $x = 0$. Initial point $x_0 = 0.5$

conclude that the equilibrium at $x = 0$ is unstable. Next, $f'(-2) = -1$. We calculate $f''(-2) = 2$ and $f'''(-2) = 0$ so that $S(-2) = -12 < 0$. Hence, $x = -2$ is an asymptotically stable equilibrium.

Remark 5.3 Analysing cob-web diagrams (or otherwise) we observe that we can provide a further fine-tuning of the stability. Clearly, if $f'(x^*) < 0$, then the solution behaves in an oscillatory way around x^* and if $f'(x^*) > 0$, it is monotonic. Indeed, consider (in a neighbourhood of x^* where $f'(x) < 0$) $f(x) - f(x^*) = f(x) - x^* = f'(\xi)(x - x^*)$, where ξ is between x^* and x . Since $f' < 0$, $f(x) > x^*$ if $x < x^*$ and $f(x) < x^*$ if $x > x^*$, which means that each

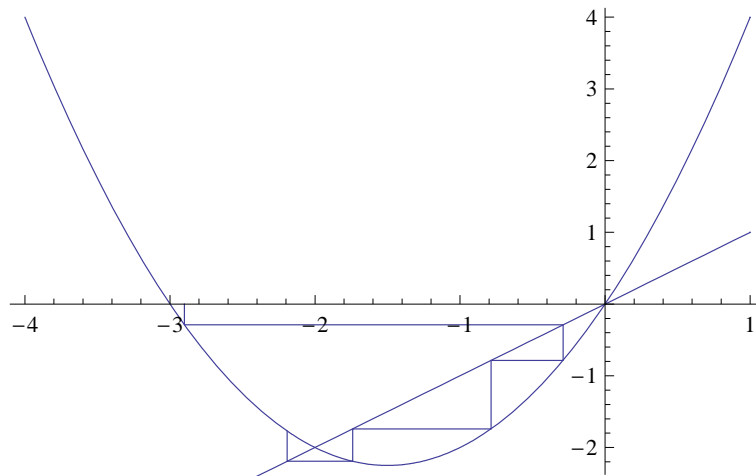


Fig. 5.3. Stable character of the equilibrium $x = -2$. Initial point $x_0 = -2.9$

iteration move the point to the other side of x^* . If $|f'| < 1$ over this interval, then $f^n(x)$ converge to x^* in an oscillatory way, while if $|f'| > 1$, the iterations will move away from the interval, also in an oscillatory way.

Based on on this observation, we may say that the equilibrium is oscillatory unstable or stable if $f'(x^*) < -1$ or $-1 < f'(x^*) < 0$, respectively, and monotonically stable or unstable depending on whether $0 < f'(x^*) < 1$ or $f'(x^*) > 1$, respectively.

Periodic points and cycles

Definition 5.3 Let b be in the domain of f . Then:

- (i) b is called a periodic point of f if $f^k(b) = b$ for some $k \in \mathbb{N}$. The periodic orbit of b , $O(b) = \{b, f(b), f^2(b), \dots, f^{k-1}(b)\}$ is called a k -cycle.
- (ii) b is called eventually k -periodic if, for some integer m , $f^m(b)$ is a k -periodic point.

Example 5.4 The Tent Map revisited. Consider

$$x(n + 1) = T^2x(n)$$

where we have

$$T^2(x) = \begin{cases} 4x & \text{for } 0 \leq x \leq 1/4, \\ 2(1 - 2x) & \text{for } 1/4 < x \leq 1/2, \\ 2x - 1 & \text{for } 1/2 < x \leq 3/4, \\ 4(1 - x) & \text{for } 3/4 < x \leq 1. \end{cases}$$

There are four equilibrium points, 0, 0.4, 2/3 and 0.8, two of which are equilibria of T . Hence $\{0, 4, 0.8\}$ is the only 2-cycle of T . $x^* = 0.8$ is not stable. Calculation for T^3 shows that $\{2/7, 4/7, 6/7\}$ is a 3-cycle. There is a famous theorem by Šarkowski (rediscovered by Li and Yorke) that if a map has a 3-cycle, then it has k -cycles for arbitrary k . This is one of symptoms of chaotic behaviour.

Definition 5.4 Let b be a k -periodic point of f . Then b is said to be:

- (i) stable if it is a stable fixed point of f^k ;
- (ii) asymptotically stable if it is an asymptotically stable fixed point of f^k ;
- (iii) unstable if it is an unstable fixed point of f^k .

It follows that if b is k -periodic, then every point of its k -cycle $\{x(0) = b, x(1) = f(b), \dots, x(k - 1) = f^{k-1}(b)\}$ is

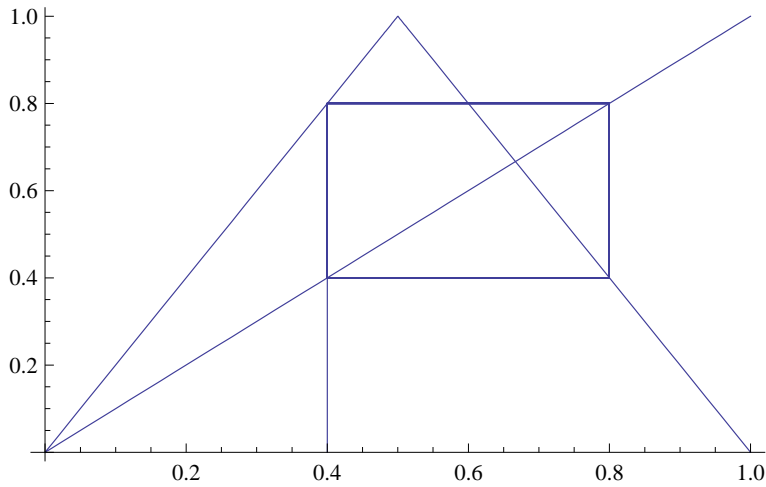


Fig. 5.4. 2-cycle for the tent map

also k -periodic. This follows from $f^k(f^r(b)) = f^r(f^k(b)) = f^r(b)$, $r = 0, 1, \dots, k - 1$. Moreover, each such point possesses the same stability property as b . Here, the stability of b means that $|f^{nk}(x) - b| < \epsilon$ for all n , provided x is close enough to b . To prove the statement, we have to show that for any ϵ there is δ such that $|f^{nk}(x) - f^r(b)| < \epsilon$ for any fixed $r = 0, 1, \dots, k - 1$ and $n \in \mathbb{N}$, if $|x - f^r(b)| < \delta$. Let us take arbitrary $\epsilon > 0$. From continuity of f (at thus of f^k), there is δ_1 such that $|x - f^r(b)| < \delta_1$ implies, by $f^{k+r}(b) = f^r(f^k(b)) = f^r(b)$, that

$$|f^k(x) - f^r(b)| = |f^k(x) - f^{k+r}(b)| < \epsilon. \quad (5.25)$$

With the same ϵ , using continuity of f^r we find δ_2 such that $|f^r(z) - f^r(b)| < \epsilon$, provided $|z - b| < \delta_2$. For this δ_2 , we

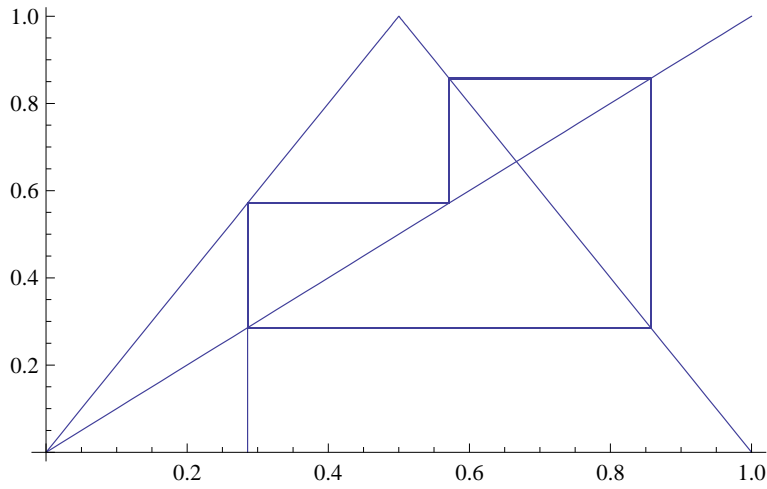


Fig. 5.5. 3-cycle for the tent map

find δ_3 such that if $|y - b| < \delta_3$, then $|f^{nk}(y) - b| < \delta_2$ for any n . Hence, for $|y - b| < \delta_3$, taking $z = f^{nk}(y)$, we obtain

$$|f^{r+nk}(y) - f^r(b)| < \epsilon \quad (5.26)$$

for any n . On the other hand, for this δ_3 we find δ_4 such that if $|x - f^r(b)| < \delta_4$, then $|f^{k-r}(x) - f^k(b)| = |f^{k-r}(x) - b| < \delta_3$ and, using $y = f^{k-r}(x)$ in (5.26), we obtain

$$|f^{(n+1)k}(x) - f^r(b)| < \epsilon \quad (5.27)$$

for any $n \geq 1$. Taking $|x - f^r(b)| < \delta_5 = \min\{\delta_4, \delta_1\}$ and combining (5.25) with (5.27), we get

$$|f^{nk}(x) - f^r(b)| < \epsilon,$$

for any $n \geq 1$.

The definition together with Theorem 5.2 yield the following classification of stability of k -cycles.

Theorem 5.6 *Let $O(b) = \{x(0) = b, x(1) = f(b), \dots, x(k-1) = f^{k-1}(b)\}$ be a k -cycle of a continuously differentiable function f . Then*

(i) *The k -cycle $O(b)$ is asymptotically stable if*

$$|f'(x(0))f'(x(1)) \dots f'(x(k-1))| < 1.$$

(ii) *The k -cycle $O(b)$ is unstable if*

$$|f'(x(0))f'(x(1)) \dots f'(x(k-1))| > 1.$$

Proof. Follow from Theorem 5.2 by the Chain Rule applied to f^k . \square

The Logistic Equation and Bifurcations Consider the logistic equation

$$x(n+1) = \mu x(n)(1-x(n)), \quad x \in [0, 1], \mu > 0 \quad (5.28)$$

which arises from iterating $F_\mu(x) = \mu x(1-x)$. To find equilibrium point, we solve

$$F_\mu(x^*) = x^*$$

which gives $x^* = 0, (\mu - 1)/\mu$.

We investigate stability of each point separately.

(a) For $x^* = 0$, we have $F'_\mu(0) = \mu$ and thus $x^* = 0$ is asymptotically stable for $0 < \mu < 1$ and unstable for $\mu > 1$. To investigate the stability for $\mu = 1$, we find $F''_\mu(0) = -2 \neq 0$ and thus $x^* = 0$ is unstable in this case. However, instability comes from negative

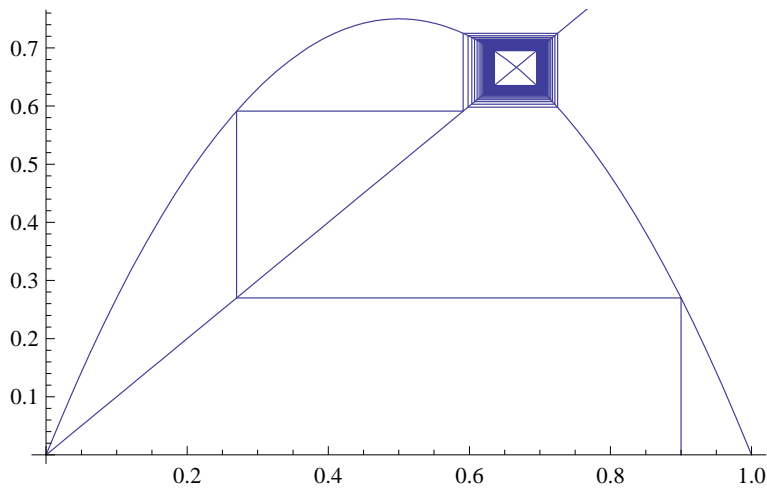


Fig. 5.6. Asymptotically stable equilibrium $x = 2/3$ for $\mu = 3$.

values of x which we discarded from the domain. If we restrict our attention to the domain $[0, 1]$, then $x^* = 0$ is stable. Such points are called *semi-stable*.

- (b) The equilibrium point $x^* = (\mu - 1)/\mu$ belongs to the domain $[0, 1]$ only if $\mu > 1$. Here, $F'((\mu - 1)/\mu) = 2 - \mu$ and $F''((\mu - 1)/\mu) = -2\mu$. Thus, using Theorems 5.2 and 5.3 we obtain:

- (i) x^* is asymptotically stable if $1 < \mu \leq 3$,
- (ii) x^* is unstable if $3 < \mu$.

We observe further that for $1 < \mu < 2$ the population approaches the carrying capacity monotonically from below. However, for $2 < \mu \leq 3$ the population can go over the carrying capacity but eventually stabilizes around it.

What happens for $\mu = 3$? Consider 2-cycles. We have $F_\mu^2(x) = \mu^2 x(1-x)(1-\mu x(1-x))$ so that we are looking for solutions to

$$\mu^2 x(1-x)(1-\mu x(1-x)) = x$$

We can re-write this equation as

$$x(\mu^3 x^3 - 2\mu^3 x^2 + \mu^2(1+\mu)x + (1-\mu^2)) = 0.$$

To simplify the considerations, we observe that any equilibrium is also a 2-cycle (and any k -cycle for that matter). Thus, we can divide this equation by x and $x - (\mu - 1)/\mu$, getting

$$\mu^2 x^2 - \mu(\mu + 1)x + \mu + 1 = 0.$$

Solving this quadratic equation, we obtain 2-cycle

$$\begin{aligned} x(0) &= \frac{(1+\mu) - \sqrt{(\mu-3)(\mu+1)}}{2\mu} \\ x(1) &= \frac{(1+\mu) + \sqrt{(\mu-3)(\mu+1)}}{2\mu}. \end{aligned} \quad (5.29)$$

Clearly, these points determine 2-cycle provided $\mu > 3$ (in fact, for $\mu = 3$ these two points collapse into the equilibrium point $x^* = 2/3$). Thus, we see that when the parameter μ passes through $\mu = 3$, the stable equilibrium becomes unstable and bifurcates into two 2-cycles.

The stability of 2-cycles can be determined by Theorem 5.5. We have $F'(x) = \mu(1-2x)$ so the 2-cycle is stable provided

$$-1 < \mu^2(1-2x(0))(1-2x(1)) < 1.$$

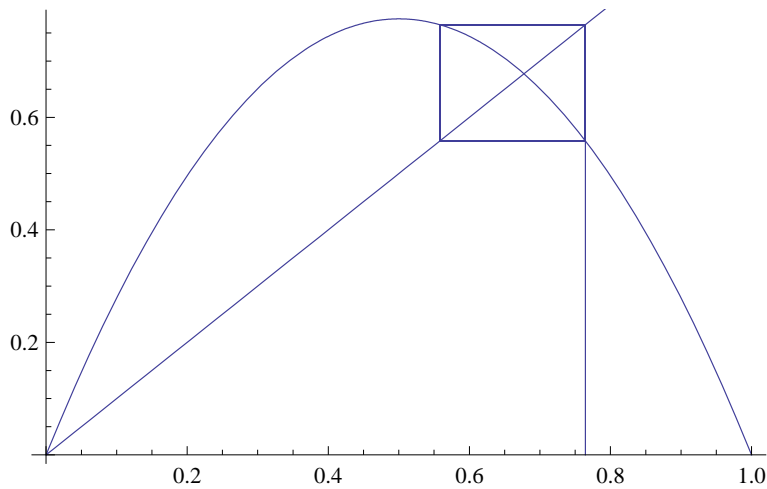


Fig. 5.7. 2-cycle for $x \approx 0.765$ and $\mu = 3.1$.

Using Viète's formulae we find that the above yields

$$-1 < \mu^2 + 2\mu + 4 < 1$$

and solving this we see that this is satisfied if $\mu < -1$ or $\mu > 3$ and $1 - \sqrt{6} < \mu < 1 + \sqrt{6}$ which yields $3 < \mu < 1 + \sqrt{6}$.

In similar fashion we can determine that for $\mu_1 = 1 + \sqrt{6}$ the 2-cycle is still attracting but becomes unstable for $\mu > \mu_1$.

Remark 5.4 To find 4-cycles, we solve $F_\mu^4(x)$. However, in this case algebra becomes unbearable and one should resort to a computer. It turns out that there is 4-cycle when $\mu > 1 + \sqrt{6}$ which is attracting for $1 + \sqrt{6} < \mu < 3.544090 \dots =: \mu_2$. When $\mu = \mu_2$, then 2²-cycle bifurcates into a 2³-cycle,

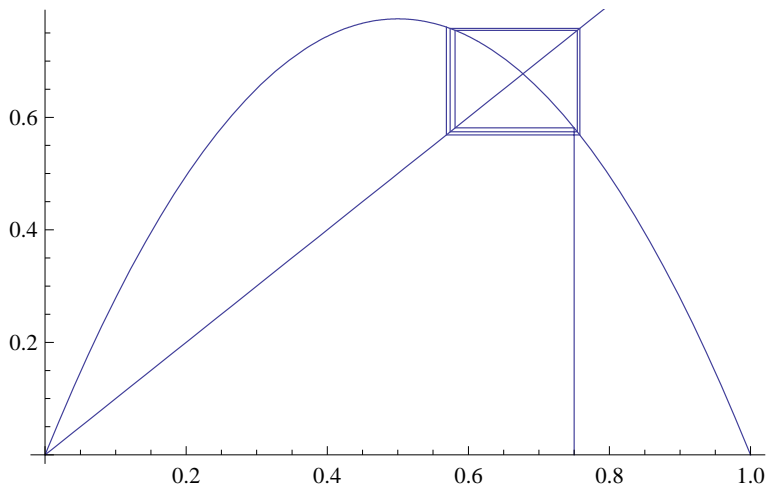


Fig. 5.8. Asymptotic stability of the 2-cycle for $x \approx 0.765$ and $\mu = 3.1$.

which is stable for $\mu_2 \leq \mu \leq \mu_3 := 3.564407..$ Continuing, we obtain a sequence of numbers $(\mu_n)_{n \in \mathbb{N}}$ such that the 2^n -cycle bifurcates into 2^{n+1} -cycle passing through μ_n . In this particular case, $\lim_{n \rightarrow \infty} \mu_n = \mu_\infty = 3.57\dots$. A remarkable observation is

Theorem 5.7 (Feigenbaum, 1978) For sufficiently smooth families F_μ of mapping of an interval into itself, the number

$$\delta = \lim_{n \rightarrow \infty} \frac{\mu_n - \mu_{n-1}}{\mu_{n+1} - \mu_n} = 4.6692016\dots$$

in general does not depend on the family of maps, provided they have single maximum.

This theorem expresses the fact that the bifurcation dia-

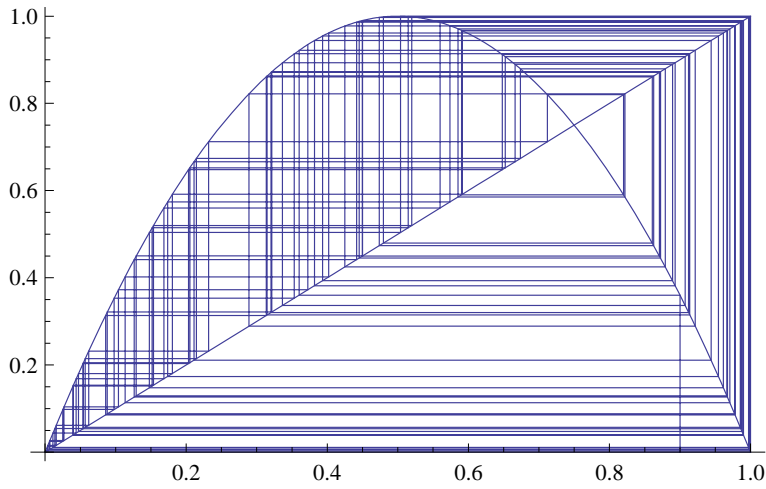


Fig. 5.9. Chaotic orbit for $x = 0.9$ and $\mu = 4$.

grams for such maps are equivalent to the bifurcation diagram of a unique mapping for which it is exactly self-similar.

What happens for μ_∞ ? Here we find a densely interwoven region with both periodic and chaotic orbits. In particular, a 3-cycle appears and, as we mentioned earlier, period 3 implies existence of orbits of any period. We can easily prove that 3-cycles appear if $\mu = 4$. Consider first $F_4(x)$. We have $F_4(0) = F_4(1) = 0$ and $F_4(0.5) = 1$. This shows that $F_4^2(0.5) = F_4(1) = 0$. From the Darboux property, there are $a_1 \in (0, 0.5)$ and $a_2 \in (0.5, 1)$ such that $F_4(a_i) = 0.5$ and $F_4^2(a_i) = 1$. Thus we have graph with two peaks at 1 and attaining zero in between. This shows that $F_4^2(x) = x$ has four solutions, two of which are (unsta-

ble) equilibria and two are (unstable) 2-cycles. Repeating the argument there is $b_1 \in (0, a_1)$ such that $F_4(b_1) = a_1$ (since the graph is steeper than that of $y = x$) and thus $F_4^3(b_1) = F_4^2(a_1) = 1$. Similarly, we get 3 other points in which $F_4^3 = 1$ and clearly $F_4^3(a_i) = F_4^3(0.5) = 0$. This means that $y = x$ meets $F_4^3(x)$ at 8 points, two of which are equilibria (2-cycles are not 3-cycles). So, we obtain two 3-cycles.

The Beverton-Holt-Hassell equation We conclude with a brief description of stability of equilibrium points for the Hassell equation.

Let us recall the equation

$$x(n+1) = f(x_n, R_0, b) = \frac{R_0 x_n}{(1+x_n)^b}.$$

Writing

$$x^*(1+x^*)^b = R_0 x^*$$

we find steady state $x^* = 0$ and we observe that if $R_0 \leq 1$, then this is the only steady state (at least for positive values of x). If $R_0 > 1$, there is another steady state given by

$$x^* = R_0^{1/b} - 1.$$

Evaluating the derivative, we have

$$f'(x^*, R_0, b) = \frac{R_0}{(1+x^*)^b} - \frac{R_0 b x^*}{(1+x^*)^{b+1}} = 1 - b + \frac{b}{R_0^{1/b}}$$

Clearly, with $R_0 > 1$, we always have $f' < 1$, so for the monotone stability we must have

$$1 - b + \frac{b}{R_0^{1/b}} > 0$$

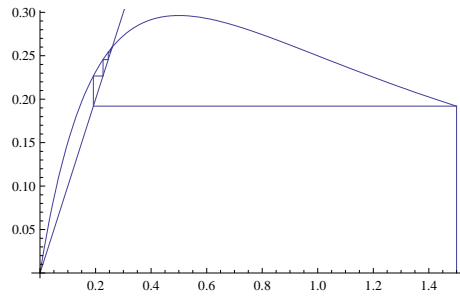


Fig. 5.10. Monotonic stability of the equilibrium for the Beverton-Holt model with $b = 3$ and $R_0 = 2$; see Eqn (5.30).

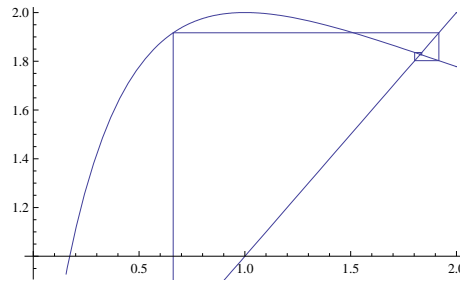


Fig. 5.11. Oscillatory stability of the equilibrium for the Beverton-Holt model with $b = 2$ and $R_0 = 8$; see Eqn (5.31).

and for oscillatory stability

$$-1 < 1 - b + \frac{b}{R_0^{1/b}} < 0.$$

Solving this inequalities, we obtain that the borderlines

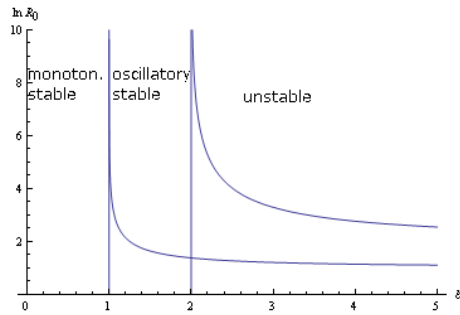


Fig. 5.12. Regions of stability of the Beverton-Holt model described by (5.30) and (5.31)

between different behaviours are given by

$$R_0 = \left(\frac{b}{b-1} \right)^b \quad (5.30)$$

and

$$R_0 = \left(\frac{b}{b-2} \right)^b. \quad (5.31)$$

Let us consider existence of 2-cycles. The second iteration of the map

$$Hx = \frac{R_0 x}{(1+x)^b}$$

is given by

$$H(H(x)) = \frac{R_0^2 x (1+x)^{b^2-b}}{((1+x)^b + R_0 x)^b}$$

so that 2-cycles can be obtained from $H(H(x)) = x$ which can be rewritten as

$$x R_0^2 (1+x)^{b^2-b} = x ((1+x)^b + R_0 x)^b,$$

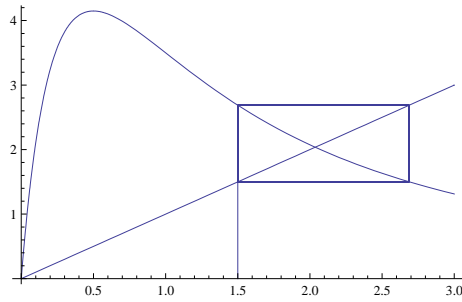


Fig. 5.13. 2-cycles for the Beverton-Holt model with $b = 3$ and $R_0 = 28$; see Eqn (5.31).

or, discarding the trivial equilibrium $x = 0$ and taking the b th root:

$$(1+x)R_0^{\frac{2}{b}} = (1+x)^b + R_0x.$$

Introducing the change of variables $z = 1 + x$, we see that we have to investigate existence of positive roots of

$$f(z) = z^b - z^{b-1}R_0^{\frac{2}{b}} + R_0z - R_0.$$

Clearly we have $f(R_0^{\frac{1}{b}}) = 0$ as any equilibrium of H is also an equilibrium of H^2 . First let us consider $1 < b < 2$ (the case $b = 1$ yields explicit solution (see Example ??) whereas the case $b = 2$ can be investigated directly and is referred to the tutorial problems).

We have

$$f'(z) = bz^{b-1} - (b-1)z^{b-2}R_0^{\frac{2}{b}} + R_0$$

and

$$f''(z) = (b-1)z^{b-3}(bz + (2-b)R_0^{\frac{2}{b}})$$

and we see that $f'' > 0$ for all $z > 0$. Furthermore, $f(0) = -R_0 < 0$. Hence, the region Ω bounded from the left by the axis $z = 0$ and lying above the graph of f for $z > 0$ is convex. Thus, the z axis, being transversal to the axis $z = 0$ cuts the boundary of Ω in exactly two points, one being $(0, 0)$ and the other $(R_0^{\frac{1}{b}}, 0)$. Hence, there are no additional equilibria of H^2 and therefore H does not have 2-cycles for $b \leq 2$.

Let us consider $b > 3$ (the case $b = 3$ is again referred to tutorials). In this case f has exactly one inflection point

$$z_i = \frac{b-2}{b} R_0^{\frac{2}{b}}$$

The fact that the equilibrium $x^* = R_0^{\frac{1}{b}} - 1$ loses stability at $R_0 = (b/b-2)^b$ suggests that a 2-cycle can appear when R_0 increases passing through this point. Let us first discuss the stable region $R_0 \leq (b/b-2)^b$. Then

$$z_i \leq \frac{b}{b-2} < 1,$$

that is, the inflection point occurs in the nonphysical region $x = z - 1 < 0$. For $z = 1$ we have $f(1) = 1 - R_0^{\frac{2}{b}} < 0$ and we can argue as above, using the line $z = 1$ instead of the axis $z = 0$. Thus, when the equilibrium $x^* = R_0^{\frac{1}{b}} - 1$ is stable, there are no 2-cycles. Let us consider the case with $R_0 > (b/b-2)^b$. At the equilibrium we find

$$\begin{aligned} f'(R_0^{\frac{1}{b}}) &= bR_0^{\frac{b-1}{b}} - (b-1)R_0^{\frac{b-2}{b}} R_0^{\frac{2}{b}} + R_0 \\ &= bR_0^{\frac{b-1}{b}} - (b-2)R_0 = R_0(bR_0^{-\frac{1}{b}} - (b-2)) \end{aligned}$$

and $f'(R_0^{\frac{1}{b}}) > 0$ provided $R_0 > (b/b-2)^b$. So, f takes negative values for $z > R_0^{\frac{1}{b}}$ but, on the other hand, $f(z)$

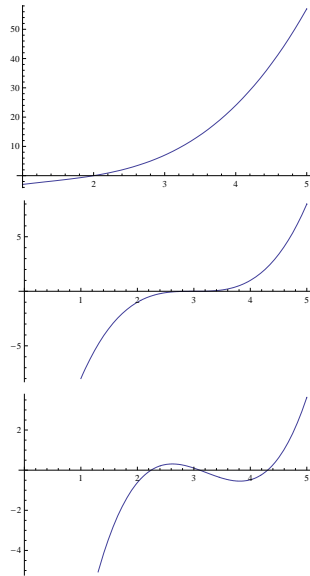


Fig. 5.14. Function f for $b = 3$ and, from top to bottom, $R_0 = 8, 27, 30$. Notice the emergence of 2-cycles represented here by new zeros of f besides $z = \sqrt[3]{R_0}$.

tends to $+\infty$ for $z \rightarrow \infty$ and therefore there must be $z^* > R_0^{\frac{1}{b}}$ for which $f(z^*) = 0$. Since $R_0^{\frac{1}{b}} - 1$ and 0 were the only equilibria of H , z^* must give a 2-cycle.

With much more, mainly computer aided, work we can establish that, as with the logistic equation, we obtain period doubling and transition to chaos.

Experimental results are in quite good agreement with the model. Most models fell into the stable region. It

is interesting to note that laboratory populations are usually less stable than the field ones. This is because scramble for resources is confined and more homogeneous and low density-independent mortality (high R_0). Also, it is obvious that high reproductive ratio R_0 and highly over-compensating density dependence (large b) are capable of provoking periodic or chaotic fluctuations in population density. This can be demonstrated mathematically (before the advent of mathematical theory of chaos it was assumed that these irregularities are of stochastic nature) and is observed in the fluctuations of the population of the Colorado beetle.

The question whether chaotic behaviour do exist in ecology is still an area of active debate. Observational time series are always finite and inherently noisy and it can be argued that regular models can be found to fit these data. However, several laboratory host-parasitoid systems do seem to exhibit chaos as good fits were obtained between the data and chaotic mathematical models.

6

From discrete to continuous models and back

Unless a given phenomenon occurs indeed at well defined and evenly spaced time intervals, usually it is up to us whether we describe it using difference or differential equations. Each choice has its advantages and disadvantages in the modelling process, however, both are closely intertwined. Indeed, as we have seen, continuous models are obtained using the same principles as corresponding discrete models. In fact, a discrete model, represented by a difference equation, is an intermediate step in deriving a relevant differential equation. Furthermore, since most interesting differential equations cannot be solved explicitly, we have to resort to numerical methods to deliver a testable solution and numerics involves discretization of the differential equations which usually leads to a difference equation which often different from the one we began with.

Thus, an important question is whether, under reasonable circumstances, discrete and continuous models are equivalent in the sense that they give the same solutions (or at least, solutions with the same qualitative features) and

whether there is a correspondence between continuous and discrete models of the same type.

6.1 Discretizing differential equations

There are several ways of discretization of differential equations. We shall discuss two commonly used methods.

6.1.1 The Euler method

The first one is the standard in numerical analysis practice of replacing the derivative by the difference quotient:

$$\frac{df}{dt} \approx \frac{f(t + \Delta t) - f(t)}{\Delta t}.$$

For instance, for the exponential growth equation

$$N' = rN,$$

this discretization gives

$$N(t + \Delta t) \approx N(t) + rN(t)\Delta t$$

or, denoting for a fixed t , $n(k) = N(t + k\Delta t)$

$$\begin{aligned} n(k + 1) &\approx N(t + (k + 1)\Delta t) = N(t + k\Delta t + \Delta t) \\ &\approx N(t + k\Delta t) + rN(t + k\Delta t)\Delta t \approx n(k) + trn(k)\Delta t \end{aligned}$$

we get a difference equation giving us (approximate) value of N at $t + k\Delta t$, $k = 1, 2, \dots$ provided the initial value at $k = 0$ is given. Note, however, that here we do not have any guarantee that at any time step $n(k) = N(t + k\Delta t)$.

