J. Banasiak
School of Mathematical Sciences
University of KwaZulu-Natal, Durban, South Africa

INTRODUCTION TO MATHEMATICAL METHODS IN POPULATION THEORY

Springer

Berlin Heidelberg New York
Hong Kong London
Milan Paris Tokyo

Contents

1

Introduction

Why did I write this book?

I see, and teach, mathematical biology and, in particular, population theory, as a part of mathematical modelling; that is, translating the rules of nature into mathematical formulae (most often equations), applying mathematical methods to analyse them and then trying to understand the implications of the obtained results for the original problems. There are many textbooks on mathematical biology or population theory, ranging from excellent to weak. Most of them have, in my opinion, a certain drawback: their main focus is on the biological content, while the mathematics is often ornamental or, at best, sketchy and relegated to appendices or to more advanced textbooks. Also, many textbooks devoted to population dynamics and mathematical biology use very sophisticated mathematics but it is very difficult to find precise references linking this level of sophistication with standard undergraduate education. My experience as an external examiner, referee and editor dealing with courses, theses and articles on mathematical biology, is that postgraduate students and young researchers entering the field make appalling mistakes trying work on biologically relevant problems with hopelessly limited mathematical skills. When I started teaching mathematical biology to mathematicians, I found this situation quite frustrating. Thus, I decided to design a course which will be focused more on developing and presenting mathematics inspired by biology than on biology itself. Following this, I have tried to write these notes in such a way that the lecturer using them will be able to present mathematical results which are relevant to biological application but with full mathematical rigour without having to look for proper statements and their proofs elsewhere or to prove them him/herself. Thus, the book mostly is addressed to students and researchers primarily interested in mathematical methods of analysing biological models and in developing these methods for their own sake, as a mathematical theory, and not as just a tool for a particular biological application. In other words, the main topic of this book is mathematics inspired by biology. Such books do exists, but mostly at a much higher level, and thus I hope that this volume will fill some existing gap in the literature and find readers also outside the mathematical circles.

What this book is about?

Population dynamics is a fast growing discipline spanning biology and mathematics. However, in many existing textbooks the authors do not provide any mathematically sound introduction to it focusing, instead, on apparently more accessible biological content. Even though the analysis of it requires quite sophisticated mathematical tools, even if the biology of the model is simples, the explanation of the involved mathematics is relegated to appendices or more advanced texts. We hope that this book will provide some bridge between introductory texts in population dynamics and more advanced texts. Hence, the book is predominantly concerned with mathematical analysis of known models, which have been drawn from a wide range of very good books in the field, and then with the interpretation of the obtained mathematical results. Usually we will not go deeper into the construction of models. However, within the space limitation, we will try show that the discussed models are not purely phenomenological; that is, constructed just to fit the data, but that they can be derived from more fundamental principles.

As we said, mathematical population dynamics deals with techniques and methods of solving and/or analysing time behaviour of models describing populations of living entities, ranging from cells or bacteria (or even genes) through plants, animals to humans. The description can be stochastic or deterministic at the individual, population, or community level. In this book we shall focus on deterministic models. We begin with unstructured single population models in discrete and continuous time and go through exponential, logistic, Ricker, Gompertz and Allee models. Next we analyse discrete and continuous structured models discussing, in particular, Leslie and Usher matrices, general projection matrices, Lotka's renewal equation and extending some of these results to the nonlinear cases. Mathematically, we will discuss eigenvalue-eigenvector method of solving discrete and continuous systems of equations, spectral decomposition and the Perron-Frobenius theory leading to asynchronous exponential growth property. In the nonlinear case we will focus on stability results beginning from the one dimensional theory, through the stability by linearization in both discrete and continuous case and the Lyapunov functions method. We provide basics of phase-plane analysis for classical problems and show its applicability to analysis of travelling waves. Next we shall provide a brief introduction to dynamics in infinite dimensional spaces. Here we discuss the birth-and-death problems and provide its solution using the generating function method. Next we introduce the McKendrick model, formally deriving it from the Leslie model. We discuss some explicitly solvable cases and provide general well-posedness theory in the linear case. Staying with the linear case, we discuss the asynchronous exponential growth property using the Laplace transform technique (following the method in the lecture notes [15] which are not widely available). Next we move to nonlinear models, beginning with the so called linear chain trickery. Further, we introduce the concept of a semigroup and, using the semigroup approach, we provide theory of solvability of models of age structured epidemiology by casting them into the form of abstract semilinear equation. The book is concluded with several appendices which provide some longer proofs of results used in the text.

The book is based on lectures given by the author at the Honours (4th year) level at the University of KwaZulu-Natal in Durban (Mathematical Biology), 4th year level at the Technical University of Łódź (Introduction to Mathematical Ecology), for postgraduate students at the SA Centre for Epidemiological Modelling and Analysis in Stellenbosch and at the African Institute for Mathematical Sciences in 2012 (Introduction to Population Theory) and at the 2011 CIMPA-UNESCO School on Modelling and Simulation in Population Biology in Muizenburg (SA). It is written mostly for senior undergraduate and postgraduate students with working knowledge of advanced calculus and linear algebra, who are interested in applications of mathematics in biology. The main body of the it can be used for a course for such an audience. However, selected parts of the book touch some aspects of functional analysis and, being of interest for senior postgraduate students and researchers, do not have to be included in the basic course. It is my hope that the the book will equip the students with mathematical grammar of the language of population dynamics and will allow them to enter into the research field without too many disasters.

2

Unstructured models

1 Mathematical modelling

Population dynamics is about analysis of models arising in real life and thus it is related to a large field of mathematical modelling. Therefore we begin by briefly describing the process of mathematical modelling (that not only applies to biology or population theory).

First, we must have a phenomenon of interest that one wants to describe or, more importantly, to explain and make predictions about. Observing this phenomenon, we to make hypotheses about the quantities that are most relevant to the problem and about the relations between them. In this way we can devise a possible mechanism that can explain the phenomenon. At this stage we have to decide how to quantify the observations; that is, for instance, whether time should be continuous or discrete or whether we will consider discrete or continuously distributed attributes of individuals, such as age or position in space. This choice is not always obvious or unique but one needs to decide on a particular approach before one begins to build a model. Sometimes it is worthwhile to build several models with various configurations and then compare the results delivered by each of them.

The main purpose of building the model is to quantify a description of the mechanism driving the phenomenon of interest; that is, write it as mathematical equations which can be mathematically analysed. Once the equations are solved or analysed, it is necessary to interpret the solution, or any other information extracted from the equations, as statements about the original problem so that they can be tested against observations. Ideally, the model also leads to predictions which, if verified, serve as a further validation of the model. It is important to realize that modelling is usually an iterative procedure as it is very difficult to achieve a proper balance between the simplicity and meaningfulness of the model. Indeed, usually the first model is built using the Ockham razor, or parsimony, principle which states that among possible equations satisfying the requirements of the model, we select the simplest one. However, often such a model turns out to be over-simplified so that there is insufficient agreement between the actual experiment and the results predicted from the model. This indicates that in the modeling process we must have overlooked some important features of the described phenomenon and we haves to return to the first step of modelling process and construct a new equations which, again, should be the simplest one that allows to cater for the enlarged set of requirements. On the other hand, sometimes the model is over complicated to yield itself to any meaningful analysis and then again we have to construct a simpler model which can be analysed but which, nevertheless, still reflects the main features of the phenomenon under investigation.

This first step in modelling is the most creative but also the most difficult, often involving a concerted effort of specialists in many diverse fields. Hence, as we said earlier, 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, on analysing and solving the derived equations, interpreting their solutions in the language of the original problem and reflecting on whether the answers seem reasonable.

In all cases discussed in this book, the model is a representation of a process; that is, it describes a change in the states of some system in time. This description could be discrete and continuous. The former corresponds to the situation in which we observe a system at 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 typically will be modelled by difference equations. On the other hand, in the continuous cases we treat time as a continuum allowing for observations of the system at any time. In such a case the model expresses relations between the rates of change of various quantities rather than between the states at various times, as was the case in discrete time modelling. Then, since the rates of change are given by derivatives, the model is represented by differential equations.

Furthermore, models can describe the evolution of a given system either in a non-interacting environment, such as a population of bacteria in a Petri dish, or else engaged in interactions with other systems. In the first case the model consists of a single equation and we say that the model is one-dimensional, while in the second case we have to deal with several (sometimes infinitely many) equations describing the interactions; then the model is said to be multidimensional. We observe that often we try to describe a real population interacting with its environment by a single equation. Then the impact of the environment on the population is reflected in the parameters of the equation which are supposed to contain an averaged information of the environment. While such an approach is often successful and useful, especially for short term predictions, one always should be aware of its limitations.

2 Discrete models

We begin our course with discrete time models of single populations in a non-interactive environment; that is, as explained in the introduction, when the impact of the environment on the observed population is modelled by parameters appearing in the equations. For instance, in many models described below we recognize that there is a reaction of the environment to overpopulation. However, we do not model this reaction explicitly but by modifying the vital parameters of the populations, such as the birth and death rates, where the environment is represented by a certain parameter, for instance the carrying capacity in the logistic equation.

2.1 Background

Many plants and animals breed only during a short, well-defined, breeding season. Also, often the adult population dies soon after breeding. Such populations are ideal for modelling using discrete time. Let us consider a few typical examples, [6].

(i) So called monocarpic plants flower once and then die. Such plants may be annual but, for instance, bamboos grow vegetatively for 20 years and then flower and die.

(ii) Animals with such a life cycle are called semelparous.

a) Insects typically die after lying eggs but their life-cycle may range from several days (e.g. house flies) to 13–17 years (cicads).

b) Similar life cycle is observed in some species of fish, such as the Pacific salmon or European eel. For instance, the latter lives 10-15 years in freshwater European lakes, migrates to the Sargasso Sea, spawns and dies.

c) Some mammals, such as marsupials, ovulate once per year and produce a single litter. There occurs abrupt and total mortality of males after mating. The births in the population are synchronized to within a day or two, which is related to the environment with predictable 'bloom' of insects in a given season of the year.

(iii) On the other hand, a species is called iteroparous if it is characterized by multiple reproductive cycles over the course of its lifetime. Such populations can be modelled by difference equations if the breeding only occurs during short, regularly spaced breeding periods. It is typical for birds. For instance, females of the Greater Snow Geese lay eggs between 8th–20th of June (peak occurs at 12th–17th of June) and practically all eggs hatch between 8th and 13th of July.

From the modelling point of view it is important to distinguish between populations in which all adults die immediately after the breeding season and populations in which individuals can survive several generations. As we said earlier,

- monocarpic plants, cicads, most insects and some fish species have non-overlapping generations,
- in birds' populations, such as the Snow Geese we observe overlapping generations.

In the first case as well as in the cases with overlapping generations, but with no significant differences between generations, we often can model the population by a single, first order in time, equation. Occurrence of significant differences between generations leads to the so-called structured models which will be discussed later.

2.2 Exponential growth

Possibly the simplest way to describe a population is to provide information of how its size (number of individuals) changes in time. If we take the census of adults in a population in fixed time intervals, say, immediately after the breeding season, then we obtain a sequence of numbers

$$P(0), P(1), \ldots, P(k)$$

where $P(k)$ is the number of adults just after the k-th breeding season. This approach makes sense in the models described above; that is, if the population has well defined breeding seasons.

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

$$P(k+1) = R(P(k))P(k), \quad k = 0, 1, \ldots, \tag{2.2.1}$$

where $R(P(k))$ is the (density dependent) net growth rate. Then the simplest case of (2.2.1) is when $R(P) = R_0$, that is,

$$P(k+1) = R_0 P(k), \quad k = 0, 1, \ldots \tag{2.2.2}$$

The exponential (or Malthusian) equation (2.2.2) well describes populations which are completely homogeneous with characteristics of the environment and individual members constant over time. Moreover, the ratio of females to males remains the same in each breeding season. This allows to describe the population by only giving the number of females. Further

- Each member of the population produces on average the same number of offspring;
- Each member has an equal chance of dying and death only can occur after giving birth;
- Age differences between members of the population can be ignored;
- The population is isolated - there is no immigration or emigration.

Under these assumptions the net growth rate R_0 is given as

$$R_0 = 1 + \beta - \mu.$$

Here β is the average number of offspring per member of the population in each season; it is called the fecundity or fertility. The constant μ is the probability that an individual will die after giving birth and before the next census; it is called the per capita death rate. Thus, (2.2.2) can be written as

$$P(k+1) = P(k) + \beta P(k) - \mu P(k), \tag{2.2.3}$$

which expresses the basic principle of population modelling in discrete time:

> The number of individuals in the $k+1$th census equals the population in the kth census + the total number of births in the population between the censuses − the total number of deaths in the population between the censuses.

Remark 2.1. Note that, strictly speaking, the model (2.2.3) describes the situation in which all neonates survive till the breeding season. It is an unrealistic assumption but the model allows for some flexibility. We can accommodate the situation in which any individual can die between breeding seasons by modifying (2.2.3)

$$P(k+1) = P(k) + \beta(1-\mu_0)P(k) - \mu_0(1-\mu_1)P(k) - \mu_1 P(k), \tag{2.2.4}$$

where μ_0 is the probability that an individual will die from natural causes before the breeding season while μ_1 is the per capita death rate due to giving birth. The constant $\beta(1-\mu_0)$ is the (effective) per capita birth rate. It is a very important parameter from the population point of view since for the survival of the population it is not only important what is the natural fertility of a female; that is, how many births she can give each season, but also how many of the survive on average till the next breeding season. We note that (2.2.3) will describe a population with no overlapping generations if $\mu_1 = 1$; that is, if all adults die just after giving birth.

It is important to recognize that mathematically this model is the same (2.2.2) the interpretation of R_0, however, changes. This becomes important when we try to validate the model using experimental data.

Changes due to migrations can be incorporated but the structure of the equation may not persist – e.g. they are not necessarily proportional to the total population.

Clearly,

$$P(k) = R_0^k P(0), \qquad k = 0, 1, 2 \ldots \tag{2.2.5}$$

and if $R_0 < 1$, then the population decreases towards extinction, but with $R_0 > 1$ it grows indefinitely. Such behaviour over long periods of time is not observed in any population so it is clear that the model is over-simplified and requires corrections. However, it is often valid over short time intervals and can bring some demographical insights.

2.3 The death rate μ and the average lifespan of an individual

Using the concept of the death rate we can derive an important characteristic of a population – the expected life span of its individuals. Denote by $p(k)$ the probability that an individual, born at $k = 0$ is alive at time k. In order to be alive at time k, it had to be alive at time $k-1$ and could not die between $k-1$ and k. Assuming that the probability of dying is constant in time and using the conditional probability formula we arrive at

$$p(k) = (1-\mu)p(k-1), \qquad p(0) = 1,$$

where μ is the probability that an individual dies, for whatever reason, between the censuses (in model (2.2.4), $\mu = \mu_0(1-\mu_1) + \mu_1$). Thus

$$p(k) = (1-\mu)^k. \tag{2.2.6}$$

The average lifespan L is the expected duration of life. To find it we observe that to die exactly at time k, an individual must be alive at time $k-1$ and die in the interval $(k-1, k]$, which occurs with probability $\mu p(k-1) = \mu(1-\mu)^{k-1}$. It also can be explained as follows: in the first year a proportion μ of the population dies and $1-\mu$ survives, then after the second year a proportion μ of them die; that is, $\mu(1-\mu)$ fraction of the initial population dies before the 2nd census while $(1-\mu)^2$ survives beyond the second birthday, etc. Thus, the average life span is given by

$$L = \mu \sum_{k=1}^{\infty} k(1-\mu)^{k-1} = -\mu \frac{d}{d\mu} \frac{1-\mu}{1-(1-\mu)} = \frac{1}{\mu}, \tag{2.2.7}$$

where we used

$$\sum_{k=1}^{\infty} kz^{k-1} = \frac{d}{dz}\sum_{k=1}^{\infty} z^k = \frac{d}{dz}\frac{z}{1-z}$$

for $z = 1 - \mu$.

2.4 Basic nonlinear models

In real populations, some of the β offspring produced by each adult will not survive to be counted as adults in the next census. Let us denote by $S(P)$ the *survival rate*; that is, the fraction of newborns that survives till the next breeding season.

If we consider, for simplicity, populations with no overlapping generations, then the Malthusian equation will is replaced by

$$P(k+1) = \beta S(P(k))P(k), \quad k = 0, 1, \ldots \tag{2.2.8}$$

which may be alternatively written as

$$P(k+1) = R(P(k))P(k) = f(P(k)), \quad k = 0, 1, \ldots \tag{2.2.9}$$

where $R(P)$ is the (effective) per capita production of a population of size P. If $S(P)$ is independent of P, then (2.2.8) is the same as (2.2.4) with $S = 1 - \mu_0$; that is, with an adjusted birth rate, and with $\mu_1 = 1$. However, in general we have a density dependent S which leads to nonlinear equations. However, before introducing basic examples, we discuss typical types of behaviour in such populations.

The survival rate S in (2.2.8) reflects the intraspecific (within-species) competition for some resource (typically, food or space) which is in short supply. The three main (idealized) forms of intraspecific competition are, [6],

- *No competition*; then $S(P) = 1$ for all P.
- *Contest competition*: here there is a finite number of units of resource. Each individual who obtains one of these units survives to breed, and produces β offspring in the subsequent generations; all others die without producing offspring. Thus $S(P) = 1$ for $P \leq P_c$ and $S(P) = P_c/P$ for $P > P_c$ for some critical value P_c.
- *Scramble competition*: here each individual is assumed to get equal share of a limited resource. If this amount is sufficient for survival to breeding, then all individuals survive and produce β offspring each; if its is not sufficient, then all of them die. Thus, $S(P) = 1$ for $P \leq P_C$ and $S(P) = 0$ if $P > P_C$ for a critical value P_C (P_C is different from P_c).

These ideal situations do not occur in real populations: real data are not easily classified in terms of the contest or scramble competition. Threshold density is not usually seen, zero survival is unrealistic, at least for large populations. Thus, a similar classification is done on the basis of asymptotic behaviour of $S(P)$ (or $f(P)$) as $P \to \infty$.

1. Contest competition corresponds to *exact compensation*:

$$\lim_{P\to\infty} f(P) = c \tag{2.2.10}$$

   for some constant c (or $S(P) \sim cP^{-1}$ for large P). This describes the situation if the increased mortality compensates exactly any increase in numbers and thus only the predetermined number of individuals in the population can survive.

2. The other case is when

$$S(P) \sim c/P^b, \quad P \to \infty. \tag{2.2.11}$$

   Here we have

Under-compensation if $0 < b < 1$; that is, when the increased mortality less than compensates for the increase for increase in numbers;

Over-compensation if $b > 1$.

In general, if $b \approx 1$, then we say that there is a contest, and a scramble if b is large. Indeed, in the first case, $f(P)$ eventually levels-out at a nonzero level for large populations which indicates that the population stabilizes by rejecting too many newborns. On the other hand, for $b > 1$, $f(P)$ tends to zero for large populations which indicates that the resources are over-utilized leading to eventual extinction.

We note that the linear case, $S(P) = 1 - \mu_0 \geq 0$, does not really fit into this description which is designed to model the response of the environment to the increasing population. However, we see that the case $1 - \mu_0 < 1/\beta$ is over-compensatory, as the population eventually dies out. On the other hand, the case $1 - \mu_0 > 1/\beta$ is under-compensatory as the population grows indefinitely.

We introduce most typical nonlinear models.

Beverton-Holt type models

Let us look at the model (2.2.9)

$$P(k+1) = R(P(k))P(k), \quad k = 0, 1, \ldots,$$

where $R(P(k)) = \beta S(P(k))$. To exhibit compensatory behaviour, we should have $PS(P) \approx \textit{const}$. Also, for small P, $S(P)$ should be approximately 1 as we expect very small intra-species competition so that the growth should be exponential with the effective birth rate given by fertility β. A simple function of this form is

$$S(P) = \frac{1}{1 + aP}$$

leading to

$$P(k+1) = \frac{\beta P(k)}{1 + aP(k)}. \tag{2.2.12}$$

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 $P(k) = K$ for some k, then $P(k+m) = K$ for all $m \geq 0$, then

$$K(1 + aK) = \beta K \tag{2.2.13}$$

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

$$P(k+1) = \frac{\beta P(k)}{1 + \frac{\beta - 1}{K} P(k)}. \tag{2.2.14}$$

As we said earlier, this model is compensatory.

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

$$P(k+1) = \frac{\beta P(k)}{(1 + aP(k))^b}. \tag{2.2.15}$$

It exhibits all types of compensatory behaviour, depending on b. For $b > 1$ the models describes *scramble* competition, while for $b = 1$ we have contest.

Substitution $x(k) = aP(k)$ reduces the number of parameters giving

$$x(k+1) = \frac{\beta x(k)}{(1 + x(k))^b} \tag{2.2.16}$$

which will be analysed later.

If the justification of the Beverton-Holt equation given in this section seems not very convincing, an alternative derivations will be presented in Subsection 3.2 and also in Subsection 5.2.

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 it more informative to write the difference equation in the form

$$P(k+1) = P(k) + R(P(k))P(k), \tag{2.2.17}$$

so that the increase in the population is given by $R(P) = \beta S(P)$. Here we assume for simplicity that no adults die (such deaths can be incorporated by subtracting the term $\mu P(k)$ from the right hand side of the equation but this will not change the structure of the equation). Below we will show another way of incorporating death into the equation.

As with the Beverton-Holt model, we will use the Ockham razor principle to write down the equation. Indeed, the function R can have different forms but must satisfy the requirements:

(a) Due to overcrowding, $R(P)$ must decrease as P increases until P equals the carrying capacity K; then $R(K) = 0$ so that, as above, $P = K$ stops changing.

(b) Since for P much smaller than K there is small intra-species competition, we should observe an exponential growth of the population so that $R(P) \approx \beta$ as $P \to 0$; here R_0 is called the unrestricted birth rate of the population.

Constants β and K are usually determined experimentally.

The simplest function satisfying these requirements is a linear function which, to satisfy (a) and (b), must be chosen as

$$R(P) = -\frac{\beta}{K}P + \beta.$$

Substituting this formula into (2.2.17) yields the so-called discrete logistic equation

$$P(k+1) = P(k) + \beta P(k)\left(1 - \frac{P(k)}{K}\right), \tag{2.2.18}$$

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

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

$$P(k+1) = \beta P(k)\left(1 - \frac{P(k)}{K}\right). \tag{2.2.19}$$

Both equations are mathematically equivalent. Indeed,

$$P(k) + \beta P(k)\left(1 - \frac{P(k)}{K}\right) = (1+\beta)P(k)\left(1 - \frac{P(k)}{\frac{1+\beta}{\beta}K}\right)$$

which is the right hand side of (2.2.19) with just different constants. Conversely, the right hand side of (2.2.19) can be transformed to (2.2.18) provided $\beta > 1$.

While the above arguments may seem to be of *bunny-out-of-the-hat* type, it could be justified by generalizing (2.2.3). Indeed, assume that the mortality μ is not constant but equals

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

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 (2.2.3) can be written as

$$P(k+1) = (1+\beta-\mu_0)P(k)\left(1 - \frac{P(k)}{\frac{1+\beta-\mu_0}{\mu_1}}\right) \tag{2.2.20}$$

which is (2.2.19), with β modified as $\beta - \mu_0$ and the carrying capacity $K = 1 + \beta - \mu_0/\mu_1$. A generalization of this equation, obtained by a more flexible formula for the death rate,

$$\mu = \mu_0 + \mu_1 P^\theta, \quad \theta > 0,$$

is the *Bernoulli equation*

$$P(k+1) = P(k) + \beta P(k)\left(1 - \left(\frac{P(k)}{K}\right)^\theta\right), \qquad (2.2.21)$$

By substitution

$$x(k) = \frac{\beta}{1+\beta}\frac{P(k)}{K}, \qquad \gamma = 1 + \beta$$

we can reduce (2.2.18) to a simpler form

$$x(k+1) = \gamma x(k)(1 - x(k)) \qquad (2.2.22)$$

The problem with the discrete logistic equation is that large (close to K) populations can become negative in the next step. For instance, for

$$P(k+1) = P(k) + 4P(k)(1 - P(k))$$

beginning with $P(0) = 0.9$, we obtain $P(1) = 1.26$ and $P(2) = -0.0504$.

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 $P = K$ as a stationary population. Thus, if we happen to hit exactly K, then the population survives but if we even marginally overshot, the population becomes extinct.

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

$$S(P(k)) = \left(1 - \frac{P(k)}{K}\right)_+. \qquad (2.2.23)$$

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. We shall discuss other ways to circumvent this difficulty in the following subsection.

Here, to conclude, we only observe that the logistic equation, especially with S given by (2.2.23), is an extreme example of the scramble competition.

Ricker equation

Here we will try to find a model in which large values of $P(k)$ produce very small, but still positive, values of $P(k+1)$. Thus, a population well over the carrying capacity crashes to a very low levels, but survives. Let us find a way in which this can be modelled. Consider the per capita population change

$$\frac{\Delta P}{P} = R(P).$$

First we note that it is impossible for R to be less than -1 - this would mean that an individual could die more than once. We also need a decreasing R which is non-zero ($= \beta$) at 0. One such function can be recovered from the Beverton-Holt model, another simple choice is an exponential shifted down by 1; that is,

$$\frac{\Delta P}{P} = ae^{-bP} - 1,$$

which leads to

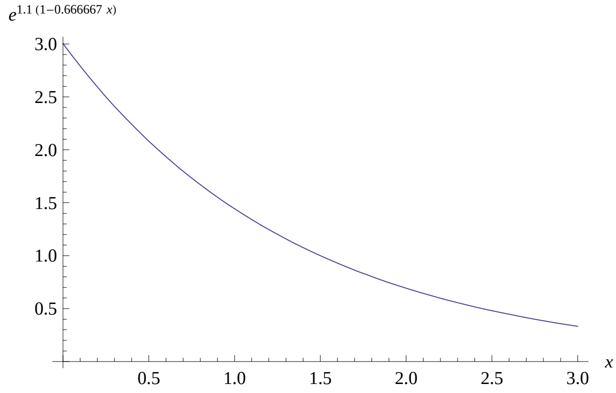

Fig. 2.1. The function $f(x) = e^{r(1-x/K)}$ for $r = 1.1$ and $K = 1.5$.

$$P(k+1) = aP(k)e^{-bP(k)}. \tag{2.2.24}$$

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

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

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

$$P(k+1) = P(k)e^{\xi(1-\frac{P(k)}{K})}. \tag{2.2.25}$$

We note that if $P(k) > K$, then $P(k+1) < P(k)$ and if $P(k) < K$, then $P(k+1) > P(k)$. The intrinsic growth rate β is given by $\beta = e^{\xi} - 1$ but, using the Maclaurin formula, for small ξ we have $\beta \approx \xi$.

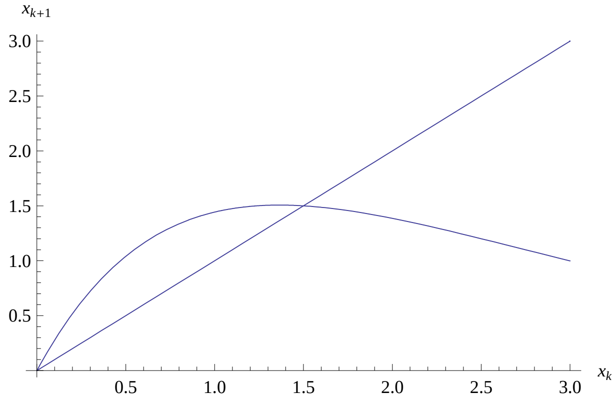

Fig. 2.2. The relation $x(k+1) = x(k)e^{\xi(1-x(k)/K)}$. The intersection of the graph of the function and the diagonal gives the carrying capacity K.

3 Discrete nonlinear population models from first principles

3.1 The Ricker equation in a continuous space framework, [3, 20].

The way we have introduced the Ricker equation may seem completely *ad hoc*. It follows, however, that it can be derived from a set of assumptions describing a uniformly distributed population in which the reproductive success of an individual is adversely affected by other individuals staying in its close proximity. To do this, first we must describe how to model the situation in which one may have various numbers of neighbours in a fixed neighbourhood. The simplest way to do this is to use the so-called *Poisson distribution*.

Interlude – Poisson distribution.

Let us first assume that there are n individuals in a population and let p be the probability that one of them happens to be in your neighbourhood. Under usual assumption of independence of individuals occurring in the neighbourhood, the probability of having exactly r neighbours is given by the binomial formula

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

Average number of neighbours is thus $\omega = np$. This is intuitively clear and also can be checked by direct calculation

$$\omega = \sum_{r=0}^{n} r \binom{n}{r} p^r (1-p)^{n-r} = np \sum_{r=0}^{n-1} \binom{n-1}{r-1} p^{r-1} (1-p)^{n-r} = np.$$

If we consider a potentially infinite population so that n grows to infinity in such a way that the average number of neighbours ω stays constant (so p goes to zero), then the probability of having exactly r neighbours is given by

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

which is called the *Poisson distribution*.

Returning to our problem, we consider a population the size of which at time k is given by $P(k)$. The standard growth equation is (2.2.1)

$$P(k+1) = R(P(k))P(k), \quad k = 0,1,\ldots, \tag{2.3.26}$$

where R gives the average number of offspring in the cycle. As mentioned before, we assume that the number of offspring per individual decreases with the number of individuals living in its region, say, a disc D of area s. A simple possible model is that the number of offspring per capita is given by bc^r where $b \geq 0, 0 < c < 1$, where r is the number of neighbours in D. If we assume that the population is uniformly distributed in an environment with area A, then the average number of individuals in D at time k is given by $sP(k)/A$ and, using the Poisson distribution, the probability of having r neighbours in D is

$$\frac{(sP(k))^r e^{-\frac{sP(k)}{A}}}{A^r r!}$$

and the average number of offspring per individual is

$$R(P(k)) = be^{-\frac{sP(k)}{A}} \sum_{r=0}^{\infty} \frac{(csP(k))^r}{A^r r!} = be^{-\frac{s(1-c)P(k)}{A}}.$$

Hence we obtain the Ricker model

$$P(k+1) = bP(k)e^{-\frac{s(1-c)P(k)}{A}}.$$

Comparing this expression with (2.2.25) we see that the carrying capacity K can be expressed as

$$K = \frac{A\ln b}{s(1-c)}.$$

Since $R(P)$ tends to zero faster than any power of P, we see that the Ricker model describes a scramble competition.

3.2 Ricker and Beverton-Holt models via site-based approach, [3].

There is a general framework which allows to derive most of the single species unstructured models by assigning the probability distribution of resource sites' occupation. Let us consider a habitat consisting of N discrete resource sites. At time n a population of $P(n)$ individuals is distributed and then reproduce. Let h_k be the proportion of sites accommodating k individuals. In general, it will be a function of both $P(n)$ and N. Once on site, the individuals reproduce and the success of reproduction; that is, the number of offspring, denoted by $\phi(k)$, only depends on the number of individuals at the site. Then the difference equation governing the growth of the population is

$$P(n+1) = N\sum_{k=0}^{\infty} h_k\phi(k). \tag{2.3.27}$$

To be able to make use of this equation, we must specify the site occupation function h_k and the maternity function $\phi(k)$.

We consider two types of the site occupation function: uniform and preferential. In what follows we assume a large population with a large number of sites, so that we can write the expected occupation to equal $P(k)/N$.

Uniform distribution. If individuals are uniformly distributed then, as in the above section, the probability of finding k individuals at any given site is Poisson distributed:

$$h_k = \frac{\left(\frac{P(n)}{N}\right)^k e^{-\frac{P(n)}{N}}}{k!}$$

and (2.3.27) can be written as

$$P(n+1) = Ne^{-\frac{P(n)}{N}}\sum_{k=0}^{\infty}\frac{(P(n))^k}{N^k k!}\phi(k). \tag{2.3.28}$$

Preferential distribution. To shorten notation, denote $\omega = P(k)/N$. Assume that the sites are allocated at random a value, say $t \in \mathbb{R}_+$, representing e.g. accessibility. We assume that the random variable defined as the value of the site has a continuous probability density f. Then, we assume that the average occupation of a site with value t is $t\omega$. If t was known then, as before, the number h_k of occupants of the site would be given by Poisson distribution $(\omega t)^k e^{-\omega t}/k!$. Since, however, the value of t is not known, h_k will be given by

$$h_k = \int_0^{\infty} f(t)\frac{(\omega t)^k e^{-\omega t}}{k!}dt. \tag{2.3.29}$$

The function $f(t)$ is the probability density of the number of sites with value t and must be selected for each particular case. Quite often it is assumed to be Gamma distributed; that is,

$$f(t) = \frac{\lambda^\lambda}{\Gamma(\lambda)}t^{\lambda-1}e^{-\lambda t}$$

where λ is a positive parameter. Then

$$h_k = \int_0^{\infty} f(t)\frac{(\omega t)^k e^{-\omega t}}{k!}dt = \frac{\lambda^\lambda}{k!\Gamma(\lambda)\omega^\lambda}\int_0^{\infty} t^{\lambda+k-1}e^{-t\frac{\lambda+\omega}{\omega}}dt \tag{2.3.30}$$

$$= \frac{\lambda^\lambda\omega^k}{k!\Gamma(\lambda)(\lambda+\omega)^{\lambda+k}}\int_0^{\infty} s^{\lambda+k-1}e^{-s}ds = \frac{\lambda^\lambda\omega^k\Gamma(k+\lambda)}{k!\Gamma(\lambda)(\lambda+\omega)^{\lambda+k}} \tag{2.3.31}$$

which is the so called negative binomial distribution. The formula for the growth of the population is then given by

$$P(n+1) = N\frac{\lambda^\lambda}{\Gamma(\lambda)(\lambda+P(n)/N)^\lambda}\sum_{k=0}^{\infty}\frac{(P(n)/N)^k\Gamma(k+\lambda)}{k!(\lambda+P(n)/N)^k}\phi(k). \tag{2.3.32}$$

The next step is to specify the offspring outcome at each site.

Scramble competition. Let as assume that each site contains resources to support one individual. Then

$$\phi_k = \begin{cases} b & \text{if} k = 1, \\ 0 \text{ otherwise}, \end{cases}$$

where b is the number of offspring produced by a site containing only one individual. Substituting this into (2.3.28), we obtain

$$P(n+1) = bP(n)e^{-\frac{P(n)}{N}}, \tag{2.3.33}$$

which is the Ricker model. On the other hand, the negative binomial distribution gives

$$P(n+1) = b\frac{\lambda^{\lambda+1}P(n)}{\Gamma(\lambda)(\lambda+P(n)/N)^{\lambda+1}}, \tag{2.3.34}$$

where we used $\Gamma(\lambda+1) = \lambda\gamma(\lambda)$. Since $\lambda > 0$, this is the Hassel model (2.2.14). Note, that since we have scramble competition, we cannot get here the basic Beverton-Holt model which is compensatory and thus describes a contest competition.

Contest competition. Again we assume that each site can support one individual but, in contrast to the scramble competition, if there are more individuals at the site, only one emerges victorious and the others perish. Thus, the function ϕ_k is given by

$$\phi_k = \begin{cases} b \text{ if } k \geq 1, \\ 0 \text{ if } k = 0, \end{cases}$$

where, as before, b is the number of offspring produced by one individual. Then the uniform distribution gives

$$P(n+1) = bNe^{-\frac{P(n)}{N}}\sum_{k=1}^{\infty}\frac{(P(n))^k}{N^k k!} = bN\left(1-e^{-\frac{P(n)}{N}}\right), \tag{2.3.35}$$

which is the so-called Skellam model (not discussed in this book).

Let us consider the negative binomial distribution. We have

$$P(n+1) = bN\frac{\lambda^\lambda}{\Gamma(\lambda)(\lambda+P(n)/N)^\lambda}\sum_{k=1}^{\infty}\frac{(P(n)/N)^k\Gamma(k+\lambda)}{k!(\lambda+P(n)/N)^k}. \tag{2.3.36}$$

Now, using the fact that $\Gamma(k+\lambda) = \lambda(\lambda+1)\cdot\ldots\cdot(\lambda+k-1)\Gamma(\lambda)$ and denoting $z = (P(n)/N)/(\lambda+P(n)/N)$, we get

$$\sum_{k=0}^{\infty}\frac{\Gamma(k+\lambda)}{k!}z^k = \Gamma(\lambda)\sum_{k=0}^{\infty}\frac{\lambda(\lambda+1)\cdot\ldots\cdot(\lambda+k-1)}{k!}z^k = \Gamma(\lambda)(1-z)^{-\lambda}.$$

Now,

$$1-z = 1-\frac{P(n)/N}{\lambda+P(n)/N} = \frac{\lambda}{\lambda+P(n)/N}$$

and thus (2.3.36) can be written as

$$\begin{aligned} P(n+1) &= bN\frac{\lambda^\lambda}{\Gamma(\lambda)(\lambda+P(n)/N)^\lambda}\left(\Gamma(\lambda)\lambda^{-\lambda}(\lambda+P(n)/N)^\lambda - \Gamma(\lambda)\right) \\ &= bN\left(1-\frac{\lambda^\lambda}{(\lambda+P(n)/N)^\lambda}\right). \end{aligned}$$

Questions:
i) Think about the interpretation of λ in the context of this model.
ii) If the model represents a pest population, what interventions could you undertake to decrease the its size?
iii) For what λ the above equation gives to the Beverton-Holt model.

If $\lambda = 1$, the above equation corresponds to the Beverton-Holt model (2.2.14). Indeed, in this case

$$P(n+1) = bN\left(1 - \frac{1}{(1+P(n)/N)}\right) = \frac{bP(n)}{1+P(n)/N}.$$

We see that the carrying capacity is given by $K = N(b-1)$; that is, it is proportional to the number of sites as well as to the per capita birth rate above the simple reproduction.

3.3 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 the intrinsic growth rate in 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 the population changes so that $\Delta P/P < 0$ for $0 < P < L$ and $P > K$ and $\Delta P/P > 0$ for $L < P < K$. If

$$\Delta P/P = R(P),$$

then the resulting difference equation is

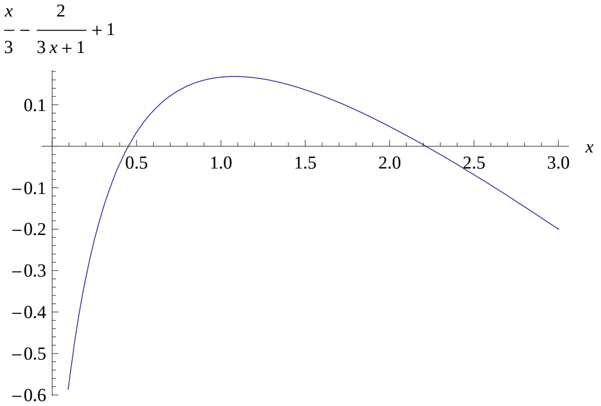

Fig. 2.3. The function $1 - \frac{x}{K} - \frac{A}{1+Bx}$

$$P(k+1) = P(k) + P(k)R(P(k))$$

and the required properties can be obtained by taking $R(P) \leq 0$ for $0 < P < L$ and $P > K$ and $R(P) \geq 0$ for $L < P < K$. To avoid negative populations we have to assume that $PR(P) + P \geq 0$, that is, $R(P) \geq -1$ for $0 \leq P \leq L$.

A simple model like that is offered by choosing $R(P) = r(L-P)(P-K)$, where r is a parameter satisfying $0 < r < 1/LK$ so that

$$P(k+1) = P(k)(1 + r(L - P(k))(P(k) - K)). \tag{2.3.37}$$

A more detailed analysis of this model is referred to Section 6.6

An Allee type model by multiple time scales

Another model of this type, which can be justified by modelling looking of a mating partner or introducing a generalist predator (that is, preying also on other species). Let us consider the latter, while the former model

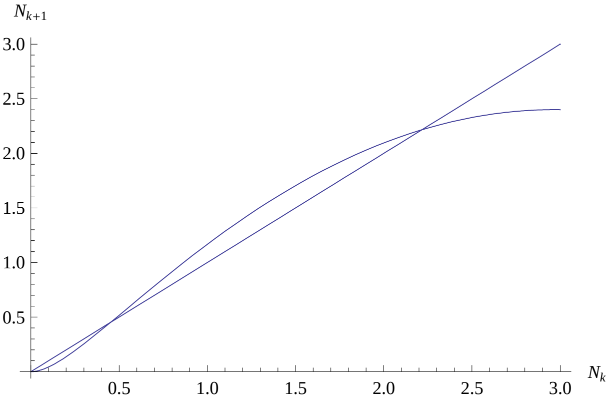

Fig. 2.4. The relation $P(k+1) = P(k) + P(k)R(P(k)$

in continuous time is discussed in a subsection of Section 6.2. Assume that $N(k)$ describes a population of prey and $H(k)$ describes the population of active predators at time k. The total population of predators P is supposed to be constant in time and is subdivided into $H(k)$ active (hungry) predators and $S(k)$ satiated predators

$$P = H(k) + S(k).$$

We assume that the prey population without the predator follows the logistic law with carrying capacity C and unrestricted growth rate $\lambda = \beta - \mu$. To model the predation process is modelled by the *law of mass action*. It is the simplest and thus most often used way of describing nonlinear interactions and thus it warrants a closer look. In the current context, it amounts to saying that one hungry predator in the unit time can eat on average a fraction ξ of the prey population. Thus, H predators will consume ξNH prey. In real modelling the the determining of the coefficient ξ is of paramount importance: it includes the probability of a predator meeting a prey, efficacy of killing etc. Clearly, the model has many shortcomings, such as taking account interactions between predators pursuing the same pray or finite capacity of the predator's stomach, but it has been successfully used in many a situation. With this understanding we write down the equation for the prey population as

$$N(k+1) = N(k) + \lambda N(k)\left(1 - \frac{N(k)}{C}\right) - \xi N(k)H(k).$$

We assume that the life cycle of predators is much longer than that of prey and we neglect the vital processes in the predator population. Thus, the evolution of active predators is governed by the equation

$$\begin{aligned} H(k+1) &= H(k) + \sigma S(k) - \theta N(k)H(k) - \zeta H(k) \\ &= H(k) + \sigma(P - H(k)) - \theta N(k)H(k) - \zeta H(k). \end{aligned}$$

Here σ is the proportion the satiated predators becomes hungry again (so that, as in (2.2.7), $1/\sigma$ is the average duration of satiation). The second term gives the number of active predators becoming satiated due to catching the prey from the prey population (note that we used a coefficient different than ξ – while ξNH is the total amount of prey killed in one cycle, this amount can feed more or (less) than ξNH predators. The last terms gives the number of active predators becoming inactive due to predation on other species. If we introduce new variables as fractions of C and P, respectively, by $N = xC$ and $H = yP$, then we obtain the system

$$\begin{aligned} x(k+1) &= x(k) + \lambda x(k)(1 - x(k)) - \xi P x(k)y(k), \\ y(k+1) &= y(k) + \sigma(1 - y(k)) - \theta C x(k)y(k) - \zeta y(k). \end{aligned} \tag{2.3.38}$$

Note that the second equation only describes the changes in the number of predators due to hunting. If we assume that the reproductive cycle of prey is, say, one year (e.g. antelopes), it is clear that the predator will hunt many times during this period. In other words, the state of y will change many times when k changes to $k+1$. It is then plausible to introduce another time scale for predators, say, n and write

$$y(n+1) = y(n) + \sigma(1 - y(n)) - \theta C x(k) y(n) - \zeta y(n). \tag{2.3.39}$$

Note that here $x(k)$ is treated as a constant. Assuming that y quickly settles close to the equilibrium determined by solving

$$\bar{y} = \bar{y} + \sigma(1 - \bar{y}) - \theta C x(k)\bar{y} - \zeta\bar{y};$$

that is,

$$\bar{y} = \frac{\sigma}{\sigma + \theta C x(k) + \zeta}, \tag{2.3.40}$$

Substitution this into the first equation of (2.3.38) gives

$$x(k+1) = x(k) + \lambda x(k)(1 - x(k)) - \frac{\sigma\xi P x(k)}{\sigma + \zeta + \theta C x(k)} \tag{2.3.41}$$

which, returning to the old variable $x(k) = N(k)/C$

$$N(k+1) = N(k)\left(1 + \lambda\left(1 - \frac{N(k)}{C} - \frac{A}{1 + BN(k)}\right)\right) \tag{2.3.42}$$

where $A = \sigma\xi P/\lambda(\sigma + \zeta)$ and $B = \theta/(\sigma + \zeta)$.