6.1.2 The time-one map

The second method is based on the observation that solutions of autonomous differential equations display the so-called semigroup property: if $x(t, x_0)$ is the solution to the Cauchy problem

$$x' = g(x), \quad x(0) = x_0, \quad (6.1)$$

then

$$x(t_1 + t_2, x_0) = x(t_1, x(t_2, x_0)).$$

In other words, the process can be stopped and any time and re-started again using the final state of the first time interval as the initial state of the next time interval without changing the final output. The semigroup property sometimes is referred to as the causality property.

Using this property, we can write

$$x((n+1)\Delta t, x_0) = x(\Delta t, x(n\Delta t, x_0)). \quad (6.2)$$

This amounts to saying that the solution after $n+1$ time steps can be obtained as the solution after one time step with initial condition given as the solution after n time steps. In other words, denoting $x(n) = x(n\Delta t, x_0)$ we have

$$x(n+1) = f_{\Delta t}(x(n))$$

where by f_{Δ} we denoted the operation of getting solution of the Cauchy problem (6.1) at $t = \Delta t$ with the initial condition which appear as its argument.

We note that, contrary to the Euler method, the time-one map method is exact that is $x(n+1) = x(n\Delta, x_0)$ but its drawback is that we have to now the solution of (6.1) and thus its practical value is limited. We shall discuss these two methods on two examples.

In further discussion for simplicity we shall take $\Delta t = 1$.

6.1.3 Discrete and continuous exponential growth

Let us consider the Cauchy problem for the equation of exponential growth.

$$N' = rN, \quad N(0) = N_0$$

having the solution

$$N(t) = N_0 e^{rt}.$$

As we have seen above, the Euler discretization gives

$$n(k+1) - n(k) = rn(k)$$

with the solution

$$n(k) = (1+r)^k N_0$$

and, contrary to the remark made at the end of Subsection 6.1.1, for this model the Euler discretization gives a perfect agreement with the discrete model. However, one must remember to re-scale the growth rate from discrete to continuous using the formula $R_0 = 1 + r$.

On the other hand, consider the time-one discretization which amounts to assuming that we take census of the population in evenly spaced time moments $t_0 = 0, t_1 = 1, \dots, t_k = k, \dots$ so that

$$N(k) = e^{rk} N_0 = (e^r)^k N_0.$$

Comparing this equation with (1.17), we see that it corresponds to the discrete model with intrinsic growth rate

$$R_0 = e^r.$$

Thus we can state that if we observe a continuously growing population in discrete time intervals and the observed (discrete) intrinsic growth rate is R_0 , then the real (continuous) growth rate is given by $r = \ln(1 + R_0)$. However, the qualitative features are preserved as in the Euler discretization.

6.1.4 Logistic growth in discrete and continuous time

Consider the logistic differential equation

$$y' = ay(1 - y), \quad y(0) = y_0. \quad (6.3)$$

Euler discretization (with $\Delta t = 1$) gives

$$y(n+1) = y(n) + ay(n)(1 - y(n)) = (1+a)y(n) \left(1 - \frac{y(n)}{\frac{1+a}{a}}\right), \quad (6.4)$$

which is a discrete logistic equation. We have already solved (6.3) and we know that its solutions monotonically converge to the equilibrium $y = 1$. However, if we plot solutions to (6.4) with, say, $a = 4$, we obtain the picture presented in Fig. 6.1. Hence, in general it seems unlikely that we can use the Euler discretization as an approximation to the continuous model.

Let us, however, write down the complete Euler scheme:

$$y(n+1) = y(n) + a\Delta t y(n)(1 - y(n)), \quad (6.5)$$

where $y(n) = y(n\Delta t)$ and $y(0) = y_0$. Then

$$y(n+1) = (1 + a\Delta t)y(n) \left(1 - \frac{a\Delta t}{1 + a\Delta t}y(n)\right).$$

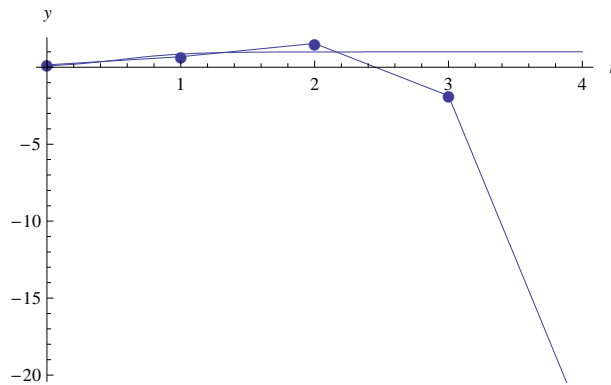


Fig. 6.1. Comparison of solutions to (6.3) and (6.4) with $a = 4$.

Substitution

$$x(n) = \frac{a\Delta t}{1 + a\Delta t}y(n) \tag{6.6}$$

reduces (6.5) to

$$x(n + 1) = \mu x(n)(1 - x(n)). \tag{6.7}$$

Thus, the parameter μ which controls the long time behaviour of solutions to the discrete equation (6.7) depends on Δt and, by choosing a suitably small Δt we can get solutions of (6.7) to mimic the behaviour of solutions to (6.3). Indeed, by taking $1 + a\Delta t < 3$ we obtain convergence of solutions $x(n)$ to the equilibrium

$$x = \frac{a\Delta t}{1 + a\Delta t}$$

which, reverting (6.6), gives the discrete approximation $y(n)$ which converges to 1, as the solution to (6.3). However,

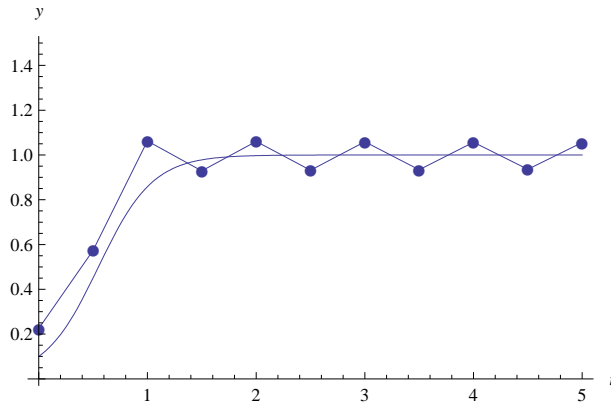


Fig. 6.2. Comparison of solutions to (6.3) with $a = 4$ and (6.7) with $\mu = 3$ ($\Delta t = 0.5$).

as seen on Fig 6.2, this convergence is not monotonic which shows that the approximation is rather poor. This can be remedied by taking $1+a\Delta t < 2$ in which case the qualitative features of $y(t)$ and $y(n)$ are the same, see Fig. 6.3).

We note that above problems can be also solved by introducing the so-called non-standard difference schemes which consists in replacing the derivatives and/or nonlinear terms by more sophisticated expressions which, though equivalent when the time step goes to 0 produce, nevertheless, qualitatively different discrete picture. In the case of the logistic equation such a non-standard scheme can be constructed replacing y^2 not by $y^2(n)$ but by $y(n)y(n + 1)$.

$$y(n + 1) = y(n) = a\Delta t(y(n) - y(n)y(n + 1)).$$

In general, such a substitution yields an implicit scheme but in our case the resulting recurrence can be solved for

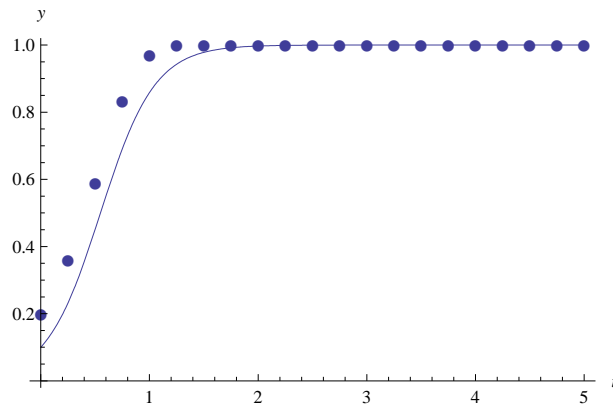


Fig. 6.3. Comparison of solutions to (6.3) with $a = 4$ and (6.7) with $\mu = 2$ ($\Delta t = 0.25$).

$y(n + 1)$ producing

$$y(n + 1) = \frac{(1 + a\Delta t)y(n)}{1 + a\Delta ty(n)}$$

and we recognize the Beverton-Holt-Hassel equation with $R_0 = 1 + a\Delta t$ (and $K = 1$).

Consider now the logistic equation

$$N' = rN \left(1 - \frac{N}{K} \right).$$

The first type of discretization immediately produces the discrete logistic equation (1.24)

$$N_{k+1} = N_k + rN_k \left(1 - \frac{N_k}{K} \right),$$

solutions of which, as we shall see later, behave in a dramatically different way than those of the continuous equation.

This is in contrast to the exponential growth equation discussed earlier.

To use the time-one map discretization, we re-write (4.14) as

$$N(t) = \frac{N_0 e^{rt}}{1 + \frac{e^{rt}-1}{K} N_0}.$$

which, upon denoting $e^r = R_0$ gives the time-one map

$$N(1, N_0) = \frac{N_0 R_0}{1 + \frac{R_0-1}{K} N_0},$$

which, according to the discussion above, yields the Beverton-Holt model

$$N_{k+1} = \frac{N_k R_0}{1 + \frac{R_0-1}{K} N_k},$$

with the discrete intrinsic growth rate related to the continuous one in the same way as in the exponential growth equation.

6.1.5 Discrete models of seasonally changing population

So far we have considered models in which laws of nature are independent of time. In many real processes we have to take into account phenomena which depend on time such as seasons of the year. The starting point of modelling is as before the balance equation. If we denote by $B(t)$, $D(t)$, $E(t)$ and $I(t)$ rates of birth, death, emigration and immigration, so that e.g. the number of births in time interval $[t_1, t_2]$ equals $\int_{t_1}^{t_2} B(s) ds$. Then, the change in the size of the popu-

lation in this interval is

$$N(t_2) - N(t_1) = \int_{t_1}^{t_2} (B(s) - D(s) + I(s) - E(s)) ds,$$

or, in differential form

$$\frac{dN(t)}{dt} = B(t) - D(t) + I(t) - E(t).$$

Processes of birth, death and emigration are often proportional to the size of the population and thus it makes sense to introduce *per capita* coefficients so that $B(t) = b(t)N(t)$, $D(t) = d(t)N(t)$, $E(t) = e(t)N(t)$. Typically, it would be unreasonable to assume that immigration is proportional to the number of the target population (possibly rather to the inverse unless we consider processes like gold rush), so that we leave $I(t)$ unchanged and thus write the rate equation as

$$\frac{dN(t)}{dt} = (b(t) - d(t) + e(t))N(t) + I(t). \quad (6.8)$$

This equation provides good description of small populations in which birth and death coefficients are not influenced by the size of the population.

Our interest is in populations in which the coefficients change periodically e.g. with seasons of the year. We start with closed populations; that is we do not consider emigration and immigration. Then we define $\lambda(t) = b(t) - d(t)$ to be the net growth rate of the population and assume that it is a periodic function with period T . Under this assumption we introduce the average growth rate of the population

by

$$\bar{\lambda} = \frac{1}{T} \int_0^T \lambda(t) dt. \quad (6.9)$$

Thus, let us consider the initial value problem

$$\frac{dN(t)}{dt} = \lambda(t)N(t), \quad N(t_0) = N_0, \quad (6.10)$$

where $\lambda(t)$ is a continuous periodic function with period T . Clearly, the solution is given by

$$N(t) = N_0 e^{\int_{t_0}^t \lambda(s) ds}. \quad (6.11)$$

It would be tempting to believe that a population with periodically changing growth rate also changes in a periodic way. However, we have

$$r(t+T) := \int_{t_0}^{t+T} \lambda(s) ds = \int_{t_0}^t \lambda(s) ds + \int_t^{t+T} \lambda(s) ds = r(t) + \int_0^T \lambda(s) ds = r(t) + \bar{\lambda}T$$

so that

$$N(t+T) = N(t)e^{\bar{\lambda}T}$$

and we do not have periodicity in the solution. However, we may provide a better description of the evolution. Let us try to find what is ‘missing’ in the function r so that it is not periodic. Assume that $\tilde{r}(t) = r(t) + \phi(t)$, where ϕ is as yet an unspecified function, is periodic hence

$$\tilde{r}(t+T) = r(t+T) + \phi(t+T) = r(t) + \bar{\lambda}T + \phi(t+T) = \tilde{r}(t) + \bar{\lambda}T + \phi(t+T) - \phi(t)$$

thus

$$\phi(t+T) = \phi(t) - \bar{\lambda}T.$$

This shows that $\psi = \phi'$ is a periodic function. To reconstruct ϕ from its periodic derivative, first we assume that the average of ψ is zero. Then $F(t) = \int_{t_0}^t \psi(s) ds$ is periodic. Indeed, $F(t+T) = \int_{t_0}^{t+T} \psi(s) ds = F(t) + \int_t^{t+T} \psi(s) ds = F(t) + \int_0^T \psi(s) ds = F(t)$. Next, if the average of ψ is $\bar{\psi}$, then $\psi - \bar{\psi}$ has zero average. Indeed,

$$\int_{t_0}^{t_0+T} (\psi(s) - \bar{\psi}) ds = T\bar{\psi} - T\bar{\psi} = 0$$

Hence

$$\int_{t_0}^t \psi(s) ds = g(t) + (t - t_0)\bar{\psi}$$

where $g(t)$ is a periodic function. Returning to function ϕ , we see that

$$\psi(t) = g(t) + c(t - t_0)$$

for some constant c and periodic function g . As we are interested in the simplest representation, we put $g(t) = 0$ and so $\psi(t)$ becomes a linear function and

$$-\bar{\lambda}T = \phi(t+T) - \phi(t) = c(t+T-t_0) - c(t-t_0)$$

and so $c = \bar{\lambda}$.

Using this result we write

$$N(t) = N_0 e^{\int_{t_0}^t \lambda(s) ds} = N_0 e^{\bar{\lambda}(t-t_0)} Q(t)$$

where

$$Q(t) = e^{\int_{t_0}^t \lambda(s) ds - \bar{\lambda}(t-t_0)} \quad (6.12)$$

is a periodic function.

In particular, if we observe the population in discrete time intervals of the length of the period T , we get

$$N(k) = N(t_0 + kT) = N_0 e^{\bar{\lambda}T} Q(t_0 + kT) = N_0 e^{\bar{\lambda}kT} Q(t_0) = N_0 [e^{\bar{\lambda}T}]^k,$$

which is the expected difference equation with growth rate given by $e^{\bar{\lambda}T}$.

6.2 A comparison of stability results for differential and difference equations

Let us consider a phenomenon in a static environment which can be described in both continuous and discrete time. In the first case we have an (autonomous) differential equation

$$y' = f(y), \quad y(0) = y_0, \quad (6.13)$$

and in the second case a difference equation

$$y(n+1) = g(y(n)), \quad y(0) = y_0. \quad (6.14)$$

In all considerations of this section we assume that both f and g are sufficiently regular functions so as not to have any problems with existence, uniqueness etc.

First we note that while in both cases y is the number of individuals in the population, the equations (6.13) and (6.14) refer to two different aspects of the process. In fact, while (6.13) describes the (instantaneous) rate of change of the population's size, (6.14) give the size of the population after each cycle. To be more easily comparable, (6.14)

should be written as

$$y(n+1)-y(n) = -y(n)+g(y(n)) =: \bar{f}(y(n)), \quad y(0) = y_0, \quad (6.15)$$

which would describe the rate of change of the population size per unit cycle. However, difference equations typically are written and analysed in the form (6.14).

Let us recall the general result describing dynamics of (6.13). As mentioned above, we assume that f is at least a Lipschitz continuous function on \mathbb{R} and the solutions exist for all t . An equilibrium solution is any solution $y(t) \equiv y$ satisfying $f(y) = 0$.

Theorem 6.1 *(i) If y_0 is not an equilibrium point, then $y(t)$ never equals an equilibrium point.*

(ii) All non-stationary solutions are either strictly decreasing or strictly increasing functions of t .

(iii) For any $y_0 \in \mathbb{R}$, the solution $y(t)$ either diverges to $+\infty$ or $-\infty$, or converges to an equilibrium point, as $t \rightarrow \infty$.

From this theorem it follows that if f has several equilibrium points, then the stationary solutions corresponding to these points divide the (t, y) plane into strips such that any solution remains always confined to one of them. If we look at this from the point of phase space and orbits, first we note that the phase space in the 1 dimensional case is the real line \mathbb{R} , divided by equilibrium points and thus orbits are open segments (possibly stretching to infinity) between equilibrium points.

Furthermore, we observe that if $f(y) > 0$, then the solution $y(t)$ is increasing at any point t when $y(t) = y$; con-

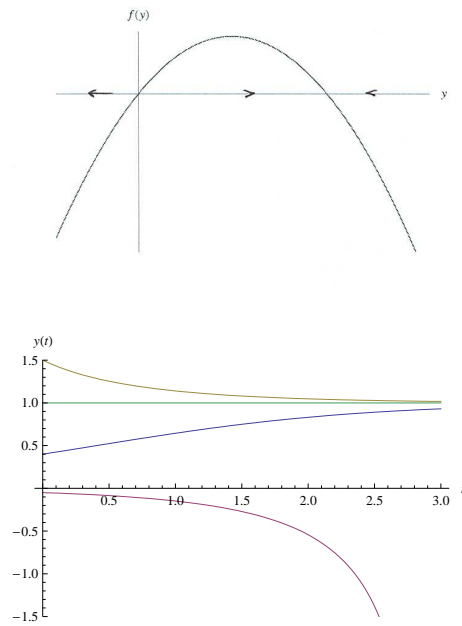


Fig. 6.4. Monotonic behaviour of solutions to (6.13) depends on the right hand side f of the equation.

versely, $f(y) < 0$ implies that the solution $y(t)$ is decreasing when $y(t) = y$. This also implies that any equilibrium point y^* with $f'(y^*) < 0$ is asymptotically stable and with $f'(y^*) > 0$ is unstable; there are no stable, but not asymptotically stable, equilibria.

If we look now at the difference equation (6.14), then at first we note some similarities. Equilibria are defined as

$$g(y) = y,$$

(or $\bar{f}(y) = 0$) and, as in the continuous case we compared f with zero, in the discrete case we compare $g(x)$ with x : $g(y) > y$ means that $y(n+1) = g(y(n)) > y(n)$ so that the iterates are increasing while if $g(x) < x$, then they are decreasing. Also, stability of equilibria is characterized in a similar way: if $|g'(y^*)| < 1$, then y^* asymptotically stable and if $|g'(y^*)| > 1$, then y^* unstable. In fact, if $g'(y^*) > 0$, then we have exact equivalence: y^* is stable provided $\bar{f}'(y^*) < 0$ and unstable if $\bar{f}'(y^*) > 0$. Indeed, in such a case, if we start on a one side of an equilibrium y^* , then no iteration can overshoot this equilibrium as for, say $y < y^*$ we have $f(y) < f(y^*) = y^*$. Thus, as in the continuous case, the solutions are confined to intervals between successive equilibria.

However, similarities end here as the dynamics of difference equation is much richer than that of the corresponding differential equation as the behaviour of the solution near an equilibrium is also governed the sign of g itself.

First, contrary to Theorem 6.1 (i), solutions can reach an equilibrium in a finite time, as demonstrated in Example 6.1.

In differential equations, an equilibrium cannot be reached in finite time. Difference equations do not share this property. This leads to the definition:

Definition 6.1 *A point x in the domain of f is said to be an eventual equilibrium of (5.4) if there is an equilibrium point x^* of (5.4) and a positive integer r such that $x^* = f^r(x)$ and $f^{r-1}(x) \neq x^*$.*

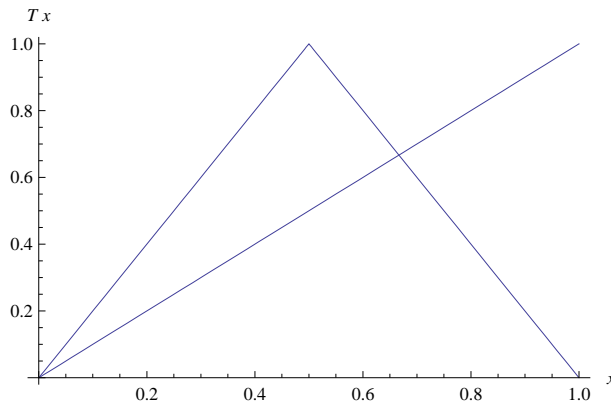


Fig. 6.5. The tent map

Example 6.1 The Tent Map. Consider

$$x(n+1) = Tx(n)$$

where

$$T(x) = \begin{cases} 2x & \text{for } 0 \leq x \leq 1/2, \\ 2(1-x) & \text{for } 1/2 < x \leq 1. \end{cases}$$

There are two equilibrium points, 0 and $2/3$. Looking for eventual equilibria is not as simple. Taking $x(0) = 1/8$, we find $x(1) = 1/4$, $x(2) = 1/2$, $x(3) = 1$ and $x(4) = 0$, and hence $1/8$ (as well as $1/4, 1/2$ and 1) are eventual equilibria. It can be checked that all points of the form $x = n/2^k$, where $n, k \in \mathbb{N}$ satisfy $0 < n/2^k < 1$ are eventual equilibria.

Further, recalling Remark 5.3, we see that if $-1 < g'(y^*) < 0$, then the solution can overshoot the equilibrium creating damped oscillations towards equilibrium, whereas any re-

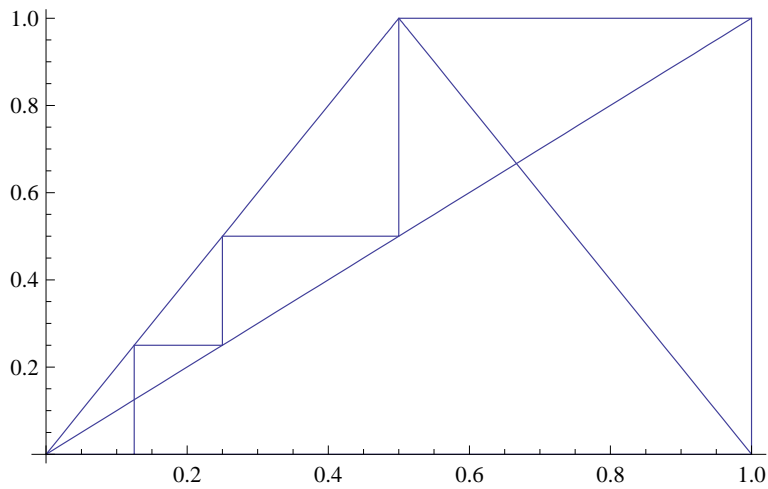


Fig. 6.6. Eventual equilibrium $x = 1/8$ for the tent map.

versal of the direction of motion is impossible in autonomous scalar differential equations. Also, as we have seen, difference equations may have periodic solutions which are precluded from occurring in the continuous case. Finally, no chaotic behaviour can occur in scalar differential equations (partly because they do not admit periodic solutions abundance of which is a signature of chaos). In fact, it can be proved that chaos in differential equations may occur only if the dimension of the state space exceeds 3.

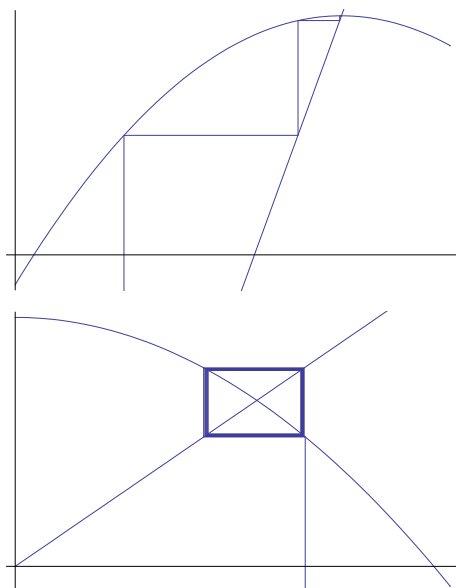


Fig. 6.7. Change of the type of convergence to the equilibrium from monotonic if $0 < g'(y^*) < 1$ to oscillatory for $-1 < g'(y^*) < 0$.

7

Simultaneous systems of equations and higher order equations

7.1 Systems of equations

7.1.1 Why systems?

Two possible generalizations of the first order scalar equation

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

are: a differential equation of a higher order

$$y^{(n)} = F(t, y', y'', \dots, y^{(n-1)}) = 0, \quad (7.1)$$

(where, for simplicity, we consider only equations solved with respect to the highest derivative), or a system of first order equations, that is,

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}) \quad (7.2)$$

where,

$$\mathbf{y}(t) = \begin{pmatrix} y_1(t) \\ \vdots \\ y_n(t) \end{pmatrix},$$

and

$$\mathbf{f}(t, \mathbf{y}) = \begin{pmatrix} f_1(t, y_1, \dots, y_n) \\ \vdots \\ f_n(t, y_1, \dots, y_n) \end{pmatrix},$$

is a nonlinear function of t and \mathbf{y} . It turns out that, at least from the theoretical point of view, there is no need to consider these two cases separately as any equation of a higher order can be always written as a system (the converse, in general, is not true). To see how this can be accomplished, we introduce new unknown variables $z_1(t) = y(t)$, $z_2(t) = y'(t)$, $z_n = y^{(n-1)}(t)$ so that $z_1'(t) = y'(t) = z_2(t)$, $z_2'(t) = y''(t) = z_3(t)$, ... and (7.1) converts into

$$\begin{aligned} z_1' &= z_2, \\ z_2' &= z_3, \\ &\vdots \\ z_n' &= F(t, z_1, \dots, z_n) \end{aligned}$$

Clearly, solving this system, we obtain simultaneously the solution of (7.1) by taking $y(t) = z_1(t)$.

7.1.2 Linear systems

At the beginning we shall consider only systems of first order differential equations that are solved with respect to the derivatives of all unknown functions. The systems we deal with in this section are linear, that is, they can be

Simultaneous systems of equations and higher order equations

written as

$$\begin{aligned}y_1' &= a_{11}y_1 + a_{12}y_2 + \dots + a_{1n}y_n + g_1(t), \\ \vdots & \quad \quad \quad \vdots, \\ y_n' &= a_{n1}y_1 + a_{n2}y_2 + \dots + a_{nn}y_n + g_n(t),\end{aligned}\tag{7.3}$$

where y_1, \dots, y_n are unknown functions, a_{11}, \dots, a_{nn} are constant coefficients and $g_1(t), \dots, g_n(t)$ are known continuous functions. If $g_1 = \dots = g_n = 0$, then the corresponding system (7.3) is called the associated homogeneous system. The structure of (7.3) suggest that a more economical way of writing it is to use the vector-matrix notation. Denoting $\mathbf{y} = (y_1, \dots, y_n)$, $\mathbf{g} = (g_1, \dots, g_n)$ and $\mathcal{A} = \{a_{ij}\}_{1 \leq i, j \leq n}$, that is

$$\mathcal{A} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix},$$

we can write (7.3) in a more concise way as

$$\mathbf{y}' = \mathcal{A}\mathbf{y} + \mathbf{g}.\tag{7.4}$$

Here we have n unknown functions and the system involves first derivative of each of them so that it is natural to consider (7.4) in conjunction with the following initial conditions

$$\mathbf{y}(t_0) = \mathbf{y}^0,\tag{7.5}$$

or, in the expanded form,

$$y_1(t_0) = y_1^0, \dots, y_n(t_0) = y_n^0,\tag{7.6}$$

where t_0 is a given argument and $\mathbf{y}^0 = (y_1^0, \dots, y_n^0)$ is a given vector.

As we noted in the introduction, systems of first order equations are closely related to higher order equations. In particular, any n th order linear equation

$$y^{(n)} + a_{n-1}y^{(n-1)} + \dots + a_1y' + a_0y = g(t) \quad (7.7)$$

can be written as a linear system of n first order equations by introducing new variables $z_1 = y$, $z_2 = y' = z_1'$, $z_3 = y'' = z_2'$, \dots , $z_n = y^{(n-1)} = z_{n-1}'$ so that $z_n' = y^{(n)}$ and (7.7) turns into

$$\begin{aligned} z_1' &= z_2, \\ z_2' &= z_3, \\ &\vdots \\ z_n' &= -a_{n-1}z_n - a_{n-2}z_{n-1} - \dots - a_0z_1 + g(t). \end{aligned}$$

Note that if (7.7) was supplemented with the initial conditions $y(t_0) = y_0, y'(t_0) = y_1, \dots, y^{(n-1)}(t_0) = y_{n-1}$, then these conditions will become natural initial conditions for the system as $z_1(t_0) = y_0, z_2(t_0) = y_1, \dots, z_n(t_0) = y_{n-1}$. Therefore, all the results we shall prove here are relevant also for n th order equations.

In some cases, especially when faced with simple systems of differential equations, it pays to revert the procedure and to transform a system into a single, higher order, equation rather than to apply directly a heavy procedure for full systems. We illustrate this remark in the following example.

Example 7.1 Consider the system

$$\begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2, \\ y_2' &= a_{21}y_1 + a_{22}y_2. \end{aligned} \quad (7.8)$$

Simultaneous systems of equations and higher order equations

Firstly, note that if either a_{12} or a_{21} equal zero, then the equations are uncoupled, e.g., if $a_{12} = 0$, then the first equation does not contain y_2 and can be solved for y_1 and this solution can be inserted into the second equation which then becomes a first order nonhomogeneous equation for y_2 .

Assume then that $a_{12} \neq 0$. We proceed by eliminating y_2 from the first equation. Differentiating it, we obtain

$$y_1'' = a_{11}y_1' + a_{12}y_2',$$

so that, using the second equation,

$$y_1'' = a_{11}y_1' + a_{12}(a_{21}y_1 + a_{22}y_2).$$

To get rid of the remaining y_2 , we use the first equation once again obtaining

$$y_2 = a_{12}^{-1}(y_1' - a_{11}y_1), \quad (7.9)$$

$$y_1'' = (a_{11} + a_{22})y_1' + (a_{12}a_{21} - a_{22}a_{11})y_1$$

which is a second order linear equation. If we are able to solve it to obtain y_1 , we use again (7.9) to obtain y_2 .

However, for larger systems this procedure becomes quite cumbersome unless the matrix \mathcal{A} of coefficients has a simple structure.

7.1.3 Algebraic properties of systems

In this subsection we shall prove several results related to the algebraic structure of the set of solutions to

$$\mathbf{y}' = \mathcal{A}\mathbf{y}, \quad \mathbf{y}(t_0) = \mathbf{y}^0. \quad (7.10)$$

An extremely important rôle here is played by the uniqueness of solutions. In Section 4.2 we discussed Picard's theorem, Theorem 4.2, that dealt with the existence and uniqueness of solution to the Cauchy problem

$$y' = f(t, y), \quad y(t_0) = y_0$$

where y and f were scalar valued functions. It turns out that this theorem can be easily generalized to the vector case, that is to the case where f is a vector valued function $\mathbf{f}(t, \mathbf{y})$ of a vector valued argument \mathbf{y} . In particular, it can be applied to the case when $\mathbf{f}(t, \mathbf{y}) = \mathcal{A}\mathbf{y} + \mathbf{g}(t)$. Thus, we can state

Theorem 7.1 *Let $\mathbf{g}(t)$ be a continuous function from \mathbb{R} to \mathbb{R}^n . Then there exists one and only one solution of the initial value problem*

$$\mathbf{y}' = \mathcal{A}\mathbf{y} + \mathbf{g}(t), \quad \mathbf{y}(t_0) = \mathbf{y}^0. \quad (7.11)$$

Moreover, this solution exists for all $t \in \mathbb{R}$.