Example 2.2. Using the results of Subsection 6.6 to guide our choice, we consider the following coefficients: $\lambda = 1.2, \sigma = \zeta = 0.01, P = C = 1, \theta = 0.1$ and $\xi = 3$ so that (2.3.38) takes the form

$$\begin{aligned} x(k+1) &= x(k) + 1.2x(k)(1 - x(k)) - 3x(k)y(k), \\ y(k+1) &= y(k) + 0.01(1 - y(k)) - 0.1x(k)y(k) - 0.01y(k). \end{aligned} \tag{2.3.43}$$

Then the approximation (2.3.41) is given by

$$x(k+1) = x(k) + 1.2x(k)(1 - x(k)) - \frac{0.03x(k)}{0.02 + 0.1x(k)}. \tag{2.3.44}$$

Plotting the graph of

$$f(x) = 1.2x(1 - x) - \frac{0.03x}{0.02 + 0.1x} \tag{2.3.45}$$

we get a graph of the required shape. Finally, we provide a numerical illustration that, in this case, the solution to (2.3.44) indeed is an approximation to the x-component of the solution to (2.3.43).

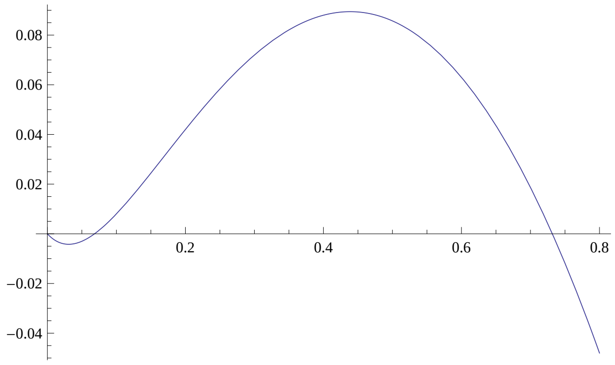

Fig. 2.5. Function f of (2.3.45): we see three equilibria, as required by the Allee model.

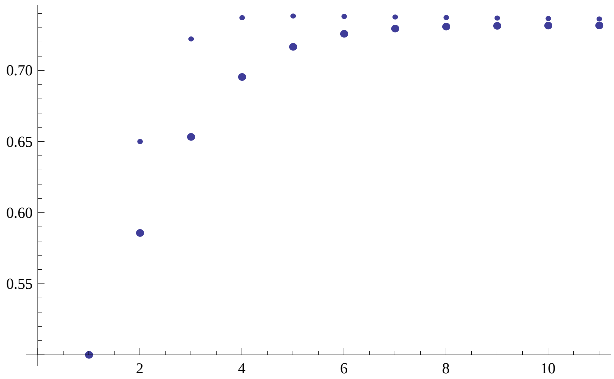

Fig. 2.6. Comparison of the solution to (2.3.44) (large dots) with the x-component of the solution to (2.3.43) (small dots). The initial conditions in both cases were $x(1) = 0.5$ and $y(1) = 0.1$.

3.4 Some explicitly solvable nonlinear population models

We complete this chapter by presenting some nonlinear models which can be explicitly solved by appropriate substitutions. An important role is played, however, by the formula for solutions of linear equations.

Solvability of linear difference equations

The simplest difference equations are these defining geometric and arithmetic progressions:

$$x(n+1) = ax(n),$$

and

$$y(n+1) = y(n) + a,$$

respectively, where a is a constant. The solutions of equations are known to be

$$x(n) = a^n x(0),$$

and

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

We shall consider the generalization of both these equations: the general first order difference equation,

$$x(n+1) = a(n)x(n) + g(n) \tag{2.3.46}$$

with an initial condition $x(0) = x_0$. Calculating first few iterates, we obtain

$$\begin{aligned} x(1) &= a(0)x(0) + g(0), \\ x(2) &= a(1)x(1) + g(1) = a(1)a(0)x(0) + a(1)g(0) + g(1), \\ x(3) &= a(2)x(2) + g(2) = a(2)a(1)a(0)x(0) + a(2)a(1)g(0) + a(2)g(1) + g(2), \\ x(4) &= a(3)x(3) + g(3) \\ &= a(3)a(2)a(1)a(0)x(0) + a(3)a(2)a(1)g(0) + a(3)a(2)g(1) + a(3)g(2) + g(3). \end{aligned}$$

At this moment we have enough evidence to conjecture that the general form of the solution could be

$$x(n) = x(0)\prod_{k=0}^{n-1} a(k) + \sum_{k=0}^{n-1} g(k) \prod_{i=k+1}^{n-1} a(i) \tag{2.3.47}$$

where we adopted the convention that $\prod\limits_{n}^{n-1} = 1$. Similarly, to simplify notation, we agree to put $\sum\limits_{k=j+1}^{j} = 0$. Then

$$\begin{aligned}
x(n+1) &= a(n)x(n) + g(n)\\
&= a(n)\left(x(0)\prod_{k=0}^{n-1} a(k) + \sum_{k=0}^{n-1} g(k) \prod_{i=k+1}^{n-1} a(i)\right) + g(n)\\
&= x(0)\prod_{k=0}^{n} a(k) + a(n)\sum_{k=0}^{n-1} g(k) \prod_{i=k+1}^{n-1} a(i) + g(n)\\
&= x(0)\prod_{k=0}^{n} a(k) + \sum_{k=0}^{n-1} g(k) \prod_{i=k+1}^{n} a(i) + g(n) \prod_{i=n+1}^{n} a(i)\\
&= x(0)\prod_{k=0}^{n} a(k) + \sum_{k=0}^{n} g(k) \prod_{i=k+1}^{n} a(i)
\end{aligned}$$

which proves that (2.3.47) is valid for all $n \in \mathbb{N}$.

Two special cases

There are two special cases of (2.3.46) that appear in many applications. In the first, the equation is given by

$$x(n+1) = ax(n) + g(n), \tag{2.3.48}$$

with the value $x(0)$ given. In this case $\prod\limits_{k=k_1}^{k_2} a(k) = a^{k_2-k_1+1}$ and (2.3.47) takes the form

$$x(n) = a^n x(0) + \sum_{k=0}^{n-1} a^{n-k-1} g(k). \tag{2.3.49}$$

The second case is a simpler form of (2.3.48), given by

$$x(n+1) = ax(n) + g, \tag{2.3.50}$$

with g independent of n. In this case the sum in (2.3.49) can be evaluated in an explicit form giving

$$x(n) = \begin{cases} a^n x(0) + g\frac{a^n - 1}{a-1} \text{ if } a \neq 1,\\ x(0) + gn. \end{cases} \tag{2.3.51}$$

The Hassel–Beverton–Holt model

We recall that the Beverton–Holt equation, Eq. (2.2.15), can be simplified to

$$x(n+1) = \frac{\beta x(n)}{(1+x(n))^b}. \tag{2.3.52}$$

While for general b this equation can display very rich dynamics, for $b = 1$ it can be solved explicitly. So, let us consider

$$x(n+1) = \frac{\beta x(n)}{1+x(n)}. \tag{2.3.53}$$

The substitution $y(n) = 1/x(n)$ converts (2.3.53) into

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

Using (2.3.51), we find

$$y(n) = \frac{1}{\beta}\frac{\beta^{-n}-1}{\beta^{-1}-1} + \beta^{-n}y_0 = \frac{1-\beta^n}{\beta^n(1-\beta)} + \beta^{-n}y_0$$

if $\beta \neq 1$ and

$$y(n) = n + y_0$$

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

The logistic equation

In general, the discrete logistic equation does not admit closed form solutions and also displays a very rich dynamics. However, some special cases can be solved by an appropriate substitution. We will look at two such cases. First consider

$$x(n+1) = 2x(n)(1-x(n)) \tag{2.3.54}$$

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

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

so that $y(n+1) = 2(y(n))^2$. We see that if $y_0 = 0$, then $y(n) = 0$ for all $n \geq 1$. Furthermore, the solution $y(n)$ for $n \geq 1$ does not change if we change the sign of y_0. Thus, we can take $|y_0| > 0$ as the initial condition. Then $y(n) > 0$ for $n \geq 1$ and we can take the logarithm of both sides getting, for $n \geq 1$, $\ln y(n+1) = 2\ln y(n) + \ln 2$ which, upon substitution $z(n) = \ln y(n)$, becomes the inhomogeneous linear equation $z(n+1) = 2z(n) + \ln 2$. Using (2.3.51), we find the solution to be $z(n) = 2^n z_0 + \ln 2(2^n - 1)$. Hence

$$y(n) = e^{z(n)} = e^{2^n \ln|y_0|}e^{\ln 2(2^n-1)} = y_0^{2^n}2^{2^n-1},$$

where we dropped the absolute value bars as we rise y_0 to even powers. Thus

$$x(n) = \frac{1}{2} - \left(\frac{1}{2} - x_0\right)^{2^n} 2^{2^n-1}.$$

We note that for $x_0 = 1/2$ we have $x(n) = 1/2$ for all n, so that we obtain a constant solution. In other words, $x = 1/2$ is an equilibrium point of (2.5.83).

Another particular logistic equation which can be solved by substitution is

$$x(n+1) = 4x(n)(1-x(n)). \tag{2.3.55}$$

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

$$x(n) = \sin^2 y(n) \tag{2.3.56}$$

which yields

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

This gives the family of equations

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

However, bearing in mind that our aim is to find $x(n)$ given by (2.3.56) and using the periodicity and symmetry of the function $\sin^2$, we can discard $k\pi$ as well as the minus sign and focus on $y(n+1) = 2y(n)$. This is a geometric progression and we get $y(n) = C2^n$, where $C \in \mathbb{R}$ is arbitrary, 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, despite the fact that there is an explicit formula for the solution, the dynamics generated by (2.3.55) is very irregular (chaotic).

4 Continuous in time single species unstructured models

At a 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 change continuously (and even in a 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.

In this section we shall introduce continuous models derivation of which parallels the derivation of discrete models above. We commence with exponential growth.

Let $P(t)$ denote the size of a population of a given isolated species at time t and let Δt be a small time interval. As in the discrete case, the population at time $t + \Delta t$ can be expressed as

$$P(t+\Delta t) - P(t) = \text{number of births in } \Delta t - \text{number of deaths in } \Delta t.$$

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, so that introducing birth and death rates β and μ, respectively, we obtain

$$P(t+\Delta t) - P(t) = \beta(t, P(t))P(t)\Delta t - \mu(t, P(t))P(t)\Delta t. \tag{2.4.57}$$

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

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

If we fix Δt and take it as a unit time interval and drop the dependence on t, then the above equation is exactly (2.2.9) with $R(P) = 1 + r(P)$. Here, however, we assume that the change happens continuously, so dividing by Δt and passing with $\Delta t \to 0$ we arrive at the continuous in time counterpart of (2.2.9):

$$\frac{dP}{dt} = r(t, P)P. \tag{2.4.58}$$

To proceed, we have to specify the form of r.

4.1 Exponential growth

As before, the simplest possible $r(t, P)$ is a constant and in fact such a model is used in a short-term population forecasting. So let us assume that $r(t, P(t)) = r$ so that

$$\frac{dP}{dt} = rP. \tag{2.4.59}$$

which has a general solution given by

$$P(t) = P(t_0)e^{r(t-t_0)}, \tag{2.4.60}$$

where $P(t_0)$ 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 $P(t_0) = P(1965) = 3.34 \times 10^9$ with $r = 0.02$, and (2.4.60) takes the form

$$P(t) = 3.34 \times 10^9 e^{0.02(t-1965)}. \tag{2.4.61}$$

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

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

thus

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

and

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

This gives a good agreement with the estimated value of the Earth population in 2000, which was 6070 billion (though we already observe that is is an overestimate). We see that it also agrees relatively well 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 ≈ 200000 billion. To realize what it means, let us recall that the Earth total surface area 510072000 square kilometers, 70.8% of which is covered by water, thus we have only 148940000 square kilometers to our disposal and there will be only 0.7447m^2 ($86.3\,\text{cm} \times 86.3\,\text{cm}$) per person. Therefore we can only hope that this model is not valid for all times. Indeed, as for discrete models, it is observed that the linear model for the population growth often is in good agreement with observations 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 it growth.

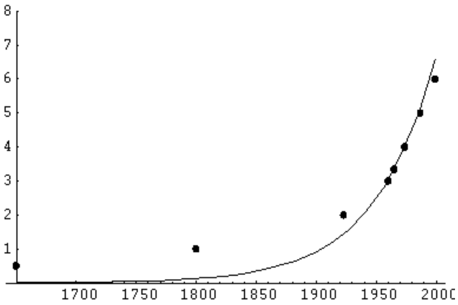

Fig 1.1. Comparison of actual population figures (points) with those obtained from equation (2.4.61).

4.2 Logistic equation

Again, the simplest way to take this into account is to take $r(t,P) = r(K-P)$ and then we obtain the so-called *continuous logistic model*

$$\frac{dP}{dt} = rP\left(1 - \frac{P}{K}\right), \tag{2.4.62}$$

which proved to be one of the most successful models for describing a single species population. Alternatively, as in the discrete case, we can obtain (2.4.62) by taking in (2.4.57) constant birth rate β but introduce density dependent mortality rate

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

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

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

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

$$\frac{dP}{dt} = (\beta - \mu_0)P - \mu_1 P^2$$

which is another form of (2.4.62).

A more general form of this equation is obtained by taking $\mu(P) = \mu_0 + \mu_1 P^\theta$ for some positive constant θ which leads to a continuous Bernoulli equation

$$\frac{dP}{dt} = (\beta - \mu_0)P - \mu_1 P^{\theta+1} \tag{2.4.63}$$

Let us focus on the logistic equation (2.4.62). Since the right-hand side does not contain t, it is a separable equation which, unlike its discrete counterpart, can be solved explicitly.

Let us start with some qualitative features. The constant in time solutions, corresponding to (2.2.13), are obtained by solving (2.4.63) with the right-hand side equal to zero. This gives $P = 0$ and $P = K$. In other words, the constant functions $P(t) = 0$ and $P(t) = K$ are equilibria. We shall discuss them in more detail in Section 6.1. We shall focus on solutions with the initial condition $P(t_0) > 0$. Then, from general theory (see Lemma 2.5) if $P(t_0) < K$, then $P(t)$ stays between 0 and K. Similarly, if $P(t_0) > K$, then the solution stays always above K. With this information, we can proceed with solving the related Cauchy problem

$$\begin{aligned} \frac{dP}{dt} &= rP\left(1 - \frac{P}{K}\right), \\ P(t_0) &= P_0 \end{aligned} \tag{2.4.64}$$

Separating variables and integrating we obtain

$$\frac{K}{r}\int\limits_{P_0}^{P} \frac{ds}{(K-s)s} = t - t_0.$$

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

$$\begin{aligned} \frac{K}{r}\int\limits_{P_0}^{P} \frac{ds}{(K-s)s} &= \frac{1}{r}\int\limits_{P_0}^{P} \left(\frac{1}{s} + \frac{1}{K-s}\right) ds \\ &= \frac{1}{r}\ln\frac{P}{P_0}\left|\frac{K-P_0}{K-P}\right|. \end{aligned}$$

From the considerations preceding (2.4.64), if $P_0 < K$, then $P(t) < K$ for any t, and if $P_0 > K$, then $P(t) > K$ for all $t > 0$. Therefore $(K - P_0)/(K - P(t))$ is always positive and

$$r(t - t_0) = \ln \frac{P}{P_0} \frac{K - P_0}{K - P}.$$

Exponentiating, we get

$$e^{r(t-t_0)} = \frac{P(t)}{P_0} \frac{K - P_0}{K - P(t)}$$

or

$$P_0(K - P(t))e^{r(t-t_0)} = P(t)(K - P_0).$$

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

$$P(t)\left(P_0 e^{r(t-t_0)} + K - P_0\right) = P_0 K e^{r(t-t_0)},$$

thus finally

$$P(t) = \frac{P_0 K}{P_0 + (K - P_0)e^{-r(t-t_0)}}. \tag{2.4.65}$$

Let us examine (2.4.65) to see whether we obtained the population's behaviour predicted by qualitative analysis (which helps to ensure that we havn't made any mistake solving the equation). First observe that we have

$$\lim_{t\to\infty} P(t) = K,$$

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

$$\frac{dP}{dt} = \frac{rP_0K(K - P_0)e^{-r(t-t_0)}}{(P_0 + (K - P_0)e^{-r(t-t_0)})^2}$$

thus, if $P_0 < K$, the population monotonically increases, whereas if we start with the population which is larger then the capacity of the habitat, then such a population will decrease until it reaches K. Also

$$\frac{d^2P}{dt^2} = r\frac{d}{dt}(P(K - P)) = P'(K - 2P) = P(K - P)(K - 2P)$$

from which it follows that, if we start from $P_0 < K$, then the population curve is convex down for $P < K/2$ and convex up for $P > K/2$. Thus, as long as the population is small (less then 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 $r = 0.02, K = 10$ and $t_0 = 0$, resulting in the following function:

$$P(t) = \frac{10P_0}{P_0 + (10 - P_0)e^{-0.2t}}.$$

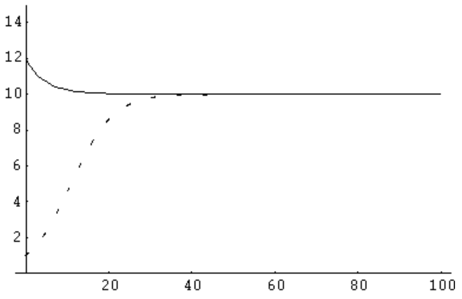

Fig 2.3 Logistic curves with $P_0 < K$ (dashed line) and $P_0 > K$ (solid line) for $K = 10$ and $r = 0.02$.

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 $r = 0.029$. Then the logistic equation for the growth of the Earth population will read

$$P(t) = \frac{P_0(10.76 \times 10^9)}{P_0 + ((10.76 \times 10^9) - P_0)e^{-0.029(t-t_0)}}.$$

We use this function with the value $P_0 = 3.34 \times 10^9$ at $t_0 = 1965$. The comparison is shown on Fig. 2.4.

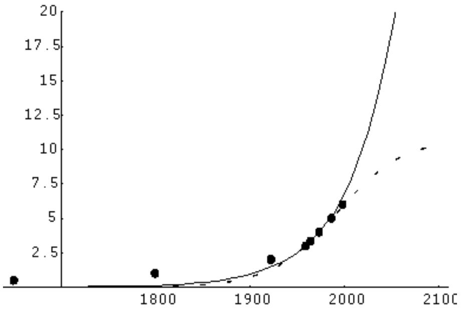

Fig 2.4 Human population on Earth. Comparison of observational data (points), exponential growth (solid line) and logistic growth (dashed line).

4.3 Continuous Allee model

The argument used in deriving discrete Allee models can be used to derive equations describing a similar behaviour in continuous time. In this way we can obtain the simplest model

$$\frac{dP(t)}{dt} = r(L - P(t))(P(t) - K))P(t). \tag{2.4.66}$$

To derive a more complex model corresponding to (2.3.42), we provide an analogue of the quasi steady state argument in continuous time. However, for diversity, we consider a different model in which females search for a mate, see e.g. [24].

An Allee model from a 'female searching for a mate' model

We consider a spatially homogeneous population inhabiting a certain area. We assume the constant one-to-one sex ratio so that we are not going to model explicitly the male population. The density of the female population is denoted by P.

As mentioned earlier, in many cases the population models, in which only the increase in the density of the population has a negative impact on the demography of the population, are inadequate. For instance, it has been observed that often the sparsity of the population may have a detrimental effect. Earlier we presented an example of such a population in which the Allee effect was a result of being preyed upon by a fast generalist predator. Here we will present another model of this type, describing a population of females who have to look for a mate. The total population density is subdivided as

$$P = P_1 + P_2,$$

where P_1 denotes the density of females who recently have mated and P_2 denotes the density of females who are searching for a mate. We assume that females reproduce in a very short time after mating. Then the population can be described by a typical mass action coupled model

$$\begin{aligned}\frac{dP_1}{dt} &= \beta P_1 - (\mu + \nu P)P_1 - \sigma P_1 + \xi P P_2,\\ \frac{dP_2}{dt} &= -(\mu + \lambda + \nu P)P_2 + \sigma P_1 - \xi P P_2.\end{aligned} \tag{2.4.67}$$

Here β denotes the per capita reproduction rate of recently mated females, $\mu + \nu P$ denotes the per capita mortality rate of recently mated females, $\mu+\lambda+\nu P$ denotes the per capita mortality rate of females searching for a mate, σ denotes the rate at which the females switch from the reproductive stage to the searching stage and ξP denotes the per capita rate at which a searching female finds one out of P potential mates. Note the increased mortality rate of the searching females, which is attributed to the fact that such females have to leave their shelters and travel, increasing thus risk of being, say, killed by predators.

Identifying different time scales is a little different than in the discrete case. Here, the ratio of the time scales will appear as a small parameter in the system. However, the first step in both cases is adimensionalization of the system; that is, finding typical time and size scales. Arguing as in the discrete case, Section 2.3, since μ is the natural mortality rate, $1/\mu$ is the average life span of individuals without external influence. Hence it is natural to measure time in the units of the average life span and to do this, we introduce new time $s = \mu t$; that is, the average life span of an individual is of order 1. Using this time, we obtain $\frac{d}{dt} = \mu \frac{d}{ds}$ and the system can be written as

$$\begin{aligned}\mu \dot{P}_1 &= \beta P_1 - (\mu + \nu P)P_1 - \sigma P_1 + \xi P P_2,\\ \mu \dot{P}_2 &= -(\mu + \lambda + \nu P)P_2 + \sigma P_1 - \xi P P_2,\end{aligned}$$

where now $\dot{}$ denotes the differentiation with respect to s. Similarly, thinking about the population without searching females, $P = P_1$, we see that the carrying capacity K can be taken as

$$K = \frac{\beta - \mu}{\nu},$$

where we assume $\beta - \mu > 0$. Then we have

$$\begin{aligned}\mu \dot{P}_1 &= (\beta - \mu)P_1 \left(1 - \frac{P}{K}\right) - \sigma P_1 + \xi P P_2,\\ \mu \dot{P}_2 &= -(\mu + \lambda + \nu P)P_2 + \sigma P_1 - \xi P P_2.\end{aligned} \tag{2.4.68}$$

Taking the carrying capacity as our reference population size and setting $P_1 = xK$ and $P_2 = yK$, we obtain our system in dimensionless form,

$$\begin{aligned}\mu \dot{x} &= (\beta - \mu)x(1 - (x + y)) - \sigma x + \xi K y(x + y),\\ \mu \dot{y} &= -(\mu + \lambda + \nu K(x + y))y + \sigma x - \xi K y(x + y).\end{aligned}$$

Let us denote $\varepsilon = \frac{\mu}{\sigma}$. Arguing as with μ, we see that $\frac{1}{\sigma}$ is the average time of satiation after mating; that is, the average time a female stays in the first population. Thus

$$\varepsilon = \frac{\frac{1}{\sigma}}{\frac{1}{\mu}};$$

that is, ε is the ratio of average time of satiation to the average life span. Hence, in many cases ϵ can be considered to be a very small parameter. Denoting further $R_0 = \frac{\beta}{\mu}$, we can write our system in the form

$$\begin{aligned}\dot{x} &= (R_0 - 1)x(1 - (x + y)) + \frac{\xi K}{\mu} y(x + y) - \frac{1}{\epsilon} x,\\ \dot{y} &= -\left(1 + \frac{\lambda + \nu K(x + y)}{\mu}\right) y - \frac{\xi K}{\mu} y(x + y) + \frac{1}{\epsilon} x,\end{aligned} \tag{2.4.69}$$

supplemented by the initial conditions

$$x(0) = \overset{\circ}{x}, \qquad y(0) = \overset{\circ}{y}\,. \tag{2.4.70}$$

Adding equations in (2.4.69) and denoting $z = x + y$, we obtain

$$\begin{aligned}
\dot{z} &= x(R_0 - 1)(1 - z) - \left(1 + \frac{\lambda + \nu K z}{\mu}\right)(z - x),\\
\epsilon \dot{x} &= \epsilon(R_0 - 1)x(1 - z) + \epsilon\frac{\xi K}{\mu}(z - x)z - x\\
z(0) &= \overset{\circ}{z} = \overset{\circ}{x} + \overset{\circ}{y}, \qquad x(0) = \overset{\circ}{x},
\end{aligned} \tag{2.4.71}$$

The so-called quasi steady state approximation amounts to saying that since ϵ is very small, we can set it equal to zero. This converts the second equation of (2.4.71) into an algebraic equation the solution of which (here $x = 0$) is substituted into the first equation creating a simpler dynamical system which, under certain conditions, provides a reasonable approximation of the original system. We shall proceed with the quasi steady approximation in a purely formal manner; the justification of the presented steps is based on the Tikhonov theorem, see [1].

Hence, setting $\epsilon = 0$ in the second equation of (2.4.71), getting $x = 0$ and thus the first equation becomes

$$\dot{z} = -\left(1 + \frac{\lambda + \nu K z}{\mu}\right)z, \qquad z(0) = \overset{\circ}{z}\,. \tag{2.4.72}$$

While it can be proved that the solution of (2.4.72), together with $x = 0$, provide a good approximation of (2.4.71), they describe rather uninteresting dynamics in which the population P_1 disappears and the total population has only one equilibrium, equal to zero, and hence cannot display the Allee effect. A biological reason for this is that our assumption of ϵ means that the mated females quickly return to active life, but cannot easily find a new mate. This creates a significant imbalance between P_1 and P_2, with P_1 becoming small. Since only the P_1 population produces offspring, the population could become extinct. We can surmise that for a balanced population, the rate at which a searching female finds a mate should be comparable with the rate she rests after reproduction. In other words, a female should be able to find a mate soon after she is ready for reproduction. Thus, a good candidate for another small parameter is ξ. We also note that the parameters R_0, K, ν refer to the demography of the whole population and therefore they should not have any relation to σ. Another parameter which could be related to σ is λ – it is not unnatural to consider the additional death rate due to searching for a mate to have the same order as σ. We shall look at this case later.

Thus, writing $\xi/\mu = \xi\sigma/\mu\sigma = \xi/\sigma\epsilon = \bar{\xi}/\epsilon$, we consider

$$\begin{aligned}
\dot{z} &= (z - y)(R_0 - 1)(1 - z) - \left(1 + \frac{\lambda + \nu K z}{\mu}\right)y,\\
\dot{y} &= -\left(1 + \frac{\lambda + \nu K z}{\mu}\right)y - \frac{\bar{\xi}K}{\epsilon}yz + \frac{1}{\epsilon}(z - y),\\
z(0) &= \overset{\circ}{z}, \qquad y(0) = \overset{\circ}{y}\,.
\end{aligned} \tag{2.4.73}$$

The right hand side of the first equation can be simplified as follows

$$(z - y)(R_0 - 1)(1 - z) - \left(1 + \frac{\lambda + \nu K z}{\mu}\right)y = (R_0 - 1)z(1 - z) - \frac{\beta + \lambda}{\mu}y$$

so that we finally consider

$$\begin{aligned}
\dot{z} &= (R_0 - 1)z(1 - z) - \frac{\beta + \lambda}{\mu}y,\\
\varepsilon \dot{y} &= -\epsilon\left(1 + \frac{\lambda + \nu K z}{\mu}\right)y - \bar{\xi}Kyz + z - y,\\
z(0) &= \overset{\circ}{z}, \qquad y(0) = \overset{\circ}{y}\,.
\end{aligned} \tag{2.4.74}$$

Using the quasi steady state approximation and setting $\epsilon = 0$, we get

$$y = \frac{z}{1 + \bar{\xi} K z}. \tag{2.4.75}$$

Hence, approximating equation for z is given by

$$\dot{z} = (R_0 - 1)z(1 - z) - \frac{\beta + \lambda}{\mu} \frac{z}{1 + \bar{\xi} z}. \tag{2.4.76}$$

Returning to the original notation, we find that (2.4.76) can be written as

$$\frac{dP(t)}{dt} = \lambda P(t) \left(1 - \frac{P(t)}{C} - \frac{A}{1 + BP(t)}\right), \tag{2.4.77}$$

where

$$\lambda = \mu(R_0 - 1) = \beta - \mu, \quad C = K = \frac{\beta - \mu}{\nu}, \quad A = \frac{\beta + \lambda}{\mu(R_0 - 1)} = \frac{\beta + \lambda}{\beta - \mu}, \quad B = \frac{\bar{\xi}}{K} = \frac{\nu \bar{\xi}}{\beta - \mu}.$$

As we shall see in Section 6.2, there is a range of parameters for which (2.4.77) describes an Allee dynamics. The accuracy of the approximation is illustrated on Fig. 2.7. As we noted earlier, one can argue that the additional death rate λ, corresponding to the searching of mates, could be of the same order as σ. Then (2.4.69) can be written as

$$\begin{aligned} \dot{x} &= (R_0 - 1)x(1 - (x + y)) + \frac{\bar{\xi} K}{\epsilon} y(x + y) - \frac{1}{\epsilon} x, \\ \dot{y} &= -\left(1 + \frac{\nu K}{\mu}(x + y)\right) y - \frac{\bar{\lambda} + \bar{\xi} K(x + y)}{\epsilon} y + \frac{1}{\epsilon} x, \end{aligned} \tag{2.4.78}$$

with

$$x(0) = \overset{\circ}{x}, \qquad y(0) = \overset{\circ}{y}, \tag{2.4.79}$$

where $\lambda = \bar{\lambda}\sigma$ so that $\bar{\lambda}/\epsilon = \lambda/\mu$. If we multiply both equations by ϵ and let $\epsilon = 0$ we find

$$\begin{aligned} &\bar{\xi} K y(x + y) - x = 0, \\ -(&\bar{\lambda} + \bar{\xi} K(x + y))y + x = 0 \end{aligned} \tag{2.4.80}$$

which has the unique solution $(x, y) = (0, 0)$. It is not exactly the situation we encountered earlier, but it can be indeed proved that the solution to (2.4.78) converges exponentially to $(0, 0)$ as $\epsilon \to 0$. This is not an unreasonable result - the increasing death rate in the N_2 population combined with the fast rate of transfer of females from N_1 to N_2 drives the population to zero even faster than in the first case, when the reason for the extinction was the non-breeding N_1 part of the population becoming too large.

5 From discrete to continuous models and back.

5.1 Discretization of continuous models

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 corresponding differential equation. The question arises whether, under reasonable circumstances, discrete and continuous models are equivalent in the sense that they give the same solutions (or at least, the same qualitative features of the solution) and whether there is one-to-one correspondence between continuous and discrete models.

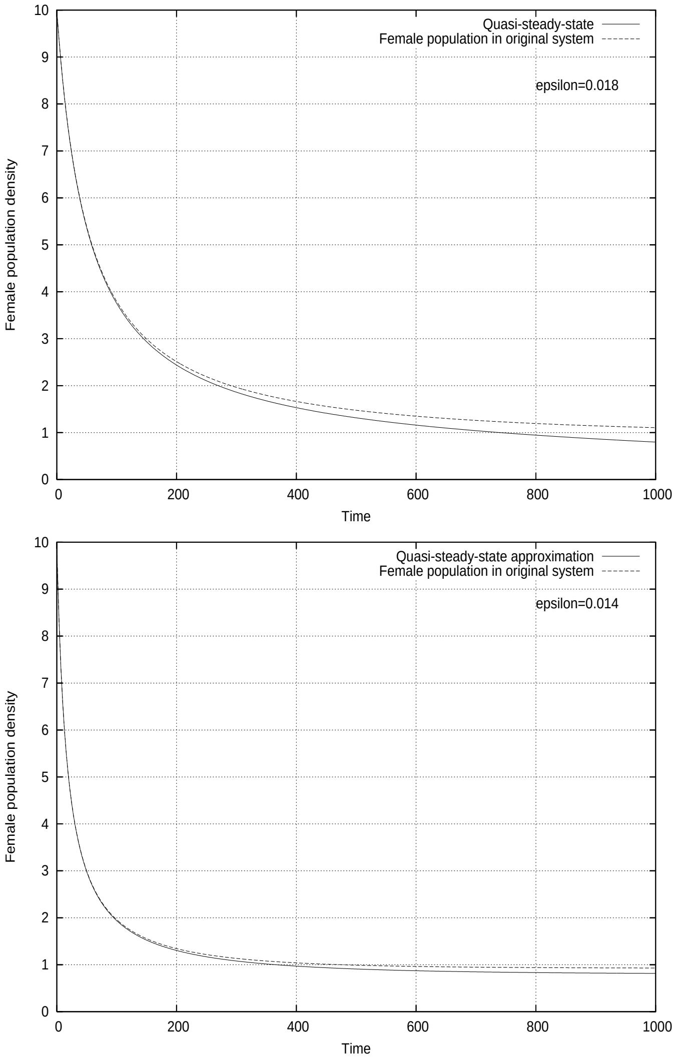

Fig. 2.7. Comparison of the total female population z, given by (2.4.74) with the approximation provided by (2.4.76) for $\epsilon = 0.018$ (top) and $\epsilon = 0.014$ (bottom).

There are several ways of discretization of differential equations. We shall use two most commonly used. The first one is similar to standard numerical analysis practice of replacing the derivative by a difference quotient:

$$\frac{df}{dt} \approx \frac{f(t+\Delta t) - f(t)}{\Delta t}.$$

Another one is based on the observation that solutions of autonomous equations display the so-called semigroup property: Denote by $x(t, x_0)$ the solution to the equation

$$x' = g(x), \qquad x(0) = x_0,$$

then

$$x(t_1 + t_2, x_0) = x(t_1, x(t_2, x_0)).$$

Thus,

$$x((n+1)\Delta t, x_0) = x(\Delta t, x(n\Delta t, x_0)). \tag{2.5.81}$$

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 f is an operation of getting solution of the Cauchy problem at Δt with initial condition as its argument. In further applications we shall take $\Delta t = 1$.

Exponential growth

Let us start with the exponential growth

$$P' = rP, \qquad P(0) = P_0$$

having the solution

$$P(t) = P_0 e^{rt}.$$

Let us denote by $(p(k))_{k\geq 0}$ the solution of the discrete equation obtained by the Euler discretization

$$p(k+1) - p(k) = rp(k)$$

with the solution

$$p(k) = (1+r)^k P_0.$$

Clearly, this solution does not coincide with the solution $P_0 e^{rt}$ for any value $t = 1, 2, \ldots$. However, qualitatively these solutions are similar as both grow exponentially and one can be transformed to the other by rescaling the growth rate. In other words the discrete Malthusian process with growth rate R_0, see (2.2.5), coincides with the continuous Malthusian growth with the growth rate r at any time $t \in \mathbb{P}$ if, instead of $R_0 = 1 + r$ as above, we take $R_0 = e^r$.

On the other hand, consider the second discretization, which amounts to assuming that we take census of the population in evenly spaced time moments $t_0 = 0, t_1 = 1, \ldots, t_k = k, \ldots$ so that

$$p(k) = P(k) = e^{rk} P_0 = (e^r)^k P_0. \tag{2.5.82}$$

Comparing this equation with (2.2.5), we see that it corresponds to the discrete model with intrinsic growth rate

$$R_0 = e^r,$$

as before. However, with this discretization we do not need any rescaling.

Thus we can state that if we observe a continuously growing population in discrete unit time intervals and the observed (discrete) intrinsic growth rate is R_0, then the real (continuous) growth rate is given by $r = \ln R_0$.

Logistic growth

Consider now the logistic equation

$$P' = rP\left(1-\frac{P}{K}\right).$$

Euler scheme for the logistic equation

The first discretization immediately produces the discrete logistic equation (2.2.18)

$$p(k+1) = p(k) + rp(k)\left(1-\frac{p(k)}{K}\right),$$

solutions of which, as we shall see, behave in a dramatically different way that those of the continuous equation, unlike the exponential growth equation.

We shall work with the simplified logistic differential equation

$$Y' = aY(1-Y), \qquad Y(0) = y_0. \tag{2.5.83}$$

We know that for, say, $a = 4$, the dynamics of the corresponding difference equation

$$y(n+1) = y(n) + 4y(n)(1-y(n)) \tag{2.5.84}$$

is chaotic and thus the latter cannot be used for numerical calculations of (2.5.83) as the solutions to (2.5.83) are monotonic. This is shown in Fig. 2.8. Let us, however, write down the complete Euler scheme:

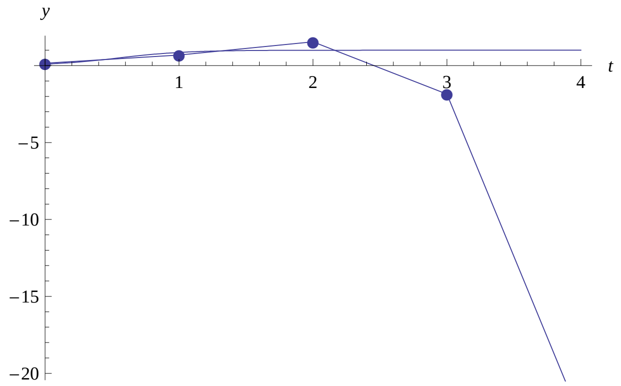

Fig. 2.8. Comparison of solutions to (2.5.83) with $a = 4$ and (2.5.84).

$$y(n+1) = y(n) + a\Delta t y(n)(1-y(n)), \tag{2.5.85}$$

where $y(n) := y(n\Delta t)$ and $y(0) = y_0$. Then

$$y(n+1) = (1+a\Delta t)y(n)\left(1-\frac{a\Delta t}{1+a\Delta t}y(n)\right).$$

Substitution

$$x(n) = \frac{a\Delta t}{1+a\Delta t}y(n) \tag{2.5.86}$$

reduces (2.5.85) to

$$x(n+1) = \gamma x(n)(1-x(n)), \tag{2.5.87}$$

where $\gamma = 1+a\Delta t$. Thus, the parameter γ, which controls the long time behaviour of solutions to the discrete equation (2.5.87), depends on Δt and, by choosing a suitably small Δt we can get solutions of (2.5.87) to mimic the behaviour of solutions to (2.5.83). Indeed, by taking $1 + a\Delta t \leq 3$ we obtain the convergence of solutions $x(n)$ to the equilibrium

$$x = 1\frac{a\Delta t}{1 + a\Delta t},$$

see Section 6.8. Reverting (2.5.86), we get the discrete approximation $y(n)$ which converges to 1, as the solution to (2.5.83). However, as seen on Fig 2.9, this convergence is not monotonic which shows that the approximation is rather poor. This can be remedied by taking $1 + a\Delta t \leq 2$ in which case the qualitative features of $y(t)$ and $y(n)$ are the same, see Fig. 2.10. This fact is proved in more detail in Remark 2.18.

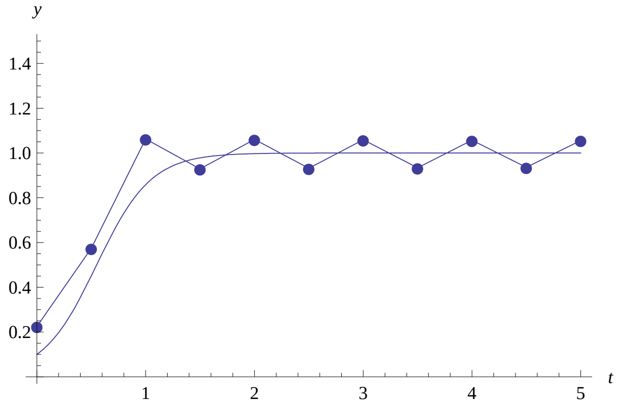

Fig. 2.9. Comparison of solutions to (2.5.83) with $a = 4$ and (2.5.87) with $\gamma = 3$ ($\Delta t = 0.5$).

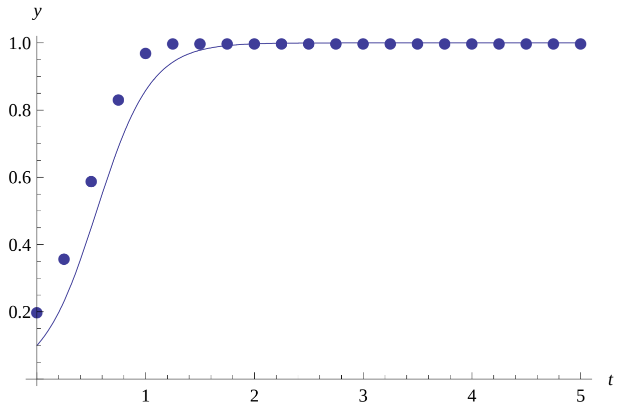

Fig. 2.10. Comparison of solutions to (2.5.83) with $a = 4$ and (2.5.87) with $\gamma = 2$ ($\Delta t = 0.25$).

These features of the discrete logistic model can, to a certain extent, be explained by interpreting it as a game between the population and the environment, in which the response of the environment to the population size $y(n)$ comes only after the full time step, resulting in the population of $y(n + 1)$. It is then natural to expect that the system is more likely to lose stability if the response times are long. On the other hand, in the continuous logistic model the responses are instantaneous, resulting in its monotonic and smooth behaviour.