One of the important implications of this theorem is that if \mathbf{y} is a non-trivial, that is, not identically equal to zero, solution to the homogeneous equation

$$\mathbf{y}' = \mathcal{A}\mathbf{y}, \quad (7.12)$$

then $\mathbf{y}(t) \neq 0$ for any t . In fact, as $\mathbf{y}^* \equiv 0$ is a solution to (7.12) and by definition $\mathbf{y}^*(\bar{t}) = 0$ for any \bar{t} , the existence of other solution satisfying $\mathbf{y}(\bar{t}) = 0$ for some \bar{t} would violate Theorem 7.1.

Let us denote by \mathbf{X} the set of all solutions to (7.12). Due to linearity of differentiation and multiplication by \mathcal{A} , it is easy to see that \mathbf{X} is a vector space. Moreover

Theorem 7.2 *The dimension of \mathbf{X} is equal to n .*

Proof. We must exhibit a basis of \mathbf{X} that contains exactly n elements. Thus, let $\mathbf{z}_j(t)$, $j = 1, \dots, n$ be solutions of special Cauchy problems

$$\mathbf{y}' = \mathcal{A}\mathbf{y}, \quad \mathbf{y}(0) = \mathbf{e}_j, \quad (7.13)$$

where $\mathbf{e}_j = (0, 0, \dots, 1, \dots, 0)$ with 1 at j th place is a versor of the coordinate system. To determine whether the set $\{\mathbf{z}_1(t), \dots, \mathbf{z}_n(t)\}$ is linearly dependent, we ask whether from

$$c_1\mathbf{z}_1(t) + \dots + c_n\mathbf{z}_n(t) = 0,$$

it follows that $c_1 = \dots = c_n = 0$. If the linear combination vanishes for any t , then it must vanish in particular for $t = 0$. Thus, using the initial conditions $\mathbf{z}_j(0) = \mathbf{e}_j$ we see that we would have

$$c_1\mathbf{e}_1 + \dots + c_n\mathbf{e}_n = 0,$$

but since the set $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ is a basis in \mathbb{R}^n , we see that necessarily $c_1 = \dots = c_n = 0$. Thus $\{\mathbf{z}_1(t), \dots, \mathbf{z}_n(t)\}$ is linearly independent and $\dim \mathbf{X} \geq n$. To show that $\dim \mathbf{X} = n$ we must show that \mathbf{X} is spanned by $\{\mathbf{z}_1(t), \dots, \mathbf{z}_n(t)\}$, that is, that any solution $\mathbf{y}(t)$ can be written as

$$\mathbf{y}(t) = c_1\mathbf{z}_1(t) + \dots + c_n\mathbf{z}_n(t)$$

for some constants c_1, \dots, c_n . Let $\mathbf{y}(t)$ be any solution to (7.12) and define $\mathbf{y}^0 = \mathbf{y}(0) \in \mathbb{R}^n$. Since $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ is a basis \mathbb{R}^n , there are constants c_1, \dots, c_n such that

$$\mathbf{y}^0 = c_1\mathbf{e}_1 + \dots + c_n\mathbf{e}_n.$$

Consider

$$\mathbf{x}(t) = c_1 \mathbf{z}_1(t) + \dots + c_n \mathbf{z}_n(t).$$

Clearly, $\mathbf{x}(t)$ is a solution to (7.12), as a linear combination of solutions, and $\mathbf{x}(0) = c_1 \mathbf{e}_1 + \dots + c_n \mathbf{e}_n = \mathbf{y}^0 = \mathbf{y}(0)$. Thus, $\mathbf{x}(t)$ and $\mathbf{y}(t)$ are both solutions to (7.12) satisfying the same initial condition and therefore $\mathbf{x}(t) = \mathbf{y}(t)$ by Theorem 7.1. Hence,

$$\mathbf{y}(t) = c_1 \mathbf{z}_1(t) + \dots + c_n \mathbf{z}_n(t).$$

and the set $\{\mathbf{z}_1(t), \dots, \mathbf{z}_n(t)\}$ is a basis for \mathbf{X} . ■

Next we present a convenient way of determining whether solutions to (7.12) are linearly independent.

Theorem 7.3 *Let $\mathbf{y}_1, \dots, \mathbf{y}_k$ be k linearly independent solutions of $\mathbf{y}' = \mathcal{A}\mathbf{y}$ and let $t_0 \in \mathbb{R}$ be an arbitrary number. Then, $\{\mathbf{y}_1(t), \dots, \mathbf{y}_k(t)\}$ form a linearly independent set of functions if and only if $\{\mathbf{y}_1(t_0), \dots, \mathbf{y}_k(t_0)\}$ is a linearly independent set of vectors in \mathbb{R}^n .*

Proof. If $\{\mathbf{y}_1(t), \dots, \mathbf{y}_k(t)\}$ are linearly dependent functions, then there exist constants c_1, \dots, c_k , not all zero, such that for all t

$$c_1 \mathbf{y}_1(t) + \dots + c_k \mathbf{y}_k(t) = \mathbf{0}.$$

Taking this at a particular value of t , $t = t_0$, we obtain that

$$c_1 \mathbf{y}_1(t_0) + \dots + c_k \mathbf{y}_k(t_0) = \mathbf{0},$$

with not all c_i vanishing. Thus the set $\{\mathbf{y}_1(t_0), \dots, \mathbf{y}_k(t_0)\}$ is a set of linearly dependent vectors in \mathbb{R}^n .

Conversely, suppose that $\{\mathbf{y}_1(t_0), \dots, \mathbf{y}_k(t_0)\}$ is a linearly dependent set of vectors. Then for some constants

$$c_1 \mathbf{y}_1(t_0) + \dots + c_n \mathbf{y}_k(t_0) = 0,$$

where not all c_i are equal to zero. Taking these constants we construct the function

$$\mathbf{y}(t) = c_1 \mathbf{y}_1(t) + \dots + c_n \mathbf{y}_k(t),$$

which is a solution to (7.12) as a linear combination of solutions. However, since $\mathbf{y}(t_0) = 0$, by the uniqueness theorem we obtain that $\mathbf{y}(t) = 0$ for all t so that $\{\mathbf{y}_1(t), \dots, \mathbf{y}_k(t)\}$ is a linearly dependent set of functions. ■

Remark 7.1 *To check whether a set of n vectors of \mathbb{R}^n is linearly independent, we can use the determinant test: $\{\mathbf{y}_1, \dots, \mathbf{y}_k\}$ is linearly independent if and only if*

$$\det\{\mathbf{y}_1, \dots, \mathbf{y}_k\} = \begin{vmatrix} y_1^1 & \dots & y_1^n \\ \vdots & & \vdots \\ y_n^1 & \dots & y_n^n \end{vmatrix} \neq 0.$$

If $\{\mathbf{y}_1(t), \dots, \mathbf{y}_k(t)\}$ is a set of solution of the homogeneous system, then the determinant

$$\det\{\mathbf{y}_1(t), \dots, \mathbf{y}_k(t)\} = \begin{vmatrix} y_1^1(t) & \dots & y_1^n(t) \\ \vdots & & \vdots \\ y_n^1(t) & \dots & y_n^n(t) \end{vmatrix}$$

is called wronskian. The theorems proved above can be rephrased by saying that the wronskian is non-zero if it is constructed with independent solutions of a system of equations and, in such a case, it is non-zero if and only if it is non-zero at some point.

Example 7.2 Consider the system of differential equations

$$\begin{aligned}y_1' &= y_2, \\y_2' &= -y_1 - 2y_2,\end{aligned}\tag{7.14}$$

or, in matrix notation

$$\mathbf{y}' = \begin{pmatrix} 0 & 1 \\ -1 & -2 \end{pmatrix} \mathbf{y}.$$

Let us take two solutions:

$$\mathbf{y}^1(t) = (y_1^1(t), y_2^1(t)) = (\phi(t), \phi'(t)) = (e^{-t}, -e^{-t}) = e^{-t}(1, -1)$$

and

$$\mathbf{y}^2(t) = (y_1^2(t), y_2^2(t)) = (\psi(t), \psi'(t)) = (te^{-t}, (1-t)e^{-t}) = e^{-t}(t, 1-t).$$

To check whether these are linearly independent solutions to the system and thus whether they span the space of all solutions, we use Theorem 7.3 and check the linear dependence of vectors $\mathbf{y}^1(0) = (1, -1)$ and $\mathbf{y}^2(0) = (0, 1)$. Using e.g. the determinant test for linear dependence we evaluate

$$\begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = 1 \neq 0,$$

thus the vectors are linearly independent. Consequently, all solutions to (7.14) can be written in the form

$$\mathbf{y}(t) = C_1 \begin{pmatrix} e^{-t} \\ -e^{-t} \end{pmatrix} + C_2 \begin{pmatrix} te^{-t} \\ (1-t)e^{-t} \end{pmatrix} = \begin{pmatrix} (C_1 + C_2t)e^{-t} \\ (C_2 - C_1 - C_2t)e^{-t} \end{pmatrix}.$$

Assume now that we are given this $\mathbf{y}(t)$ as a solution to the system. The system is equivalent to the second order equation

$$y'' + 2y' + y = 0\tag{7.15}$$

under identification $y(t) = y_1(t)$ and $y'(t) = y_2(t)$. How can we recover the general solution to (7.15) from $\mathbf{y}(t)$? Remembering that y solves (7.15) if and only if $\mathbf{y}(t) = (y_1(t), y_2(t)) = (y(t), y'(t))$ solves the system (7.14), we see that the general solution to (7.15) can be obtained by taking first components of the solution of the associated system (7.14). We also note the fact that if $\mathbf{y}^1(t) = (y_1^1(t), y_2^1(t)) = (y^1(t), \frac{dy^1}{dt}(t))$ and $\mathbf{y}^2(t) = (y_1^2(t), y_2^2(t)) = (y^2(t), \frac{dy^2}{dt}(t))$ are two linearly independent solutions to (7.14), then $y^1(t)$ and $y^2(t)$ are linearly independent solutions to (7.15). In fact, otherwise we would have $y^1(t) = Cy^2(t)$ for some constant C and therefore also $\frac{dy^1}{dt}(t) = C\frac{dy^2}{dt}(t)$ so that the wronskian, having the second column as a scalar multiple of the first one, would be zero, contrary to the assumption that $\mathbf{y}^1(t)$ and $\mathbf{y}^2(t)$ are linearly independent.

7.1.4 The eigenvalue-eigenvector method of finding solutions

We start with a brief survey of eigenvalues and eigenvectors of matrices. Let \mathcal{A} be an $n \times n$ matrix. We say that a number λ (real or complex) is an *eigenvalue* of \mathcal{A} if there exist a non-zero solution of the equation

$$\mathcal{A}\mathbf{v} = \lambda\mathbf{v}. \quad (7.16)$$

Such a solution is called an *eigenvector* of \mathcal{A} . The set of eigenvectors corresponding to a given eigenvalue is a vector subspace. Eq. (7.16) is equivalent to the homogeneous system $(\mathcal{A} - \lambda\mathcal{I})\mathbf{v} = \mathbf{0}$, where \mathcal{I} is the identity matrix, therefore λ is an eigenvalue of \mathcal{A} if and only if the determinant of \mathcal{A}

satisfies

$$\det(\mathcal{A} - \lambda \mathcal{I}) = \begin{vmatrix} a_{11} - \lambda & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} - \lambda \end{vmatrix} = 0. \quad (7.17)$$

Evaluating the determinant we obtain a polynomial in λ of degree n . This polynomial is also called the characteristic polynomial of the system (7.3) (if (7.3) arises from a second order equation, then this is the same polynomial as the characteristic polynomial of the equation). We shall denote this polynomial by $p(\lambda)$. From algebra we know that there are exactly n , possibly complex, roots of $p(\lambda)$. Some of them may be multiple, so that in general $p(\lambda)$ factorizes into

$$p(\lambda) = (\lambda_1 - \lambda)^{n_1} \dots (\lambda_k - \lambda)^{n_k}, \quad (7.18)$$

with $n_1 + \dots + n_k = n$. It is also worthwhile to note that since the coefficients of the polynomial are real, then complex roots appear always in conjugate pairs, that is, if $\lambda_j = \xi_j + i\omega_j$ is a characteristic root, then so is $\bar{\lambda}_j = \xi_j - i\omega_j$. Thus, eigenvalues are roots of the characteristic polynomial of \mathcal{A} . The exponent n_i appearing in the factorization (7.18) is called the *algebraic multiplicity* of λ_i . For each eigenvalue λ_i there corresponds an eigenvector \mathbf{v}_i and eigenvectors corresponding to distinct eigenvalues are linearly independent. The set of all eigenvectors corresponding to λ_i spans a subspace, called the *eigenspace* corresponding to λ_i which we will denote by E_{λ_i} . The dimension of E_{λ_i} is called the *geometric multiplicity* of λ_i . In general, algebraic and geometric multiplicities are different with geometric multiplicity being at most equal to the algebraic one. Thus, in partic-

ular, if λ_i is a single root of the characteristic polynomial, then the eigenspace corresponding to λ_i is one-dimensional.

If the geometric multiplicities of eigenvalues add up to n , that is, if we have n linearly independent eigenvectors, then these eigenvectors form a basis for \mathbb{R}^n . In particular, this happens if all eigenvalues are single roots of the characteristic polynomial. If this is not the case, then we do not have sufficiently many eigenvectors to span \mathbb{R}^n and if we need a basis for \mathbb{R}^n , then we have to find additional linearly independent vectors. A procedure that can be employed here and that will be very useful in our treatment of systems of differential equations is to find solutions to equations of the form $(\mathcal{A} - \lambda_i \mathcal{I})^k \mathbf{v} = 0$ for $1 < k \leq n_i$, where n_i is the algebraic multiplicity of λ_i . Precisely speaking, if λ_i has algebraic multiplicity n_i and if

$$(\mathcal{A} - \lambda_i \mathcal{I}) \mathbf{v} = 0$$

has only $\nu_i < n_i$ linearly independent solutions, then we consider the equation

$$(\mathcal{A} - \lambda_i \mathcal{I})^2 \mathbf{v} = 0.$$

It follows that all the solutions of the preceding equation solve this equation but there is at least one more independent solution so that we have at least $\nu_i + 1$ independent vectors (note that these new vectors are no longer eigenvectors). If the number of independent solutions is still less than n_i , we consider

$$(\mathcal{A} - \lambda_i \mathcal{I})^3 \mathbf{v} = 0,$$

and so on, till we get a sufficient number of them. Note, that to make sure that in the step j we select solutions that

are independent of the solutions obtained in step $j - 1$ it is enough to find solutions to $(\mathcal{A} - \lambda_i \mathcal{I})^j \mathbf{v} = 0$ that satisfy $(\mathcal{A} - \lambda_i \mathcal{I})^{j-1} \mathbf{v} \neq 0$.

Now we show how to apply the concepts discussed above to solve systems of differential equations. Consider again the homogeneous system

$$\mathbf{y}' = \mathcal{A}\mathbf{y}. \quad (7.19)$$

Our goal is to find n linearly independent solutions of (7.19). We have seen that solutions of the form $e^{\lambda t}$ play a basic rôle in solving first order linear equations so let us consider $\mathbf{y}(t) = e^{\lambda t} \mathbf{v}$ for some vector $\mathbf{v} \in \mathbb{R}^n$. Since

$$\frac{d}{dt} e^{\lambda t} \mathbf{v} = \lambda e^{\lambda t} \mathbf{v}$$

and

$$\mathcal{A}(e^{\lambda t} \mathbf{v}) = e^{\lambda t} \mathcal{A}\mathbf{v}$$

as $e^{\lambda t}$ is a scalar, $\mathbf{y}(t) = e^{\lambda t} \mathbf{v}$ is a solution to (7.19) if and only if

$$\mathcal{A}\mathbf{v} = \lambda \mathbf{v}. \quad (7.20)$$

Thus $\mathbf{y}(t) = e^{\lambda t} \mathbf{v}$ is a solution if and only if \mathbf{v} is an eigenvector of \mathcal{A} corresponding to the eigenvalue λ .

Thus, for each eigenvector \mathbf{v}^j of \mathcal{A} with eigenvalue λ_j we have a solution $\mathbf{y}^j(t) = e^{\lambda_j t} \mathbf{v}^j$. By Theorem 7.3 these solutions are linearly independent if and only if the eigenvectors \mathbf{v}^j are linearly independent in \mathbb{R}^n . Thus, if we can find n linearly independent eigenvectors of \mathcal{A} with eigenvalues $\lambda_1, \dots, \lambda_n$ (not necessarily distinct), then the general solution of (7.19) is of the form

$$\mathbf{y}(t) = C_1 e^{\lambda_1 t} \mathbf{v}^1 + \dots + C_n e^{\lambda_n t} \mathbf{v}^n. \quad (7.21)$$

Distinct real eigenvalues

The simplest situation is of course if the characteristic polynomial $p(\lambda)$ has n distinct roots, that is, all roots are single and in this case the eigenvectors corresponding to different eigenvalues (roots) are linearly independent, as we mentioned earlier. However, this can also happen if some eigenvalues are multiple ones but the algebraic and geometric multiplicity of each is the same. In this case to each root of multiplicity n_1 there correspond n_1 linearly independent eigenvectors.

Example 7.3 *Find the general solution to*

$$\mathbf{y}' = \begin{pmatrix} 1 & -1 & 4 \\ 3 & 2 & -1 \\ 2 & 1 & -1 \end{pmatrix} \mathbf{y}.$$

To obtain the eigenvalues we calculate the characteristic polynomial

$$\begin{aligned} p(\lambda) &= \det(\mathcal{A} - \lambda\mathcal{I}) = \begin{vmatrix} 1 - \lambda & -1 & 4 \\ 3 & 2 - \lambda & -1 \\ 2 & 1 & -1 - \lambda \end{vmatrix} \\ &= -(1 + \lambda)(1 - \lambda)(2 - \lambda) + 12 + 2 - 8(2 - \lambda) + (1 - \lambda) - 3(1 + \lambda) \\ &= -(1 + \lambda)(1 - \lambda)(2 - \lambda) + 4\lambda - 4 = (1 - \lambda)(\lambda - 3)(\lambda + 2), \end{aligned}$$

so that the eigenvalues of \mathcal{A} are $\lambda_1 = 1$, $\lambda_2 = 3$ and $\lambda_3 = -2$. All the eigenvalues have algebraic multiplicity 1 so that they should give rise to 3 linearly independent eigenvectors.

(i) $\lambda_1 = 1$: we seek a nonzero vector \mathbf{v} such that

$$(\mathcal{A} - 1\mathcal{I})\mathbf{v} = \begin{pmatrix} 0 & -1 & 4 \\ 3 & 1 & -1 \\ 2 & 1 & -2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Thus

$$-v_2 + 4v_3 = 0, \quad 3v_1 + v_2 - v_3 = 0, \quad 2v_1 + v_2 - 2v_3 = 0$$

and we get $v_2 = 4v_3$ and $v_1 = -v_3$ from the first two equations and the third is automatically satisfied. Thus we obtain the eigenspace corresponding to $\lambda_1 = 1$ containing all the vectors of the form

$$\mathbf{v}^1 = C_1 \begin{pmatrix} -1 \\ 4 \\ 1 \end{pmatrix}$$

where C_1 is any constant, and the corresponding solutions

$$\mathbf{y}^1(t) = C_1 e^t \begin{pmatrix} -1 \\ 4 \\ 1 \end{pmatrix}.$$

(ii) $\lambda_2 = 3$: we seek a nonzero vector \mathbf{v} such that

$$(\mathcal{A} - 3\mathcal{I})\mathbf{v} = \begin{pmatrix} -2 & -1 & 4 \\ 3 & -1 & -1 \\ 2 & 1 & -4 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Hence

$$-2v_1 - v_2 + 4v_3 = 0, \quad 3v_1 - v_2 - v_3 = 0, \quad 2v_1 + v_2 - 4v_3 = 0.$$

Solving for v_1 and v_2 in terms of v_3 from the first two equations gives $v_1 = v_3$ and $v_2 = 2v_3$. Consequently, vectors of the form

$$\mathbf{v}^2 = C_2 \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$$

Simultaneous systems of equations and higher order equations

are eigenvectors corresponding to the eigenvalue $\lambda_2 = 3$ and the function

$$\mathbf{y}^2(t) = e^{3t} \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$$

is the second solution of the system.

(iii) $\lambda_3 = -2$: We have to solve

$$(\mathcal{A} + 2\mathcal{I})\mathbf{v} = \begin{pmatrix} 3 & -1 & 4 \\ 3 & 4 & -1 \\ 2 & 1 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Thus

$$3v_1 - v_2 + 4v_3 = 0, \quad 3v_1 + 4v_2 - v_3 = 0, \quad 2v_1 + v_2 + v_3 = 0.$$

Again, solving for v_1 and v_2 in terms of v_3 from the first two equations gives $v_1 = -v_3$ and $v_2 = v_3$ so that each vector

$$\mathbf{v}^3 = C_3 \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$$

is an eigenvector corresponding to the eigenvalue $\lambda_3 = -2$. Consequently, the function

$$\mathbf{y}^3(t) = e^{-2t} \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$$

is the third solution of the system. These solutions are linearly independent since the vectors $\mathbf{v}^1, \mathbf{v}^2, \mathbf{v}^3$

are linearly independent as eigenvectors corresponding to distinct eigenvalues. Therefore, every solution is of the form

$$\mathbf{y}(t) = C_1 e^t \begin{pmatrix} -1 \\ 4 \\ 1 \end{pmatrix} + C_2 e^{3t} \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + C_3 e^{-2t} \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}.$$

Distinct complex eigenvalues

If $\lambda = \xi + i\omega$ is a complex eigenvalue, then also its complex conjugate $\bar{\lambda} = \xi - i\omega$ is an eigenvalue, as the characteristic polynomial $p(\lambda)$ has real coefficients. Eigenvectors \mathbf{v} corresponding to a complex eigenvalue λ will be complex vectors, that is, vectors with complex entries. Thus, we can write

$$\mathbf{v} = \begin{pmatrix} v_1^1 + iv_1^2 \\ \vdots \\ v_n^1 + iv_n^2 \end{pmatrix} = \begin{pmatrix} v_1^1 \\ \vdots \\ v_n^1 \end{pmatrix} + i \begin{pmatrix} v_1^2 \\ \vdots \\ v_n^2 \end{pmatrix} = \Re \mathbf{v} + i \Im \mathbf{v}.$$

Since $(\mathcal{A} - \lambda \mathcal{I})\mathbf{v} = \mathbf{0}$, taking complex conjugate of both sides and using the fact that matrices \mathcal{A} and \mathcal{I} have only real entries, we see that

$$\overline{(\mathcal{A} - \lambda \mathcal{I})\mathbf{v}} = (\mathcal{A} - \bar{\lambda} \mathcal{I})\bar{\mathbf{v}} = \mathbf{0}$$

so that the complex conjugate $\bar{\mathbf{v}}$ of \mathbf{v} is an eigenvector corresponding to the eigenvalue $\bar{\lambda}$. Since $\lambda \neq \bar{\lambda}$, as we assumed that λ is complex, the eigenvectors \mathbf{v} and $\bar{\mathbf{v}}$ are linearly independent and thus we obtain two linearly independent complex valued solutions

$$\mathbf{z}^1(t) = e^{\lambda t} \mathbf{v}, \quad \mathbf{z}^2(t) = e^{\bar{\lambda} t} \bar{\mathbf{v}} = \overline{\mathbf{z}^1(t)}.$$

Since the sum and the difference of two solutions are again

solutions, by taking

$$\mathbf{y}^1(t) = \frac{\mathbf{z}^1(t) + \mathbf{z}^2(t)}{2} = \frac{\mathbf{z}^1(t) + \overline{\mathbf{z}^1(t)}}{2} = \Re \mathbf{z}^1(t)$$

and

$$\mathbf{y}^2(t) = \frac{\mathbf{z}^1(t) - \mathbf{z}^2(t)}{2i} = \frac{\mathbf{z}^1(t) - \overline{\mathbf{z}^1(t)}}{2i} = \Im \mathbf{z}^1(t)$$

we obtain two real valued (and linearly independent) solutions. To find explicit formulae for $\mathbf{y}^1(t)$ and $\mathbf{y}^2(t)$, we write

$$\begin{aligned} \mathbf{z}^1(t) &= e^{\lambda t} \mathbf{v} = e^{\xi t} (\cos \omega t + i \sin \omega t) (\Re \mathbf{v} + i \Im \mathbf{v}) \\ &= e^{\xi t} (\cos \omega t \Re \mathbf{v} - \sin \omega t \Im \mathbf{v}) + i e^{\xi t} (\cos \omega t \Im \mathbf{v} + \sin \omega t \Re \mathbf{v}) \\ &= \mathbf{y}^1(t) + i \mathbf{y}^2(t) \end{aligned}$$

Summarizing, if λ and $\bar{\lambda}$ are single complex roots of the characteristic equation with complex eigenvectors \mathbf{v} and $\bar{\mathbf{v}}$, respectively, then we can use two real linearly independent solutions

$$\begin{aligned} \mathbf{y}^1(t) &= e^{\xi t} (\cos \omega t \Re \mathbf{v} - \sin \omega t \Im \mathbf{v}) \\ \mathbf{y}^2(t) &= e^{\xi t} (\cos \omega t \Im \mathbf{v} + \sin \omega t \Re \mathbf{v}) \end{aligned} \quad (7.22)$$

Example 7.4 Solve the initial value problem

$$\mathbf{y}' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & 1 \end{pmatrix} \mathbf{y}, \quad \mathbf{y}(0) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

The characteristic polynomial is given by

$$\begin{aligned} p(\lambda) &= \det(\mathcal{A} - \lambda\mathcal{I}) = \begin{vmatrix} 1 - \lambda & 0 & 0 \\ 0 & 1 - \lambda & -1 \\ 0 & 1 & 1 - \lambda \end{vmatrix} \\ &= (1 - \lambda)^3 + (1 - \lambda) = (1 - \lambda)(\lambda^2 - 2\lambda + 2) \end{aligned}$$

so that we have eigenvalues $\lambda_1 = 1$ and $\lambda_{2,3} = 1 \pm i$.

It is immediate that

$$\mathbf{v} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

is an eigenvector corresponding to $\lambda_1 = 1$ and thus we obtain a solution to the system in the form

$$\mathbf{y}^1(t) = e^t \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Let us take now the complex eigenvalue $\lambda_2 = 1 + i$. We have to solve

$$(\mathcal{A} - (1 + i)\mathcal{I})\mathbf{v} = \begin{pmatrix} -i & 0 & 0 \\ 0 & -i & -1 \\ 0 & 1 & -i \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Thus

$$-iv_1 = 0, \quad -iv_2 - v_3 = 0, \quad v_2 - iv_3 = 0.$$

The first equation gives $v_1 = 0$ and the other two yield $v_2 = iv_3$ so that each vector

$$\mathbf{v}^2 = C_2 \begin{pmatrix} 0 \\ i \\ 1 \end{pmatrix}$$

210 *Simultaneous systems of equations and higher order equations*

is an eigenvector corresponding to the eigenvalue $\lambda_2 = 1+i$.

Consequently, we obtain a complex valued solution

$$\mathbf{z}(t) = e^{(1+i)t} \begin{pmatrix} 0 \\ i \\ 1 \end{pmatrix}.$$

To obtain real valued solutions, we separate \mathbf{z} into real and imaginary parts:

$$\begin{aligned} e^{(1+i)t} \begin{pmatrix} 0 \\ i \\ 1 \end{pmatrix} &= e^t (\cos t + i \sin t) \left(\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + i \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right) \\ &= e^t \left(\cos t \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} - \sin t \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + i \sin t \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + i \cos t \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right) \\ &= e^t \begin{pmatrix} 0 \\ -\sin t \\ \cos t \end{pmatrix} + i e^t \begin{pmatrix} 0 \\ \cos t \\ \sin t \end{pmatrix}. \end{aligned}$$

Thus, we obtain two real solutions

$$\begin{aligned} \mathbf{y}^1(t) &= e^t \begin{pmatrix} 0 \\ -\sin t \\ \cos t \end{pmatrix} \\ \mathbf{y}^2(t) &= e^t \begin{pmatrix} 0 \\ \cos t \\ \sin t \end{pmatrix} \end{aligned}$$

and the general solution to our original system is given by

$$\mathbf{y}(t) = C_1 e^t \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + C_2 e^t \begin{pmatrix} 0 \\ -\sin t \\ \cos t \end{pmatrix} + C_3 e^t \begin{pmatrix} 0 \\ \cos t \\ \sin t \end{pmatrix}.$$

We can check that all these solutions are independent as their initial values

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix},$$

are independent. To find the solution to our initial value problem we set $t = 0$ and we have to solve for C_1, C_2 and C_3 the system

$$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = C_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + C_3 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix}.$$

Thus $C_1 = C_2 = C_3 = 1$ and finally

$$\mathbf{y}(t) = e^t \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + e^t \begin{pmatrix} 0 \\ -\sin t \\ \cos t \end{pmatrix} + e^t \begin{pmatrix} 0 \\ \cos t \\ \sin t \end{pmatrix} = e^t \begin{pmatrix} 1 \\ \cos t - \sin t \\ \cos t + \sin t \end{pmatrix}.$$

Multiple eigenvalues

If not all roots of the characteristic polynomial of \mathcal{A} are distinct, that is, there are multiple eigenvalues of \mathcal{A} , then it may happen that \mathcal{A} has less than n linearly independent eigenvectors. Precisely, let us suppose that an $n \times n$ matrix \mathcal{A} has only $k < n$ linearly independent solutions. Then, the differential equation $\mathbf{y}' = \mathcal{A}\mathbf{y}$ has only k linearly independent solutions of the form $e^{\lambda t}\mathbf{v}$. Our aim is to find additional $n - k$ independent solutions. We approach this problem by introducing an abstract framework for solving systems of differential equations.

Recall that for a single equation $y' = ay$, where a is a constant, the general solution is given by $y(t) = e^{at}C$,

Simultaneous systems of equations and higher order equations

where C is a constant. In a similar way, we would like to say that the general solution to

$$\mathbf{y}' = \mathcal{A}\mathbf{y},$$

where \mathcal{A} is an $n \times n$ matrix, is $\mathbf{y} = e^{\mathcal{A}t}\mathbf{v}$, where \mathbf{v} is any constant vector in \mathbb{R}^n . The problem is that we do not know what it means to evaluate the exponential of a matrix. However, if we reflect for a moment that the exponential of a number can be evaluated as the power (Maclaurin) series

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \dots + \frac{x^k}{k!} + \dots,$$

where the only involved operations on the argument x are additions, scalar multiplications and taking integer powers, we come to the conclusion that the above expression can be written also for a matrix, that is, we can define

$$e^{\mathcal{A}} = \mathcal{I} + \mathcal{A} + \frac{1}{2}\mathcal{A}^2 + \frac{1}{3!}\mathcal{A}^3 + \dots + \frac{1}{k!}\mathcal{A}^k + \dots \quad (7.23)$$

It can be shown that if \mathcal{A} is a matrix, then the above series always converges and the sum is a matrix. For example, if we take

$$\mathcal{A} = \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix} = \lambda\mathcal{I},$$

then

$$\mathcal{A}^k = \lambda^k\mathcal{I}^k = \lambda^k\mathcal{I},$$

and

$$\begin{aligned}
 e^{\mathcal{A}} &= \mathcal{I} + \lambda\mathcal{I} + \frac{\lambda^2}{2}\mathcal{I} + \frac{\lambda^3}{3!}\mathcal{I} + \dots + \frac{\lambda^k}{k!} + \dots \\
 &= \left(1 + \lambda + \frac{\lambda^2}{2} + \frac{\lambda^3}{3!} + \dots + \frac{\lambda^k}{k!} + \dots\right)\mathcal{I} \\
 &= e^{\lambda\mathcal{I}}.
 \end{aligned} \tag{7.24}$$

Unfortunately, in most cases finding the explicit form for $e^{\mathcal{A}}$ directly is impossible.

Matrix exponentials have the following algebraic properties

$$(e^{\mathcal{A}})^{-1} = e^{-\mathcal{A}}$$

and

$$e^{\mathcal{A}+\mathcal{B}} = e^{\mathcal{A}}e^{\mathcal{B}} \tag{7.25}$$

provided the matrices \mathcal{A} and \mathcal{B} commute: $\mathcal{A}\mathcal{B} = \mathcal{B}\mathcal{A}$.

Let us define a function of t by

$$e^{t\mathcal{A}} = \mathcal{I} + t\mathcal{A} + \frac{t^2}{2}\mathcal{A}^2 + \frac{t^3}{3!}\mathcal{A}^3 + \dots + \frac{t^k}{k!}\mathcal{A}^k + \dots \tag{7.26}$$

It follows that this function can be differentiated with respect to t by termwise differentiation of the series, as in the scalar case, that is,

$$\begin{aligned}
 \frac{d}{dt}e^{t\mathcal{A}} &= \mathcal{A} + t\mathcal{A}^2 + \frac{t^2}{2!}\mathcal{A}^3 + \dots + \frac{t^{k-1}}{(k-1)!}\mathcal{A}^k + \dots \\
 &= \mathcal{A} \left(\mathcal{I} + t\mathcal{A} + \frac{t^2}{2!}\mathcal{A}^2 + \dots + \frac{t^{k-1}}{(k-1)!}\mathcal{A}^{k-1} + \dots \right) \\
 &= \mathcal{A}e^{t\mathcal{A}} = e^{t\mathcal{A}}\mathcal{A},
 \end{aligned}$$

proving thus that $y(t) = e^{t\mathcal{A}}\mathbf{v}$ is a solution to our system of equations for any constant vector \mathbf{v} .

Simultaneous systems of equations and higher order equations

As we mentioned earlier, in general it is difficult to find directly the explicit form of $e^{t\mathcal{A}}$. However, we can always find n linearly independent vectors \mathbf{v} for which the series $e^{t\mathcal{A}}\mathbf{v}$ can be summed exactly. This is based on the following two observations. Firstly, since $\lambda\mathcal{I}$ and $\mathcal{A} - \lambda\mathcal{I}$ commute, we have by (7.24) and (7.25)

$$e^{t\mathcal{A}}\mathbf{v} = e^{t(\mathcal{A}-\lambda\mathcal{I})}e^{t\lambda\mathcal{I}}\mathbf{v} = e^{\lambda t}e^{t(\mathcal{A}-\lambda\mathcal{I})}\mathbf{v}.$$

Secondly, if $(\mathcal{A} - \lambda\mathcal{I})^m\mathbf{v} = \mathbf{0}$ for some m , then

$$(\mathcal{A} - \lambda\mathcal{I})^r\mathbf{v} = \mathbf{0}, \quad (7.27)$$

for all $r \geq m$. This follows from

$$(\mathcal{A} - \lambda\mathcal{I})^r\mathbf{v} = (\mathcal{A} - \lambda\mathcal{I})^{r-m}[(\mathcal{A} - \lambda\mathcal{I})^m\mathbf{v}] = \mathbf{0}.$$

Consequently, for such a \mathbf{v}

$$e^{t(\mathcal{A}-\lambda\mathcal{I})}\mathbf{v} = \mathbf{v} + t(\mathcal{A} - \lambda\mathcal{I})\mathbf{v} + \dots + \frac{t^{m-1}}{(m-1)!}(\mathcal{A} - \lambda\mathcal{I})^{m-1}\mathbf{v}.$$

and

$$e^{t\mathcal{A}}\mathbf{v} = e^{\lambda t}e^{t(\mathcal{A}-\lambda\mathcal{I})}\mathbf{v} = e^{\lambda t} \left(\mathbf{v} + t(\mathcal{A} - \lambda\mathcal{I})\mathbf{v} + \dots + \frac{t^{m-1}}{(m-1)!}(\mathcal{A} - \lambda\mathcal{I})^{m-1}\mathbf{v} \right). \quad (7.28)$$

Thus, to find all solutions to $\mathbf{y}' = \mathcal{A}\mathbf{y}$ it is sufficient to find n independent vectors \mathbf{v} satisfying (7.27) for some scalars λ . To check consistency of this method with our previous consideration we observe that if $\lambda = \lambda_1$ is a single eigenvalue of \mathcal{A} with a corresponding eigenvector \mathbf{v}^1 , then $(\mathcal{A} - \lambda_1\mathcal{I})\mathbf{v}^1 = \mathbf{0}$, thus m of (7.27) is equal to 1. Consequently, the sum in (7.28) terminates after the first term and we obtain

$$\mathbf{y}_1(t) = e^{\lambda_1 t}\mathbf{v}^1$$

in accordance with (7.21). From our discussion of eigenvalues and eigenvectors it follows that if λ_i is a multiple eigenvalue of \mathcal{A} of algebraic multiplicity n_i and the geometric multiplicity is less than n_i , that is, there is less than n_i linearly independent eigenvectors corresponding to λ_i , then the missing independent vectors can be found by solving successively equations $(\mathcal{A} - \lambda_i \mathcal{I})^k \mathbf{v} = \mathbf{0}$ with k running at most up to n_i . Thus, we have the following algorithm for finding n linearly independent solutions to $\mathbf{y}' = \mathcal{A}\mathbf{y}$:

- (i) Find all eigenvalues of \mathcal{A} ;
- (ii) If λ is a single real eigenvalue, then there is an eigenvector \mathbf{v} so that the solution is given by

$$\mathbf{y}(t) = e^{\lambda t} \mathbf{v} \quad (7.29)$$

- (iii) If λ is a single complex eigenvalue $\lambda = \xi + i\omega$, then there is a complex eigenvector $\mathbf{v} = \Re \mathbf{v} + i \Im \mathbf{v}$ such that two solutions corresponding to λ (and $\bar{\lambda}$) are given by

$$\begin{aligned} \mathbf{y}^1(t) &= e^{\xi t} (\cos \omega t \Re \mathbf{v} - \sin \omega t \Im \mathbf{v}) \\ \mathbf{y}^2(t) &= e^{\xi t} (\cos \omega t \Im \mathbf{v} + \sin \omega t \Re \mathbf{v}) \end{aligned} \quad (7.30)$$

- (iv) If λ is a multiple eigenvalue with algebraic multiplicity k (that is, λ is a multiple root of the characteristic equation of multiplicity k), then we first find eigenvectors by solving $(\mathcal{A} - \lambda \mathcal{I})\mathbf{v} = \mathbf{0}$. For these eigenvectors the solution is again given by (7.29) (or (7.30), if λ is complex). If we found k independent eigenvectors, then our work with this eigenvalue is finished. If not, then we look for vectors that satisfy $(\mathcal{A} - \lambda \mathcal{I})^2 \mathbf{v} = \mathbf{0}$ but $(\mathcal{A} - \lambda \mathcal{I})\mathbf{v} \neq \mathbf{0}$. For these

Simultaneous systems of equations and higher order equations

vectors we have the solutions

$$e^{t\mathcal{A}}\mathbf{v} = e^{\lambda t}(\mathbf{v} + t(\mathcal{A} - \lambda\mathcal{I})\mathbf{v}).$$

If we still do not have k independent solutions, then we find vectors for which $(\mathcal{A} - \lambda\mathcal{I})^3\mathbf{v} = \mathbf{0}$ and $(\mathcal{A} - \lambda\mathcal{I})^2\mathbf{v} \neq \mathbf{0}$, and for such vectors we construct solutions

$$e^{t\mathcal{A}}\mathbf{v} = e^{\lambda t} \left(\mathbf{v} + t(\mathcal{A} - \lambda\mathcal{I})\mathbf{v} + \frac{t^2}{2}(\mathcal{A} - \lambda\mathcal{I})^2\mathbf{v} \right).$$

This procedure is continued till we have k solutions (by the properties of eigenvalues we have to repeat this procedure at most k times).

If λ is a complex eigenvalue of multiplicity k , then also $\bar{\lambda}$ is an eigenvalue of multiplicity k and we obtain pairs of real solutions by taking real and imaginary parts of the formulae presented above.

Remark 7.2 *Once we know that all solutions must be of the form (7.28) with the degree of the polynomial being at most equal to the algebraic multiplicity of λ , we can use the method of undetermined coefficients to find the solutions. Namely, if λ is an eigenvalue of multiplicity k , then we can look for a solutions in the form*

$$\mathbf{y}(t) = e^{\lambda t}(\mathbf{a}_0 + \mathbf{a}_1 t + \dots + \mathbf{a}_{k-1} t^{k-1})$$

where unknown vectors $\mathbf{a}_0, \dots, \mathbf{a}_{k-1}$ are to be determined by inserting $\mathbf{y}(t)$ into the equation and solving the resulting simultaneous systems of algebraic equations.

Example 7.5 *Find three linearly independent solutions of*

the differential equation

$$\mathbf{y}' = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} \mathbf{y}.$$

To obtain the eigenvalues we calculate the characteristic polynomial

$$\begin{aligned} p(\lambda) &= \det(\mathcal{A} - \lambda\mathcal{I}) = \begin{vmatrix} 1 - \lambda & 1 & 0 \\ 0 & 1 - \lambda & 0 \\ 0 & 0 & 2 - \lambda \end{vmatrix} \\ &= (1 - \lambda)^2(2 - \lambda) \end{aligned}$$

so that $\lambda_1 = 1$ is eigenvalue of multiplicity 2 and $\lambda_2 = 2$ is an eigenvalue of multiplicity 1.

(i) $\lambda = 1$: We seek all non-zero vectors such that

$$(\mathcal{A} - 1\mathcal{I})\mathbf{v} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

This implies that $v_2 = v_3 = 0$ and v_1 is arbitrary so that we obtain the corresponding solutions

$$\mathbf{y}^1(t) = C_1 e^t \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

However, this is only one solution and $\lambda_1 = 1$ has algebraic multiplicity 2, so we have to look for one

Simultaneous systems of equations and higher order equations

more solution. To this end we consider

$$\begin{aligned}(\mathcal{A} - \mathcal{I})^2 \mathbf{v} &= \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}\end{aligned}$$

so that $v_3 = 0$ and both v_1 and v_2 arbitrary. The set of all solutions here is a two-dimensional space spanned by

$$\begin{pmatrix} v_1 \\ v_2 \\ 0 \end{pmatrix} = v_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + v_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

We have to select from this subspace a vector that is not a solution to $(\mathcal{A} - \lambda \mathcal{I})\mathbf{v} = \mathbf{0}$. Since for the later the solutions are scalar multiples of the vector $(1, 0, 0)$ we see that the vector $(0, 1, 0)$ is not of this form and consequently can be taken as the second independent vector corresponding to the eigenvalue $\lambda_1 = 1$. Hence

$$\begin{aligned}\mathbf{y}^2(t) &= e^t (\mathcal{I} + t(\mathcal{A} - \mathcal{I})) \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = e^t \left(\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + t \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right) \\ &= e^t \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + te^t \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = e^t \begin{pmatrix} t \\ 1 \\ 0 \end{pmatrix}\end{aligned}$$

(ii) $\lambda = 2$: We seek solutions to

$$(\mathcal{A} - 2\mathcal{I})\mathbf{v} = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

This implies that $v_1 = v_2 = 0$ and v_3 is arbitrary so that the corresponding solutions are of the form

$$\mathbf{y}^3(t) = C_3 e^{2t} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Thus we have found three linearly independent solutions.

Fundamental solutions and nonhomogeneous problems

Let us suppose that we have n linearly independent solutions $\mathbf{y}^1(t), \dots, \mathbf{y}^n(t)$ of the system $\mathbf{y}' = \mathcal{A}\mathbf{y}$, where \mathcal{A} is an $n \times n$ matrix, like the ones constructed in the previous paragraphs. Let us denote by $\mathcal{Y}(t)$ the matrix

$$\mathcal{Y}(t) = \begin{pmatrix} y_1^1(t) & \dots & y_1^n(t) \\ \vdots & & \vdots \\ y_n^1(t) & \dots & y_n^n(t) \end{pmatrix},$$

that is, the columns of $\mathcal{Y}(t)$ are the vectors \mathbf{y}^i , $i = 1, \dots, n$. Any such matrix is called a *fundamental matrix* of the system $\mathbf{y}' = \mathcal{A}\mathbf{y}$.

We know that for a given initial vector \mathbf{y}^0 the solution is given by

$$\mathbf{y}(t) = e^{t\mathcal{A}}\mathbf{y}^0$$

on one hand, and, by Theorem 7.2, by

$$\mathbf{y}(t) = C_1\mathbf{y}^1(t) + \dots + C_n\mathbf{y}^n(t) = \mathcal{Y}(t)\mathbf{C},$$

on the other, where $\mathbf{C} = (C_1, \dots, C_n)$ is a vector of constants to be determined. By putting $t = 0$ above we obtain the equation for \mathbf{C}

$$\mathbf{y}^0 = \mathcal{Y}(0)\mathbf{C}$$

Since \mathcal{Y} has independent vectors as its columns, it is invertible, so that

$$\mathbf{C} = \mathcal{Y}^{-1}(0)\mathbf{y}^0.$$

Thus, the solution of the initial value problem

$$\mathbf{y}' = \mathcal{A}\mathbf{y}, \quad \mathbf{y}(0) = \mathbf{y}^0$$

is given by

$$\mathbf{y}(t) = \mathcal{Y}(t)\mathcal{Y}^{-1}(0)\mathbf{y}^0.$$

Since $e^{t\mathbf{A}}\mathbf{y}^0$ is also a solution, by the uniqueness theorem we obtain explicit representation of the exponential function of a matrix

$$e^{t\mathbf{A}} = \mathcal{Y}(t)\mathcal{Y}^{-1}(0). \quad (7.31)$$

Let us turn our attention to the non-homogeneous system of equations

$$\mathbf{y}' = \mathcal{A}\mathbf{y} + \mathbf{g}(t). \quad (7.32)$$

The general solution to the homogeneous equation ($\mathbf{g}(t) \equiv 0$) is given by

$$\mathbf{y}_h(t) = \mathcal{Y}(t)\mathbf{C},$$

where $\mathcal{Y}(t)$ is a fundamental matrix and \mathbf{C} is an arbitrary vector. Using the technique of variation of parameters, we will be looking for the solution in the form

$$\mathbf{y}(t) = \mathcal{Y}(t)\mathbf{u}(t) = u_1(t)\mathbf{y}^1(t) + \dots + u_n(t)\mathbf{y}^n(t) \quad (7.33)$$

where $\mathbf{u}(t) = (u_1(t), \dots, u_n(t))$ is a vector-function to be determined so that (7.33) satisfies (7.32). Thus, substituting (7.33) into (7.32), we obtain

$$\mathcal{Y}'(t)\mathbf{u}(t) + \mathcal{Y}(t)\mathbf{u}'(t) = \mathcal{A}\mathcal{Y}(t)\mathbf{u}(t) + \mathbf{g}(t).$$

Since $\mathcal{Y}(t)$ is a fundamental matrix, $\mathcal{Y}'(t) = \mathcal{A}\mathcal{Y}(t)$ and we find

$$\mathcal{Y}(t)\mathbf{u}'(t) = \mathbf{g}(t).$$

As we observed earlier, $\mathcal{Y}(t)$ is invertible, hence

$$\mathbf{u}'(t) = \mathcal{Y}^{-1}(t)\mathbf{g}(t)$$

and

$$\mathbf{u}(t) = \int^t \mathcal{Y}^{-1}(s)\mathbf{g}(s)ds + \mathbf{C}.$$

Finally, we obtain

$$\mathbf{y}(t) = \mathcal{Y}(t)\mathbf{C} + \mathcal{Y}(t) \int^t \mathcal{Y}^{-1}(s)\mathbf{g}(s)ds \quad (7.34)$$

This equation becomes much simpler if we take $e^{t\mathcal{A}}$ as a fundamental matrix because in such a case $\mathcal{Y}^{-1}(t) = (e^{t\mathcal{A}})^{-1} = e^{-t\mathcal{A}}$, that is, to calculate the inverse of $e^{t\mathcal{A}}$ it is enough to replace t by $-t$. The solution (7.34) takes then the form

$$\mathbf{y}(t) = e^{t\mathcal{A}}\mathbf{C} + \int e^{(t-s)\mathcal{A}}\mathbf{g}(s)ds. \quad (7.35)$$

Example 7.6 Find the general solution to

$$\begin{aligned} y_1' &= 5y_1 + 3y_2 + 2te^{2t}, \\ y_2' &= -3y_1 - y_2 + 4. \end{aligned}$$

Simultaneous systems of equations and higher order equations

Writing this system in matrix notation, we obtain

$$\mathbf{y}' = \mathcal{A}\mathbf{y} + \mathbf{g}(t)$$

with

$$\mathcal{A} = \begin{pmatrix} 5 & 3 \\ -3 & -1 \end{pmatrix}$$

and

$$\mathbf{g}(t) = \begin{pmatrix} 2te^{2t} \\ 4 \end{pmatrix}.$$

We have to find $e^{t\mathcal{A}}$. The first step is to find two independent solutions to the homogeneous system. The characteristic polynomial is

$$p(\lambda) = \begin{vmatrix} 5 - \lambda & 3 \\ -3 & -1 - \lambda \end{vmatrix} = \lambda^2 - 4\lambda + 4 = (\lambda - 2)^2$$

We have double eigenvalue $\lambda = 2$. Solving

$$(\mathcal{A} - 2\mathcal{I})\mathbf{v} = \begin{pmatrix} 3 & 3 \\ -3 & -3 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

we obtain $v_1 = -v_2$ so that we obtain the eigenvector $\mathbf{v}^1 = (1, -1)$ and the corresponding solution

$$\mathbf{y}^1(t) = C_1 e^{2t} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Since $\lambda_1 = 2$ has algebraic multiplicity 2, we have to look for another solution. To this end we consider

$$\begin{aligned} (\mathcal{A} - 2\mathcal{I})^2 \mathbf{v} &= \begin{pmatrix} 3 & 3 \\ -3 & -3 \end{pmatrix} \begin{pmatrix} 3 & 3 \\ -3 & -3 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \end{aligned}$$

so that v_1 and v_2 arbitrary. We must simply select a vector linearly independent of \mathbf{y}^1 – to make things simple we can take

$$\mathbf{y}^2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

so that the second solution is given as

$$\begin{aligned} \mathbf{y}^2(t) &= e^{2t} (\mathcal{I} + t(\mathcal{A} - \mathcal{I})) \begin{pmatrix} 1 \\ 0 \end{pmatrix} = e^{2t} \left(\begin{pmatrix} 1 \\ 0 \end{pmatrix} + t \begin{pmatrix} 3 & 3 \\ -3 & -3 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right) \\ &= e^{2t} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + te^{2t} \begin{pmatrix} 3 \\ -3 \end{pmatrix} = e^{2t} \begin{pmatrix} 1+3t \\ -3t \end{pmatrix} \end{aligned}$$

Thus, the fundamental matrix is given by

$$\mathcal{Y}(t) = e^{2t} \begin{pmatrix} 1 & 1+3t \\ -1 & -3t \end{pmatrix}$$

with

$$\mathcal{Y}(0) = \begin{pmatrix} 1 & 1 \\ -1 & 0 \end{pmatrix}.$$

The discriminant of $\mathcal{Y}(0)$ is equal to 1 and we immediately obtain

$$\mathcal{Y}^{-1}(0) = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}.$$

so that

$$e^{t\mathcal{A}} = \mathcal{Y}(t)\mathcal{Y}^{-1}(0) = e^{2t} \begin{pmatrix} 1+3t & 3t \\ -3t & 1-3t \end{pmatrix}.$$

Thus

$$e^{-t\mathcal{A}} = e^{-2t} \begin{pmatrix} 1-3t & -3t \\ 3t & 1+3t \end{pmatrix}$$

Simultaneous systems of equations and higher order equations

and

$$e^{-tA}\mathbf{g}(t) = e^{-2t} \begin{pmatrix} 1-3t & -3t \\ 3t & 1+3t \end{pmatrix} \begin{pmatrix} 2te^{2t} \\ 4 \end{pmatrix} = \begin{pmatrix} 2t - 6t^2 - 12te^{-2t} \\ 6t^2 + 4e^{-2t} + 12te^{-2t} \end{pmatrix}.$$

To find the particular solution, we integrate the above, getting

$$\int e^{-tA}\mathbf{g}(t)dt = \begin{pmatrix} \int(2t - 6t^2 - 12te^{-2t})dt \\ \int(6t^2 + 4e^{-2t} + 12te^{-2t})dt \end{pmatrix} = \begin{pmatrix} t^2 - 2t^3 + 3(2t + 1)e^{-2t} \\ 2t^3 - (6t + 5)e^{-2t} \end{pmatrix},$$

and multiply the above by e^{tA} to obtain

$$e^{2t} \begin{pmatrix} 1+3t & 3t \\ -3t & 1-3t \end{pmatrix} \begin{pmatrix} t^2 - 2t^3 + 3(2t + 1)e^{-2t} \\ 2t^3 - (6t + 5)e^{-2t} \end{pmatrix} = \begin{pmatrix} (t^2 + t^3)e^{2t} + 3 \\ -t^3e^{2t} - 5 \end{pmatrix}.$$

Therefore, the general solution is given by

$$\mathbf{y}(t) = e^{2t} \begin{pmatrix} 1+3t & 3t \\ -3t & 1-3t \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} + \begin{pmatrix} (t^2 + t^3)e^{2t} + 3 \\ -t^3e^{2t} - 5 \end{pmatrix}$$

where C_1 and C_2 are arbitrary constants.

7.2 Second order linear equations

Second order equations occur very often in practice so that it is useful to specify the general theory of systems for this particular case.

$$\frac{d^2y}{dt^2} + a_1 \frac{dy}{dt} + a_0y = f(t) \quad (7.36)$$

where a_1, a_0 are real constants and f is a given continuous function. As before in what follows we shall abbreviate $d^2y/dt^2 = y''$ and $dy/dt = y'$.

As we mentioned earlier, (7.36) can be written as an

equivalent system of 2 first order equations by introducing new variables $y_1 = y$, $y_2 = y' = y_1'$,

$$\begin{aligned} y_1' &= y_2, \\ \vdots & \quad \quad \quad \vdots \\ y_2' &= -a_1 y_2 - a_0 y_1 + f(t). \end{aligned}$$

Note that if (7.36) was supplemented with the initial conditions $y(t_0) = y^0, y'(t_0) = y^1$, then these conditions will become natural initial conditions for the system as $y_1(t_0) = y^0, y_2(t_0) = y^1$.

Let us first recall the theory for first order linear equations, specified to the case of a constant coefficient a :

$$y' + ay = f(t). \quad (7.37)$$

By (2.11), the general solution to (7.37) is given by

$$y(t) = Ce^{-at} + e^{-at} \int e^{as} f(s) ds,$$

where the first term is the general solution of the homogeneous ($f \equiv 0$) version of (7.37) and the second is a particular solution to (7.37). This suggests that a sensible strategy for solving (7.36) is to look first for solutions to the associated homogeneous equation

$$\frac{d^2 y}{dt^2} + a_1 \frac{dy}{dt} + a_0 y = 0. \quad (7.38)$$

Let us denote by y_0 the general solution to (7.38), that is, y_0 is really a class of functions depending on two constants. Next, let y_p be a particular solution of (7.36) and consider

$y(t) = y_p(t) + z(t)$. Then

$$\begin{aligned} y'' + a_1y' + a_0y &= y_p'' + a_1y_p' + a_0y_p + z'' + a_1z' + a_0z \\ &= f(t) + z'' + a_1z' + a_0z, \end{aligned}$$

that is, y is a solution to (7.36) if and only if z is any solution to (7.38) or, in other words, if and only if z is the general solution to (7.38), $z = y_c$.

Accordingly, we shall first develop methods for finding general solutions to homogeneous equations.

7.2.1 Homogeneous equations

Let us consider the homogeneous equation (7.38)

$$\frac{d^2y}{dt^2} + a_1 \frac{dy}{dt} + a_0y = 0. \quad (7.39)$$

Since the space of solutions of the corresponding 2×2 homogeneous system

$$\begin{aligned} y_1' &= y_2, \\ \vdots &\quad \vdots \\ y_2' &= -a_1y_2 - a_0y_1. \end{aligned} \quad (7.40)$$

is two-dimensional, the space of solutions to (7.39) is also two-dimensional, that is, there are two independent solutions of (7.39) $y_1(t), y_2(t)$ such that any other solution is given by

$$y(t) = C_1y_1(t) + C_2y_2(t).$$

How can we recover the general solution to (7.39) from the solution $\mathbf{y}(t)$ of the system? A function $y(t)$ solves (7.39) if and only if $\mathbf{y}(t) = (y_1(t), y_2(t)) = (y(t), y'(t))$

solves the system (7.40), we see that the general solution to (7.39) can be obtained by taking first components of the solution of the associated system (7.40). We note once again that if $\mathbf{y}^1(t) = (y_1^1(t), y_2^1(t)) = (y^1(t), \frac{dy^1}{dt}(t))$ and $\mathbf{y}^2(t) = (y_1^2(t), y_2^2(t)) = (y^2(t), \frac{dy^2}{dt}(t))$ are two linearly independent solutions to (7.40), then $y^1(t)$ and $y^2(t)$ are linearly independent solutions to (7.39). In fact, otherwise we would have $y^1(t) = Cy^2(t)$ for some constant C and therefore also $\frac{dy^1}{dt}(t) = C\frac{dy^2}{dt}(t)$ so that the wronskian, having the second column as a scalar multiple of the first one, would be zero, contrary to the assumption that $\mathbf{y}^1(t)$ and $\mathbf{y}^2(t)$ are linearly independent.

To find explicit formulae for two linearly independent particular solutions to (7.39) we write the equation for the characteristic polynomial of (7.40):

$$\begin{vmatrix} -\lambda & 1 \\ -a_0 & -a_1 - \lambda \end{vmatrix} = 0$$

that is

$$\lambda^2 + a_1\lambda + a_0 = 0,$$

which is also called the characteristic polynomial of (7.39). This is a quadratic equation in λ which is zero when $\lambda = \lambda_1$ or $\lambda = \lambda_2$ with

$$\lambda_{1,2} = \frac{-a_1 \pm \sqrt{\Delta}}{2}$$

where the discriminant $\Delta = a_1^2 - 4a_0$.

If $\Delta > 0$, then $\lambda_1 \neq \lambda_2$, and we obtain two different solutions $y_1 = e^{\lambda_1 t}$ and $y_2 = e^{\lambda_2 t}$. Thus

$$y(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}$$

with two arbitrary constants is the sought general solution to (7.39). If $\Delta < 0$, then λ_1 and λ_2 are complex conjugates: $\lambda_1 = \xi + i\omega$, $\lambda_2 = \xi - i\omega$ with $\xi = -a_1/2$ and $\omega = -\sqrt{-\Delta}/2$. Since in many applications it is undesirable to work with complex functions, we shall express the solution in terms of real functions. Using the Euler formula for the complex exponential function, we obtain

$$\begin{aligned} y(t) &= C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t} = C_1 e^{\xi t} (\cos \omega t + i \sin \omega t) + C_2 e^{\xi t} (\cos \omega t - i \sin \omega t) \\ &= (C_1 + C_2) e^{\xi t} \cos \omega t + i(C_1 - C_2) e^{\xi t} \sin \omega t. \end{aligned}$$

If as the constants C_1 and C_2 we take complex conjugates $C_1 = (A - iB)/2$ and $C_2 = (A + iB)/2$ with arbitrary real A and B , then we obtain y as a combination of two real functions with two arbitrary real coefficients

$$y(t) = A e^{\xi t} \cos \omega t + B e^{\xi t} \sin \omega t.$$

We have left behind the case $\lambda_1 = \lambda_2$ (necessarily real). In this case we have only one function $e^{\lambda_1 t}$ with one arbitrary constant C_1 so that $y(t) = C_1 e^{\lambda_1 t}$ is not the general solution to (7.39). Using the theory for systems, we obtain the other solution in the form

$$y_2(t) = t e^{\lambda_1 t}$$

with $\lambda_1 = -a_1/2$. Thus the general solution is given by

$$y(t) = (C_1 + C_2 t) e^{-ta_1/2}.$$

Summarizing, we have the following general solutions corresponding to various properties of the roots of the charac-

characteristic polynomial λ_1, λ_2 .

$$\begin{aligned} y(t) &= C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t} && \text{if } \lambda_1 \neq \lambda_2, \lambda_1, \lambda_2 \text{ real,} \\ y(t) &= C_1 e^{\xi t} \cos \omega t + C_2 e^{\xi t} \sin \omega t && \text{if } \lambda_{1,2} = \xi \pm i\omega, \\ y(t) &= (C_1 + C_2 t) e^{-t\lambda} && \text{if } \lambda_1 = \lambda_2 = \lambda. \end{aligned}$$

7.2.2 Nonhomogeneous equations

At the beginning of this section we have shown that to find the general solution to

$$\frac{d^2 y}{dt^2} + a_1 \frac{dy}{dt} + a_0 y = f(t) \quad (7.41)$$

we have to find the general solution to the homogeneous version (7.39) and then just one particular solution to the full equation (7.41). In the previous subsection we have presented the complete theory for finding the general solution to homogeneous equations. Here we shall discuss two methods of finding solutions to nonhomogeneous equation. We start with the so-called variation of parameters method that is very general but sometimes rather cumbersome to apply. The second method, of judicious guessing, can be applied for special right-hand sides only, but then it gives the solution really quickly.

Variation of parameters The method of variations of parameters was introduced for systems of equations, specifying it for second order equations would be, however, quite cumbersome. Thus, we shall derive it from scratch. Let

$$y_0(t) = C_1 y_1(t) + C_2 y_2(t)$$

be the general solution to the homogeneous version of (7.41). We are looking for a solution to (7.41) in the form

$$y(t) = u(t)y_1(t) + v(t)y_2(t), \quad (7.42)$$

that is, we allow the arbitrary parameters C_1 and C_2 to depend on time. To determine $v(t)$ and $u(t)$ so that (7.42) is a solution to (7.41), we substitute $y(t)$ to the equation. Since there is only one equation, this will give one condition to determine two functions, giving some freedom to pick up the second condition is such a way that the resulting equation becomes the easiest. Let us work it out. Differentiating (7.42) we have

$$y' = uy_1' + vy_2' + u'y_1 + v'y_2,$$

and

$$y'' = u'y_1' + v'y_2' + uy_1'' + vy_2'' + u''y_1 + v''y_2 + u'y_1' + v'y_2'.$$

We see that there appear second order derivatives of the unknown functions and this is something we would like to avoid, as we are trying to simplify a second order equation. For the second order derivatives not to appear we simply require that the part of y' containing u' and v' to vanish, that is,

$$u'y_1 + v'y_2 = 0.$$

With this, we obtain

$$\begin{aligned} y' &= uy_1' + vy_2', \\ y'' &= u'y_1' + v'y_2' + uy_1'' + vy_2''. \end{aligned}$$

Substituting these into (7.41) we obtain

$$\begin{aligned} & u'y'_1 + v'y'_2 + uy''_1 + vy''_2 + a_1(uy'_1 + vy'_2) + a_0(uy_1 + vy_2) \\ &= u(y''_1 + a_1y'_1 + a_0y_1) + v(y''_2 + a_1y'_2 + a_0y_2) + u'y'_1 + v'y'_2 \\ &= f(t). \end{aligned}$$

Since y_1 and y_2 are solutions of the homogeneous equation, first two terms in the second line vanish and for y to satisfy (7.41) we must have

$$u'y'_1 + v'y'_2 = f(t).$$

Summarizing, to find u and v such that (7.42) satisfies (7.41) we must solve the following system of equations

$$u'y_1 + v'y_2 = 0, \quad (7.43)$$

$$u'y'_1 + v'y'_2 = f(t) \quad (7.44)$$

System (7.44) is to be solved for u' and v' and the solution integrated to find u and v .

Remark 7.3 System (7.44) can be solved by determinants. The main determinant

$$W(t) = \begin{vmatrix} y_1(t) & y_2(t) \\ y'_1(t) & y'_2(t) \end{vmatrix} = y_1(t)y'_2(t) - y_2(t)y'_1(t) \quad (7.45)$$

is the wronskian and plays an important rôle in the general theory of differential equations. Here we shall only note that clearly for (7.44) to be solvable, $W(t) \neq 0$ for all t which is ensured by y_1 and y_2 being linearly independent which, as we know, must be the case if y_0 is the general solution to the homogeneous equation, see Remark 7.1.