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 by 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 $y(n+1)$ 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 $\beta = 1 + a\Delta t$ (and $K = 1$). We have seen in Section 3.4 that $(y(n))_{n\in\mathbb{N}}$ monotonically converges to an equilibrium and, as we shall see below, it exactly follows the solution of the continuous logistic equation. In the spirit of the game interpretation of the model, discussed above, this stability can be attributed to the fact that the environment response is based on the input $y(n)y(n+1)$ combining the previous and the current time instants, in contrast to $(y(n))^2$ in the case of the Euler discretization above.

To use the time-one map discretization, we re-write (2.4.65) as

$$P(t) = \frac{P_0 e^{rt}}{1 + \frac{e^{rt}-1}{K}P_0}.$$

which, upon denoting $e^r = \beta$, gives the time-one map

$$P(1, P_0) = \frac{P_0\beta}{1 + \frac{\beta-1}{K}P_0},$$

which, according to the discussion above, yields the Beverton-Holt model

$$p(k+1) = \frac{\beta p(k)}{1 + \frac{\beta-1}{K}p(k)},$$

with the discrete intrinsic growth rate related to the continuous one in the same way as in the exponential growth equation.

5.2 Discrete equations in continuous time models

In the previous section we have seen that often it is difficult to describe processes occurring in continuous time using difference equations which, on the other hand, usually are easier to handle. Here we shall describe two situations in which it is possible. First we discuss models with periodic coefficients. They yield to a discrete time description, provided one is satisfied with only knowing the state of the system in time instants, corresponding to the period of the coefficients. The second case concerns hybrid models, in which we have a continuous repetitive process interspersed with instantaneous events at evenly spaced time intervals.

Discrete models of seasonally changing populations

So far we have considered models in which the laws of nature are independent of time. In most real processes we have to take into account phenomena which depend on time, such as the seasons of the year. Here we use the same modelling principles as in Section 4, but with time dependent birth and death coefficients $\beta(t)$ and $\mu(t)$. We also consider emigration, which also is supposed to be proportional to the total population, and immigration, which is just a given flux of individuals into the system. Then, instead of (2.4.59), we have

$$N'(t) = (\beta(t) - \mu(t))N(t) - e(t)N(t) + c(t), \tag{2.5.88}$$

where e is a (time dependent) per capita emigration rate and c is the global immigration rate.

Closed systems.

Here we are interested in populations in which the coefficients change periodically with the same period, e.g., with the seasons of the year. As we shall see, contrary to naive expectations, in general this assumption does not yield periodic solutions. We start with a closed population, that is, we do not consider emigration or immigration processes. As in Section 4, we define $r(t) = \beta(t) - \mu(t)$ to be the net growth rate of the population and assume that it is a periodic function with a period T. Under this assumption, we introduce the average growth rate of the population by

$$\bar{r} = \frac{1}{T}\int\limits_{0}^{T} r(t)dt. \tag{2.5.89}$$

Hence, let us consider the initial value problem

$$P'(t) = r(t)P(t), \qquad P(t_0) = P_0, \tag{2.5.90}$$

the solution of which is given by

$$P(t) = P_0 e^{\int\limits_{t_0}^{t} r(s)ds}. \tag{2.5.91}$$

Since, by the periodicity of r,

$$\int\limits_{t_0}^{t+T} r(s)ds = \int\limits_{t_0}^{t} r(s)ds + \int\limits_{t}^{t+T} r(s)ds = \int\limits_{t_0}^{t} r(s)ds + \bar{r}T,$$

we have

$$P(t+T) = P(t)e^{\bar{r}T}$$

and hence the solution is not periodic. However, we may provide a better description of the evolution by identifying a periodic component in it. In other words, let us try to find what is 'missing' in the function $R(t) := \int_{t_0}^{t} r(s)ds$ so that it is not periodic. We observe that

$$R(t+T) = \int\limits_{t_0}^{t+T} r(s)ds = \int\limits_{t_0}^{t} r(s)ds + \int\limits_{t}^{t+T} r(s)ds = R(t) + \bar{r}T,$$

so that

$$R(t+T) - \bar{r}(t+T-t_0) = R(t) - \bar{r}(t-t_0),$$

and therefore the function $R(t)$, complemented by $-\bar{r}(t-t_0)$, becomes periodic.

Using this result, we can write

$$P(t) = P_0 e^{\int\limits_{t_0}^{t} r(s)ds} = P_0 e^{\bar{r}(t-t_0)}Q(t), \tag{2.5.92}$$

where

$$Q(t) = e^{\int\limits_{t_0}^{t} r(s)ds - \bar{r}(t-t_0)} \tag{2.5.93}$$

is a periodic function satisfying $Q(t_0) = 1$.

In particular, if we observe the population in discrete time intervals of the length T, we get

$$p(k) := P(t_0 + kT) = P_0 e^{\bar{r}kT} Q(t_0) = P_0[e^{\bar{r}T}]^k,$$

which is the exponential discrete model with the growth rate given by $e^{\bar{r}T}$.

Remark 2.3. What we presented above is a simple case of the Floquet theory, see [10, 13], which deals with systems of linear equations with periodic coefficients. The number $e^{\bar{r}T}$ is called the Floquet multiplier while the exponent $\bar{r}$ is called the Floquet exponent.

Since the function Q in (2.5.92) is periodic and continuous, it is bounded. Hence the solutions $P(t)$ are bounded if the Floquet exponent is non-positive and tend to the stationary point $P = 0$ if the Floquet exponent is negative.

Open systems.

Consider next an open population described by

$$P'(t) = r(t)P(t) + c(t), \tag{2.5.94}$$

where $r(t) = \beta(t) - \mu(t) - e(t)$ and $c(t)$ are continuous and periodic functions with period T. Let the constant $\bar{r}$ and the periodic function $Q(t)$ be defined as in (2.5.89) and (2.5.93). Using the integrating factor, we find the general solution to (2.5.94) to be

$$P(t) = e^{\int\limits_{t_0}^{t} r(s)ds} P(t_0) + e^{\int\limits_{t_0}^{t} r(s)ds} \int\limits_{t_0}^{t} e^{-\int\limits_{t_0}^{u} r(s)ds} c(u)du. \tag{2.5.95}$$

If there is a periodic solution of period T, say $\bar{P}$, there should be an initial condition P_o satisfying $P_o = \bar{P}(t_0) = \bar{P}(t_0 + T)$. Using (2.5.95), we obtain

$$\begin{aligned} P_o = \bar{P}(t_0) &= e^{\int\limits_{t_0}^{t_0+T} r(s)ds} \bar{P}(t_0) \\ &+ e^{\int\limits_{t_0}^{t_0+T} r(s)ds} \int\limits_{t_0}^{t_0+T} e^{-\int\limits_{t_0}^{u} r(s)ds} c(u)du. \end{aligned}$$

For simplicity we assume $\bar{r} \neq 0$ (see [10] for a discussion of this case in full generality) . By (2.5.93),

$$\bar{P}(t_0) = \frac{e^{\bar{r}T}}{1 - e^{\bar{r}T}} \int\limits_{t_0}^{t_0+T} \frac{e^{-\bar{r}(u-t_0)}c(u)}{Q(u)} du.$$

Let us define $\hat{P}(t) = \bar{P}(t + T)$, where $\bar{P}(t_0) = P_o$. Then

$$\begin{aligned} \hat{P}'(t) &= \bar{P}'(t + T) = r(t + T)\bar{P}(t + T) + c(t + T) \\ &= r(t)\hat{P}(t) + c(t) \end{aligned}$$

and, since $\hat{P}(t_0) = \bar{P}(t_0 + T) = \bar{P}(t_0) = P_o$, the uniqueness of solutions of linear differential equations yields

$$\bar{P}(t + T) = \hat{P}(t) = \bar{P}(t)$$

for any $t \in \mathbb{R}$. Hence $\bar{P}$ is periodic.

We know that the general solution to an inhomogeneous linear equation can be expressed as a sum of an arbitrary solution of the inhomogeneous equation and the general solution of the homogeneous equation. Hence, since $\bar{P}$ is a solution of the inhomogeneous equation (2.5.94) and the general solution of its homogeneous version is given (2.5.92), we have

$$P(t) = Ke^{\bar{r}(t-t_0)}Q(t) + \bar{P}(t),$$

where $K = P(t_0) - P_o$. Finally

$$P(t) = (P(t_0) - P_o)e^{\bar{r}(t-t_0)}Q(t) + \bar{P}(t). \tag{2.5.96}$$

This formula yields, in particular, that if $\bar{r} < 0$, then

$$\lim_{t\to\infty} (P(t) - \bar{P}(t)) = 0,$$

that is, with negative average growth rate, an arbitrary solution is asymptotically periodic.

Hybrid models

In many cases the observed process appears as a combination of two different ones: one occurring continually and the other in discrete time intervals. For instance, breeding happens in evenly spaced intervals but an unrelated death may happen any time. Similarly to the periodic cases, such combined processes can be described by discrete equations. We shall consider two such models, one leading to the Beverton-Holt equation, while the other providing yet another derivation of the Ricker equation from more basic principles.

The Hassel-Beverton-Holt equation model.

Consider a population which reproduces once a year and the reproductive season is of a negligible length. Let $P(n)$ be the population size in the nth year immediately after the reproductive season. During the year, outside the reproductive season, the population only is subjected to mortality and then the size p of the population obeys the equation

$$p' = -\mu(p)p = -(\mu_0 + \mu_1 p)p, \quad p(0) = P(n),\ \mu_0, \mu_1 > 0,$$

where t denotes the time that has elapsed since the end of the previous reproductive season (with one year as the time unit). Here, the death process is split into two parts: death from natural causes (aging) represented by the death rate μ_0 and death due to overcrowding, the per capita rate of which is given by $\mu_1 p$. Thus, $p(1)$ gives the population after one year, immediately before the reproductive season. Then, after the reproductive season, the population enters the next year with $P(n+1) = \beta p(1)$ individuals, where β is the birth rate. To solve the above equation, we separate variables and, integrating, we obtain

$$-t = \int\limits_{P(n)}^{p(t)} \frac{ds}{s(\mu_0 + \mu_1 s)} = \frac{1}{\mu_0} \ln \frac{p(t)(\mu_0 + \mu_1 P(n))}{(\mu_0 + \mu_1 p(t))P(n)}.$$

Exponentiating and solving for $p(t)$ gives

$$p(t) = \frac{\mu_0 P(n)}{\mu_0 e^{\mu_0 t} + \mu_1 P(n)(e^{\mu_0 t} - 1)}$$

and, using the reproduction law,

$$P(n+1) = \beta p(1) = \frac{\beta \mu_0 P(n)}{\mu_0 e^{\mu_0} + \mu_1 P(n)(e^{\mu_0} - 1)}.$$

We can rewrite it as

$$P(n+1) = \frac{\beta e^{-\mu_0} P(n)}{1 + \frac{\mu_1(1 - e^{-\mu_0})}{\mu_1} P(n)}. \tag{2.5.97}$$

Question: Find the carrying capacity in this model and the condition on β that ensure the survival of the population. Further, find the continuous logistic equation corresponding to this model.

The Ricker equation from a hybrid model.

We consider a model describing the size of a salmon population at the end of each spawning cycle. We adopt the notation from the previous paragraph. However, the construction of the model is slightly different. For simplicity, we assume that the breeding occurs at yearly intervals, numbered $n = 0, 1, 2, \ldots$. The time between breeding seasons will be denoted by $t \in [0, 1]$. We assume that the adults live for a short time $t \in [0, \tau]$, with $\tau << 1$, after each breeding. This is the case for Pacific salmon (but not Atlantic salmon). So, let $P(n)$ denotes the population of adult salmon immediately after the and of the spawning season n and thus at the beginning of the season $n+1$. Let us look at this season. During spawning the adult salmon produce larvae which become adult by the end of the cycle; that is, at $t = 1$. The size of the population of larvae is denoted by $p_n(t)$ with $0 \leq t \leq 1$. Let us assume that the number of produced larvae is proportional to the adult population at the end of the previous season; that is,

$$p_n(0) = \beta P(n).$$

Not all of these larvae become adult. The prevalent cause during $0 \leq t \leq \tau$ is cannibalism by the adult salmon. Again, a simple assumption is that the rate the larvae are eaten is proportional to the adult population and to the larvae population (law of mass action) but here the number of adults by $P(n)$ (we assume that during this short time the death of natural causes is negligible). Thus, for $t \in [0, \tau]$,

$$\frac{dp_n(t)}{dt} = -\lambda P(n) p_n(t), \tag{2.5.98}$$

where $\lambda > 0$ is a constant. Using the initial condition above, we find

$$p_n(t) = \beta P(n) e^{-\lambda P(n) t}. \tag{2.5.99}$$

Even when all adults are gone, not all larvae survive till the next breeding season. Again, the simple assumption is that a fraction γ of them will become adults. Thus

$$P(n+1) = \gamma p_n(\tau) = \gamma \beta P(n) e^{-\lambda \tau P(n)} \tag{2.5.100}$$

which we recognize as the Ricker equation (2.2.24).

Let us modify the equation to cater for the Atlantic salmon. In this case, the adults do not die shortly after spawning. Assume instead that they die at the constant rate μ (per capita per year). In such a case, equation (2.5.99) is no longer correct as we cannot assume that the number $P(n)$ is constant. Let $P_n(t)$ be the amount of adults alive at time $t \in [0, 1]$ in the n-th cycle. Using the assumption, we find that

$$\frac{dP_n(t)}{dt} = -\mu P_n(t), \qquad P_n(0) = P(n)$$

so that

$$P_n(t) = e^{-\mu t} P(n). \tag{2.5.101}$$

We no longer can separate the natural death of larvae from the death through cannibalism. Hence, using the law of mass action and the exponential death rate γ as before, we rewrite (2.5.99) as

$$\frac{dp_n(t)}{dt} = -\lambda P(n) p_n(t) e^{-\mu t} - \gamma p_n(t), \qquad p_n(0) = \beta P(n). \tag{2.5.102}$$

This is a separable equation whose solution is

$$p_n(t) = \beta P(n) e^{-\frac{\lambda}{\mu} P(n)(1 - e^{-\mu t})} e^{-\gamma t}. \tag{2.5.103}$$

Thus, using the fact that at the beginning of the $n+1$ cycle the number of adults is given by the number of surviving larvae (changing into adults) and the surviving adults (note that the surviving adults give birth to larvae which are not adults!), we have

$$P(n+1) = p_n(1) + P_n(1)$$

we find

$$P(n+1) = \beta P(n) e^{-\frac{\lambda}{\mu} P(n)(1-e^{-\mu})} e^{-\gamma} + e^{-\mu} P(n) = e^{-\mu} P(n) (e^{-\frac{\lambda}{\mu} P(n)(1-e^{-\mu})} + 1) = e^{-\mu} P(n) (A e^{-BP(n)} + 1)$$

for appropriate constants A and B.

Question: Find the carrying capacity of the environment and conditions ensuring the survival of the population in both cases.

6 Qualitative theory for a single equation

In most cases it is impossible to find an explicit solution to a given differential, or difference, equation. However, the power of mathematics lies in the fact that one often can deduce properties of solutions and answer some relevant questions just by analyzing the right hand side of the equation.

6.1 Equilibria of first order equations

One of the typical problems in the theory of differential and difference equations is to determine whether the system is stable, that is, whether if we allow it to run for a sufficiently long time (which, in the case of difference equations, means many iterations), it will eventually settle at some state, which should be an equilibrium.

In both difference and differential equations, by *equilibria* or *stationary solutions* we understand solutions which are constant with respect to the independent variable. Since, however, in the differential equation

$$x' = f(x) \tag{2.6.1}$$

the right hand side describes the rate of change of a given quantity, whereas in the difference equation

$$x(n+1) = f(x(n)) \tag{2.6.2}$$

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.

As we will see below, finding equilibria of equations is considerably easier than solving them. Thus, knowing that the system will converge to a particular equilibrium allows us to regard this equilibrium as an approximation of solutions originating in its neighbourhood.

In the next subsections we will make these notions precise.

6.2 Stability of equilibria of autonomous differential equations

We recall that the word autonomous refers to the fact that f in (2.6.1) does not explicitly depend on time. To have anything to talk about, we must ensure that (2.6.1) has solutions different from the equilibrium solutions. This is settled by the Picard–Lindelöf theorem which asserts that if f is sufficiently regular, for instance, differentiable on $\mathbb{R}$, then the initial value problem

$$x' = f(x), \qquad x(t_0) = x_0 \tag{2.6.3}$$

has exactly one solution defined for t on some interval (t_{min}, t_{max}) containing t_0. Furthermore, if the solution is bounded at the endpoints, then it can be extended to a larger interval. It is possible that the solution only is defined on a finite interval. However, if we can show that the solution is bounded on each finite interval of its existence, then it is defined on $\mathbb{R}$. In other words, if a solution to (2.6.3) with differentiable f is defined only on an interval with a finite endpoint, then it must be unbounded at this endpoint.

For further discussion we fix attention by assuming that f is an everywhere defined function satisfying all assumptions of the Picard–Lindelöf theorem on $\mathbb{R}$.

In many problems it is important to emphasize the dependence of the solution on the initial conditions. Thus we introduce the notion of the flow $x(t, t_0, x_0)$ of (2.6.1), which is the solution of the Cauchy problem (2.6.3). Here we only use $t_0 = 0$ and then we write $x(t, 0, x_0) = x(t, x_0)$.

If (2.6.1) has a stationary solution $x(t) \equiv x^*$ that, by definition, is constant in time, then such a solution satisfies $x'(t) \equiv 0$ and consequently

$$f(x^*) = 0. \tag{2.6.4}$$

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^*. If we now consider a function defined by $x(t) \equiv x^*$, then $x'(t) \equiv 0$. Consequently,

$$0 \equiv x'(t) \equiv (x^*)' = f(x^*)$$

and such a function is a stationary solution. Summarizing, equilibrium points are solutions to the algebraic equation (2.6.4) and, treated as constant functions, they are (the only) stationary, or equilibrium, solutions to (2.6.1). Therefore usually we will not differentiate between these terms.

Next we give a definition of stability of an equilibrium.

Definition 2.4. *1. The equilibrium x^* is stable if for given $\epsilon > 0$ there is $\delta > 0$ such that for any x_0 $|x_0 - x^*| < \delta$ implies $|x(t, x_0) - x^*| < \epsilon$ for all $t > 0$. If x^* is not stable, then it is called unstable.*

2. A point x^* is called attracting if there is $\eta > 0$ such that $|x_0 - x^*| < \eta$ implies $\lim\limits_{t\to\infty} x(t, x_0) = x^*$. If $\eta = \infty$, then x^* is called a global attractor or a globally attracting equilibrium.

3. The equilibrium x^* is called asymptotically stable if it is both stable and attracting. If x^* is globally attracting, then it is said to be a globally asymptotically stable equilibrium.

Equilibrium points play another important role for differential equations – they are the only limit points of bounded solutions as $t \to \pm\infty$. To make this precise, we begin with the following lemma.

Lemma 2.5. *If x_0 is not an equilibrium point of (2.6.1), then $x(t, x_0)$ is never equal to an 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 stationary solution, given by $x(t) \equiv x^*$. 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 Picard theorem. □

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 having the property that any solution always remains confined to one of them. We shall formulate and prove a theorem that strengthens this observation.

Theorem 2.6. *Let $x(t, x_0)$ be a non-stationary solution of (2.6.1) with $x_0 \in \mathbb{R}$ and let $I_{max} = (t_-, t_+)$ 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 to $-\infty$, or converges to an equilibrium point, as $t \to t_\pm$. 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_* an equilibrium point of f. This means that a non-stationary solution $x(t, x_0)$ reaches an equilibrium in finite time, which contradicts Lemma 2.5. Thus, if $x(t, x_0)$ 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 $t \to t_\pm$. Let us focus on the right end point t_+ of I_{max}. If $x(t, x_0)$ converges as $t \to t_+$, then $t_+ = \infty$, by the property of the maximal interval of existence. Thus

$$\lim_{t\to\infty} x(t, x_0) = \bar{x}.$$

Without compromising generality, we further assume that $x(t, x_0)$ is an increasing function. If $\bar{x}$ is not an equilibrium point then, by continuity, we can use the intermediate value property to claim that the values of $x(t, x_0)$ must fill the interval $[x_0, \bar{x})$. This interval cannot contain any equilibrium point as the existence of such points would violate the Picard-Lindelöf theorem. Thus, for any $x \leq \bar{x}$, $f(x)$ is strictly positive and hence, separating variables and integrating, we obtain

$$t(x) - t(x_0) = \int\limits_{x_0}^{x} \frac{ds}{f(s)}. \tag{2.6.5}$$

Passing with t to infinity (since $t(\bar{x}) = \infty$), we see that the left hand side becomes infinite and so

$$\int\limits_{x_0}^{\bar{x}} \frac{ds}{f(s)} = \infty.$$

By assumption, the interval of integration is finite so that the only way the integral could become infinite is if $1/f(s) = \infty$, that is, $f(s) = 0$, for some $s \in [x_0, \bar{x}]$. The only such point can be $s = \bar{x}$, thus $\bar{x}$ is an equilibrium point. □

Remark 2.7. We note that Eq. (2.6.5) is of independent interest 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 let $x(t, x_0)$ be increasing for $t > 0$. Then $\lim_{t\to t_+} x(t, x_0) = \infty$ so that, by (2.6.5),

$$t_+ - t(x_0) = \int\limits_{x_0}^{\infty} \frac{ds}{f(s)}$$

and, in particular, we see that if $1/f$ is integrable at $+\infty$ (precisely, if the improper integral above exists), then the maximal interval of existence is finite and we have the 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 $f(s)$ does not grow faster than s as $s \to \infty$. This occurs, e.g., if the derivative of f bounded on $\mathbb{R}$. On the other hand, If $f(s)$ behaves, say, as s^2 for large s, then the integral on the right hand side is finite and thus $t_{max} < \infty$.