Example 7.7 Find the solution to

$$y'' + y = \tan t$$

on the interval $-\pi/2 < t < \pi/2$ satisfying the initial conditions $y(0) = 1$ and $y'(0) = 1$.

Step 1.

General solution to the homogeneous equation

$$y'' + y = 0$$

is obtained by finding the roots of the characteristic equation

$$\lambda^2 + 1 = 0.$$

We have $\lambda_{1,2} = \pm i$ so that $\xi = 0$ and $\omega = 1$ and we obtain two independent solutions

$$y_1(t) = \cos t, \quad y_2(t) = \sin t.$$

Step 2.

To find a solution to the nonhomogeneous equations we first calculate wronskian

$$W(t) = \begin{vmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{vmatrix} = 1.$$

Solving (7.44), we obtain

$$u'(t) = -\sin t \tan t, \quad v'(t) = \cos t \tan t.$$

Then

$$\begin{aligned} u(t) &= -\int \sin t \tan t dt = -\int \frac{\sin^2 t}{\cos t} dt = -\int \frac{1 - \cos^2 t}{\cos t} dt \\ &= \int \cos t dt - \int \frac{dt}{\cos t} = \sin t - \int \frac{dt}{\cos t} \\ &= \sin t - \ln |\sec t + \tan t| = \sin t - \ln(\sec t + \tan t), \end{aligned}$$

where the absolute value bars can be dropped as for $-\pi/2 < t < \pi t \sec t + \tan t > 0$. Integrating the equation for v we find

$$v(t) = -\cos t$$

and a particular solution the non-homogeneous equation can be taken to be

$$\begin{aligned} y_p(t) &= u(t)y_1(t) + v(t)y_2(t) = \cos t(\sin t - \ln(\sec t + \tan t)) + \sin t(-\cos t) \\ &= -\cos t \ln(\sec t + \tan t). \end{aligned}$$

Note that we have taken the constants of integration to be zero in each case. This is allowed as we are looking for particular integrals and we are free to pick up the simplest particular solution.

Thus, the general solution to the non-homogeneous equation is

$$y(t) = C_1 \cos t + C_2 \sin t - \cos t \ln(\sec t + \tan t).$$

Step 3.

To solve the initial value problem we must find the derivative of y :

$$y'(t) = -C_1 \sin t + C_2 \cos t + \sin t \ln(\sec t + \tan t) - 1$$

so that we obtain

$$1 = y(0) = C_1, \quad 1 = y'(0) = C_2 - 1,$$

hence $C_1 = 1$ and $C_2 = 2$. Therefore

$$y(t) = \cos t + 2 \sin t - \cos t \ln(\sec t + \tan t).$$

Judicious guessing

The method of judicious guessing, called also the method of undetermined coefficients, is based on the observation

that for some functions the operations performed on the left-hand side of the differential equation, that is, taking derivatives, multiplying by constants and addition, does not change the form of the function. To wit, the derivative of a polynomial is a polynomial, the derivative of an exponential function is an exponential function and, in general the derivative of the product of an exponential function and a polynomial is again of the same form. Trigonometric functions $\sin t$ and $\cos t$ are included into this class by Euler's formulae $\sin t = \frac{e^{it} - e^{-it}}{2i}$ and $\cos t = \frac{e^{it} + e^{-it}}{2}$. Thus, if the right-hand side is of this form, then it makes sense to expect that the same of the solution. Let us test this hypothesis on the following example.

Example 7.8 *Find a particular solution to*

$$y'' - 2y' - 3y = 3t^2.$$

The right-hand side is a polynomial of the second degree so we will look for a solution amongst polynomials. To decide polynomial of what degree we should try we note that if we try polynomials of zero or first degree then the left-hand side will be at most of this degree, as the differentiation lowers the degree of a polynomial. Thus, the simplest candidate appears to be a polynomial of second degree

$$y(t) = At^2 + Bt + C,$$

where A, B, C are coefficients to be determined. Inserting this polynomial into the equation we get

$$y'' - 2y' - 3y = 2A - 2B - 3C - (4A + 3B)t - 3At^2 = 3t^2,$$

from which we obtain the system

$$\begin{aligned} -3A &= 3, \\ -4A - 3B &= 0, \\ 2A - 2B - 3C &= 0. \end{aligned}$$

Solving this system, we obtain $A = -1$, $B = 4/3$ and $C = -14/9$ so that the solution is

$$y(t) = -t^2 - \frac{4}{3}t - \frac{14}{9}.$$

Unfortunately, there are some pitfalls in this method, as shown in the following example.

Example 7.9 Find a particular solution to

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

Using our method, we take $y(t) = Ae^{-t}$ but inserting it into the equation we find that

$$y'' - 2y' - 3y = Ae^{-t} + 2Ae^{-t} - 3Ae^{-t} = 0 \neq e^{-t},$$

so that no choice of the constant A can turn $y(t)$ into the solution of our equation. The reason for this is that e^{-t} is a solution to the homogeneous equation what could be ascertained directly by solving the characteristic equation $\lambda^2 - 2\lambda - 3 = (\lambda + 1)(\lambda - 3)$. A way of this trouble is to consider $y(t) = Ate^{-t}$ so that $y' = Ae^{-t} - Ate^{-t}$ and $y'' = -2Ae^{-t} + Ate^{-t}$ and

$$\begin{aligned} y'' - 2y' - 3y &= -2Ae^{-t} + Ate^{-t} - 2(Ae^{-t} - Ate^{-t}) - 3Ate^{-t} \\ &= -4e^{-t}, \end{aligned}$$

which agrees with e^{-t} if $A = -\frac{1}{4}$. Thus we have a particular solution

$$y(t) = -\frac{1}{4}te^{-t}.$$

In general, it can be proved that the following procedure always produces the solution to

$$y'' + a_1y' + a_0y = t^m e^{at} \quad (7.46)$$

where $a_0 \neq 0$ and m is a non-negative integer.

- I. When a is not a root of the characteristic equation $\lambda^2 + a_1\lambda + a_0 = 0$, then we use

$$y(t) = e^{at}(A_m t^m + A_{m-1} t^{m-1} + \dots + A_0); \quad (7.47)$$

- II. If a is a single root of the characteristic equation, then use (7.47) multiplied by t and if a is a double root, then use (7.47) multiplied by t^2 .

Remark 7.4 Note that if $a_0 = 0$, then (7.46) is reducible to a first order equation by methods of Subsection A2.2.5.

Also, equations with right-hand sides of the form

$$y'' + a_1y' + a_0y = f_1(t) + f_2(t) \dots + f_n(t), \quad (7.48)$$

can be handled as if $y_i(t)$ is a particular solution to

$$y'' + a_1y' + a_0y = f_i(t), \quad i = 1, \dots, n,$$

then the sum $y_p(t) = y_1(t) + y_2(t) \dots + y_n(t)$ is a particular solution to (7.48) as may be checked by direct substitution.

Example 7.10 Find a particular solution of

$$y'' + 4y = 32t \cos 2t - 8 \sin 2t.$$

Let us first find the characteristic roots. From the equation $\lambda^2 + 4 = 0$ we find $\lambda = \pm 2i$. Next we convert the RHS of the equation to the exponential form. Since $\cos 2t = (e^{i2t} + e^{-i2t})/2$ and $\sin 2t = (e^{i2t} - e^{-i2t})/2i$, we obtain

$$32t \cos 2t - 8 \sin 2t = (16t + 4i)e^{i2t} + (16t - 4i)e^{-i2t}.$$

In both cases we have the exponent being a single root of the characteristic equation so that we will be looking for solutions in the form $y_1(t) = t(At + B)e^{i2t}$ and $y_2(t) = t(Ct + D)e^{-i2t}$. For y_1 we obtain $y_1'(t) = (2At + B)e^{i2t} + 2it(At + B)e^{i2t}$ and $y_1''(t) = 2Ae^{i2t} + 4i(2At + B)e^{i2t} - 4t(At + B)e^{i2t}$ so that inserting these into the equation we obtain

$$2Ae^{i2t} + 4i(2At + B)e^{i2t} - 4(At^2 + Bt)e^{i2t} + 4t(At + B)e^{i2t} = (16t + 4i)e^{i2t}$$

which gives $2A + 4iB = 4i$ and $8iA = 16$. Thus $A = -2i$ and $B = 2$. Similarly, $C = 2i$ and $D = 2$ and we obtain the particular solution in the form

$$y(t) = t(-2it + 2)e^{i2t} + t(2it + 2)e^{-i2t} = 4t^2 \sin 2t + 4t \cos t,$$

where we used Euler's formula to convert exponential into trigonometric functions once again.

7.2.3 Applications

7.2.3.1 The mixing problem

In Subsection 2.6 we have derived the system

$$\begin{aligned} \frac{dx_1}{dt} &= r_1 + p_2 \frac{x_2}{V} - p_1 \frac{x_1}{V} \\ \frac{dx_2}{dt} &= p_1 \frac{x_1}{V} - (R_2 + p_2) \frac{x_2}{V}. \end{aligned} \tag{7.49}$$

describing mixing of components in two containers. Here, x_1 and x_2 are the amount of dye in vats 1 and 2, respectively. We re-write these equations using concentrations $c_1 = x_1/V$ and $c_2 = x_2/V$, getting

$$\begin{aligned}\frac{dc_1}{dt} &= \frac{r_1}{V} + \frac{p_2}{V}c_2 - \frac{p_1}{V}c_1 \\ \frac{dc_2}{dt} &= \frac{p_1}{V}c_1 - \frac{R_2 + p_2}{V}c_2.\end{aligned}\quad (7.50)$$

We solve this equations for numerical values of the flow rates $r_1/V = 0.01$, $p_1/V = 0.04$, $p_2/V = 0.03$ and $(R_2 + p_2)/V = 0.05$,

$$\begin{aligned}\frac{dc_1}{dt} &= 0.01 - 0.04c_1 + 0.03c_2 \\ \frac{dc_2}{dt} &= 0.04c_1 - 0.05c_2,\end{aligned}\quad (7.51)$$

and assume that at time $t = 0$ there was no dye in either vat, that is, we put

$$c_1(0) = 0, \quad c_2(0) = 0.$$

To practice another technique, we shall solve this system by reducing it to a second order equations, as described in Example 7.1. Differentiating the first equation and using the second we have

$$\begin{aligned}c_1'' &= -0.04c_1' + 0.03c_2' \\ &= -0.04c_1' + 0.03(0.04c_1 - 0.05c_2) \\ &= -0.04c_1' + 0.03\left(0.04c_1 - 0.05 \cdot \frac{100}{3}(c_1' - 0.01 + 0.04c_1)\right)\end{aligned}$$

so that, after some algebra,

$$c_1'' + 0.09c_1' + 0.008c_1 = 0.005.$$

We have obtained second order non-homogenous equation with constant coefficients. To find the characteristic roots we solve the quadratic equation

$$\lambda^2 + 0.09\lambda + 0.008 = 0$$

getting $\lambda_1 = -0.08$ and $\lambda_2 = -0.01$. Thus, the space of solutions of the homogeneous equations is spanned by $e^{-0.01t}$ and $e^{-0.08t}$. The right hand side is a constant and since zero is not a characteristic root, we can look for a solution to the nonhomogeneous problem in the form $y_p(t) = A$, which immediately gives $y_p(t) = 5/8$ so that the general solution of the non-homogeneous equation for c_1 is given by

$$c_1(t) = C_1 e^{-0.08t} + C_2 e^{-0.01t} + \frac{5}{8},$$

where C_1 and C_2 are constants whose values are to be found from the initial conditions.

Next we find c_2 by solving the first equation with respect to it, so that

$$\begin{aligned} c_2 &= \frac{100}{3} (c_1' + 0.04c_1 - 0.01) \\ &= \frac{100}{3} (-0.08C_1 e^{-0.08t} - 0.01C_2 e^{-0.01t} \\ &\quad + 0.04 \left(C_1 e^{-0.08t} + C_2 e^{-0.01t} + \frac{5}{8} \right) - 0.01) \end{aligned}$$

and

$$c_2(t) = -\frac{4}{3} C_1 e^{-0.08t} + C_2 e^{-0.01t} + 0.5.$$

Finally, we use the initial conditions $c_1(0) = 0$ and $c_2(0) = 0$

240 *Simultaneous systems of equations and higher order equations*

to get the system of algebraic equations for C_1 and C_2

$$\begin{aligned}C_1 + C_2 &= -\frac{5}{8}, \\ \frac{4}{3}C_1 - C_2 &= \frac{1}{2}.\end{aligned}$$

From these equations we find $C_1 = -3/56$ and $C_2 = -4/7$.

Hence

$$\begin{aligned}c_1(t) &= -\frac{3}{56}e^{-0.08t} - \frac{4}{7}e^{-0.01t} + \frac{5}{8} \\ c_2(t) &= \frac{1}{14}e^{-0.08t} - \frac{4}{7}e^{-0.01t} + \frac{1}{2}.\end{aligned}$$

From the solution formulae we obtain that $\lim_{t \rightarrow \infty} c_1(t) = \frac{5}{8}$ and $\lim_{t \rightarrow \infty} c_2(t) = \frac{1}{2}$. This means that the concentrations approach the steady state concentration as t becomes large. This is illustrated in Figures 2.10 and 2.11

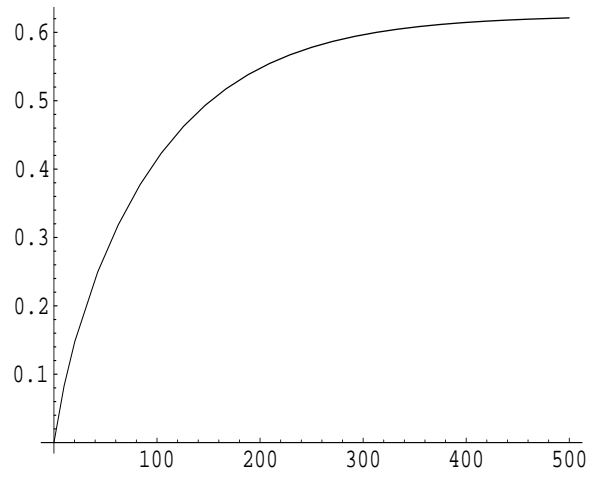


Fig 2.10 Approach to the steady-state of the concentration c_1 .

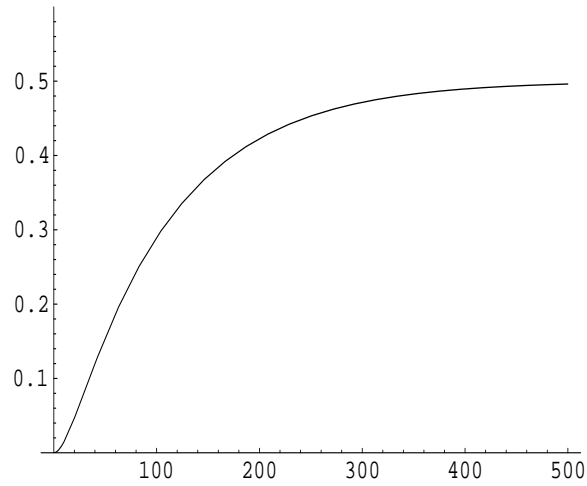


Fig 2.11 Approach to the steady state of the concentration c_2 .

7.2.3.2 *Forced oscillations*

Second order equations appear in many applications involving oscillations occurring due to the existence of an elastic force in the system. The reason for this is that the elastic force, at least for small displacements, is proportional to the displacement so that according to Newton's second law

$$my'' = -ky$$

where k is a constant. In general, there is a damping force (due to the resistance of the medium) and some external force, and then the full equation for oscillation reads

$$y'' + cy' + ky = F(t). \quad (7.52)$$

We shall discuss in detail a particular example of this equation describing the so-called *forced free vibrations*. In this case we have

$$y'' + \omega_0^2 y = \frac{F_0}{m} \cos \omega t, \quad (7.53)$$

where we denoted $\omega_0^2 = k/m$ and introduced a special periodic force $F(t) = F_0 \cos \omega t$ with constant magnitude F_0 and period ω .

The characteristic equation is $\lambda^2 + \omega_0^2 = 0$ so that we have imaginary roots $\lambda_{1,2} = \pm i\omega_0$ and the general solution to the homogeneous equations is given by

$$y_0(t) = C_1 \cos \omega_0 t + C_2 \sin \omega_0 t.$$

The frequency ω_0 is called the natural frequency of the system. The case $\omega \neq \omega_0$ gives a particular solution in the form

$$y_p(t) = \frac{F_0}{m(\omega_0^2 - \omega^2)} \cos \omega t$$

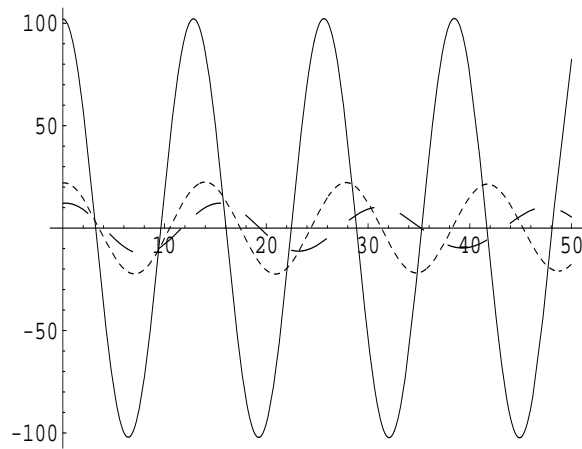


Fig 2.8 Forced free vibrations: non-resonance case with $\omega_0 = 0.5$ and $\omega = 0.4$ (dashed line), $\omega = 0.45$ (dotted line) and $\omega = 0.49$ (solid line). Note the increase in amplitude of vibrations as the frequency of the external force approaches the natural frequency of the system

so that the general solution is given by

$$y(t) = C_1 \cos \omega_0 t + C_2 \sin \omega_0 t + \frac{F_0}{m(\omega_0^2 - \omega^2)} \cos \omega t, \quad (7.54)$$

that is the solution is obtained as a sum of two periodic motions, as shown in Figure 2.8. Though there is nothing unusual here, we can sense that a trouble is brewing – if the the natural frequency of the system is close to the frequency of the external force, then the amplitude of vibrations can become very large, because the denominator in the last term in (7.54) is very small. Let us find out what happens if $\omega_0 = \omega$. In this case we convert $F(t) = F_0 \cos \omega t = F_0 \cos \omega_0 t = F_0(e^{i\omega_0 t} + e^{-i\omega_0 t})/2$ and look for the particular

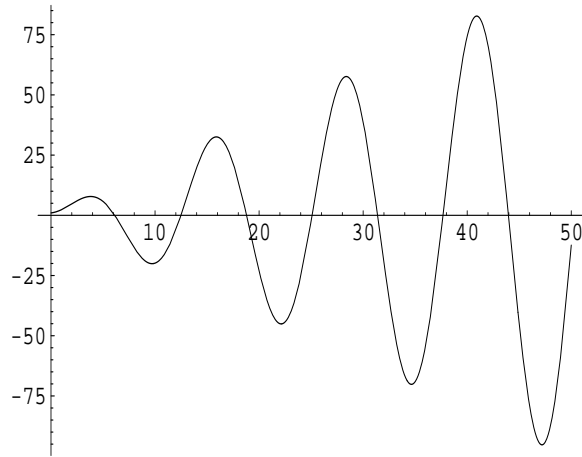


Fig 2.9 Forced free vibrations: the resonance case. The amplitude of vibrations increases to infinity.

solution in the form $y_p(t) = t(Ae^{i\omega_0 t} + Be^{-i\omega_0 t})$. We obtain $y'_p(t) = Ae^{i\omega_0 t} + Be^{-i\omega_0 t} + ti\omega_0(Ae^{i\omega_0 t} - Be^{-i\omega_0 t})$ and

$$y''_p(t) = i2(Ae^{i\omega_0 t} - Be^{-i\omega_0 t}) - t\omega_0^2(Ae^{i\omega_0 t} + Be^{-i\omega_0 t}).$$

Inserting these into the equation and comparing coefficients we find that $A = B = -F_0/4i\omega_0$ so that

$$y_p(t) = \frac{F_0}{2\omega_0} \frac{e^{i\omega_0 t} - e^{-i\omega_0 t}}{2i} = \frac{F_0}{2\omega_0} t \sin \omega_0 t$$

and the general solution is given by

$$y(t) = C_1 \cos \omega_0 t + C_2 \sin \omega_0 t + \frac{F_0}{2\omega_0} t \sin \omega_0 t.$$

A graph of such a function is shown in Figure 2.9. The important point of this example is that even small force can

induce very large oscillations in the system if its frequency is equal or even only very close to the natural frequency of the system. This phenomenon is called the *resonance* and is responsible for a number of spectacular collapses of constructions, like the collapse of Tacoma Bridge in the USA (oscillations induced by wind) and Broughton suspension bridge in England (oscillations introduced by soldiers marching in cadence).

8

Qualitative theory of differential and difference equations

8.1 Introduction

In this chapter we shall consider the system of differential equations

$$\mathbf{x}' = \mathbf{f}(t, \mathbf{x}) \quad (8.1)$$

where, in general,

$$\mathbf{x}(t) = \begin{pmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{pmatrix},$$

and

$$\mathbf{f}(t, \mathbf{x}) = \begin{pmatrix} f_1(t, x_1, \dots, x_n) \\ \vdots \\ f_n(t, x_1, \dots, x_n) \end{pmatrix}.$$

is a nonlinear function of \mathbf{x} . Our main focus will be on autonomous systems of two equations with two unknowns

$$\begin{aligned} x_1' &= f_1(x_1, x_2), \\ x_2' &= f_2(x_1, x_2). \end{aligned} \quad (8.2)$$

Unfortunately, even for such a simplified case there are no known methods of solving (8.2) in general form. Though it is, of course, disappointing, it turns out that knowing exact solution to (8.2) is not really necessary. For example, let $x_1(t)$ and $x_2(t)$ denote the populations, at time t , of two species competing amongst themselves for the limited food and living space in some region. Further, suppose that the rates of growth of $x_1(t)$ and $x_2(t)$ are governed by (8.2). In such a case, for most purposes it is irrelevant to know the population sizes at each time t but rather it is important to know some qualitative properties of them. Specifically, the most important questions biologists ask are:

- (i) Do there exist values ξ_1 and ξ_2 at which the two species coexist in a steady state? That is to say, are there numbers ξ_1 and ξ_2 such that $x_1(t) \equiv \xi_1$ and $x_2(t) \equiv \xi_2$ is a solution to (8.2)? Such values, if they exist, are called *equilibrium points* of (8.2).
- (ii) Suppose that the two species are coexisting in equilibrium and suddenly a few members of one or both species are introduced to the environment. Will $x_1(t)$ and $x_2(t)$ remain close to their equilibrium values for all future times? Or may be these extra few members will give one of the species a large advantage so that it will proceed to annihilate the other species?
- (iii) Suppose that x_1 and x_2 have arbitrary values at $t = 0$. What happens for large times? Will one species ultimately emerge victorious, or will the struggle for existence end in a draw?

Mathematically speaking, we are interested in determining the following properties of system (8.2).

Existence of equilibrium solutions. Do there exist constant vectors $\mathbf{x}^0 = (x_1^0, x_2^0)$ for which $\mathbf{x}(t) \equiv \mathbf{x}^0$ is a solution of (8.2)?

Stability. Let $\mathbf{x}(t)$ and $\mathbf{y}(t)$ be two solutions of (8.2) with initial values $\mathbf{x}(0)$ and $\mathbf{y}(0)$ very close to each other. Will $\mathbf{x}(t)$ and $\mathbf{y}(t)$ remain close for all future times, or will $\mathbf{y}(t)$ eventually diverge from $\mathbf{x}(t)$?

Long time behaviour. What happens to solutions $\mathbf{x}(t)$ as t approaches infinity. Do all solutions approach equilibrium values? If they do not approach equilibrium, do they at least exhibit some regular behaviour, like e.g. periodicity, for large times.

The first question can be answered immediately. In fact, since $\mathbf{x}(t)$ is supposed to be constant, then $\mathbf{x}'(t) \equiv 0$ and therefore \mathbf{x}^0 is an equilibrium value of (8.2) if and only if

$$\mathbf{f}(\mathbf{x}^0) \equiv \mathbf{0}, \quad (8.3)$$

that is, finding equilibrium solutions is reduced to solving a system of algebraic equations.

Example 8.1 *Find all equilibrium values of the system of differential equations*

$$\begin{aligned} x_1' &= 1 - x_2, \\ x_2' &= x_1^3 + x_2. \end{aligned}$$

We have to solve the system of algebraic equations

$$\begin{aligned} 0 &= 1 - x_2, \\ 0 &= x_1^3 + x_2. \end{aligned}$$

From the first equation we find $x_2 = 1$ and therefore $x_1^3 = -1$ which gives $x_1 = -1$ and the only equilibrium solution is

$$\mathbf{x}^0 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

8.2 The phase-plane and orbits

In this section we shall give rudiments of the "geometric" theory of differential equations. The aim of this theory is to obtain as complete a description as possible of all solutions of the system of differential equations (8.2)

$$\begin{aligned} x_1' &= f_1(x_1, x_2), \\ x_2' &= f_2(x_1, x_2), \end{aligned} \quad (8.4)$$

without solving it but by analysing geometric properties of its orbits. To explain the latter, we note that every solution $x_1(t), x_2(t)$ defines a curve in the three dimensional space (t, x_1, x_2) .

Example 8.2 *The solution $x_1(t) = \cos t$ and $x_2(t) = \sin t$ of the system*

$$\begin{aligned} x_1' &= -x_2, \\ x_2' &= x_1 \end{aligned}$$

describes a helix in the (t, x_1, x_2) space.

The foundation of the geometric theory of differential equations is the observation that every solution $x_1(t), x_2(t)$, $t_0 \leq t \leq t_1$, of (8.4) also describes a curve in the $x_1 - x_2$ plane, that is, as t runs from t_0 to t_1 , the points $(x_1(t), x_2(t))$ trace out a curve in the $x_1 - x_2$ plane. This curve is called

the *orbit*, or the *trajectory*, of the solution $\mathbf{x}(t)$ and the $x_1 - x_2$ plane is called the *phase plane* of the solutions of (8.4). Note that the orbit of an equilibrium solution reduces to a point.

Example 8.3 *The solution of the previous example, $x_1(t) = \cos t$, $x_2(t) = \sin t$ traces out the unit circle $x^2 + y^2 = 1$ when t runs from 0 to 2π , hence the unit circle is the orbit of this solution. If t runs from 0 to ∞ , then the pair $(\cos t, \sin t)$ traces out this circle infinitely often.*

Example 8.4 *Functions $x_1(t) = e^{-t} \cos t$ and $x_2(t) = e^{-t} \sin t$, $-\infty < t < \infty$, are a solution of the system*

$$\begin{aligned}x_1' &= -x_1 - x_2, \\x_2' &= x_1 - x_2.\end{aligned}$$

Since $r^2(t) = x_1^2(t) + x_2^2(t) = e^{-2t}$, we see that the orbit of this solution is a spiral traced towards the origin as t runs towards ∞ .

One of the advantages of considering the orbit of the solution rather than the solution itself is that it is often possible to find the orbit explicitly without prior knowledge of the solution. Let $x_1(t), x_2(t)$ be a solution of (8.4) defined in a neighbourhood of a point \bar{t} . If e.g. $x_1'(\bar{t}) \neq 0$, then we can solve $x_1 = x_1(t)$ getting $t = t(x_1)$ in some neighbourhood of $\bar{x} = x_1(\bar{t})$. Thus, for t near \bar{t} , the orbit of the solution $x_1(t), x_2(t)$ is given as the graph of $x_2 = x_2(t(x_1))$. Next, using the chain rule and the inverse function theorem

$$\frac{dx_2}{dx_1} = \frac{dx_2}{dt} \frac{dt}{dx_1} = \frac{x_2'}{x_1'} = \frac{f_2(x_1, x_2)}{f_1(x_1, x_2)}.$$

Thus, the orbits of the solution $x_1 = x_1(t), x_2(t) = x_2(t)$ of (8.4) are the solution curves of the first order scalar equation

$$\frac{dx_2}{dx_1} = \frac{f_2(x_1, x_2)}{f_1(x_1, x_2)} \tag{8.5}$$

and therefore to find the orbit of a solution there is no need to solve (8.4); we have to solve only the single first-order scalar equation (8.5).

Example 8.5 *The orbits of the system of differential equations*

$$\begin{aligned} x_1' &= x_2^2, \\ x_2' &= x_1^2. \end{aligned} \tag{8.6}$$

are the solution curves of the scalar equation $dx_2/dx_1 = x_2^2/x_1^2$. This is a separable equation and it is easy to see that every solution is of the form $x_2 = (x_1^3 + c)^{1/3}$, c constant. Thus, the orbits are the curves $x_2 = (x_1^3 + c)^{1/3}$ whenever $x_2 = x_1 \neq 0$ as then $x_1' = x_2' \neq 0$ and the procedure described above can be applied, see the example below.

Example 8.6 *A solution curve of (8.5) is an orbit of (8.4) if and only if $x_1' \neq 0$ and $x_2' \neq 0$ simultaneously along the solution. If a solution curve of (8.5) passes through an equilibrium point of (8.4), where $x_1'(\bar{t}) = 0$ and $x_2'(\bar{t}) = 0$ for some \bar{t} , then the entire solution curve is not an orbit but rather it is a union of several distinct orbits. For example, consider the system of differential equations*

$$x_1' = x_2(1 - x_1^2 - x_2^2), \tag{8.7}$$

$$x_2' = -x_1(1 - x_1^2 - x_2^2). \tag{8.8}$$

The solution curves of the scalar equation

$$\frac{dx_2}{dx_1} = -\frac{x_1}{x_2}$$

are the family of concentric circles $x_1^2 + x_2^2 = c^2$. Observe however that to get the latter equation we should have assumed $x_1^2 + x_2^2 = 1$ and that each point of this circle is an equilibrium point of (8.8). Thus, the orbits of (8.8) are the circles $x_1^2 + x_2^2 = c^2$ for $c \neq 1$ and each point of the unit circle.

Similarly, the full answer for the system (8.6) of the previous example is that $x_2 = (x_1^3 + c)^{1/3}$ are orbits for $c \neq 0$ as then neither solution curve passes through the only equilibrium point $(0, 0)$. For $c = 0$ the solution curve $x_2 = x_1$ consists of the equilibrium point $(0, 0)$ and two orbits $x_2 = x_1$ for $x_1 > 0$ and $x_1 < 0$.

Note that in general it is impossible to solve (8.5) explicitly. Hence, usually we cannot find the equation of orbits in a closed form. Nevertheless, it is still possible to obtain an accurate description of all orbits of (8.4). In fact, the system (8.4) provides us with an explicit information about how fast and in which direction solution is moving at each point of the trajectory. In fact, as the orbit of the solution $(x_1(t), x_2(t))$ is a curve of which $(x_1(t), x_2(t))$ is a parametric description, $(x_1'(t), x_2'(t)) = (f_1(x_1, x_2), f_2(x_1, x_2))$ is the tangent vector to the orbit at the point (x_1, x_2) showing, moreover, the direction at which the orbit is traversed. In particular, the orbit is vertical at each point (x_1, x_2) where $f_1(x_1, x_2) = 0$ and $f_2(x_1, x_2) \neq 0$ and it is horizontal at each point (x_1, x_2) where $f_1(x_1, x_2) \neq 0$ and $f_2(x_1, x_2) = 0$. As we noted earlier, each point (x_1, x_2) where $f_1(x_1, x_2) = 0$

and $f_2(x_1, x_2) = 0$ gives an equilibrium solution and the orbit reduces to this point.

8.3 Qualitative properties of orbits

Let us consider the initial value problem for the system (8.2):

$$\begin{aligned}x_1' &= f_1(x_1, x_2), \\x_2' &= f_2(x_1, x_2) \\x_1(t_0) &= x_1^0, \quad x_2(t_0) = x_2^0,\end{aligned}\tag{8.9}$$

As we have already mentioned in Subsection 7.1.3, Picard's theorem, Theorem 4.2, can be generalized to systems. Due to the importance of it for the analysis of orbits, we shall state it here in full.