Remark 2.8. Theorem 2.6 shows that for scalar differential equations with regular right hand sides, the distinction between different properties of an equilibrium made in Definition 2.4 is superfluous. Indeed, if an equilibrium x^* is stable, then solutions originating close to it stay close to it. However, by Theorem 2.6, these solutions are monotonic. Hence, the solutions are closer to x^* than their initial conditions are. In particular, they must be bounded and, being monotonic, they must converge as $t \to \infty$. From the proof of Theorem 2.6 it follows that the limit point must be the equilibrium x^*. This implies that x^* is attracting and hence asymptotically stable. Also, by monotonicity of solutions, any attracting equilibrium must be stable and thus asymptotically stable.

Remark 2.9. Theorem 2.6 usually is used in the following weaker form. Let f be continuously differentiable function. Then the equilibrium x^* is stable provided $f'(x^*) < 0$ and unstable provided $f'(x^*) > 0$. The proof is obvious–if $f'(x^*) < 0$, then $f'(x) < 0$ in some neighbourhood of x^*, by continuity of f'. Thus, $f > 0$ to the left and $f < 0$ to the right of x^* and any solution originating in such a left neighbourhood of x^* is increasing and must converge to x^*. Similarly, any solution originating in such a right neighbourhood of x^* is decreasing and also must converge to x^*. Thus x^* is asymptotically stable. An analogous argument shows that $f'(x^*) > 0$ means that x^* is unstable. However, Theorem 2.6 is much more general and allows to ascertain stability or instability in the so called nonhyperbolic cases when $f'(x^*)$ by considering the sign of f to the left and to the right of x^*.

Application to the logistic equation

Consider the Cauchy problem for the logistic equation

$$y' = y(1 - y), \qquad y(0) = y_0. \tag{2.6.6}$$

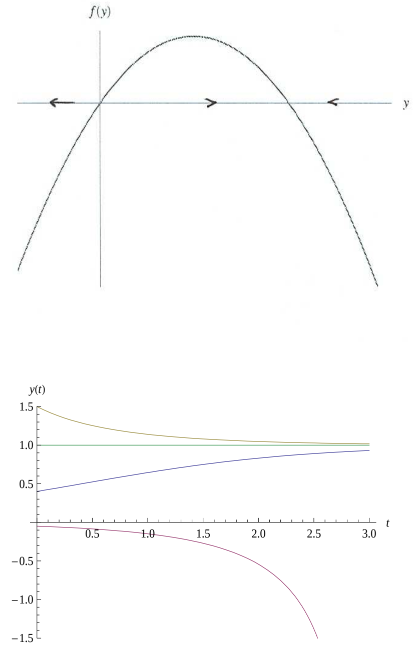

Fig. 2.11. Monotonic behaviour of solutions to (2.6.3) depends on the right hand side f of the equation.

We have solved this problem in Subsection 4.2. Let us now get as much information as possible about the solutions to this problem without actually solving it. First, 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, only one solution of (2.6.6) passes through each point (t_0, y_0). However, f is not a globally Lipschitz function, so that this solution may be defined only on a finite time interval.

The equilibrium points are found solving $y(1-y) = 0$, hence $y \equiv 0$ and $y \equiv 1$ are the only stationary solutions. Moreover, $f(y) < 0$ for $y < 0$ and $y > 1$ and $f(y) > 0$ for $0 < y < 1$. Hence, from Lemma 2.5, it follows that the solutions starting from $y_0 < 0$ will stay strictly negative, those starting from $0 < y_0 < 1$ will stay in this interval and those with $y_0 > 1$ will be larger than 1, for all times of their respective existence, as they cannot cross the equilibrium solutions. Then, from Theorem 2.6, we see that the solutions with negative initial condition are decreasing and therefore tend to $-\infty$ if time increases. In fact, they blow up in finite time since, by integrating the equation, we obtain

$$t(y) = \int\limits_{y_0}^{y} \frac{d\eta}{\eta(1-\eta)}$$

and we see, passing with y to $-\infty$ on the right hand side, that we obtain a finite time of the blow up.

Next, solutions with $0 < y_0 < 1$ are bounded and thus they are defined for all times. They are increasing and thus they must converge to the larger equilibrium, that is, $\lim_{t\to\infty} y(t, y_0) = 1$. Finally, if we start with $y_0 > 1$, then $y(t, y_0)$ decreases and thus is bounded from below, satisfying again $\lim_{t\to\infty} y(t, y_0) = 1$. The shape of the solution curves can be determined as in Subsection 4.2. By differentiating Eq. 2.6.6 with respect to time, we obtain

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

Since for each solution (apart from the stationary ones), y' has a fixed sign, we see that an inflection point can exist only for solutions starting at $y_0 \in (0,1)$ and it occurs at $y = 1/2$, where 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 the same picture as when solving the equation but with much less work.

Application to the Allee model

Let us consider the model (2.4.77),

$$\frac{dP}{dt} = \lambda P\left(1 - \frac{P}{C} - \frac{A}{1+BP}\right),$$

$\lambda, C, A, B > 0$, and finally prove that it indeed describes a behaviour required from the Allee model. Let us recall that for this, the equation must have three equilibria, 0 and, say, $0 < L < K$ such that if the size of the population P satisfies $0 < P < L$, then P decreases to 0 and if $L < P < K$, then P increases to K. In the terminology of this section, 0 and K should be asymptotically stable equilibria of (2.4.77) and L should be its unstable equilibrium.

Since (2.4.77) is difficult to solve explicitly (though it is possible as it is a separable equation), we use Theorem 2.6 to analyse it. The equilibria are solutions to

$$f(P) := P\left(1 - \frac{P}{C} - \frac{A}{1+BP}\right) = 0. \tag{2.6.7}$$

Clearly, $P \equiv 0$ is an equilibrium so, in particular, any solution originating from $P(0) = P_0 > 0$ satisfies $P(t) > 0$. We see that

$$f'(P) = 1 - \frac{2P}{C} - \frac{A}{(1+BP)^2} \tag{2.6.8}$$

and since $f'(0) = 1 - A$ we obtain that if $A > 1$, then $P = 0$ is an asymptotically stable equilibrium. By analysing the second derivative we can also state that if $A = 1$ and $BC < 1$, then $P = 0$ is semi-stable, that is, it attracts trajectories originating from positive initial conditions but this case is not relevant in studying the Allee type behaviour. Now we can focus on the other equilibria. For (2.4.77) to describe an Allee model first we must show that

$$1 - \frac{P}{C} - \frac{A}{1+BP} = 0 \tag{2.6.9}$$

has two positive solutions. It could be done directly but then the calculations become little messy so that we follow a more elegant approach of [24] and use the above equation to define a function $A(P)$ by

$$A(P) = \frac{1}{C}(C-P)(1+BP)$$

and analyse it. It is an inverted parabola satisfying $A(0) = 1$. $A(P)$ takes its maximum at the point P^*, where

$$A'(P) = -\frac{1}{C} + B - \frac{2B}{C}P = 0.$$

This gives

$$P^* = \frac{BC-1}{2B}$$

with the maximum

$$A^* = \frac{(BC+1)^2}{4BC}.$$

Now, the nonzero equilibria of (2.4.77) are the points at which the horizontal line $A = \mathit{const}$ cuts the the graph of $A(P)$, see Fig. 2.12. First, we note that if $BC < 1$, then the stationary point P^* is negative and thus there is a positive and a negative solution for $0 < A < 1$, a negative and 0 solution for $A = 1$, two negative solutions if $1 < A < A^*$, one (double) negative solution if $A = A^*$ and no solutions if $A > A^*$. If $BC = 1$, then we have one positive, one negative solution for $0 < A < 1$, double 0 solution for $A = A^* = 1$ and no solutions for $A > 1$. Thus, in none case with $BK \leq 1$ we can expect the Allee type behaviour. Let us focus then on the case $BK > 1$. Since $A > 0$, we have the following cases

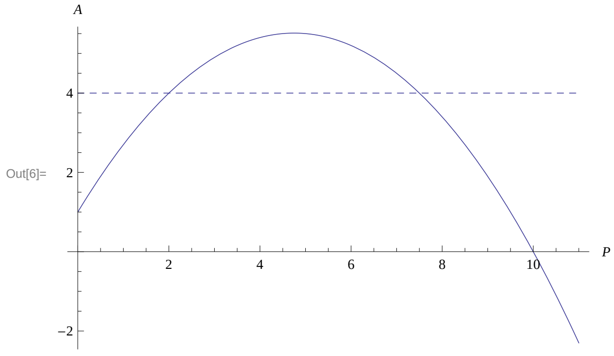

Fig. 2.12. The equilibria as a function of A.

(a) If $0 < A < 1$, then there are two solutions to (2.6.9), but only one is positive while the other is negative;

(b) If $A = 1$, then there is one 0 and one positive solution to (2.6.9);

(c) If $1 < A < A^*$, then there are two distinct positive solutions to (2.6.9);

(d) If $A = A^*$, then there is a double positive solution to (2.6.9);

(e) If $A > A^*$, then there are no solutions to (2.6.9).

To determine the stability of the equilibria, we re-write (2.4.77) in the following form

$$\begin{aligned}\frac{dP}{dt} &= \lambda P\left(1 - \frac{P}{C} - \frac{A}{1+BP}\right) = \frac{\lambda BP}{C(1+BP)}\left(-P^2 + P\frac{BC-1}{B} + \frac{C(1-A)}{B}\right)\\ &= \frac{\lambda BP}{C(1+BP)}(P-L)(K-P). \qquad (2.6.10)\end{aligned}$$

Using the results of the first part of this section and Theorem 2.6 we can describe the dynamics of (2.4.77) as follows. Let $BC > 1$. Then

(i) For $0 < A < 1$, there is one negative, L, and two nonnegative equilibria of (2.4.77), 0 and K. Zero is unstable and K is asymptotically stable;

(ii) At $A = 1$, the negative equilibrium L merges with 0. Zero becomes semi-stable (unstable for positive trajectories) and K is asymptotically stable;

(iii) For $1 < A < A^*$, there are three nonnegative equilibria, 0 and $0 < L < K$. 0 becomes a stable equilibrium, L is unstable and K is asymptotically stable negative equilibrium L merges with 0. Zero becomes semi-stable (unstable for positive trajectories) and K is asymptotically stable;

(iv) At $A = A^*$, there are two nonnegative equilibria, 0 and double $L = K$. 0 is stable and $L = K$ becomes semistable attracting trajectories from the right and repelling those from the left.

(v) For $A > A^*$, there is only one equilibrium at 0 which is globally attracting.

If $BC \leq 1$, then we cannot have two positive equilibria so that the Allee effect cannot occur in this case. However, to complete analysis, we note that if $0 < BC \leq 1$ then the only case in which there is a positive equilibrium K is for $0 < A < 1$ and in this case K is asymptotically stable while 0 is unstable. For all other cases the only biologically relevant equilibrium is 0 and it is stable if $1 < A$, semistable (attracting positive trajectories) if $A = 1$ and $BC < 1$ and stable if $A = 1 = BC$. Summarizing, (2.4.77) describes the Allee effect if and only if

$$BC > 1 \quad \text{and} \quad 1 < A < \frac{(BC+1)^2}{4BC}. \qquad (2.6.11)$$

In any other case with a positive equilibrium the dynamics described by (2.4.77) is similar to the dynamics described by the logistic equation.

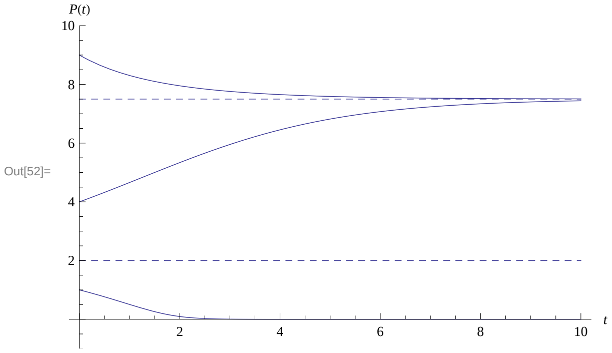

Fig. 2.13. Trajectories $P(t)$ of (2.4.77) for various initial conditions. Here $A = 4$, $C = 10$, $B = 2$, $L = 2$ (lower dashed line), $K = 7.5$ (upper dashed line).

Another way of looking at the problem is to consider the number and stability of the equilibria as a function of a parameter. This approach is known as the *bifurcation theory*. Here we focus on the case $BC > 1$ and we select the parameter A, which can be regarded as representing the extra mortality, over the mortality due to the overcrowding characteristic for the logistic model. Then, for small $A \in (0,1)$, 0 is an unstable equilibrium and K is stable, as in the logistic model. When A moves through 1, a new positive equilibrium L 'bifurcates' from 0 and the latter changes from being repelling to being attractive; K stays attractive and we are in the 'Allee region'. Finally, when A moves across A^*, K vanishes and 0 becomes globally attractive – large mortality drives the population to extinction. The Allee phenomenon is of concern in many practical applications. For instance, if we try to eradicate a pest whose population can be modelled by an Allee type equation, then it is enough to create conditions if which the size of the population will be below L; the population will then die out without any external intervention. Similarly, if by overhunting or overfishing we drive a population below L, then it will become extinct even if we stop its exploitation.

6.3 Equilibrium points of difference equations

Consider the autonomous first order difference equation

$$x(n+1) = f(x(n)), \qquad n \in \mathbb{N}_0, \tag{2.6.12}$$

with the initial condition x_0. In what follows we shall assume that f is at least continuous. It is clear that the solution to (2.6.12) is given by iterations

$$x(n) = f(f(\dots f(x_0))) = f^n(x_0) \tag{2.6.13}$$

and henceforth we will be using both notations.

A point x^* in the domain of f is said to be an *equilibrium point* of (2.6.2) if it is a fixed point of f, that is, if $f(x^*) = x^*$. In other words, the constant sequence $(x^*, x^*, \dots)$ is a stationary solution of (2.6.2). As in the case of differential equations, here also we shall not differentiate between these concepts.

Example 2.10. Consider the logistic equation

$$x(n+1) = 3x(n)(1 - x(n)). \tag{2.6.14}$$

The equation for the equilibrium points is $x = 3x(1-x)$, which gives $x_0 = 0$ and $x_1 = 2/3$. Clearly, if $x_0 = 0$, then $x_n = 0$ for any $n \in \mathbb{N}$. Similarly, if $x_0 = 2/3$, then $x_1 = 3 \cdot (2/3) \cdot (1 - 2/3) = 2/3$ and, by iteration, $x_n = 2/3$ for any $n \in \mathbb{N}$.

Graphically, an equilibrium is the x-coordinate of a point, where the graph of f intersects the diagonal $y = x$. This is the basis of the cobweb method of finding and analysing equilibria, described in the next subsection.

Definition 2.11. *1. 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 $|f^n(x) - x^*| \geq \epsilon$.)*

2. A point x^* is called attracting if there is $\eta > 0$ such that $|x_0 - x^*| < \eta$ implies $\lim\limits_{n\to\infty} f^n(x_0) = x^*$. If $\eta = \infty$, then x^* is called a global attractor or globally attracting.

3. The point x^* is called an asymptotically stable equilibrium if it is stable and attracting. If $\eta = \infty$, then x^* is said to be a globally asymptotically stable equilibrium.

Remark 2.12. We note an important difference between discrete and continuous time systems. In Lemma 2.5 we showed that no non-equilibrium solution to a differential equation with Lipschitz right hand side can reach the equilibrium in finite time. On the other hand, such behaviour is possible for solutions to difference equations, see Example 2.17. The points, which lie on a trajectory which reaches an equilibrium in finite time are called *eventual equilibria*.

6.4 The cobweb diagrams

We describe an important graphical method for analysing the stability of equilibrium (and periodic) points of (2.6.2). 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 $x(0) = x_0$, we pinpoint the value $x(1)$ by drawing a vertical line through $x(0)$ so that it also intersects the graph of f at $(x(0), x(1))$. Next, we 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 any $x(n)$. This is illustrated in Fig. 2.14, where we presented several steps of drawing the cobweb diagram for the logistic equation (2.6.14) 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 for analysing stability.

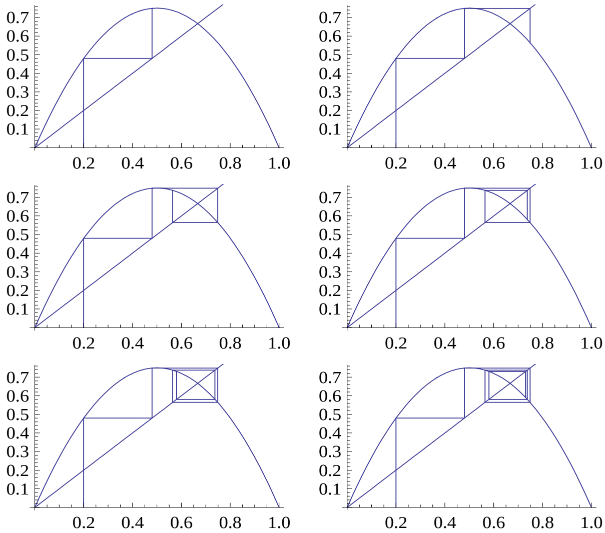

Fig. 2.14. Cobweb diagram of a logistic difference equation

6.5 Analytic criterion for stability

Theorem 2.13. *Let x^* be an isolated equilibrium point of the difference equation*

$$x(n+1) = f(x(n)), \tag{2.6.15}$$

where f is continuously differentiable in some neighbourhood of 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 functions, [7]. Let $x_0 \in J$. We have

$$|x(1) - x^*| = |f(x_0) - f(x^*)|$$

and, by the mean value theorem, for some $\xi \in [x_0, x^*]$,

$$|f(x_0) - f(x^*)| = |f'(\xi)||x_0 - x^*|.$$

Hence, $|x(1) - x^*| = |f(x_0) - f(x^*)| \leq M|x_0 - x^*|$. Since $M < 1$, the inequality 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 $\delta = \epsilon$. Then $|x(n) - x^*| < \epsilon$ for $n > 0$ provided $|x_0 - x^*| < \delta$ (since $M < 1$). Furthermore $x(n) \to x^*$ and $n \to \infty$ 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 $J = (x^* - \epsilon, x^* + \epsilon)$ we have $|f'(x)| \geq M > 1$. Take an arbitrary $\delta > 0$ smaller than ϵ and choose x satisfying $|x - x^*| < \delta$. Again using the mean value theorem, we get $|f(x) - x^*| = |f'(\xi)||x - x^*|$ 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, $f^2(x)$ is further away from x^* than $f(x)$. If $f^2(x)$ is still in J, we continue the procedure till $|f^n(x) - x^*| \geq M^n|x - x^*| > \epsilon$ for some n. □

Equilibrium x^* with $|f'(x^*)| \neq 1$ is called *hyperbolic.*

What happens if the equilibrium point x^* is not hyperbolic? To simplify the considerations, we assume that f is at least three times continuously differentiable in a neighbourhood of x^*. First, let us reflect on the geometry of the situation. 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, if 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, while 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 f is convex (or concave) close to x^* and the graph of f will be (locally) either entirely above or entirely below the line $y = x$. Therefore the picture is 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 $y = 0$. 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 $f'''(x^*) < 0$, while the latter if $f'''(x^*)$. Summarizing, we have:

Theorem 2.14. *Let x^* be an isolated equilibrium with $f'(x^*) = 1$ and let f be at least three times continuously differentiable in a neighbourhood of x^*. 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 $g(x) = -x + 2x^*$, that is, if g is a linear function giving an equilibrium at $x = x^*$ with $f'(x^*) = -1$, then the iterations starting from $x_0 \neq x^*$ produce a solution taking on only two values oscillating around x^*. Thus, if $-1 < f'(x^*) < 0$, then f passes from below the line $y = -x+2x^*$ to above as x increases. Hence, the stability follows from the fact that subsequent iterations oscillate around x^* getting closer to x^* with each iteration. If, on the contrary, $f'(x^*) < -1$, then 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 $y = -x + 2x^*$, only produces a 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 also an equilibrium of $g(x) := f(f(x))$ and it is a stable equilibrium of f if and only if it is stable for g. This statement follows from the 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. Since $g'(x) = f'(f(x))f'(x)$ with $g'(x^*) = 1$, we can apply Theorem 2.14 to the function g. The second derivative of g is given by

$$g''(x) = f''(f(x))[f'(x)]^2 + f'(f(x))f''(x)$$

and, since $f(x^*) = x^*$ and $f'(x^*) = -1$, we have $g''(x^*) = 0$. Using again the chain rule, we find

$$g'''(x^*) = -2f'''(x^*) - 3[f''(x^*)]^2.$$

Hence, we can write

Theorem 2.15. *Suppose that at an equilibrium point x^* we have $f'(x^*) = -1$. Define $S(x^*) = -f'''(x^*) - 3(f''(x^*))^2/2$. Then x^* is asymptotically stable if $S(x^*) < 0$ and unstable if $S(x^*) > 0$.*

Example 2.16. 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 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.