Theorem 8.1 *If each of the functions $f_1(x_1, x_2)$ and $f_2(x_1, x_2)$ have continuous partial derivatives with respect to x_1 and x_2 . Then the initial value problem (8.9) has one and only one solution $\mathbf{x}(t) = (x_1(t), x_2(t))$, for every $\mathbf{x}^0 = (x_1^0, x_2^0) \in \mathbb{R}^2$ defined at least for t in some neighborhood of t_0 .*

Firstly, we prove the following result.

Lemma 8.1 *If $\mathbf{x}(t)$ is a solution to*

$$\mathbf{x}' = \mathbf{f}(\mathbf{x}),\tag{8.10}$$

then for any c the function $\hat{\mathbf{x}}(t) = \mathbf{x}(t+c)$ also satisfies this equation.

Proof. Define $\tau = t + c$ and use the chain rule for \hat{x} . We get

$$\frac{d\hat{\mathbf{x}}(t)}{dt} = \frac{d\mathbf{x}(t+c)}{dt} = \frac{d\mathbf{x}(\tau)}{d\tau} \frac{d\tau}{dt} = \frac{d\mathbf{x}(\tau)}{d\tau} = \mathbf{f}(\mathbf{x}(\tau)) = \mathbf{f}(\mathbf{x}(t+c)) = \mathbf{f}(\hat{\mathbf{x}}(t)).$$

Example 8.7 For linear systems the result follows directly as $\mathbf{x}(t) = e^{t\mathbf{A}}\mathbf{v}$ for arbitrary vector \mathbf{v} , so that $\hat{\mathbf{x}}(t) = \mathbf{x}(t+c) = e^{(t+c)\mathbf{A}}\mathbf{v} = e^{t\mathbf{A}}e^{c\mathbf{A}}\mathbf{v} = e^{t\mathbf{A}}\mathbf{v}'$ for some other vector \mathbf{v}' so that $\hat{\mathbf{x}}(t)$ is again a solution.

Proposition 8.1 Suppose that a solution $\mathbf{y}(t)$ of (8.9) approaches a vector \mathbf{v} as $t \rightarrow \infty$. Then \mathbf{v} is an equilibrium point of (8.9).

Proof. $\lim_{t \rightarrow \infty} \mathbf{y}(t) = \mathbf{v}$ is equivalent to $\lim_{t \rightarrow \infty} y_i(t) = v_i$, $i = 1, \dots, n$. This implies $\lim_{t \rightarrow \infty} y_i(t+h) = v_i$ for any fixed h . Using the mean value theorem we have

$$y_i(t+h) - y_i(t) = h y_i'(\tau) = h f_i(y_1(\tau), \dots, y_n(\tau)),$$

where $\tau \in [t, t+h]$. If $t \rightarrow \infty$, then also $\tau \rightarrow \infty$ and passing to the limit in the above equality, we obtain

$$0 = v_i - v_i = h f_i(v_1, \dots, v_n), \quad i = 1, \dots, n,$$

so that \mathbf{v} is an equilibrium point. ■

We shall now prove two properties of orbits that are crucial to analyzing system (8.2).

Theorem 8.2 Assume that the assumptions of Theorem 8.1 are satisfied. Then

(i) there exists one and only one orbit through every point

$\mathbf{x}^0 \in \mathbb{R}^2$. In particular, if the orbits of two solutions $\mathbf{x}(t)$ and $\mathbf{y}(t)$ have one point in common, then they must be identical.

(ii) Let $\mathbf{x}(t)$ be a solution to (8.2). If for some $T > 0$ and some t_0 we have $\mathbf{x}(t_0 + T) = \mathbf{x}(t_0)$, then $\mathbf{x}(t + T) = \mathbf{x}(t)$ for all t . In other words, if a solution $\mathbf{x}(t)$ returns to its starting value after a time $T > 0$, then it must be periodic (that is, it must repeat itself over every time interval of length T).

Proof. ad (i) Let \mathbf{x}^0 be any point in \mathbb{R}^2 . Then from Theorem 8.1 we know that there is a solution of the problem $\mathbf{x}' = \mathbf{f}(\mathbf{x}), \mathbf{x}(0) = \mathbf{x}^0$ and the orbit of this solution passes through \mathbf{x}^0 from the definition of the orbit. Assume now that there is another orbit passing through \mathbf{x}^0 , that is, there is a solution $\mathbf{y}(t)$ satisfying $\mathbf{y}(t_0) = \mathbf{x}^0$ for some t_0 . From Lemma 8.1 we know that $\hat{\mathbf{y}}(t) = \mathbf{y}(t + t_0)$ is also a solution. However, this solution satisfies $\hat{\mathbf{y}}(0) = \mathbf{y}(t_0) = \mathbf{x}^0$, that is, the same initial condition as $\mathbf{x}(t)$. By the uniqueness part of Theorem 8.1 we must then have $\mathbf{x}(t) = \hat{\mathbf{y}}(t) = \mathbf{y}(t + t_0)$ for all t for which the solutions are defined. This implies that the orbits are identical. In fact, if ξ is an element of the orbit of \mathbf{x} , then for some t' we have $\mathbf{x}(t') = \xi$. However, we have also $\xi = \mathbf{y}(t' + t_0)$ so that ξ belongs to the orbit of $\mathbf{y}(t)$. Conversely, if ξ belongs to the orbit of \mathbf{y} so that $\xi = \mathbf{y}(t'')$ for some t'' , then by $\xi = \mathbf{y}(t'') = \mathbf{x}(t'' - t_0)$, we see that ξ belongs to the orbit of \mathbf{x} .

ad (ii) Assume that for some numbers t_0 and $T > 0$ we have $\mathbf{x}(t_0) = \mathbf{x}(t_0 + T)$. The function $\mathbf{y}(t) = \mathbf{x}(t + T)$ is again a solution satisfying $\mathbf{y}(t_0) = \mathbf{x}(t_0 + T) = \mathbf{x}(t_0)$, thus from Theorem 8.1, $\mathbf{x}(t) = \mathbf{y}(t)$ for all t for which they are defined and therefore $\mathbf{x}(t) = \mathbf{x}(t + T)$ for all such t . ■

Example 8.8 *A curve in the shape of a figure 8 cannot be an orbit. In fact, suppose that the solution passes through the intersection point at some time t_0 , then completing the first loop, returns after time T , that is, we have $\mathbf{x}(t_0) = \mathbf{x}(t_0 + T)$. From (ii) it follows then that this solution is periodic, that is, it must follow the same loop again and cannot switch to the other loop.*

Corollary 8.1 *A solution $\mathbf{y}(t)$ of (8.2) is periodic if and only if its orbit is a closed curve in \mathbb{R}^2 .*

Proof. Assume that $\mathbf{x}(t)$ is a periodic solution of (8.2) of period T , that is $\mathbf{x}(t) = \mathbf{x}(t + T)$. If we fix t_0 , then, as t runs from t_0 to $t_0 + T$, the point $\mathbf{x}(t) = (x_1(t), x_2(t))$ traces a curve, say C , from $\xi = \mathbf{x}(t_0)$ back to the same point ξ without intersections and, if t runs from $-\infty$ to ∞ , the curve C is traced infinitely many times.

Conversely, suppose that the orbit is a closed curve (containing no equilibrium points). The point $\mathbf{x}(t)$ moves along this curve with a speed of magnitude $v(x_1, x_2) = \sqrt{f_1^2(x_1, x_2) + f_2^2(x_1, x_2)}$. The curve is closed and, since there is no equilibrium point on it, that is, f_1 and f_2 are not simultaneously zero at any point, the speed v has a non-zero minimum on it. Moreover, as the parametric description of this curve is differentiable, it has a finite length. Thus, the point $\mathbf{x}(t)$ starting from a point $\xi = \mathbf{x}(t_0)$ will traverse the whole curve in finite time, say T , that is $\mathbf{x}(t_0) = \mathbf{x}(t_0 + T)$ and the solution is periodic.

■

Example 8.9 Show that every solution $z(t)$ of the second order differential equation

$$z'' + z + z^3 = 0$$

is periodic. We convert this equation into a system: let $z = x_1$ so that

$$\begin{aligned}x_1' &= x_2, \\x_2' &= -x_1 - x_1^3.\end{aligned}$$

The orbits are the solution curves of the equation

$$\frac{dx_2}{dx_1} = -\frac{x_1 + x_1^3}{x_2},$$

so that

$$\frac{x_2^2}{2} + \frac{x_1^2}{2} + \frac{x_1^4}{4} = c^2$$

is the equation of orbits. If $c \neq 0$, then none of them contains the unique equilibrium point $(0, 0)$. By writing the above equation in the form

$$\frac{x_2^2}{2} + \left(\frac{x_1^2}{2} + \frac{1}{2}\right)^2 = c^2 + \frac{1}{4}$$

we see that for each $c \neq 0$ it describes a closed curve consisting of two branches $x_2 = \pm \frac{1}{\sqrt{2}} \sqrt{4c^2 + 1 - (x_1^2 + 1)^2}$ that stretch between $x_1 = \pm \sqrt{1 + \sqrt{4c^2 + 1}}$. Consequently, every solution is a periodic function.

8.4 An application to predator-prey models

8.4.1 Lotka-Volterra model

In this section we shall discuss the predator-prey model introduced in Section 2.6. It reads

$$\begin{aligned}\frac{dx_1}{dt} &= (r - f)x_1 - \alpha x_1 x_2, \\ \frac{dx_2}{dt} &= -(s + f)x_2 + \beta x_1 x_2\end{aligned}\quad (8.11)$$

where α, β, r, s, f are positive constants. In the predator-prey model x_1 is the density of the prey, x_2 is the density of the predators, r is the growth rate of the prey in the absence of predators, $-s$ is the growth rate of predators in the absence of prey (the population of predators dies out without the supply of the sole food source – prey). The quadratic terms account for predator–prey interaction and f represents indiscriminate killing of both prey and predators. The model was introduced in 1920s by an Italian mathematician Vito Volterra to explain why, in the period of reduced (indiscriminate) fishing, the relative number predators (sharks) significantly increased.

Let us consider first the model without fishing

$$\begin{aligned}\frac{dx_1}{dt} &= rx_1 - \alpha x_1 x_2, \\ \frac{dx_2}{dt} &= -sx_2 + \beta x_1 x_2\end{aligned}\quad (8.12)$$

Observe that there are two equilibrium solutions $x_1(t) = 0, x_2(t) = 0$ and $x_1(t) = s/\beta, x_2(t) = r/\alpha$. The first solution is not interesting as it corresponds to the total extinction. We observe also that we have two other solutions $x_1(t) = c_1 e^{rt}, x_2(t) = 0$ and $x_1(t) = 0, x_2(t) = c_2 e^{-st}$

that correspond to the situation when one of the species is extinct. Thus, both positive x_1 and x_2 semi-axes are orbits and, by Theorem 8.2 (i), any orbit starting in the first quadrant will stay there or, in other words, any solution with positive initial data will remain strictly positive for all times.

The orbits of (8.12) are the solution curves of the first order separable equation

$$\frac{dx_2}{dx_1} = \frac{x_2(-s + \beta x_1)}{x_1(r - \alpha x_2)} \quad (8.13)$$

Separating variables and integrating we obtain

$$r \ln x_2 - \alpha x_2 + s \ln x_1 - \beta x_1 = k$$

which can be written as

$$\frac{x_2^r}{e^{\alpha x_2}} \frac{x_1^s}{e^{\beta x_1}} = K. \quad (8.14)$$

Next, we prove that the curves defined by (8.14) are closed. It is not an easy task. To accomplish this we shall show that for each x_1 from a certain open interval $(x_{1,m}, x_{1,M})$ we have exactly two solutions $x_{2,m}(x_1)$ and $x_{2,M}(x_1)$ and that these two solutions tend to common limits as x_1 approaches $x_{1,m}$ and $x_{1,M}$.

First, let us define $f(x_2) = x_2^r e^{-\alpha x_2}$ and $g(x_1) = x_1^s e^{-\beta x_1}$. We shall analyze only f as g is of the same form. Due to positivity of all the coefficients, we see that $f(0) = 0$, $\lim_{x_2 \rightarrow \infty} f(x_2) = 0$ and also $f(x_2) > 0$ for $x_2 > 0$. Further

$$f'(x_2) = x_2^{r-1} e^{-\alpha x_2} (r - \alpha x_2),$$

so that f is increasing from 0 to $x_2 = r/\alpha$ where it attains global maximum, say M_2 , and then starts to decrease

monotonically to 0. Similarly, $g(0) = \lim_{x_1 \rightarrow \infty} g(x_1) = 0$ and $g(x_1) > 0$ for $x_1 > 0$ and it attains global maximum M_1 at $x_1 = s/\beta$. We have to analyze solvability of

$$f(x_2)g(x_1) = K.$$

Firstly, there are no solutions if $K > M_1M_2$, and for $K = M_1M_2$ we have the equilibrium solution $x_1 = s/\beta, x_2 = r/\alpha$. Thus, we have to consider $K = \lambda M_2$ with $\lambda < 1$. Let us write this equation as

$$f(x_2) = \frac{\lambda}{g(x_1)} M_2. \quad (8.15)$$

From the shape of the graph g we find that the equation $g(x_1) = \lambda$ has no solution if $\lambda > M_1$ but then $\lambda/g(x_1) \geq \lambda/M_1 > 1$ so that (8.15) is not solvable. If $\lambda = M_1$, then we have again the equilibrium solution. Finally, for $\lambda < M_1$ there are two solutions $x_{1,m}$ and $x_{1,M}$ satisfying $x_{1,m} < s/\beta < x_{1,M}$. Now, for x_1 satisfying $x_{1,m} < x_1 < x_{1,M}$ we have $\lambda/g(x_1) < 1$ and therefore for such x_1 equation (8.15) has two solutions $x_{2,m}(x_1)$ and $x_{2,M}(x_1)$ satisfying $x_{2,m} < r/\alpha < x_{2,M}$, again on the basis of the shape of the graph of f . Moreover, if x_1 moves towards either $x_{1,m}$ or $x_{1,M}$, then both solutions $x_{2,m}$ and $x_{2,M}$ move towards r/α , that is the set of points satisfying (8.15) is a closed curve.

Summarizing, the orbits are closed curves encircling the equilibrium solution $(s/\beta, r/\alpha)$ and are traversed in the anticlockwise direction. Thus, the solutions are periodic in time. The evolution can be described as follows. Suppose that we start with initial values $x_1 > s/\beta, x_2 < r/\alpha$, that is, in the lower right quarter of the orbit. Then the solution will move right and up till the prey population reaches

maximum $x_{1,M}$. Because there is a lot of prey, the number of predators will be still growing but then the number of prey will start decreasing, slowing down the growth of the predator's population. The decrease in the prey population will eventually bring the growth of predator's population to stop at the maximum $x_{2,M}$. From now on the number of predators will decrease but the depletion of the prey population from the previous period will continue to prevail till the population reaches the minimum $x_{1,m}$, when it will start to take advantage of the decreasing number of predators and will start to grow; this growth will, however, start to slow down when the population of predators will reach its minimum. However, then the number of prey will be increasing beyond the point when the number of predators is the least till the growing number of predators will eventually cause the prey population to decrease having reached its peak at $x_{1,M}$ and the cycle will repeat itself.

Now we are ready to provide the explanation of the observational data. Including fishing into the model, according to (8.11), amounts to changing parameters r and s to $r - f$ and $s + f$ but the structure of the system does not change, so that the equilibrium solution becomes

$$\left(\frac{s + f}{\beta}, \frac{r - f}{\alpha} \right). \quad (8.16)$$

Thus, with a moderate amount of fishing ($f < r$), in the equilibrium solution there is more fish and less sharks in comparison with no-fishing situation. Thus, if we reduce fishing, the equilibrium moves towards larger amount of sharks and lower amount of fish. Of course, this is true for equilibrium situation, which not necessarily corresponds to reality, but as the orbits are closed curves around the

20 Qualitative theory of differential and difference equations

equilibrium solution, we can expect that the amounts of fish and sharks in a non-equilibrium situation will change in a similar pattern. We can confirm this hypothesis by comparing average numbers of sharks and fish over the full cycle. For any function f its average over an interval (a, b) is defined as

$$\bar{f} = \frac{1}{b-a} \int_a^b f(t) dt,$$

so that the average numbers of fish and sharks over one cycle is given by

$$\bar{x}_1 = \frac{1}{T} \int_0^T x_1(t) dt, \quad \bar{x}_2 = \frac{1}{T} \int_0^T x_2(t) dt.$$

It turns out that these averages can be calculated explicitly. Dividing the first equation of (8.12) by x_1 gives $x_1'/x_1 = r - \alpha x_2$. Integrating both sides, we get

$$\frac{1}{T} \int_0^T \frac{x_1'(t)}{x_1(t)} dt = \frac{1}{T} \int_0^T (r - \alpha x_2(t)) dt.$$

The left-hand side can be evaluated as

$$\int_0^T \frac{x_1'(t)}{x_1(t)} dt = \ln x_1(T) - \ln x_1(0) = 0$$

on account of the periodicity of x_1 . Hence,

$$\frac{1}{T} \int_0^T (r - \alpha x_2(t)) dt = 0,$$

and

$$\overline{x_2} = \frac{r}{\alpha}. \quad (8.17)$$

In the same way,

$$\overline{x_1} = \frac{s}{\beta}, \quad (8.18)$$

so that the average values of x_1 and x_2 are exactly the equilibrium solutions. Thus, we can state that introducing fishing is more beneficial to prey than predators as the average numbers of prey increases while the average number of predators will decrease in accordance with (8.16), while reducing fishing will have the opposite effect of increasing the number of predators and decreasing the number of prey.

8.4.2 Modified Lotka-Volterra model

Interestingly enough, the Volterra model was not universally accepted by biologists and ecologists, despite being able to explain the observed data for the shark-fish population. Some researchers pointed out that the oscillatory behaviour predicted by the Lotka-Volterra model is not observed in most predator-prey systems, which rather tend to equilibrium states as time evolves. However, one has to have in mind that system (8.11) is not intended as a description of a general predator-prey system but a particular one in which there is an abundance of food for fish that allow them to grow at an exponential rate. In general, there is a competition among both food fish and predators that can be taken into account by including "overcrowding" terms into the system, as in the logistic model. This results in

the following system of differential equations.

$$\begin{aligned}x_1' &= ax_1 - bx_1x_2 - ex_1^2, \\x_2' &= -cx_2 + dx_1x_2 - fx_2^2,\end{aligned}\tag{8.19}$$

where a, b, e, c, d, f are positive constants. This system can describe the population growth of two species x_1 and x_2 in an environment of limited capacity, where the species x_2 depends on the species x_1 for its survival. Assume that $c/d > a/e$. We prove that every solution $(x_1(t), x_2(t))$ of (8.19), with $x_1(0), x_2(0) > 0$ approaches the equilibrium solution $x_1 = a/e, x_2 = 0$, as t approaches infinity. As a first step, we show that the solutions with positive initial data must stay positive, that is, the orbit of any solution originating in the first quadrant must stay in this quadrant. Otherwise the model would not correspond to reality. First, let us observe that putting $x_2(t) \equiv 0$ we obtain the logistic equation for x_1 that can be solved giving

$$x_1(t) = \frac{ax^0}{ex^0 + (a - ex^0)\exp(-at)}$$

where $x^0 \geq 0$. The orbits of these solutions is the equilibrium point $(0, 0)$, the segment $0 \leq x_1 < a/e, x_2 = 0$ for $x^0 < a/e$, the equilibrium point $(a/e, 0)$ for $x^0 = a/e$ and the segments $a/e < x_1 < \infty, x_2 = 0$ for $x^0 > a/e$. Thus, the positive x_1 -semiaxis $x_1 \geq 0$ is the union of these four orbits. Similarly, putting $x_1(t) \equiv 0$ we obtain the equation

$$x_2' = -cx_2 - fx_2^2.$$

To use the theory of Section 5, we observe that the equilibrium points of this equation are $x_2 = 0$ and $x_2 = -c/f$ so that there are no equilibria on the positive x_2 -semiaxis and

Fig.5.1 Regions described in the analysis of 8.19

$-cx_2 - fx_2^2 < 0$ for $x_2 > 0$. Therefore any solution with initial value $x_2^0 > 0$ will decrease converging to 0 and the semiaxis $x_2 > 0$ is a single orbit of (8.19). Thus, if a solution of (8.19) left the first quadrant, its orbit would cross one of the orbits the positive semiaxes consist of, which is precluded by uniqueness of orbits.

In the next step we divide the first quadrant into regions where the derivatives x_1' and x_2' are of a fixed sign. This is done by drawing lines l_1 and l_2 , as in Fig. 5, across which one of the other derivative vanishes. The line l_1 is determined by $-ex_1/b + a/b$ so that $x_1' > 0$ in region I

and $x'_1 < 0$ in regions II and III. The line l_2 is given by $x_2 = dx_1/f - c/f$ and $x'_2 < 0$ in regions I and II, and $x'_2 > 0$ in region III.

We describe the behaviour of solutions in each region in the sequence of observations.

Observation 1. *Any solution to (8.19) which starts in the region I at $t = t_0$ will remain in this region for all $t > t_0$ and ultimately approach the equilibrium $x_1 = a/e, x_2 = 0$.*

Proof. If the solution $x_1(t), x_2(t)$ leaves region I at some time $t = t^*$, then $x'_1(t^*) = 0$, since the only way to leave this region is to cross the line l_1 . Differentiation the first equation in (8.19) gives

$$x''_1 = ax'_1 - bx'_1x_2 - bx_1x'_2 - 2ex_1x'_1$$

so that at $t = t^*$ we obtain

$$x''_1(t^*) = -bx_1(t^*)x'_2(t^*).$$

Since $x'_2(t^*) < 0, x''_1(t^*) > 0$ which means that $x_1(t^*)$ is a local minimum. However, $x_1(t)$ reaches this point from region I where it is increasing, which is a contradiction. Thus, the solution $(x_1(t), x_2(t))$ stays in region I for all times $t \geq t_0$. However, any solution staying in I must be bounded and $x'_1 > 0$ and $x'_2 < 0$ so that $x_1(t)$ is increasing and $x_2(t)$ is decreasing and therefore they must tend to a finite limit. By Proposition 8.1, this limit must be an equilibrium point. The only equilibria are $(0, 0)$ and $(a/e, 0)$ and the solution cannot tend to the former as $x_1(t)$ is positive and increasing. Thus, any solution starting in region I for some time $t = t_0$ tends to the equilibrium $(a/e, 0)$ as $t \rightarrow \infty$.

Observation 2. *Any solution of (8.19) that starts in*

region III at time t_0 must leave this region at some later time.

Proof. Suppose that a solution $x_1(t), x_2(t)$ stays in region III for all time $t \geq 0$. Since the sign of both derivatives x_1' and x_2' is fixed, $x_1(t)$ decreases and $x_2(t)$ increases, thus $x_1(t)$ must tend to a finite limit. $x_2(t)$ cannot escape to infinity as the only way it could be achieved would be if also $x_1(t)$ tended to infinity, which is impossible. Thus, $(x_1(t), x_2(t))$ tend to a finite limit that, by Proposition 8.1, has to be an equilibrium. However, there are no equilibria to be reached from region III and thus the solution must leave this region at some time.

Observation 3. Any solution of (8.19) that starts in region II at time $t = t_0$ and remains in this region for all $t \geq 0$ must approach the equilibrium solution $x_1 = a/e, x_2 = 0$.

Proof. Suppose that a solution $(x_1(t), x_2(t))$ stays in region II for all $t \geq t_0$. Then both $x_1(t)$ and $x_2(t)$ are decreasing and, since the region is bounded from below, we see that this solution must converge to an equilibrium point, in this case necessarily $(a/e, 0)$.

Observation 4. A solution cannot enter region III from region II.

Proof. This case is similar to Observation 1. Indeed, if the solution crosses l_2 from II to III at $t = t^*$, then $x_2'(t^*) = 0$ but then, from the second equation of (8.19)

$$x_2''(t^*) = dx_2(t^*)x_1'(t^*) < 0$$

so that $x_2(t^*)$ is a local maximum. This is, however, impossible, as $x_2(t)$ is decreasing in region II.

Summarizing, if the initial values are in regions I or II,

then the solution tends to the equilibrium $(a/e, 0)$ as $t \rightarrow \infty$, by Observations 1,3 and 4. If the solution starts from region III, then at some point it must enter region II and we can apply the previous argument to claim again that the solution will eventually approach the equilibrium $(a/e, 0)$. Finally, if a solution starts on l_1 , it must immediately enter region I as $x'_2 < 0$ and $x'_1 < 0$ in region II (if the solution ventured into II from l_1 , then either x'_1 or x'_2 would have to be positive somewhere in II). Similarly, any solution starting from l_2 must immediately enter II. Thus, all the solution starting in the first quadrant (with strictly positive initial data) will converge to $(a/e, 0)$ as $t \rightarrow \infty$.

8.5 Stability of linear systems

8.5.1 Planar linear systems

In this section we shall present a complete description of all orbits of the linear differential system

$$\mathbf{y}' = \mathcal{A}\mathbf{y} \quad (8.20)$$

where $\mathbf{y}(t) = (y_1(t), y_2(t))$ and

$$\mathcal{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

We shall assume that \mathcal{A} is invertible, that is, $ad - bc \neq 0$. In such a case $\mathbf{y} = (0, 0)$ is the only equilibrium point of (8.20).

The phase portrait is fully determined by the eigenvalues of the matrix \mathcal{A} . Let us briefly describe all possible cases, as determined by the theory of Chapter 7. The general solution can be obtained as a linear combination of two

linearly independent solutions. To find them, we have to find first the eigenvalues of \mathcal{A} , that is, solutions to

$$(\lambda - \lambda_1)(\lambda - \lambda_2) = (a - \lambda)(d - \lambda) - bc = \lambda^2 - \lambda(d + a) + ad - bc.$$

Note that by the assumption on invertibility, $\lambda = 0$ is not an eigenvalue of \mathcal{A} . We have the following possibilities:

- a) $\lambda_1 \neq \lambda_2$. In this case each eigenvalue must be simple and therefore we have two linearly independent eigenvectors $\mathbf{v}^1, \mathbf{v}^2$. The expansion $e^{t\mathcal{A}}\mathbf{v}^i$ for $i = 1, 2$ terminates after the first term. We distinguish two cases.

◊ If λ_1, λ_2 are real numbers, then the general solution is given simply by

$$\mathbf{y}(t) = c_1 e^{\lambda_1 t} \mathbf{v}^1 + c_2 e^{\lambda_2 t} \mathbf{v}^2. \tag{8.21}$$

◊ If λ_1, λ_2 are complex numbers, then the general solution is still given by the above formula but the functions above are complex and we would rather prefer solution to be real. To achieve this, we note that λ_1, λ_2 must be necessarily complex conjugate $\lambda_1 = \xi + i\omega, \lambda_2 = \xi - i\omega$, where ξ and ω are real. It can be also proved that the associated eigenvectors \mathbf{v}^1 and \mathbf{v}^2 are also complex conjugate. Let $\mathbf{v}^1 = \mathbf{u} + i\mathbf{v}$; then the real-valued general solution is given by

$$\mathbf{y}(t) = c_1 e^{\xi t} (\mathbf{u} \cos \omega t - \mathbf{v} \sin \omega t) + c_2 e^{\xi t} (\mathbf{u} \sin \omega t + \mathbf{v} \cos \omega t). \tag{8.22}$$

This solution can be written in a more compact

form

$$\mathbf{y}(t) = e^{\xi t} (A_1 \cos(\omega t - \phi_1), A_2 \cos(\omega t - \phi_2)), \quad (8.23)$$

for some choice of constants $A_1, A_2 > 0$ and ϕ_1, ϕ_2 .

b) $\lambda_1 = \lambda_2 = \lambda$. There are two cases to distinguish.

◊ There are two linearly independent eigenvectors \mathbf{v}^1 and \mathbf{v}^2 corresponding to λ . In this case the general solution is given by

$$\mathbf{y}(t) = e^{\lambda t} (c_1 \mathbf{v}^1 + c_2 \mathbf{v}^2). \quad (8.24)$$

◊ If there is only one eigenvector, then following the discussion above, we must find a vector \mathbf{v}^2 satisfying $(\lambda I - \mathcal{A})\mathbf{v}^2 \neq 0$ and $(\lambda I - \mathcal{A})^2 \mathbf{v}^2 = 0$. However, since we are in the two-dimensional space, the latter is satisfied by any vector \mathbf{v}^2 and, since the eigenspace is one dimensional, from

$$(\lambda I - \mathcal{A})^2 \mathbf{v}^2 = (\lambda I - \mathcal{A})(\lambda I - \mathcal{A})\mathbf{v}^2 = 0$$

it follows that $(\lambda I - \mathcal{A})\mathbf{v}^2 = k\mathbf{v}^1$. Thus, the formula for $e^{\mathcal{A}t}\mathbf{v}^2$ simplifies as

$$e^{t\mathcal{A}}\mathbf{v}^2 = e^{\lambda t} (\mathbf{v}^2 + t(\lambda I - \mathcal{A})\mathbf{v}^2) = e^{\lambda t} (\mathbf{v}^2 + kt\mathbf{v}^1).$$

Thus, the general solution in this case can be written as

$$\mathbf{y}(t) = e^{\lambda t} ((c_1 + c_2 kt)\mathbf{v}^1 + c_2 \mathbf{v}^2). \quad (8.25)$$

Remark 8.1 *Before we embark on describing phase portraits, let us observe that if we change the direction of time in (8.20): $\tau = -t$ and $\mathbf{z}(\tau) = \mathbf{y}(-\tau) = \mathbf{y}(t)$, then we obtain*

$$\mathbf{z}'_{\tau} = -\mathcal{A}\mathbf{z}$$

and the eigenvalues of $-\mathcal{A}$ are precisely the negatives of the eigenvalues of \mathcal{A} . Thus, the orbits of solutions corresponding to systems governed by \mathcal{A} and $-\mathcal{A}$ or, equivalently, with eigenvalues that differ only by sign, are the same with only difference being the direction in which they are traversed.

We are now in a position to describe all possible phase portraits of (8.20). Again we have to go through several cases.

- i) $\lambda_2 < \lambda_1 < 0$. Let \mathbf{v}^1 and \mathbf{v}^2 be eigenvectors of \mathcal{A} with eigenvalues λ_1 and λ_2 , respectively. In the $y_1 - y_2$ plane we draw four half-lines l_1, l'_1, l_2, l'_2 parallel to $\mathbf{v}^1, -\mathbf{v}^1, \mathbf{v}^2$ and $-\mathbf{v}^2$, respectively, and emanating from the origin, as shown in Fig 2.1. Observe first that $\mathbf{y}(t) = ce^{\lambda_i t} \mathbf{v}^i, i = 1, 2$, are the solutions to (8.20) for any choice of a non-zero constant c and, as they are parallel to \mathbf{v}^i , the orbits are the half-lines l_1, l'_1, l_2, l_2 (depending on the sign of the constant c) and all these orbits are traced towards the origin as $t \rightarrow \infty$. Since every solution $\mathbf{y}(t)$ of (8.20) can be written as

$$\mathbf{y}(t) = c_1 e^{\lambda_1 t} \mathbf{v}^1 + c_2 e^{\lambda_2 t} \mathbf{v}^2$$

for some choice of constants c_1 and c_2 and $\lambda_1, \lambda_2 < 0$, every solution tends to $(0, 0)$ as $t \rightarrow \infty$, and so every orbit approaches the origin for $t \rightarrow \infty$. We can prove an even stronger fact – as $\lambda_2 < \lambda_1$, the second term becomes negligible for large t and therefore the tangent of the orbit of $\mathbf{y}(t)$ approaches the direction of l_1 if $c_1 > 0$ and of l'_1 if $c_1 < 0$. Thus, every orbit except that with $c_1 = 0$

Fig. 5.2 Stable node

approaches the origin along the same fixed line. Such a type of an equilibrium point is called a *stable node*. If we have $0 < \lambda_1 < \lambda_2$, then by Remark 8.1, the orbits of (8.20) will have the same shape as in case i) but the arrows will be reversed so that the origin will repel all the orbits and the orbits will be unbounded as $t \rightarrow \infty$. Such an equilibrium point is called an *unstable node*.