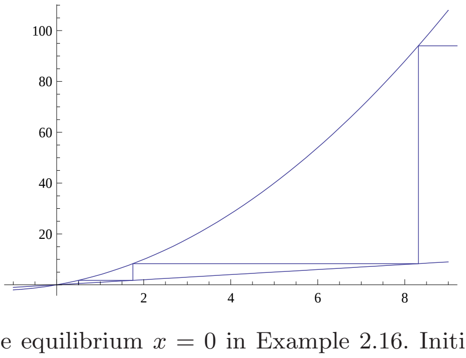

Fig. 2.15. Unstable character of the equilibrium $x = 0$ in Example 2.16. Initial point is $x_0 = 0.5$ and we observe divergence to the right.

Example 2.17. Consider the equation $x(n+1) = Tx(n)$, where

$$T(x) = \begin{cases} 2x & \text{for } 0 \leq x \leq 1/2, \\ 2(1-x) & \text{for } 1/2 < x \leq 1. \end{cases}$$

is the so-called tent map. here 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.

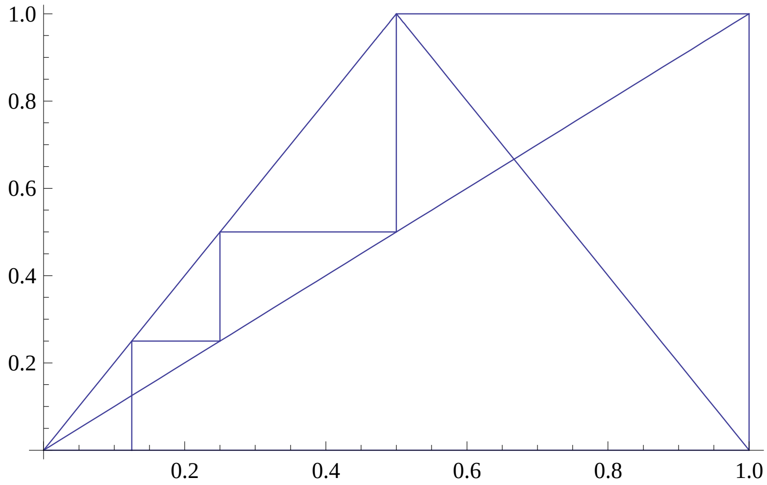

Fig. 2.16. Eventual equilibrium $x = 1/8$ for the tent map.

Remark 2.18. We can provide a fine-tuning of the notion of stability by noting that if $f'(x^*) < 0$, then the solution behaves in an oscillatory way around x^* and if $f'(x^*) > 0$, then it is monotonic. Indeed, consider (in a neighbourhood of x^* where $f'(x) < 0$) $f(x) - f(x^*) = f(x) - x^* = f'(\xi)(x - x^*)$, $\xi \in (x^*, x)$. Since $f' < 0$, $f(x) > x^*$ if $x < x^*$ and $f(x) < x^*$ if $x > x^*$, hence each iteration moves the point to the other side of x^*. If $|f'| < 1$ over this interval, then $f^n(x)$ converges 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 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.

What happens if $f'(x^*) = 0$? Clearly, if this is a local extremum, then in some neighbourhood the derivative will have a fixed sign and thus the behaviour of the iterates will be as above. Let, on the other hand, $f'(x) < 0$ is some one-sided neighbourhood of x^* and $f'(x) > 0$ on the other side. Then if the iterates start in the latter, they will stay there, converging to x^*, while if the iterates start in the former, they will begin converging to x^* in an oscillatory way until they reach the neighbourhood in which the derivative is positive and then they will converge monotonically.

6.6 Some applications

In this subsection we shall present applications of the above considerations to the problem of sustainable fishing, the biological pest control and the discrete Allee model.

Sustainable fishing

Let us consider a population of fish living in a pond, which grows according to the logistic equation (2.2.18),

$$N(k+1) = N(k) + rN(k)\left(1 - \frac{N(k)}{K}\right).$$

This equation only can be solved in some particular cases, see Section 3.4. 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 for the community. 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 logistic equation, we find

$$H(k) = rN(k)\left(1 - N(k)/K\right).$$

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 attains the absolute maximum. We note that the right hand side is a quadratic function, $f(N) = rN\left(1 - N/K\right)$, 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 harvest is

$$H = rK/4. \tag{2.6.16}$$

The considered model does not include the actual fishing – it simply says that to sustain a population we cannot harvest more fish than there are borne and indicates the size of the population giving the maximum yield but it does not specify how to organize the fishing. We can generalize the model to

$$N(k+1) = N(k) + rN(k)\left(1 - \frac{N(k)}{K}\right) - qEN(k), \tag{2.6.17}$$

where E is the so-called fishing effort, for instance the number of fishing boats at sea, and q is the fishing efficiency, that is, the fraction of the population caught by one boat in the unit time. The concept of sustainable fishing is the same as above. We should find find the amount of fish that can be caught during a year to maintain the population at a constant level and hence find the population for which the yield is optimal.

To find the possible constant population level for a given level of fishing we solve

$$N = N + rN\left(1 - \frac{N}{K}\right) - qEN \tag{2.6.18}$$

This gives $N = 0$ or

$$N^* = K\left(1 - \frac{qE}{r}\right). \tag{2.6.19}$$

The first solution is trivial and not interesting. The second solution is positive if

$$qE < r.$$

This is consistent with the first part by showing that the fishing rate cannot exceed the unrestricted net growth of the population. However, this model can furnish further information. Let us suppose that for the given fishing rate qE the population is kept at N^* given by (2.6.19). Then the yield is $Y(qE) = qEN^*$ or

$$Y(qE) = qEK\left(1 - \frac{qE}{r}\right).$$

Thus the yield is a quadratic function of the fishing rate qE and and thus its maximum can be found as in the first part of the section. Maximum of $Y(qE)$ is attained at $E = \dfrac{r}{2q}$ which gives the maximum sustainable yield as

$$Y_{max}(qE) = \frac{rK}{4},$$

as in (2.6.16). Clearly, if we increase the fishing effort, $E > r/2q$, then the yield will decrease (greed does not pay!). To understand why, let us summarize the mechanism described by the model. For a given fishing rate qE, we find that the equilibrium N^*, given by (2.6.19), is asymptotically stable provided

$$\left| \frac{d}{dN} \left(N + rN \left(1 - \frac{N}{K} \right) - qEN \right) \Bigg|_{N=N^*} \right| = \left| 1 + r - \frac{2rN^*}{K} - qE \right| < 1.$$

Using (2.6.19), this gives

$$-1 < 1 - r + qE < 1$$

or

$$r - 2 < qE < r.$$

Thus, if the fishing rate satisfies the above condition, then the fish population eventually will stabilize at N^* and N^* decreases if the fishing rate increases. The yield is the product of the fishing rate and the size of the population. Thus, if the fishing rate is too low, then though N^* is large, the product is not optimal. Similarly, overfishing results in the population settling at a lower N^* again resulting in a suboptimal yield.

Remark 2.19. One may wonder what is the significance of the condition $r - 2 < qE$ which puts an upper bound on r. A common sense would suggest that the higher net growth rate, the better the yield. To explain this, we observe that then (2.6.18) can be written as

$$N(k+1) = (1 + r - qE)\, N(k) \left(1 - \frac{N}{\frac{K(1+r-qE)}{r}} \right).$$

The results of Section 6.8 show that the positive equilibrium of a discrete logistic equation is unstable if the unrestricted growth rate is larger than 3 which in our case corresponds to

$$1 + r - qE > 3 \quad \text{iff} \quad r - 2 > qE.$$

Thus, the case $r > qE + 2$ does not give a stable fish population ensuring the sustainable constant yield.

Biological pest control

Assume that we have to deal with an insect population which invades our plantation. The insects reproduce according to the Beverton-Holt model (2.2.12),

$$P(k+1) = \frac{\beta P(k)}{1 + aP(k)} \tag{2.6.20}$$

where β is the natural fertility of insects and $a = (\beta - 1)/K$ where K is the capacity of the environment. As we know, see Section 3.4, if $\beta > 1$, then the population $P(k)$ tends a nonzero equilibrium $K = (\beta - 1)/a$; that is, there is a nonzero stable population of insects. An ecological way of eradicating the pest is to decrease the birth rate and one of the methods is to introduce a number of sterile insects into the population. We assume that S is under our control and we can keep the number of the sterile insects constant in time. Though suppressed in the model, insects reproduce sexually and thus the effective birth rate depends on the probability of finding a mate. If, say, S individuals are sterile, and $P(k)$ is the number of fertile insects, then the probability of picking a fertile insect is $P(k)/(P(k) + S)$. Thus, the Beverton-Holt model can be modified as

$$P(k+1) = \beta P(k) \frac{P(k)}{S + P(k)} \frac{1}{1 + aP(k)} = \beta P(k) f(P(k)). \tag{2.6.21}$$

To find the equilibria of (2.6.21), we solve

$$P = \beta P \frac{P}{S+P} \frac{1}{1+aP}$$

which gives $P_1 = 0$ and the simplified equation

$$1 = \beta \frac{P}{S+P} \frac{1}{1+aP}.$$

While the above equation can be solved for P^*, a faster way to find the answer is the approach used in Section 6.2 to analyse the continuous Allee model. Thus, we solve the above equation for S as a function of P^* getting

$$S(P) = \frac{(\beta - 1 - aP)P}{1+aP}. \tag{2.6.22}$$

We find that $S(0) = S((\beta-1)/a) = 0$ and the derivative is give by $S'(P) = -1 + \beta/(1+aP)^2$. Hence, the maximum in the interval $(0, (\beta-1)/a)$ is attained at $P = (\sqrt{\beta}-1)/a$; the maximum is $S_{\max} = (\sqrt{\beta}-1)^2/a$. Thus, we conclude that if $S > S_{max}$, the only equilibrium of (2.6.21) is $P_1 = 0$.

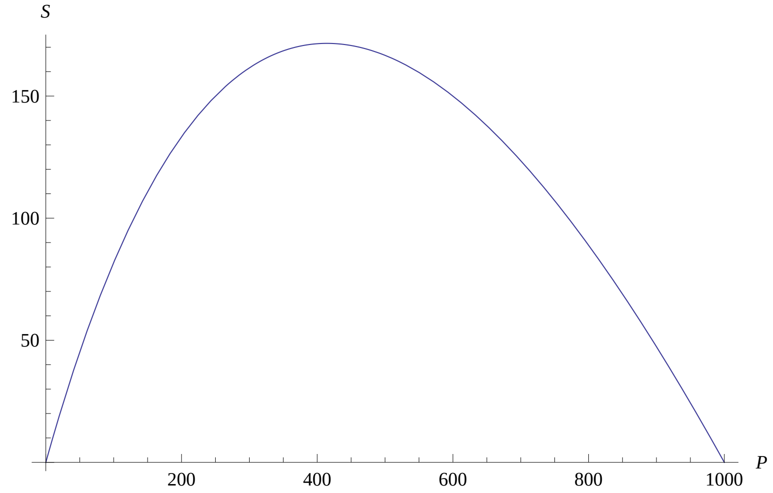

Fig. 2.17. The graph of $S(P)$, given by (2.6.22),for $\beta = 2$ and $a = 0.001$.

$$\beta \frac{d}{dP} Pf(P) = \beta \frac{P^2(1+aS) + 2PS}{((P+S)(1+aP))^2}.$$

Hence, $\beta \frac{d}{dP} Pf(P)|_{P=0} = 0$ and the equilibrium $P_1 = 0$ is asymptotically stable, for any S. It is laborious to provide explicit estimates of the derivatives of $\beta Pf(P)$ at the other two equilibria that exists for $0 < S < S_{\max}$ but still we can determine their stability by using the discussion preceding Theorem 2.14. First we notice that $\beta \frac{d}{dP} Pf(P) > 0$ for all $P > 0$ (provided $S > 0$). This means that the derivative never equals -1. Let $0 < S < S_{\max}$. The curve $\beta Pf(P)$ starts at zero below the diagonal and thus at the smaller equilibrium it must cross the diagonal from below. Hence, this equilibrium is unstable. At the second equilibrium, the curve crosses the diagonal from above and, since the curve is ascending, the derivative is between 0 and 1. Thus this equilibrium is asymptotically stable. If $S = S_{\max}$, we have a tangent point which means that the (unique) positive equilibrium is semi-stable - unstable from the left and stable from the right.

Summarizing, to eradicate the pest population we should introduce the number $S > S_{\max}$ of sterile insects. Then the population will converge to the extinction equilibrium provided we the number S of sterile insects is kept above $S_{\max}$ at each cycle (this may require our intervention as the insects die of natural causes). Otherwise, we need to drive the population of pest below the smaller equilibrium – then the population will also converge to the extinction equilibrium.

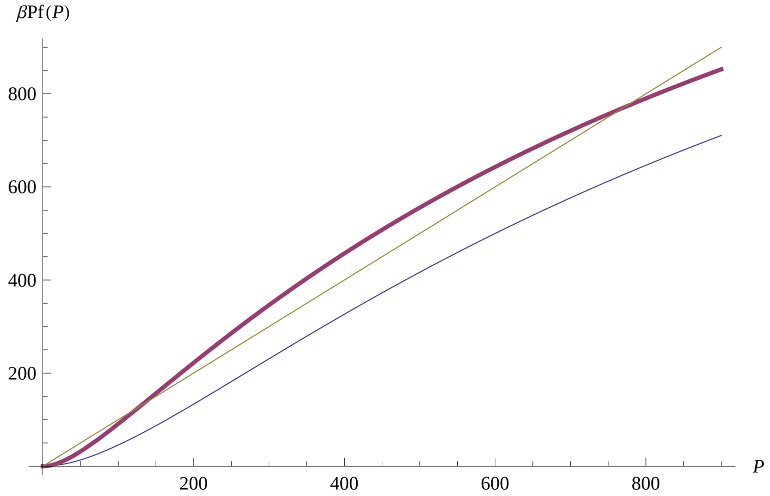

Fig. 2.18. The graphs of $\beta P f(P)$, as in (2.6.21),for $\beta = 2, a = 0.001$ with $S = 300$ (thin line) and $S = 100$ (thick line).

The discrete Allee model (2.3.42)

Let us recall the model (2.3.42)

$$N(k+1) = N(k)\left(1+\lambda\left(1-\frac{N(k)}{C}-\frac{A}{1+BN(k)}\right)\right).$$

The equilibria are determined by solving

$$N = N\left(1+\lambda\left(1-\frac{N}{C}-\frac{A}{1+BN}\right)\right)$$

and we see that the equilibria are $0, L, K$ for appropriate parameters C, A, B, exactly as in the continuous case, see Subsection 6.2 and Eqn. (2.6.7). To fix attention, we focus on the case with two positive equilibria $0 < L < K$. In contrast to (2.4.77), it is not obvious that (2.3.42) describes a population, that is, whether the solution is nonnegative for nonnegative initial condition. Here we shall prove that the interval $[0, K]$ is invariant under

$$g(N) := N\left(1+\lambda\left(1-\frac{N}{C}-\frac{A}{1+BN}\right)\right).$$

First we look at nonnegativity. We can write

$$N(k+1) - N(k) = \lambda f(N(k))$$

where f is defined in (2.6.7) and thus $N(k+1) > N(k)$ is increasing as long as $L < N(k) < K$. Hence, $N(k+1)$ is nonnegative as long as $N(k) \in (L, K)$. However, $N(k+1) < N(k)$ if $0 < N(k) < L$ and it could become negative. To avoid this, it suffices to ensure that

$$N(k) + \lambda f(N(K)) > 0$$

or

$$\lambda h(N) := \lambda\left(1-\frac{N}{C}-\frac{A}{1+BN}\right) > -1, \quad 0 < N < L.$$

Since

$$h'(N) = -\frac{1}{C} + \frac{AB}{(1+BN)^2}$$

and, in our case $A > 1, BC > 1$, we see that h has only one stationary point $\bar{N}$. Since $h(0) = 1 - A < 0$ and $h(L) = h(K) = 0$ for $0 < L < K$, we see that $L < \bar{N} < K$ and thus h is increasing on $[0, \bar{N})$. Hence, $h > -1$ on $(0, L)$ if and only if $h(0) \geq -1$ which translates into

$$A \leq \frac{1+\lambda}{\lambda}. \tag{2.6.23}$$

The stability of the equilibria 0 and L can be determined using the continuous case. Indeed, writing (2.3.42) as

$$N(k+1) = g(N(k))$$

we see that

$$g(N) = N + \lambda f(N) = N + \lambda N h(N)$$

and thus

$$g'(N) = 1 + \lambda f'(N) = 1 + \lambda h(N) + \lambda N h'(N).$$

Hence

$$g'(0) = 1 + \lambda h(0) = 1 + \lambda(1 - A)$$

which gives stability provided

$$A < 1 + \frac{2}{\lambda}.$$

However, monotonic stability coincides with the assumption (2.6.23) – otherwise we would have oscillatory convergence and consequently negative solutions. Further, since $h'(L) > 0$ by the considerations above, we immediately obtain instability of L.

To show that $N(k+1) < K$ provided $0 < N(k) < K$ and also to determine the character of stability of K requires some more work. First, observe that

$$h(N) < 1 - \frac{N}{C} =: \psi(N).$$

Since ψ is a monotonically decreasing function, having $N = C$ as the only solution, we obtain

$$L < K < C. \tag{2.6.24}$$

Indeed, $0 = h(K) < \psi(K)$ and (2.6.24) follows from monotonicity of ψ.

Next, using (2.6.10) we can write

$$\begin{aligned} g'(N) &= 1 + \lambda f'(N) = 1 + \frac{d}{dP}\left(\frac{\lambda B N}{C(1+BN)}(N-L)(K-N)\right) \\ &= 1 + \phi'(N)(N-L)(K-N) + \phi(N)(-2N + L + K), \end{aligned} \tag{2.6.25}$$

where $\phi(N) = \lambda BN/C(1+BN)$. It is easy to see that ϕ is a strictly increasing function on $[0, \infty)$ satisfying $0 \leq \phi(N) < \lambda/C$. Thus

$$g'(K) = 1 + \frac{\lambda B K}{C(1+BK)}(L - K)$$

and we obtain that $g'(K) \geq 0$ provided $\lambda(L-K)/C \geq -1$ or

$$\lambda \leq \frac{C}{K-L}. \tag{2.6.26}$$

We note that $C/(K-L) > 1$ by (2.6.24), thus the Allee effect can occur in populations with positive net growth rate. If we look closer at (2.6.25), we note that the second term in the last equality is positive for $L < N < K$ and

$$\phi(N)(-2N + L + K) > \phi(K)(L - K),$$

on this interval. Indeed, $-2N + L + K > 0$ for $L < N < (K+L)/2$ so on this interval the equality is obvious. For $(K+L)/2 \leq N \leq K$ we have

$$\phi(N)(-2N+L+K) > \phi(N)(L-K) > \phi(K)(L-K),$$

on account of monotonicity and signs of the involved functions.

Hence, provided (2.6.26) holds,

$$g'(N) \geq 1+\phi(K)(L-K) \geq 0, \quad L \leq N \leq K$$

and thus

$$g(N) \leq g(K) = K$$

for any $N \in [L, K]$. Since we proved earlier that g transforms $[0, L]$ into $[0, L]$, we see that if (2.6.11), (2.6.23) and (2.6.26) are satisfied, then $g(N) \in [0, K]$ for $N \in [0, K]$ as required.

An example of parameters satisfying these conditions is provided in Example 2.2.

6.7 Periodic points and cycles

Theorem 2.6 tells us that a solution to a scalar autonomous differential equation must be monotonic. Above we have already seen that solutions to scalar autonomous difference equations can be oscillatory. In fact, such equations may admit periodic solutions which cannot occur in the continuous case.

Definition 2.20. *Let b be a point 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, $O(b) = \{b, f(b), f^2(b), \ldots, f^{k-1}(b)\}$ 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 2.21. Consider $x(n+1) = T^2x(n)$, where T is the tent map introduced in Exercise 2.17. Then

$$T^2(x) = \begin{cases} 4x & \text{for } 0 \leq x \leq 1/4, \\ 2(1-2x) & \text{for } 1/4 < x \leq 1/2, \\ 4x-2 & \text{for } 1/2 < x \leq 3/4, \\ 4(1-x) & \text{for } 3/4 < x \leq 1. \end{cases}$$

There are four equilibrium points, $0, 2/5, 2/3$ and $4/5$, two of which are equilibria of T. Hence $\{2/5, 4/5\}$ is

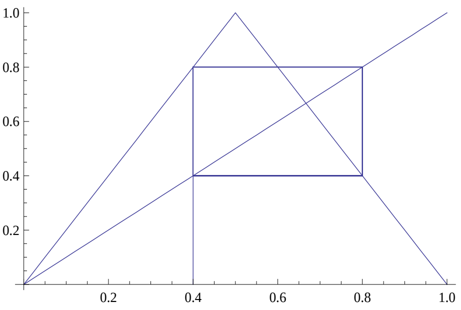

Fig. 2.19. 2-cycle for the tent map

the only 2-cycle of T.

Definition 2.22. *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.

This definition, together with Theorem 2.13, yield the following classification of the stability of k-cycles.

Theorem 2.23. *Let $O(b) = \{x_0 = b, x(1) = f(b), \ldots, x(k-1) = f^{k-1}(b)\}$ be a k-cycle of a continuously differentiable function f. Then*

(i) The k-cycle $O(b)$ is asymptotically stable if

$$|f'(x_0)f'(x(1))\ldots f'(x(k-1))| < 1.$$

(ii) The k-cycle $O(b)$ is unstable if

$$|f'(x_0)f'(x(1))\ldots f'(x(k-1))| > 1.$$

Proof. Follows from Theorem 2.13 by the Chain Rule applied to f^k. □

It follows that if b is k-periodic, then every point of its k-cycle $\{x(0) = b, x(1) = f(b), \ldots, x(k-1) = f^{k-1}(b)\}$ is also k-periodic. This follows from $f^k(f^r(b)) = f^r(f^k(b)) = f^r(b)$, $r = 0, 1, \ldots, k-1$.