- ii) $\lambda_1 = \lambda_2 = \lambda < 0$. In this case the phase portrait of (8.20) depends on whether \mathcal{A} has one or two linearly independent eigenvectors. In the latter case,

the general solution in given (see b) above) by

$$\mathbf{y}(t) = e^{\lambda t}(c_1\mathbf{v}^1 + c_2\mathbf{v}^2),$$

so that orbits are half-lines parallel to $c_1\mathbf{v}^1 + c_2\mathbf{v}^2$. These half-lines cover every direction of the $y_1 - y_2$ plane and, since $\lambda < 0$, each solution will converge to $(0, 0)$ along the respective line. Thus, the phase portrait looks like in Fig. 5.3a. If there is only one independent eigenvector corresponding to λ , then by (8.25)

$$\mathbf{y}(t) = e^{\lambda t}((c_1 + c_2kt)\mathbf{v}^1 + c_2\mathbf{v}^2)$$

for some choice of constants c_1, c_2, k . Obviously, every solution approaches $(0, 0)$ as $t \rightarrow \infty$. Putting $c_2 = 0$, we obtain two half-line orbits $c_1e^{\lambda t}\mathbf{v}^1$ but, contrary to the case i), there are no other half-line orbits. In addition, the term $c_1\mathbf{v}^1 + c_2\mathbf{v}^2$ becomes small in comparison with $c_2kt\mathbf{v}^1$ as $t \rightarrow \infty$ so that the orbits approach the origin in the direction of $\pm\mathbf{v}^1$. The phase portrait is presented in Fig. 5.3b. The equilibrium in both cases is called the *stable degenerate node*. If $\lambda_1 = \lambda_2 > 0$, then again by Remark 8.1, the picture in this case will be the same as in Fig. 5.3 a-b but with the direction of arrows reversed. Such equilibrium point is called an *unstable degenerate node*.

- iii) $\lambda_1 < 0 < \lambda_2$. As in case i), in the $y_1 - y_2$ plane we draw four half-lines l_1, l'_1, l_2, l'_2 that emanate from the origin and are parallel to $\mathbf{v}^1, -\mathbf{v}^1, \mathbf{v}^2$ and $-\mathbf{v}^2$, respectively, as shown in Fig 2.3. Any solution is

Fig 5.3 Stable degenerate node

given by

$$\mathbf{y}(t) = c_1 e^{\lambda_1 t} \mathbf{v}^1 + c_2 e^{\lambda_2 t} \mathbf{v}^2$$

for some choice of c_1 and c_2 . Again, the half-lines are the orbits of the solutions: l_1, l'_1 for $c_1 e^{\lambda_1 t} \mathbf{v}^1$ with $c_1 > 0$ and $c_1 < 0$, and l_2, l'_2 for $c_2 e^{\lambda_2 t} \mathbf{v}^2$ with $c_2 > 0$ and $c_2 < 0$, respectively. However, the direction of arrows is different on each pair of half-lines: while the solution $c_1 e^{\lambda_1 t} \mathbf{v}^1$ converges towards $(0, 0)$ along l_1 or l'_1 as $t \rightarrow \infty$, the solution $c_2 e^{\lambda_2 t} \mathbf{v}^2$ becomes unbounded moving along l_2 or




Fig 5.4 A saddle point

l'_2 , as $t \rightarrow \infty$. Next, we observe that if $c_1 \neq 0$, then for large t the second term $c_2 e^{\lambda_2 t} \mathbf{v}^2$ becomes negligible and so the solution becomes unbounded as $t \rightarrow \infty$ with asymptotes given by the half-lines l_2, l'_2 , respectively. Similarly, for $t \rightarrow -\infty$ the term $c_1 e^{\lambda_1 t} \mathbf{v}^1$ becomes negligible and the solution again escapes to infinity, but this time with asymptotes l_1, l'_1 , respectively. Thus, the phase portrait, given in Fig. 5.4, resembles a saddle near $y_1 = y_2 = 0$ and, not surprisingly, such an equilibrium point is called a *saddle*. The case $\lambda_2 < 0 < \lambda_1$ is of course

symmetric.

iv) $\lambda_1 = \xi + i\omega, \lambda_2 = \xi - i\omega$. In (8.23) we derived the solution in the form

$$\mathbf{y}(t) = e^{\xi t} (A_1 \cos(\omega t - \phi_1), A_2 \cos(\omega t - \phi_2)).$$

We have to distinguish three cases:

α) If $\xi = 0$, then

$$y_1(t) = A_1 \cos(\omega t - \phi_1), \quad y_2(t) = A_2 \cos(\omega t - \phi_2),$$

both are periodic functions with period $2\pi/\omega$ and y_1 varies between $-A_1$ and A_1 while y_2 varies between $-A_2$ and A_2 . Consequently, the orbit of any solution $\mathbf{y}(t)$ is a closed curve containing the origin inside and the phase portrait has the form presented in Fig. 5.5 a. For this reason we say that the equilibrium point of (8.20) is a *center* when the eigenvalues of \mathcal{A} are purely imaginary. The direction of arrows must be determined from the equation. The simplest way of doing this is to check the sign of y_2' when $y_2 = 0$. If at $y_2 = 0$ and $y_1 > 0$ we have $y_2' > 0$, then all the orbits are traversed in the anticlockwise direction, and conversely.

β) If $\xi < 0$, then the factor $e^{\xi t}$ forces the solution to come closer to zero at every turn so that the solution spirals into the origin giving the picture presented in Fig. 5.5 b. The orientation of the spiral must be again determined directly from the equation. Such an equilibrium point is called a *stable focus*.

γ) If $\xi > 0$, then the factor $e^{\xi t}$ forces the solution to spiral outwards creating the picture shown

Fig. 5.5 Center, stable and unstable foci

in Fig. 5. 5 c. Such an equilibrium point is called an *unstable focus*.

8.6 Stability of equilibrium solutions

In this section we shall describe how the solutions of the system (8.9) behave under small perturbations. First of all, we have to make this concept precise. Let $\mathbf{y}(t)$ be a solution of the system

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}). \quad (8.26)$$

Definition 8.1 *The solution $\mathbf{y}(t)$ of (8.26) is stable if every other solution $\mathbf{x}(t)$ that starts sufficiently close to $\mathbf{y}(t)$ will remain close to it for all times. Precisely, $\mathbf{y}(t)$ is stable if for any ϵ there is δ such that for any solution \mathbf{x} of (8.26) from*

$$\|\mathbf{x}(t_0) - \mathbf{y}(t_0)\| \leq \delta,$$

it follows

$$\|\mathbf{x}(t) - \mathbf{y}(t)\| \leq \epsilon.$$

Moreover, we say that $\mathbf{y}(t)$ is asymptotically stable, if it is stable and there is δ such that if $\|\mathbf{x}(t_0) - \mathbf{y}(t_0)\| \leq \delta$, then

$$\lim_{t \rightarrow \infty} \|\mathbf{y}(t) - \mathbf{x}(t)\| = 0.$$

The main interest in applications is to determine the stability of stationary solutions.

8.6.1 Linear systems

For linear systems the question of stability of solutions can be fully resolved. Firstly, we observe that any solution $\mathbf{y}(t)$ of the linear system

$$\mathbf{y}' = \mathcal{A}\mathbf{y} \tag{8.27}$$

is stable if and only if the stationary solution $\mathbf{x}(t) = (0, 0)$ is stable. To show this, let $\mathbf{z}(t)$ be any other solution; then $\mathbf{v}(t) = \mathbf{y}(t) - \mathbf{z}(t)$ is again a solution of (8.27). Therefore, if the null solution is stable, then $\mathbf{v}(t)$ remains close to zero for all t if $\mathbf{v}(t_0)$ is small. This, however, implies that $\mathbf{y}(t)$ will remain close to $\mathbf{z}(t)$ if $\mathbf{y}(t_0)$ is sufficiently close to $\mathbf{z}(t_0)$. A similar argument applies to the asymptotic stability. Conversely, let the null solution be unstable; then

there is a solution $\mathbf{h}(t)$ such that $\mathbf{h}(t_0)$ is small, but $\mathbf{h}(t)$ becomes very large as t approaches to infinity. For any solution $\mathbf{y}(t)$, the function $\mathbf{z}(t) = \mathbf{y}(t) + \mathbf{h}(t)$ is again a solution to (8.27) which sways away from $\mathbf{y}(t)$ for large t . Thus, any solution is unstable.

The discussion of phase-portraits for two-dimensional linear, given in the previous section allows to determine easily under which conditions $(0, 0)$ is stable. Clearly, the only stable cases are when real parts of both eigenvalues are non-positive with asymptotic stability offered by eigenvalues with strictly negative ones (the case of the centre is an example of a stable but not asymptotically stable equilibrium point).

Analogous results can be formulated for linear systems in higher dimensions. By considering formulae for solutions we ascertain that the equilibrium point is (asymptotically stable) if all the eigenvalues have negative real parts and is unstable if at least one eigenvalue has positive real part. The case of eigenvalues with zero real part is more complicated as in higher dimension we can have multiple complex eigenvalues. Here, again from the formula for solutions, we can see that if for each eigenvalue with zero real part of algebraic multiplicity k there is k linearly independent eigenvectors, the solution is stable. However, if geometric and algebraic multiplicities of at least such eigenvalue are different, then in the solution corresponding to this eigenvalue there will appear a polynomial in t which will cause the solution to be unstable.

8.6.2 Nonlinear systems

The above considerations can be used to determine stability of equilibrium points of arbitrary differential equations

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}). \quad (8.28)$$

Let us first note the following result.

Lemma 8.2 *If \mathbf{f} has continuous partial derivatives of the first order in some neighbourhood of \mathbf{y}^0 , then*

$$\mathbf{f}(\mathbf{x} + \mathbf{y}^0) = \mathbf{f}(\mathbf{y}^0) + \mathcal{A}\mathbf{x} + \mathbf{g}(\mathbf{x}) \quad (8.29)$$

where

$$\mathcal{A} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{y}^0) & \cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{y}^0) \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(\mathbf{y}^0) & \cdots & \frac{\partial f_n}{\partial x_n}(\mathbf{y}^0) \end{pmatrix},$$

and $\mathbf{g}(\mathbf{x})/\|\mathbf{x}\|$ is continuous in some neighbourhood of \mathbf{y}^0 and vanishes at $\mathbf{x} = \mathbf{y}^0$.

Proof. The matrix \mathcal{A} has constant entries so that \mathbf{g} defined by

$$\mathbf{g}(\mathbf{x}) = \mathbf{f}(\mathbf{x} + \mathbf{y}^0) - \mathbf{f}(\mathbf{y}^0) - \mathcal{A}\mathbf{x}$$

is a continuous function of \mathbf{x} . Hence, $\mathbf{g}(\mathbf{x})/\|\mathbf{x}\|$ is also continuous for $\mathbf{x} \neq \mathbf{0}$. Using now Taylor's formula for each component of \mathbf{f} we obtain

$$f_i(\mathbf{x} + \mathbf{y}^0) = f_i(\mathbf{y}^0) + \frac{\partial f_i}{\partial x_1}(\mathbf{y}^0)x_1 + \cdots + \frac{\partial f_i}{\partial x_n}(\mathbf{y}^0)x_n + R_i(\mathbf{x}), \quad i = 1, \dots, n,$$

where, for each i , the remainder R_i satisfies

$$|R_i(\mathbf{x})| \leq M(\|\mathbf{x}\|)\|\mathbf{x}\|$$

and M tends to zero is $\|\mathbf{x}\| \rightarrow 0$. Thus,

$$\mathbf{g}(\mathbf{x}) = (R_1(\mathbf{x}), \dots, R_n(\mathbf{x}))$$

and

$$\frac{\|\mathbf{g}(\mathbf{x})\|}{\|\mathbf{x}\|} \leq M(\|\mathbf{x}\|) \rightarrow 0$$

as $\|\mathbf{x}\| \rightarrow 0$ and, $\mathbf{f}(\mathbf{y}^0) = \mathbf{0}$, the lemma is proved. \blacksquare

The linear system

$$\mathbf{x}' = \mathcal{A}\mathbf{x}$$

is called the linearization of (8.28) around the equilibrium point \mathbf{y}^0 .

Theorem 8.3 *Suppose that \mathbf{f} is a twice differentiable function in some neighbourhood of the equilibrium point \mathbf{y}^0 . Then,*

- (i) *The equilibrium point \mathbf{y}^0 is asymptotically stable if all the eigenvalues of the matrix \mathcal{A} have negative real parts, that is, if the equilibrium solution $\mathbf{x}(t) = \mathbf{0}$ of the linearized system is asymptotically stable.*
- (ii) *The equilibrium point \mathbf{y}^0 is unstable if at least one eigenvalue has a positive real part.*
- (iii) *If all the eigenvalues of \mathcal{A} have non-negative real part but at least one of them has real part equal to 0, then the stability of the equilibrium point \mathbf{y}^0 of the nonlinear system (8.28) cannot be determined from the stability of its linearization.*

Proof. To prove 1) we use the variation of constants formula (7.35) applied to (8.28) written in the form of Lemma

8.2 for $\mathbf{y}(t) = \mathbf{x}(t) + \mathbf{y}^0$:

$$\mathbf{x}' = \mathbf{y}' = \mathbf{f}(\mathbf{y}) = \mathbf{f}(\mathbf{x} + \mathbf{y}^0) = \mathcal{A}\mathbf{x} + \mathbf{g}(\mathbf{x}).$$

Thus

$$\mathbf{x}(t) = e^{t\mathcal{A}}\mathbf{x}(0) + \int_0^t e^{(t-s)\mathcal{A}}\mathbf{g}(\mathbf{x}(s))ds.$$

Denoting by α' the maximum of real parts of eigenvalues of \mathcal{A} we observe that for any $\alpha > \alpha'$

$$\|e^{t\mathcal{A}}\mathbf{x}(0)\| \leq Ke^{-\alpha t}\|\mathbf{x}(0)\|, \quad t \geq 0,$$

for some constant $K \geq 1$. Note that in general we have to take $\alpha > \alpha'$ to account for possible polynomial entries in $e^{t\mathcal{A}}$. Thus, since $\alpha' < 0$, then we can take also $\alpha < 0$ keeping the above estimate satisfied. From the assumption on \mathbf{g} , for any ϵ we find $\delta > 0$ such that if $\|\mathbf{x}\| \leq \delta$, then

$$\|\mathbf{g}(\mathbf{x})\| \leq \epsilon\|\mathbf{x}\|. \quad (8.30)$$

Assuming for a moment that for $0 \leq s \leq t$ we can keep $\|\mathbf{x}(s)\| \leq \delta$, we can write

$$\begin{aligned} \|\mathbf{x}(t)\| &\leq \|e^{At}\mathbf{x}(0)\| + \int_0^t \|e^{A(t-s)}\mathbf{g}(\mathbf{x}(s))\|ds \\ &\leq Ke^{-\alpha t}\|\mathbf{x}(0)\| + K\epsilon \int_0^t e^{-\alpha(t-s)}\|\mathbf{x}(s)\|ds \end{aligned}$$

or, multiplying both sides by $e^{\alpha t}$ and setting $z(t) = e^{\alpha t}\|\mathbf{x}(t)\|$,

$$z(t) \leq K\|\mathbf{x}(0)\| + K\epsilon \int_0^t z(s)ds. \quad (8.31)$$

Using Gronwall's lemma we obtain thus

$$\|\mathbf{x}(t)\| = e^{-\alpha t} z(t) \leq K \|\mathbf{x}(0)\| e^{(K\epsilon - \alpha)t},$$

providing $\|\mathbf{x}(s)\| \leq \delta$ for all $0 \leq s \leq t$. Let us take $\epsilon \leq \alpha/2K$, then the above can be written as

$$\|\mathbf{x}(t)\| \leq K \|\mathbf{x}(0)\| e^{-\frac{\alpha t}{2}}. \quad (8.32)$$

Assume now that $\|\mathbf{x}(0)\| < \delta/K \leq \delta$ where δ was fixed for $\epsilon \leq \alpha/2K$. Then $\|\mathbf{x}(0)\| < \delta$ and, by continuity, $\|\mathbf{x}(t)\| \leq \delta$ for some time. Let $\mathbf{x}(t)$ be defined on some interval I and $t_1 \in I$ be the first time for which $\|\mathbf{x}(t)\| = \delta$. Then for $t \leq t_1$ we have $\|\mathbf{x}(t)\| \leq \delta$ so that for all $t \leq t_1$ we can use (8.32) getting, in particular,

$$\|\mathbf{x}(t_1)\| \leq \delta e^{-\frac{\alpha t_1}{2}} < \delta,$$

that is a contradiction. Thus $\|\mathbf{x}(t)\| < \delta$ if $\|\mathbf{x}(0)\| < \delta_1$ in the whole interval of existence but then, if the interval was finite, then we could extend the solution to a larger interval as the solution is bounded at the endpoint and the same procedure would ensure that the solution remains bounded by δ on the larger interval. Thus, the extension can be carried out for all the values of $t \geq 0$ and the solution exists for all t and satisfies $\|\mathbf{x}(t)\| \leq \delta$ for all $t \geq 0$. Consequently, (8.32) holds for all t and the solution $\mathbf{x}(t)$ converges exponentially to 0 as $t \rightarrow \infty$ proving the asymptotic stability of the stationary solution \mathbf{y}^0 .

Statement 2 follows from e.g. the Stable Manifold Theorem, two-dimensional version of which is discussed later.

To prove 3, it is enough to display two systems with the same linear part and different behaviour of solutions. Let

us consider

$$\begin{aligned}y_1' &= y_2 - y_1(y_1^2 + y_2^2) \\y_2' &= -y_1 - y_2(y_1^2 + y_2^2)\end{aligned}$$

with the linearized system given by

$$\begin{aligned}y_1' &= y_2 \\y_2' &= -y_1\end{aligned}$$

The eigenvalues of the linearized system are $\pm i$. To analyze the behaviour of the solutions to the non-linear system, let us multiply the first equation by y_1 and the second by y_2 and add them together to get

$$\frac{1}{2} \frac{d}{dt}(y_1^2 + y_2^2) = -(y_1^2 + y_2^2)^2.$$

Solving this equation we obtain

$$y_1^2 + y_2^2 = \frac{c}{1 + 2ct}$$

where $c = y_1^2(0) + y_2^2(0)$. Thus $y_1^2(t) + y_2^2(t)$ approaches $\mathbf{0}$ as $t \rightarrow \infty$ and $y_1^2(t) + y_2^2(t) < y_1^2(0) + y_2^2(0)$ for any $t > 0$ and we can conclude that the equilibrium point $\mathbf{0}$ is asymptotically stable.

Consider now the system

$$\begin{aligned}y_1' &= y_2 + y_1(y_1^2 + y_2^2) \\y_2' &= -y_1 + y_2(y_1^2 + y_2^2)\end{aligned}$$

with the same linear part and thus with the same eigenvalues. As above we obtain that

$$y_1^2 + y_2^2 = \frac{c}{1 - 2ct}$$

with the same meaning for c . Thus, any solution with non-zero initial condition blows up at the time $t = 1/2c$ and therefore the equilibrium solution $\mathbf{0}$ is unstable. ■

Example 8.10 Find all equilibrium solutions of the system of differential equations

$$\begin{aligned}y_1' &= 1 - y_1 y_2, \\y_2' &= y_1 - y_2^3,\end{aligned}$$

and determine, if possible, their stability.

Solving equation for equilibrium points $1 - y_1 y_2 = 0, y_1 - y_2^3 = 0$ we find two equilibria: $y_1 = y_2 = 1$ and $y_1 = y_2 = -1$. To determine their stability we have to reduce each case to the equilibrium at $\mathbf{0}$. For the first case we put $u(t) = y_1(t) - 1$ and $v(t) = y_2(t) - 1$ so that

$$\begin{aligned}u' &= -u - v - uv, \\v' &= u - 3v - 3v^2 - v^3,\end{aligned}$$

so that the linearized system has the form

$$\begin{aligned}u' &= -u - v, \\v' &= u - 3v,\end{aligned}$$

and the perturbing term is given by $\mathbf{g}(u, v) = (-uv, -3v^2 + v^3)$ and, as the right-hand side of the original system is infinitely differentiable at $(0, 0)$ the assumptions of the stability theorem are satisfied. The eigenvalues of the linearized system are given by $\lambda_{1,2} = -2$ and therefore the equilibrium solution $\mathbf{y}(t) \equiv (1, 1)$ is asymptotically stable.

For the other case we set $u(t) = y_1(t) + 1$ and $v(t) = y_2 + 1$

so that

$$\begin{aligned}u' &= u + v - uv, \\v' &= u - 3v + 3v^2 - v^3,\end{aligned}$$

so that the linearized system has the form

$$\begin{aligned}u' &= u + v, \\v' &= u - 3v,\end{aligned}$$

and the perturbing term is given by $\mathbf{g}(u, v) = (-uv, 3v^2 - v^3)$. The eigenvalues of the linearized system are given by $\lambda_1 = -1 - \sqrt{5}$ and $\lambda_2 = -1 + \sqrt{5}$ and therefore the equilibrium solution $\mathbf{y}(t) \equiv (-1, -1)$ is unstable.

Appendix 1

Methods of solving first order difference equations

The general form of a first order difference equation is

$$x(n+1) = f(n, x(n)), \quad (1.1)$$

where f is any function of two variables defined on $\mathbb{N}_0 \times \mathbb{R}$, where $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ is the set of natural numbers enlarged by 0. Eq. (1.1) is a recurrence formula and thus in difference equations we do not encounter problems related to existence and uniqueness which are often quite delicate for differential equations. However, finding an explicit formula for the solution is possible for a much more narrow class of difference equations than in the case of differential equations. We shall survey some typical cases.

A1.1 Solution of the first order linear difference equation

The general first order difference equation has the form

$$x(n+1) = a(n)x(n) + g(n), \quad (1.2)$$

where $(a_n)_{n \in \mathbb{N}}$ and $(g_n)_{n \in \mathbb{N}}$ are given sequences. It is clear that to start the recurrence we need only one initial point, so that we supplement (1.2) with the an initial condition $x(0) = x_0$. It is easy to check by induction that the solution is given by

$$x(n) = x_0 \prod_{k=0}^{n-1} a(k) + \sum_{k=0}^{n-1} g(k) \prod_{i=k+1}^{n-1} a(i) \quad (1.3)$$

where we adopted the convention that $\prod_{i=0}^{n-1} = 1$. Similarly, to simplify notation, we agree to put $\sum_{k=j+1}^j = 0$. *Special cases*

There are two special cases of (1.2) that appear in many applications. In the first, the equation is given by

$$x(n) = ax(n) + g(n), \quad (1.4)$$

with the value $x(0) = x_0$ given. In this case $\prod_{k=k_1}^{k_2} a(k) = a^{k_2 - k_1 + 1}$ and (1.3) takes the form

$$x(n) = a^n x(0) + \sum_{k=0}^{n-1} a^{n-k-1} g(k). \quad (1.5)$$

The second case is a simpler form of (1.4), given by

$$x(n) = ax(n) + g, \quad (1.6)$$

with g independent of n . In this case the sum in (1.5) can be evaluated in an explicit form giving

$$x(n) = \begin{cases} a^n x_0 + g \frac{a^n - 1}{a - 1} & \text{if } a \neq 1, \\ x(0) + gn. & \end{cases} \quad (1.7)$$

Appendix 2

Basic solution techniques in differential equations

A2.1 Some general results

In the proof of the Picard theorem, in particular, in the uniqueness part an essential role is played by the following result, known as the Gronwall lemma. It is possibly the most used auxiliary result in the theory of differential equations.

Lemma 2.1 *If $f(t)$, $g(t)$ are continuous and nonnegative for $t \in [t_0, t_0 + \alpha]$, $\alpha > 0$, and $c > 0$, then*

$$f(t) \leq c + \int_{t_0}^t f(s)g(s)ds \quad (2.1)$$

on $[t_0, t_0 + \alpha]$ implies

$$f(t) \leq c \exp \left(\int_{t_0}^t g(s)ds \right) \quad (2.2)$$

for all $[t_0, t_0 + \alpha]$.

If f satisfies (2.1) with $c = 0$, then $f(t) = 0$ on $[t_0, t_0 + \alpha]$.

A2.2 Some equations admitting closed form solutions

A2.2.1 *Separable equations*

Consider an equation that can be written in the form

$$\frac{dy}{dt} = \frac{g(t)}{h(y)}, \quad (2.3)$$

where g and h are known functions. Equations that can be put into this form are called *separable* equations. Firstly, we note that any constant function $y = y_0$, such that $1/h(y_0) = 0$, is a special solution to (2.3), as the derivative of a constant function is equal to zero. We call such solutions *stationary or equilibrium solutions*.

To find a general solution, we assume that $1/h(y) \neq 0$, that is $h(y) \neq \infty$. Multiplying then both sides of (2.3) by $h(y)$ to get

$$h(y) \frac{dy}{dt} = g(t) \quad (2.4)$$

and observe that, denoting by $H(y) = \int h(y)dy$ the antiderivative of h , we can write (2.3) in the form

$$\frac{d}{dt}(H(y(t))) = g(t),$$

that closely resembles (??). Thus, upon integration we obtain

$$H(y(t)) = \int g(t)dt + c, \quad (2.5)$$

where c is an arbitrary constant of integration. The next step depends on the properties of H : for instance, if $H : \mathbb{R} \rightarrow \mathbb{R}$ is monotonic, then we can find y explicitly for all t

as

$$y(t) = H^{-1} \left(\int g(t) dt + c \right).$$

Otherwise, we have to do it locally, around the initial values. To explain this, we solve the initial value problem for separable equation.

$$\begin{aligned} \frac{dy}{dt} &= \frac{g(t)}{h(y)}, \\ y(t_0) &= y_0, \end{aligned} \quad (2.6)$$

Using the general solution (2.5) (with definite integral) we obtain

$$H(y(t)) = \int_{t_0}^t g(s) ds + c,$$

we obtain

$$H(y(t_0)) = \int_{t_0}^{t_0} a(s) ds + c,$$

which, due $\int_{t_0}^{t_0} a(s) ds = 0$, gives

$$c = H(y(t_0)),$$

so that

$$H(y(t)) = \int_{t_0}^t g(s) ds + H(y(t_0)).$$

We are interested in the existence of the solution at least close to t_0 , which means that H should be invertible close to y_0 . From the Implicit Function Theorem we obtain that

this is possible if H is differentiable in a neighbourhood of y_0 and $\frac{\partial H}{\partial y}(y_0) \neq 0$. But $\frac{\partial H}{\partial y}(y_0) = h(y_0)$, so we are back at Picard's theorem: if $h(y)$ is differentiable in the neighbourhood of y_0 with $h(y_0) \neq 0$ (if $h(y_0) = 0$, then the equation (2.3) does not make sense at y_0 , and g is continuous, then $f(t, y) = g(t)/h(y)$ satisfies the assumptions of the theorem in some neighbourhood of (t_0, y_0)).

A2.2.2 Linear ordinary differential equations of first order

The general first order linear differential equation is

$$\frac{dy}{dt} + a(t)y = b(t). \quad (2.7)$$

Functions a and b are known continuous functions of t . Let us recall that we call this equation 'linear' because the dependent variable y appears by itself in the equation. In other words, y' and y appear in the equation only possibly multiplied by a known function and not in the form yy' , $\sin y$ or $(y')^3$.

It is not immediate how to solve (2.7), therefore we shall simplify it even further by putting $b(t) = 0$. The resulting equation

$$\frac{dy}{dt} + a(t)y = 0, \quad (2.8)$$

is called the *reduced* first order linear differential equation. We observe that the reduced equation is a separable equation and thus can be solved easily. As in Example 4.5 we obtain that if $y(t) \neq 0$ for any t , then

$$y(t) = C \exp\left(-\int a(t)dt\right), \quad C \in \mathbb{R}. \quad (2.9)$$

The Picard theorem ensures that these are all solutions of (2.8). Eq. (2.9) will be used to solve (2.7). First we multiply both sides of (2.7) by some continuous nonzero function μ (for a time being, unknown) to get the equivalent equation

$$\mu(t) \frac{dy}{dt} + \mu(t)a(t)y = \mu(t)b(t), \quad (2.10)$$

and ask the question: for which function μ the left-hand side of (2.10) is a derivative of some simple expression? We note that the first term on the left-hand side comes from

$$\frac{d\mu(t)y}{dt} = \mu(t) \frac{dy}{dt} + \frac{d\mu(t)}{dt}y,$$

thus, if we find μ in such a way that

$$\mu(t) \frac{dy}{dt} + \frac{d\mu(t)}{dt}y = \mu(t) \frac{dy}{dt} + \mu(t)a(t)y,$$

that is

$$\frac{d\mu(t)}{dt}y = \mu(t)a(t)y,$$

then we are done. Note that an immediate choice is to solve the equation

$$\frac{d\mu(t)}{dt} = \mu(t)a(t),$$

but this is a separable equation, the general solution of which is given by (2.9). Since we need only one such function, we may take

$$\mu(t) = \exp \left(\int a(t)dt \right).$$

The function μ is called an *integrating factor* of the equation (2.7). With such function, (2.7) can be written as

$$\frac{d}{dt} \mu(t)y = \mu(t)b(t),$$

thus

$$\mu(t)y = \int \mu(t)b(t)dt + c$$

where c is an arbitrary constant of integration. Finally

$$\begin{aligned} y(t) &= \frac{1}{\mu(t)} \left(\int \mu(t)b(t)dt + c \right) & (2.11) \\ &= \exp \left(- \int a(t)dt \right) \left(\int b(t) \exp \left(\int a(t)dt \right) dt + c \right). \end{aligned}$$

It is worthwhile to note that the solution consists of two parts: the general solution to the reduced equation associated with (2.7)

$$c \exp \left(- \int a(t)dt \right)$$

and, what can be checked by direct differentiation, a particular solution to the full equation.

If we want to find a particular solution satisfying $y(t_0) = y_0$, then we write (2.11) using definite integrals

$$y(t) = \exp \left(- \int_{t_0}^t a(s)ds \right) \left(\int_{t_0}^t b(s) \exp \left(\int_{t_0}^s a(r)dr \right) ds + c \right)$$

and use the fact that $\int_{t_0}^{t_0} f(s)ds = 0$ for any function f . This shows that the part of the solution satisfying the nonhomogeneous equation:

$$y_b(t) = \exp \left(- \int_{t_0}^t a(s)ds \right) \int_{t_0}^t b(s) \exp \left(\int_{t_0}^s a(r)dr \right) ds$$

takes on the zero value at $t = t_0$. Thus

$$y_0 = y(t_0) = c$$

and the solution to the initial value problem is given by

$$y(t) = y_0 \exp\left(-\int_{t_0}^t a(s) ds\right) + \exp\left(-\int_{t_0}^t a(s) ds\right) \int_{t_0}^t b(s) \exp\left(\int_{t_0}^s a(r) dr\right) ds. \quad (2.12)$$

Once again we emphasize that the first term of the formula above solves the reduced ($b(t) = 0$) equation with the desired initial value ($y(t_0) = y_0$) whereas the second solves the full equation with the initial value equal to zero.

Again, Picard's theorem shows that there are no more solutions to (2.7) than those given by (2.12).

A2.2.3 Bernoulli equation

Consider the equation

$$y'(t) = a(t)y(t) + b(t)y^\alpha(t), \quad (2.13)$$

called the Bernoulli equation. Here we assume that a and b are continuous functions on some interval I and α is a real number different from 0 and 1 (as in these two cases (2.13) becomes a linear equation).

We see that $y(t) \equiv 0$ is a solution of (2.13) if $\alpha > 0$. By the Picard theorem, the Cauchy problem for (2.13) with the initial condition $y(t_0) = y_0$, $t_0 \in I$, has a unique solution for any $y_0 \neq 0$ ($y_0 > 0$ if α is a fraction). Precisely

speaking, for $\alpha > 1$ a unique solution ($y = 0$) exists also for $y_0 = 0$ as in this case the right hand side of (2.13) is Lipschitz continuous. To find non-zero solutions of (2.13), we introduce the change of the dependent variable

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

Thus

$$z' = (1 - \alpha)y^{-\alpha}y'$$

and

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

Substituting this formula into (2.13), dividing by $(1-\alpha)^{-1}y^\alpha$ and using the previous equation we arrive at

$$z'(t) = (1 - \alpha)a(t)z(t) + (1 - \alpha)b(t) \quad (2.14)$$

which is a linear equation which can be solved by the methods introduced in the previous subsection.

A2.2.4 Equations of homogeneous type

In differential equations, as in integration, a smart substitution can often convert a complicated equation into a manageable one. For some classes of differential equations there are standard substitutions that transform them into separable equations. We shall discuss one such a class in detail.

A differential equation that can be written in the form

$$\frac{dy}{dt} = f\left(\frac{y}{t}\right), \quad (2.15)$$

where f is a function of the single variable $z = y/t$ is said

A2.2 Some equations admitting closed form solution 297

to be of *homogeneous type*. Note that in some textbooks such equations are called *homogeneous equations* but this often creates confusion as the name homogeneous equation is generally used in another context.

To solve (2.15) let us make substitution

$$y = tz \quad (2.16)$$

where z is the new unknown function. Then, by the product rule for derivatives

$$\frac{dy}{dt} = z + t \frac{dz}{dt}$$

and (2.15) becomes

$$t \frac{dz}{dt} = f(z) - z. \quad (2.17)$$

In (2.17) the variables are separable so it can be solved as in Subsection A2.2.1.

A2.2.5 Equations that can be reduced to first order equations

Some higher order equations can be reduced to equations of the first order. We shall discuss two such cases for second order equations.

Equations that do not contain the unknown function

If we have the equation of the form

$$F(y'', y', t) = 0, \quad (2.18)$$

then the substitution $z = y'$ reduces this equation to an equation of the first order

$$F(z', z, t) = 0. \quad (2.19)$$

If we can solve this equation

$$z = \phi(t, C),$$

where C is an arbitrary constant, then, returning to the original unknown function y , we obtain another first order equation

$$y' = \phi(t, C),$$

which is immediately solvable as

$$y(t) = \int \phi(t, C) dt + C_1.$$

Equations that do not contain the independent variable

Let us consider the equation

$$F(y'', y', y) = 0, \quad (2.20)$$

that does not involve the independent variable t . Such an equation can be also reduced to a first order equation, the idea, however, is a little more complicated. Firstly, we note that, as long as $y' \neq 0$, the derivative y' locally is uniquely defined by the function y ; that is, we can write $y' = g(y)$ for some function g . Indeed, the function $y = y(t)$ is locally invertible provided $y' \neq 0$ and we can write $t = t(y)$. Thus $g(y) = y'(t(y))$. Using the chain rule we obtain

$$y'' = \frac{d}{dt} y' = \frac{dg}{dy}(y) \frac{dy}{dt} = y' \frac{dg}{dy}(y) = g(y) \frac{dg}{dy}(y). \quad (2.21)$$

Substituting (2.21) into (2.20) gives a first order equation with y as an independent variable

$$F\left(g \frac{dg}{dy}, g, y\right) = 0. \quad (2.22)$$

If we solve this equation in the form $g(y) = \phi(y, C)$, then to

find y we have to solve one more first order equation with t as the independent variable

$$\frac{dy}{dt} = \phi(y, C).$$

We note that the latter is a separable equation.

The above procedure can be best explained by interpreting t as time, y as the distance travelled by a particle moving with velocity y' and acceleration y'' . If the particle does not reverse the direction of motion ($y' = 0$ at any turning point!), then velocity can be expressed as a function of the distance instead of time. This is precisely what we have done above.

A2.3 Systems of difference equations and higher order equations

A2.3.1 Homogeneous systems of difference equations

We start with the homogeneous system of difference equations

$$\begin{aligned} y_1(n+1) &= a_{11}y_1(n) + a_{12}y_2(n) + \dots + a_{1k}y_k(n), \\ &\vdots \quad \vdots \quad \ddots, \\ y_k(n+1) &= a_{k1}y_1(n) + a_{k2}y_2(n) + \dots + a_{kk}y_k(n), \end{aligned} \tag{2.23}$$

where, for $n \geq 0$, $y_1(n), \dots, y_k(n)$ are unknown sequences, a_{11}, \dots, a_{kk} are constant coefficients and $g_1(n), \dots, g_k(n)$ are known. As with systems of differential equations, we shall find it more convenient to use the matrix notation.

Denoting $\mathbf{y} = (y_1, \dots, y_k)$, $\mathcal{A} = \{a_{ij}\}_{1 \leq i, j \leq k}$, that is,

$$\mathcal{A} = \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \dots & a_{kk} \end{pmatrix},$$

(2.23) can be written as

$$\mathbf{y}(n+1) = \mathcal{A}\mathbf{y}(n). \quad (2.24)$$

Eq. (2.24) is usually supplemented by the initial condition $\mathbf{y}(0) = \mathbf{y}^0$. It is obvious, by induction, to see that the solution to (2.24) is given by

$$\mathbf{y}(n) = \mathcal{A}^n \mathbf{y}^0. \quad (2.25)$$

The problem with (2.25) is that it is rather difficult to give an explicit form of \mathcal{A}^n . We shall solve this problem in a similar way to Subsection 7.1.4, where we were to evaluate the exponential function of \mathcal{A} .

To proceed, we assume that the matrix \mathcal{A} is nonsingular. This means, in particular, that if $\mathbf{v}^1, \dots, \mathbf{v}^k$ are linearly independent vectors, then also $\mathcal{A}\mathbf{v}^1, \dots, \mathcal{A}\mathbf{v}^k$ are linearly independent. Since \mathbb{R}^k is k -dimensional, it is enough to find k linearly independent vectors \mathbf{v}^i , $i = 1, \dots, k$ for which $\mathcal{A}^n \mathbf{v}^i$ can be easily evaluated. Assume for a moment that such vectors have been found. Then, for arbitrary $\mathbf{x}^0 \in \mathbb{R}^k$ we can find constants c_1, \dots, c_k such that

$$\mathbf{x}^0 = c_1 \mathbf{v}^1 + \dots + c_k \mathbf{v}^k.$$

Precisely, let \mathcal{V} be the matrix having vectors \mathbf{v}^i as its columns, and let $\mathbf{c} = (c_1, \dots, c_k)$, then

$$\mathbf{c} = \mathcal{V}^{-1} \mathbf{x}^0. \quad (2.26)$$

A2.3 Systems of difference equations and higher order equations

Note, that \mathcal{V} is invertible as the vectors \mathbf{v}^i are linearly independent.

Thus, for arbitrary \mathbf{x}^0 we have

$$\mathcal{A}^n \mathbf{x}^0 = \mathcal{A}^n (c_1 \mathbf{v}^1 + \dots + c_k \mathbf{v}^k) = c_1 \mathcal{A}^n \mathbf{v}^1 + \dots + c_k \mathcal{A}^n \mathbf{v}^k. \quad (2.27)$$

Now, if we denote by \mathcal{A}_n the matrix whose columns are vectors $\mathcal{A}^n \mathbf{v}^1, \dots, \mathcal{A}^n \mathbf{v}^k$, then we can write

$$\mathcal{A}^n = \mathcal{A}_n \mathcal{V}^{-1} \quad (2.28)$$

Hence, the problem is to find linearly independent vectors \mathbf{v}^i , $i = 1, \dots, k$, on which powers of \mathcal{A} can be easily evaluated. As before, we use eigenvalues and eigenvectors for this purpose. Firstly, note that if \mathbf{v}^1 is an eigenvector of \mathcal{A} corresponding to an eigenvalue λ_1 , that is, $\mathcal{A} \mathbf{v}^1 = \lambda_1 \mathbf{v}^1$, then by induction

$$\mathcal{A}^n \mathbf{v}^1 = \lambda_1^n \mathbf{v}^1.$$

Therefore, if we have k linearly independent eigenvectors $\mathbf{v}^1, \dots, \mathbf{v}^k$ corresponding to eigenvalues $\lambda_1, \dots, \lambda_k$ (not necessarily distinct), then from (2.27) we obtain

$$\mathcal{A}^n \mathbf{x}^0 = c_1 \lambda_1^n \mathbf{v}^1 + \dots + c_k \lambda_k^n \mathbf{v}^k.$$

with c_1, \dots, c_k given by (2.26). Note that, as for systems of differential equations, if λ is a complex eigenvalue with eigenvector \mathbf{v} , then both $\Re(\lambda^n \mathbf{v})$ and $\Im(\lambda^n \mathbf{v})$ are real valued solutions. To find explicit expressions for them we write $\lambda = r e^{i\phi}$ where $r = |\lambda|$ and $\phi = \text{Arg} \lambda$. Then

$$\lambda^n = r^n e^{in\phi} = r^n (\cos n\phi + i \sin n\phi)$$

and

$$\begin{aligned}\Re(\lambda^n \mathbf{v}) &= r^n (\cos n\phi \Re \mathbf{v} - \sin n\phi \Im \mathbf{v}), \\ \Im(\lambda^n \mathbf{v}) &= r^n (\sin n\phi \Re \mathbf{v} + \cos n\phi \Im \mathbf{v}).\end{aligned}$$

Finally, if for some eigenvalue λ_i the number ν_i of linearly independent eigenvectors is smaller than its algebraic multiplicity n_i , then we follow the procedure described in Subsection 7.1.4, that is, we find all solutions to

$$(\mathcal{A} - \lambda_i \mathcal{I})^2 \mathbf{v} = \mathbf{0}$$

that are not eigenvectors and, if we still do not have sufficiently many independent vectors, we continue solving

$$(\mathcal{A} - \lambda_i \mathcal{I})^j \mathbf{v} = \mathbf{0}$$

with $j \leq n_i$; it can be proven that in this way we find n_i linearly independent vectors. Let \mathbf{v}^j is found as a solution to $(\mathcal{A} - \lambda_i \mathcal{I})^j \mathbf{v}^j = \mathbf{0}$ with $j \leq \nu_i$. Then, using the binomial expansion we find

$$\begin{aligned}\mathcal{A}^n \mathbf{v}^j &= (\lambda_i \mathcal{I} + \mathcal{A} - \lambda_i \mathcal{I})^n \mathbf{v}^j = \sum_{r=0}^n \lambda_i^{n-r} \binom{n}{r} (\mathcal{A} - \lambda_i \mathcal{I})^r \mathbf{v}^j \\ &= (\lambda_i^n \mathcal{I} + n \lambda_i^{n-1} (\mathcal{A} - \lambda_i \mathcal{I}) + \dots \\ &\quad + \frac{n!}{(j-1)!(n-j+1)!} \lambda_i^{n-j+1} (\mathcal{A} - \lambda_i \mathcal{I})^{j-1}) \mathbf{v}^j\end{aligned}\tag{2.29}$$

where

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

is the Newton symbol. It is important to note that (2.29) is a finite sum for any n ; it always terminates at most at the term $(\mathcal{A} - \lambda_i \mathcal{I})^{n_i-1}$ where n_i is the algebraic multiplicity of λ_i .

A2.3 Systems of difference equations and higher order equations 203

We shall illustrate these considerations by the following example.

Example 2.1 Find \mathcal{A}^n for

$$\mathcal{A} = \begin{pmatrix} 4 & 1 & 2 \\ 0 & 2 & -4 \\ 0 & 1 & 6 \end{pmatrix}.$$

We start with finding eigenvalues of \mathcal{A} :

$$p(\lambda) = \begin{vmatrix} 4-\lambda & 1 & 2 \\ 0 & 2-\lambda & -4 \\ 0 & 1 & 6-\lambda \end{vmatrix} = (4-\lambda)(16-8\lambda+\lambda^2) = (4-\lambda)^3 = 0$$

gives the eigenvalue $\lambda = 4$ of algebraic multiplicity 3. To find eigenvectors corresponding to $\lambda = 3$, we solve

$$(\mathcal{A} - 4\mathcal{I})\mathbf{v} = \begin{pmatrix} 0 & 1 & 2 \\ 0 & -2 & -4 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Thus, v_1 is arbitrary and $v_2 = -2v_3$ so that the eigenspace is two dimensional, spanned by

$$\mathbf{v}^1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{v}^2 = \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix}.$$

Therefore

$$\mathcal{A}^n \mathbf{v}^1 = 4^n \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \mathcal{A}^n \mathbf{v}^2 = 4^n \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix}.$$

To find the last vector we consider

$$\begin{aligned} (\mathcal{A} - 4\mathcal{I})^2 \mathbf{v} &= \begin{pmatrix} 0 & 1 & 2 \\ 0 & -2 & -4 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 0 & 1 & 2 \\ 0 & -2 & -4 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \end{aligned}$$

Any vector solves this equation so that we have to take a vector that is not an eigenvalue. Possibly the simplest choice is

$$\mathbf{v}^3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Thus, by (2.29)

$$\begin{aligned} \mathcal{A}^n \mathbf{v}^3 &= (4^n \mathcal{I} + n4^{n-1}(\mathcal{A} - 4\mathcal{I})) \mathbf{v}^3 \\ &= \left(\begin{pmatrix} 4^n & 0 & 0 \\ 0 & 4^n & 0 \\ 0 & 0 & 4^n \end{pmatrix} + n4^{n-1} \begin{pmatrix} 0 & 1 & 2 \\ 0 & -2 & -4 \\ 0 & 1 & 2 \end{pmatrix} \right) \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 2n4^{n-1} \\ -n4^n \\ 4^n + 2n4^{n-1} \end{pmatrix}. \end{aligned}$$

To find explicit expression for \mathcal{A}^n we use (2.28). In our case

$$\mathcal{A}_n = \begin{pmatrix} 4^n & 0 & 2n4^{n-1} \\ 0 & -2 \cdot 4^n & -n4^n \\ 0 & 4^n & 4^n + 2n4^{n-1} \end{pmatrix},$$

further

$$\mathcal{V} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 1 & 1 \end{pmatrix},$$

so that

$$\mathcal{V}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 1 \end{pmatrix}.$$

Therefore

$$\mathcal{A}^n = \mathcal{A}_n \mathcal{V}^{-1} = \begin{pmatrix} 4^n & n4^{n-1} & 2n4^{n-1} \\ 0 & 4^n - 2n4^{n-1} & -n4^n \\ 0 & n4^{n-1} & 4^n + 2n4^{n-1} \end{pmatrix}.$$

The next example shows how to deal with complex eigenvalues.

Example 2.2 Find \mathcal{A}^n if

$$\mathcal{A} = \begin{pmatrix} 1 & -5 \\ 1 & -1 \end{pmatrix}.$$

We have

$$\begin{vmatrix} 1 - \lambda & -5 \\ 1 & -1 - \lambda \end{vmatrix} = \lambda^2 + 4$$

so that $\lambda_{1,2} = \pm 2i$. Taking $\lambda_1 = 2i$, we find the corresponding eigenvector by solving

$$\begin{pmatrix} 1 - 2i & -5 \\ 1 & -1 - 2i \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix};$$

we get

$$\mathbf{v}^1 = \begin{pmatrix} 1 + 2i \\ 1 \end{pmatrix}$$

and

$$\mathbf{x}(n) = \mathcal{A}^n \mathbf{v}^1 = (2i)^n \begin{pmatrix} 1 + 2i \\ 1 \end{pmatrix}.$$

To find real valued solutions, we have to take real and imaginary parts of $\mathbf{x}(n)$. Since $i = \cos \frac{\pi}{2} + i \sin \frac{\pi}{2}$, we have by de Moivre's formula

$$(2i)^n = 2^n \left(\cos \frac{\pi}{2} + i \sin \frac{\pi}{2} \right)^n = 2^n \left(\cos \frac{n\pi}{2} + i \sin \frac{n\pi}{2} \right).$$

Therefore

$$\begin{aligned} \Re \mathbf{x}(n) &= 2^n \left(\cos \frac{n\pi}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \sin \frac{n\pi}{2} \begin{pmatrix} 2 \\ 0 \end{pmatrix} \right) \\ \Im \mathbf{x}(n) &= 2^n \left(\cos \frac{n\pi}{2} \begin{pmatrix} 2 \\ 0 \end{pmatrix} + \sin \frac{n\pi}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right). \end{aligned}$$

The initial values for $\Re \mathbf{x}(n)$ and $\Im \mathbf{x}(n)$ are, respectively,

$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 2 \\ 0 \end{pmatrix}$. Since \mathcal{A}^n is a real matrix, we have

$\Re \mathcal{A}^n \mathbf{v}^1 = \mathcal{A}^n \Re \mathbf{v}^1$ and $\Im \mathcal{A}^n \mathbf{v}^1 = \mathcal{A}^n \Im \mathbf{v}^1$, thus

$$\mathcal{A}^n \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 2^n \left(\cos \frac{n\pi}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \sin \frac{n\pi}{2} \begin{pmatrix} 2 \\ 0 \end{pmatrix} \right) = 2^n \begin{pmatrix} \cos \frac{n\pi}{2} - 2 \sin \frac{n\pi}{2} \\ \cos \frac{n\pi}{2} \end{pmatrix}$$

and

$$\mathcal{A}^n \begin{pmatrix} 2 \\ 0 \end{pmatrix} = 2^n \left(\cos \frac{n\pi}{2} \begin{pmatrix} 2 \\ 0 \end{pmatrix} + \sin \frac{n\pi}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right) = 2^n \begin{pmatrix} 2 \cos \frac{n\pi}{2} + \sin \frac{n\pi}{2} \\ \sin \frac{n\pi}{2} \end{pmatrix}.$$

To find \mathcal{A}^n we use again (2.28). In our case

$$\mathcal{A}_n = 2^n \begin{pmatrix} \cos \frac{n\pi}{2} - 2 \sin \frac{n\pi}{2} & 2 \cos \frac{n\pi}{2} + \sin \frac{n\pi}{2} \\ \cos \frac{n\pi}{2} & \sin \frac{n\pi}{2} \end{pmatrix},$$

further

$$\mathcal{V} = \begin{pmatrix} 1 & 2 \\ 1 & 0 \end{pmatrix},$$

so that

$$\mathcal{V}^{-1} = -\frac{1}{2} \begin{pmatrix} 0 & -2 \\ -1 & 1 \end{pmatrix}.$$

Therefore

$$\mathcal{A}^n = \mathcal{A}_n \mathcal{V}^{-1} = -2^{n-1} \begin{pmatrix} -2 \cos \frac{n\pi}{2} - \sin \frac{n\pi}{2} & 5 \sin \frac{n\pi}{2} \\ -\sin \frac{n\pi}{2} & -2 \cos \frac{n\pi}{2} + \sin \frac{n\pi}{2} \end{pmatrix}.$$

A2.3.2 Nonhomogeneous systems

Here we shall discuss solvability of the nonhomogeneous version of (2.23)

$$\begin{aligned} y_1(n+1) &= a_{11}y_1(n) + a_{12}y_2(n) + \dots + a_{1k}y_k(n) + g_1(n), \\ &\vdots \quad \vdots \quad \vdots, \\ y_k(n+1) &= a_{k1}y_1(n) + a_{k2}y_2(n) + \dots + a_{kk}y_k(n) + g_k(n), \end{aligned} \tag{2.30}$$

where, for $n \geq 0$, $y_1(n), \dots, y_k(n)$ are unknown sequences, a_{11}, \dots, a_{kk} are constant coefficients and $g_1(t), \dots, g_k(n)$ are known. As before, we write it using the vector-matrix notation. Denoting $\mathbf{y} = (y_1, \dots, y_k)$, $\mathbf{g} = (g_1, \dots, g_k)$ and $\mathcal{A} = \{a_{ij}\}_{1 \leq i, j \leq k}$, we have

$$\mathbf{y}(n+1) = \mathcal{A}\mathbf{y}(n) + \mathbf{g}(n). \tag{2.31}$$

Exactly as in Subsection A1.1 we find that the solution to (2.31) satisfying the initial condition $\mathbf{y}(0) = \mathbf{y}^0$ is given by

the formula

$$\mathbf{y}(n) = \mathcal{A}^n \mathbf{y}^0 + \sum_{r=0}^{n-1} \mathcal{A}^{n-r-1} \mathbf{g}(r). \quad (2.32)$$

Example 2.3 Solve the system

$$\begin{aligned} y_1(n+1) &= 2y_1(n) + y_2(n) + n, \\ y_2(n+1) &= 2y_2(n) + 1 \end{aligned}$$

with $y_1(0) = 1, y_2(0) = 0$. Here

$$\mathcal{A} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}, \quad \mathbf{g}(n) = \begin{pmatrix} n \\ 1 \end{pmatrix}, \quad \mathbf{y}^0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

We see that

$$p(\lambda) = \begin{vmatrix} 2 - \lambda & 1 \\ 0 & 2 - \lambda \end{vmatrix} = (2 - \lambda)^2,$$

so that we have double eigenvalue $\lambda = 2$. To find eigenvectors corresponding to this eigenvalue, we have to solve the system

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

so that we have one-dimensional eigenspace spanned by $\mathbf{v}^1 = (1, 0)$. To find the second linearly independent vector associated with $\lambda = 2$ we observe that

$$(\mathcal{A} - 2\mathcal{I})^2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

so that we can take $\mathbf{v}^2 = (0, 1)$. Thus, we obtain two independent solutions in the form

$$\mathbf{y}^1(n) = \mathcal{A}^n \mathbf{v}^1 = 2^n \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

and

$$\begin{aligned} \mathbf{y}^2(n) &= \mathcal{A}^n \mathbf{v}^2 = (2\mathcal{I} + (\mathcal{A} - 2\mathcal{I})^n) \mathbf{v}^2 = \left(2^n \mathcal{I} + n2^{n-1} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \right) \begin{pmatrix} 0 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} n2^{n-1} \\ 2^n \end{pmatrix}. \end{aligned}$$

Since \mathbf{v}^1 and \mathbf{v}^2 happen to be the canonical basis for \mathbb{R}^2 , that is, $\mathbf{x}^0 = (x_1^0, x_2^0) = x_1^0 \mathbf{v}^1 + x_2^0 \mathbf{v}^2$, we obtain immediately

$$\mathcal{A}^n = \begin{pmatrix} 2^n & n2^{n-1} \\ 0 & 2^n \end{pmatrix}.$$

To find the solution of the nonhomogeneous equation, we use formula (2.32). The first term is easily calculated as

$$\mathcal{A}^n \mathbf{x}^0 = \begin{pmatrix} 2^n & n2^{n-1} \\ 0 & 2^n \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2^n \\ 0 \end{pmatrix}.$$

Next,

$$\begin{aligned}
\sum_{r=0}^{n-1} \mathcal{A}^{n-r-1} \mathbf{g}(r) &= \sum_{r=0}^{n-1} \begin{pmatrix} 2^{n-r-1} & (n-r-1)2^{n-r-2} \\ 0 & 2^{n-r-1} \end{pmatrix} \begin{pmatrix} r \\ 1 \end{pmatrix} \\
&= \sum_{r=0}^{n-1} \begin{pmatrix} r2^{n-r-1} + (n-r-1)2^{n-r-2} \\ 2^{n-r-1} \end{pmatrix} \\
&= 2^n \begin{pmatrix} \frac{1}{4} \sum_{r=1}^{n-1} r2^{-r} + \frac{n-1}{4} \sum_{r=0}^{n-1} 2^{-r} \\ \frac{1}{2} \sum_{r=0}^{n-1} 2^{-r} \end{pmatrix} \\
&= 2^n \begin{pmatrix} \frac{1}{2} \left(1 - \left(\frac{1}{2}\right)^{n-1}\right) - (n-1) \left(\frac{1}{2}\right)^{n+1} + \frac{n-1}{2} \left(1 - \left(\frac{1}{2}\right)^n\right) \\ 1 - \left(\frac{1}{2}\right)^n \end{pmatrix} \\
&= 2^n \begin{pmatrix} -n \left(\frac{1}{2}\right)^n + \frac{n}{2} \\ 1 - \left(\frac{1}{2}\right)^n \end{pmatrix} \\
&= \begin{pmatrix} -n + \frac{n2^n}{2} \\ 2^n - 1 \end{pmatrix}
\end{aligned}$$

Remark 2.1 Above we used the following calculations

$$\begin{aligned}
\sum_{r=1}^{n-1} ra^r &= a(1 + a + \dots + a^{n-2}) + a^2(1 + a + \dots + a^{n-3}) + \dots + a^{n-1} \\
&= \frac{1}{1-a} (a(1 - a^{n-1}) + a^2(1 - a^{n-2}) + \dots + a^{n-1}(1 - a)) \\
&= \frac{1}{1-a} (a + a^2 + \dots + a^{n-1} - (n-1)a^n) \\
&= \frac{a(1 - a^{n-1}) - (n-1)a^n(1 - a)}{(1-a)^2}
\end{aligned}$$

Thus, the solution is given by

$$\mathbf{y}(n) = \begin{pmatrix} 2^n - n + \frac{n2^n}{2} \\ 2^n - 1 \end{pmatrix}.$$

A2.3.3 Higher order equations

Consider the linear difference equation of order k :

$$y(n+k) + a_1y(n+k-1) + \dots + a_ky(n) = g(n), \quad n \geq 0 \quad (2.33)$$

where a_1, \dots, a_k are known numbers and $g(n)$ is a known sequence. This equation determines the values of $y(N)$, $N > k$ by k preceding values of $y(r)$. Thus, it is clear that to be able to solve this equation, that is, to start the recurrence procedure, we need k initial values $y(0), y(1), \dots, y(k-1)$. Equation (2.33) can be written as a system of first order equations of dimension k . We let

$$\begin{aligned} z_1(n) &= y(n), \\ z_2(n) &= y(n+1) = z_1(n+1), \\ z_3(n) &= y(n+2) = z_2(n+1), \\ &\vdots \\ z_k(n) &= y(n+k-1) = z_{k-1}(n-1), \end{aligned} \quad (2.34)$$

hence we obtain the system

$$\begin{aligned} z_1(n+1) &= z_2(n), \\ z_2(n+1) &= z_3(n), \\ &\vdots \\ z_{k-1}(n+1) &= z_k(n), \\ z_k(n+1) &= -a_1z_1(n) - a_2z_2(n) \dots - a_kz_k(n) + g(n), \end{aligned}$$

or, in matrix notation,

$$\mathbf{z}(n+1) = \mathcal{A}\mathbf{z}(n) + \mathbf{g}(n)$$

where $\mathbf{z} = (z_1, \dots, z_k)$, $\mathbf{g}(n) = (0, 0, \dots, g(n))$ and

$$\mathcal{A} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_k & -a_{k-1} & -a_{k-2} & \dots & -a_1 \end{pmatrix}.$$

It is clear that the initial values $y(0), \dots, y(k-1)$ give the initial vector $\mathbf{z}^0 = (y(0), \dots, y(k-1))$. Next we observe that the eigenvalues of \mathcal{A} can be obtained by solving the equation

$$\begin{vmatrix} -\lambda & 1 & 0 & \dots & 0 \\ 0 & -\lambda & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_k & -a_{k-1} & -a_{k-2} & \dots & -a_1 - \lambda \end{vmatrix} \\ = (-1)^k (\lambda^k + a_1 \lambda^{k-1} + \dots + a_k) = 0,$$

that is, the eigenvalues can be obtained by finding roots of the characteristic polynomial. Consequently, solutions of higher order equations can be obtained by solving the associated first order systems but there is no need to repeat the whole procedure. In fact, to solve a $k \times k$ system we have to construct k linearly independent vectors $\mathbf{v}^1, \dots, \mathbf{v}^k$ so that solutions are given by $\mathbf{z}^1(n) = \mathcal{A}^n \mathbf{v}^1, \dots, \mathbf{z}^k(n) = \mathcal{A}^n \mathbf{v}^k$ and coordinates of each \mathbf{z}^i are products of λ_i and polynomials in n of degree strictly smaller than the algebraic multiplicity of λ_i . To obtain n_i solutions of the higher order equation corresponding to the eigenvalue λ_i , by (2.34) we take only

the first coordinates of all $\mathbf{z}^i(n)$ that correspond to λ_i . On the other hand, we must have here n_i linearly independent scalar solutions of this form and therefore we can use the set $\{\lambda_i^n, n\lambda_i^n, \dots, n^{n_i-1}\lambda_i^n\}$ as a basis for the set of solutions corresponding to λ_i , and the union of such sets over all eigenvalues to obtain a basis for the set of all solutions.

Example 2.4 Consider the Fibonacci equation (1.33), written here as

$$y(n+2) = y(n+1) + y(n) \quad (2.35)$$

to be consistent with the notation of the present chapter. Introducing new variables $z_1(n) = y(n)$, $z_2(n) = y(n+1) = z_1(n+1)$ so that $y(n+2) = z_2(n+1)$, we re-write the equation as the system

$$\begin{aligned} z_1(n+1) &= z_2(n), \\ z_2(n+1) &= z_1(n) + z_2(n). \end{aligned}$$

The eigenvalues of the matrix

$$\mathcal{A} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$$

are obtained by solving the equation

$$\begin{vmatrix} -\lambda & 1 \\ 1 & 1-\lambda \end{vmatrix} = \lambda^2 - \lambda - 1 = 0;$$

they are $\lambda_{1,2} = \frac{1 \pm \sqrt{5}}{2}$. Since the eigenvalues are distinct, we immediately obtain that the general solution of (2.35) is given by

$$y(n) = c_1 \left(\frac{1 + \sqrt{5}}{2} \right)^n + c_2 \left(\frac{1 - \sqrt{5}}{2} \right)^n.$$

314 *Basic solution techniques in differential equations*

To find the solution satisfying the initial conditions $y(0) = 1$, $y(1) = 2$ (corresponding to one pair of rabbits initially) we substitute these values and get the system of equations for c_1 and c_2

$$\begin{aligned} 1 &= c_1 + c_2, \\ 2 &= c_1 \frac{1 + \sqrt{5}}{2} + c_2 \frac{1 - \sqrt{5}}{2}, \end{aligned}$$

the solution of which is $c_1 = 1 + 3\sqrt{5}/5$ and $c_2 = -3\sqrt{5}/5$.

A2.4 Miscellaneous applications

Gambler's ruin

The characteristic equation is given by

$$\lambda^2 - \frac{1}{q}\lambda + \frac{1-q}{q} = 0$$

and the eigenvalues are $\lambda_1 = \frac{1-q}{q}$ and $\lambda_2 = 1$. Hence, if $q \neq 1/2$, then the general solution can be written as

$$p(n) = c_1 + c_2 \left(\frac{1-q}{q} \right)^n$$

and if $q = 1/2$, then $\lambda_1 = \lambda_2 = 1$ and

$$p(n) = c_1 + c_2 n.$$

To find the solution for the given boundary conditions, we denote $Q = (1-q)/q$ so that for $q \neq 1/2$

$$\begin{aligned} 1 &= c_1 + c_2, \\ 0 &= c_1 + Q^N c_2, \end{aligned}$$

from where

$$c_2 = \frac{1}{1 - Q^N}, \quad c_1 = -\frac{Q^N}{1 - Q^N}$$

and

$$p(n) = \frac{Q^n - Q^N}{1 - Q^N}.$$

Analogous considerations for $q = 1/2$ yield

$$p(n) = 1 - \frac{n}{N}.$$

For example, if $q = 1/2$ and the gambler starts with $n = 20$ rands with the target $N = 1000$, then

$$p(20) = 1 - \frac{20}{1000} = 0,98,$$

that is, his ruin is almost certain.

In general, if the gambler plays a long series of games, which can be modelled here as taking $N \rightarrow \infty$, then he will be ruined almost certainly even if the game is fair ($q = \frac{1}{2}$).

References

- Samuelson, P.A. Interaction between the Multiplier Analysis and Principle of Acceleration. *Review of Economic Statistics* 21: 75-78, May, 1939.
- Solow, R.M. A Contribution to the Theory of Economic Growth. *Quarterly Journal of Economics* 70: 65-94, February, 1956.
- Swan, T.W. Economic Growth and Capital Accumulation. *Economic Record* 32: 334-361, November, 1956.

Index

- SI* model, 30
- SIR* model, 32
- maximal interval of existence, 97
- Allee effect, 17
- Bernoulli equation, 244
- Cauchy problem, 93
- conversion period, 1
- effective interest rate, 2
- Fibonacci problem, 20
- Gronwall lemma, 238
- homogeneous type equation, 246
- law of mass action, 32
- Leslie matrix, 24
- linear differential equation, 241
- Lipschitz condition, 95
- logistic differential equation, 47, 48
- Markov processes, 29
- Peano theorem, 94
- Picard theorem, 95
- reducible second order equation, 246
- separable equation, 97, 239