Note that the stability of b means that $|f^{nk}(x) - b| < \epsilon$ for all n, provided x is close enough to b. In other words, we are ensured only that the iterates of $f^k(x)$ will stay close to b. However, from continuity of f it also follows that for $r = 1, \ldots, k-1$, $f^{nk+r}(x)$ will stay close to $f^{nk+r}(b) = f^r(b)$ if x is close enough to b.

Actually, a stronger result is valid.

Proposition 2.24. *Each such point of the k-cycle $O(b) = \{x(0) = b, x(1) = f(b), \ldots, x(k-1) = f^{k-1}(b)\}$ possesses the same stability property as b.*

Proof. The proof would be easy if f was continuously invertible. Indeed, then a preimage of any small neighbourhood of $f^r(b)$ would be a small neighbourhood of b and starting the iterates from a small neighbourhood of $f^r(b)$ would be equivalent to starting from a small neighbourhood of b. . In general, it is not possible for non-invertible functions (think about $f(x) = \sin x$ and preimage $\{x \in \mathbb{R};\ |\sin x| < \epsilon\}$ – it is unbounded). Assume that b is stable and fix $0 < r < k$. Stability of b and continuity of f^r implies that for any ϵ we can find δ so that

$$|f^r f^{(n-1)k}(y) - f^r f^{(n-1)k}(b)| = |f^{(n-1)k} f^r(y) - f^{(n-1)k} f^r(b)| = |f^{(n-1)k} f^r(y) - f^{nk} f^r(b)| < \epsilon,$$

provided $|y - b| < \delta$. Next, we can find $\delta_1 > 0$ such that from $|x - f^r(b)| < \delta_1$ it follows that

$$|f^{k-r}(x) - f^{k-r} f^r(b)| = |f^{k-r}(x) - f^k(b)| = |f^{k-r}(x) - b| < \delta.$$

Taking now $y = f^{k-r}(x)$, we have $|y - b| < \delta$, hence

$$|f^{(n-1)k} f^k(x) - f^{nk} f^r(b)| = |f^{nk}(x) - f^{nk} f^r(b)| < \epsilon,$$

provided $|x - f^r(b)| < \delta_1$, which yields stability of $f^r(b)$. Asymptotic stability can be proved in a similar way. We know that we can find δ_1 such that $|x - f^r(b)| < \delta_1$ implies $f^{kn} f^{k-r}(x) \to b$ for $n \to \infty$. The continuity of f yields

$$f^r f^{kn} f^{k-r}(x) = f^{k(n+1)}(x) \to f^r(b)$$

but this is the same as $f^{kn}(x) \to f^r(b)$. Hence $f^r(b)$ is asymptotically stable.

Finally, let b be be unstable. If $f^r(b)$ was stable, then $b = f^{k-r} f^r(b)$ would be stable, contradicting the first part of the proof. □

6.8 Dynamics of the logistic equation

Consider the logistic equation

$$x(n+1) = F_\gamma(x(n)) := \gamma x(n)(1-x(n)), \quad x \in [0,1],\ \gamma > 0. \tag{2.6.27}$$

Our aim is to investigate the properties of equilibria of (2.6.27) with respect to the parameter γ. The values of γ, for which there is a qualitative change of the properties of the equilibria are called *bifurcation points*.

To find the equilibrium points, we solve $F_\gamma(x^*) = x^*$ which gives $x^* = 0, (\gamma-1)/\gamma$.

We investigate the stability of each point separately.

(a) For $x^* = 0$, we have $F'_\gamma(0) = \gamma$ and thus $x^* = 0$ is asymptotically stable for $0 < \gamma < 1$ and unstable for $\gamma > 1$. To investigate the stability for $\gamma = 1$, we find $F''_\gamma(0) = -2\gamma \neq 0$ and thus $x^* = 0$ is unstable in this case. However, the instability comes from the negative 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 $x^* = (\gamma-1)/\gamma$ belongs to the domain $[0,1]$ only if $\gamma > 1$. Here, $F'_\gamma((\gamma-1)/\gamma) = 2-\gamma$ and $F''_\gamma((\gamma-1)/\gamma) = -2\gamma$. Thus, using Theorems 2.13 and 2.14, we obtain that x^* is asymptotically stable if $1 < \gamma \leq 3$ and it is unstable if $\gamma > 3$.

Further, by Remark 2.18, we observe that for $1 < \gamma < 2$, the population approaches the carrying capacity monotonically from below. However, for $2 < \gamma \leq 3$ the population can go over the carrying capacity but it will eventually stabilize around it. What happens for $\gamma = 3$? Consider 2-cycles. We have $F^2_\gamma(x) = \gamma^2 x(1-x)(1-\gamma x(1-x))$ so that we are looking for solutions to $\gamma^2 x(1-x)(1-\gamma x(1-x)) = x$, which can be re-written as

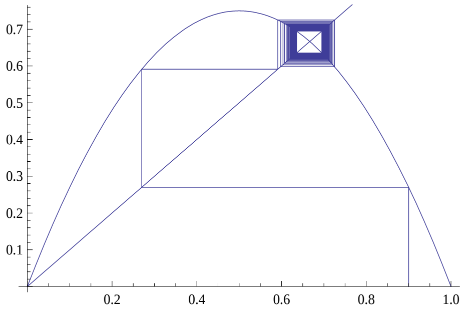

Fig. 2.20. Asymptotically stable equilibrium $x = 2/3$ for $\gamma = 3$.

$$x(\gamma^3x^3 - 2\gamma^3x^2 + \gamma^2(1+\gamma)x + (1-\gamma^2)) = 0.$$

To simplify, 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 $x - (\gamma-1)/\gamma$, getting

$$\gamma^2x^2 - \gamma(\gamma+1)x + \gamma + 1 = 0.$$

Solving this quadratic equation, we obtain a 2-cycle

$$x_\pm = \frac{(1+\gamma) \pm \sqrt{(\gamma-3)(\gamma+1)}}{2\gamma}. \tag{2.6.28}$$

Clearly, these points determine a 2-cycle provided $\gamma > 3$ (in fact, for $\gamma = 3$, these two points collapse into the equilibrium point $x^* = 2/3$. Thus, we see that when the parameter γ passes through $\gamma = 3$, the stable equilibrium becomes unstable and bifurcates into two 2-cycles.

The stability of 2-cycles can be determined by Theorem 2.23. We have $F'_\gamma(x) = \gamma(1-2x)$ so the 2-cycle is stable, provided

$$-1 < \gamma^2(1-2x_+)(1-2x_-)) < 1.$$

Using Viète's formulae we find that the above is satisfied provided $-1 < -\gamma^2+2\gamma+4 < 1$ and, upon solving this inequality, we get $\gamma < -1$ or $\gamma > 3$ and $1-\sqrt{6} < \gamma < 1+\sqrt{6}$ which yields $3 < \gamma < 1+\sqrt{6}$.

In a similar fashion, we find that for $\gamma_1 = 1+\sqrt{6}$, the 2-cycle is still attracting but becomes unstable for $\gamma > \gamma_1$.

To find 4-cycles, we solve $F^4_\gamma(x) = 0$. However, in this case the algebra becomes intractable and one should resort to a computer. It turns out that there is a 4-cycle, when $\gamma > 1+\sqrt{6}$, which is attracting for $1+\sqrt{6} < \gamma < 3.544090\ldots =: \gamma_2$. When $\gamma = \gamma_2$, the 2^2-cycle bifurcates into a 2^3-cycle which is stable for $\gamma_2 \leq \gamma \leq \gamma_3 := 3.564407\ldots$. Continuing, we obtain a sequence of numbers (γ_n), such that the 2^n-cycle bifurcates into a 2^{n+1}-cycle passing through γ_n. In this particular case, $\lim_{n\to\infty}\gamma_n = \gamma_\infty = 3.57\ldots$. A remarkable observation, made by Feigenbaum, is that for any sufficiently smooth family F_γ of mappings of an interval into itself, the number

$$\delta = \lim_{n\to\infty} \frac{\gamma_n - \gamma_{n-1}}{\gamma_{n+1}-\gamma_n} = 4.6692016\ldots,$$

in general does not depend on the family of maps, provided they have single maximum. The interested reader should consult e.g., [11] for further information on dynamics of the logistic map.

The Feigenbaum's result expresses the fact that the picture obtained for the logistic equation is to a large extent universal. What happens for γ_∞? Here we find a densely interwoven region with both periodic and very irregular orbits. In particular, a 3-cycle appears and, by the celebrated theorem by Šarkovsky, see e.g. [13], this implies the existence of orbits of any period. In fact, what we observe is the emergence of the so-called chaotic dynamics. The discussion of this topic is beyond the scope of this book and the reader is referred to e.g., [11] (discrete dynamics) and [13, 22] (continuous dynamics) for a more detailed account of it.

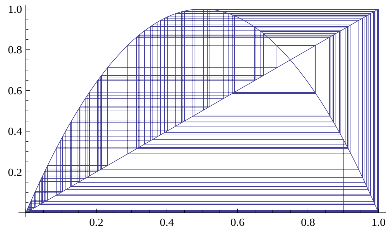

Fig. 2.21. Chaotic orbit for $x = 0.9$ and $\gamma = 4$.

6.9 Stability in the Beverton-Holt equation

We conclude with a brief description of the stability of equilibrium points for the Beverton-Holt equation which was discussed in Subsections 2.4 and 3.4. Let us recall this equation

$$x(n+1) = f(x(n),\beta,b) = \frac{\beta x(n)}{(1+x(n))^b}.$$

Writing $x^*(1+x^*)^b = \beta x^*$, we find a steady state $x^* = 0$ and we observe that if $\beta \leq 1$, then this is the only steady state (at least for positive values of x). If $\beta > 1$, then there is another steady state given by $x^* = \beta^{1/b} - 1$. Evaluating the derivative at x^*, we have

$$f'(x^*,\beta,b) = \frac{\beta}{(1+x^*)^b} - \frac{\beta b x^*}{(1+x^*)^{b+1}} = 1-b+\frac{b}{\beta^{1/b}}.$$

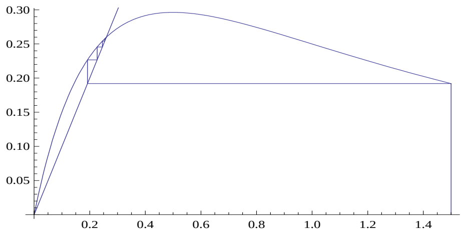

Fig. 2.22. Monotonic stability of the equilibrium for the Beverton-Holt model with $b = 3$ and $\beta = 2$; see Eqn (2.6.29).

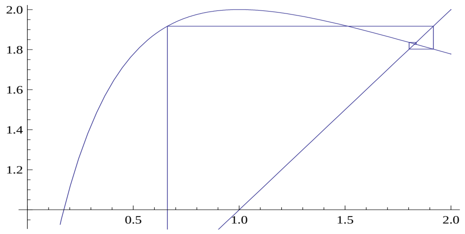

Fig. 2.23. Oscillatory stability of the equilibrium for the Beverton-Holt model with $b = 2$ and $\beta = 8$; see Eqn (2.6.29).

Clearly, with $\beta > 1$, we always have $f' < 1$. Hence, for monotone stability we must have $1 - b + b\beta^{-1/b} > 0$, whereas oscillatory stability requires $-1 < 1 - b + b\beta^{-1/b} < 0$. Solving these inequalities, we obtain that the borderlines between different types of behaviour are given by

$$\beta = \left(\frac{b}{b-1}\right)^b \text{ and } \beta = \left(\frac{b}{b-2}\right)^b. \tag{2.6.29}$$

Let us consider the existence of 2-cycles. The second iteration of the map $H(x) = \beta x/(1+x)^b$ is given by

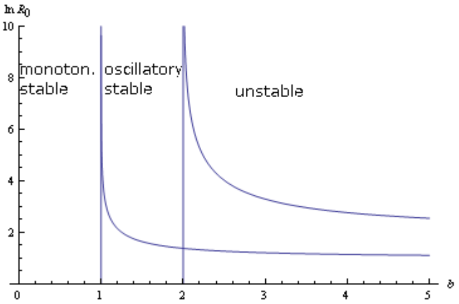

Fig. 2.24. Regions of stability of the Beverton-Holt model described by (2.6.29)

$$H(H(x)) = \frac{\beta^2 x(1+x)^{b^2-b}}{((1+x)^b + \beta x)^b},$$

so that 2-cycles can be obtained by solving $H(H(x)) = x$. This can be rewritten as

$$x\beta^2(1+x)^{b^2-b} = x((1+x)^b + \beta x)^b,$$

or, discarding $x = 0$ and taking the bth root,

$$(1+x)^{b-1}\beta^{\frac{2}{b}} = (1+x)^b + \beta x.$$

Introducing the change of variables $z = 1 + x$, we see that we have to investigate the existence of positive roots of

$$f(z) = z^b - z^{b-1}\beta^{\frac{2}{b}} + \beta z - \beta.$$

Clearly, we have $f(\beta^{\frac{1}{b}}) = 0$, since any equilibrium of H is also an equilibrium of H^2. First, let us consider $1 < b \le 2$ (the case $b = 1$ yields an explicit solution obtained in Section 3.4).

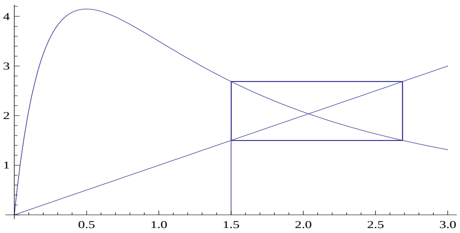

Fig. 2.25. 2-cycles for the Beverton-Holt model with $b = 3$ and $\beta = 28$; see Eqn (2.6.29).

We have $f'(z) = bz^{b-1} - (b-1)z^{b-2}\beta^{2/b} + \beta$ and $f''(z) = (b-1)z^{b-3}(bz + (2-b)\beta^{2/b})$, hence we see that $f'' > 0$ for all $z > 0$. Furthermore, $f(0) = -\beta < 0$. 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 perpendicular to the axis $z = 0$, cuts the boundary of Ω in exactly two points, one being $(0, 0)$ and the other $(\beta^{\frac{1}{b}}, 0)$. Hence, there are no additional equilibria of H^2 for $1 < b < 2$ and therefore H does not have 2-cycles for $b \le 2$. Consider next $b > 2$. Then f'' has exactly one positive root $z_i = (b-2)\beta^{\frac{2}{b}}/b$. The fact that the equilibrium $x^* = \beta^{\frac{1}{b}} - 1$ loses stability at $\beta_{crit} = (b/(b-2))^b$ suggests that a 2-cycle can appear, when β increases passing through this point.

The analysis of this case rests on the following observation.

Lemma 2.25. *Let g be a differentiable function on an interval $[\alpha, \gamma]$ such that $g(\alpha) > 0$, $g(\gamma) \ge 0$ and the only zero of g' is at $x = \alpha$ or at $x = \gamma$. Then $g(x) > 0$ for all $x \in [\alpha, \gamma)$.*

Proof. Assume there is $\alpha < x_0 < \gamma$ where $g(x_0) \le 0$. It cannot be the local minimum of g as then $g'(x_0) = 0$, contrary to the assumption. Thus, for some $x(1) > x_0$ we have $g(x(1)) < 0$ and therefore there is $x(2) \in (x(1), \gamma]$ with $g(x(2)) = 0$ and, by Rolle's theorem, there is $x(3) < \gamma$ for which $g'(x(3)) = 0$, again contradicting the assumption. □

Let us discuss the stable region $\beta \le b^b(b-2)^{-b}$. Then $z_i \le (b-2)/2 = \beta^{\frac{1}{b}}$, that is, z_i is below the original equilibrium $\beta^{\frac{1}{b}}$; it is easy to see that, when β passes through the value $R_{0,crit}$, z_i passes through $\beta^{\frac{1}{b}}$ and moves on to take on values larger than the original equilibrium. Let us evaluate the first derivative at z_i

$$f'\left(\frac{b-2}{b}\beta^{\frac{2}{b}}\right) = \beta\left(1 - \left(\frac{b-2}{b}\right)^{b-2}\beta^{\frac{(b-2)}{b}}\right).$$

Thus we see that $f'\left(\frac{b-2}{b}\beta^{\frac{2}{b}}\right) > 0$ provided $z_i < R_0^{\frac{1}{b}}$ and becomes negative as z_i moves through $R_0^{\frac{1}{b}}$. Further, we see that $f'(\beta^{\frac{1}{b}}) = \beta(b\beta^{-\frac{1}{b}} - (b-2))$ and $f'(\beta^{\frac{1}{b}}) > 0$ provided $\beta < (b(b-2))^b$, that is, for $z_i < \beta^{\frac{1}{b}}$.

Now, consider the case with $\beta < (b(b-2))^b$. Since then $f'(0) = \beta > 0, f'\left(\frac{b-2}{b}\beta^{\frac{2}{b}}\right) > 0$, $f'(\beta^{\frac{1}{b}}) > 0$, $0 < \frac{b-2}{b}\beta^{\frac{2}{b}} < \beta^{\frac{1}{b}}$ and $\frac{b-2}{b}\beta^{\frac{2}{b}}$ is the only zero of f'' on $[0, \beta^{\frac{1}{b}}]$, we can apply the lemma above (to $g = f'$ on the intervals $[0, z_i]$ and $[z_i, \beta^{\frac{1}{b}}]$). Hence, f' is positive on the interval $[0, \beta^{\frac{1}{b}}]$ and $\beta^{\frac{1}{b}}$ is the only zero of f in this interval. Consider now the interval $[\beta^{\frac{1}{b}}, \infty)$. Since $f'(\beta^{\frac{1}{b}}) > 0$ and $f(z)$ tends to $+\infty$ for $z \to \infty$, for f to have zeroes in this interval, it would have to have a local maximum, then take on 0, and then to have a local minimum before crossing the axis again to move to infinity. This would give another zero of f'' between the local extrema, but then this zero would be greater than $\beta^{\frac{1}{b}}$, which is impossible.

For $\beta = (b(b-2))^b$, the points z_i and $\beta^{\frac{1}{b}}$ coalesce and, analogously, we see that there is only one zero of f.

Let us consider the case with $\beta > (b(b-2))^b$. Then $f'(\beta^{\frac{1}{b}}) < 0$ and hence f takes on negative values for $z > \beta^{\frac{1}{b}}$. Since, however, $f(z)$ tends to $+\infty$ for $z \to \infty$, there must be $z^* > \beta^{\frac{1}{b}}$, for which $f(z^*) = 0$. Also, by $f'(\beta^{\frac{1}{b}}) < 0$, $f(z) > 0$ in a left neighbourhood of $\beta^{\frac{1}{b}}$ and hence there must be another zero $1 < z^\# < \beta^{\frac{1}{b}}$, by $f(1) = 1 - \beta^{\frac{1}{b}} < 0$. Since $\beta^{\frac{1}{b}} - 1$ and 0 were the only equilibria of H, $x^* = z^* - 1$ and $x^\# = z^\# - 1 > 0$ must give a 2-cycle.

Figure 2.26 shows, for $b = 3$, how the point z_i moves with β through the equilibrium point $z = 3$ to produce new zeros of f, giving rise to 2-cycles. With much more, mainly computer aided, work we can establish that, as with the logistic equation, we obtain period doubling and transition to chaos.

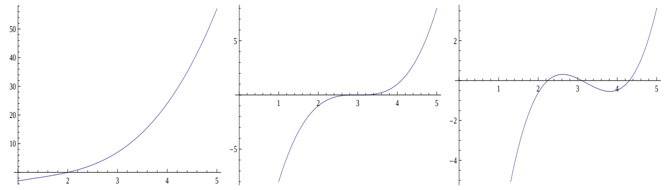

Fig. 2.26. Function f for $b = 3$ and, from left to right, $\beta = 8, 27, 30$. Notice the emergence of 2-cycles represented here by new zeros of f besides $z = \sqrt[3]{\beta}$.

Experimental results are in quite good agreement with the model, see [6]. Most models fell into the stable region. On the other hand, 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 the population density. This can be demonstrated mathematically (before the advent of the mathematical theory of chaos it was assumed that these irregularities were of stochastic nature) and it is observed in the fluctuations of the populations of the Colorado beetle.

The question of whether chaotic behaviours do exist in ecology is still an area of active debate. Observational time series are always finite and inherently noisy, thus it can be argued that always a regular models can be found to fit these data. However, in several laboratory host-parasitoid systems good fits were obtained between the data and chaotic mathematical models and therefore it is reasonable to treat these systems as chaotic.

3

Linear models with discrete structure

1 Introducing structure

The Malthusian model clearly is hugely oversimplified and its improvements may go into many directions. Earlier we discussed models still with an aggregated description but with variable and and nonlinear coefficients. Another option is to consider a relevant internal structure of the population. This could be age and related with it differentiation in birth and death rates. Other possibilities include size or geographical structure which also may impact on death and birth rates. Let us start with revisiting the classical Fibonacci's problem of rabbits.

Fibonacci's rabbits

In his famous book, *Liber abaci*, published in 1202, he 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 fix attention, we assume that we take monthly census just after the births for this month take place and the rabbits were newly born at the beginning of the experiment. Usually the problem is modelled as the initial value problem for a second order difference equation The resulting

$$y(n+2) = y(n+1) + y(n), \quad y(0) = 1, y(1) = 1. \tag{3.1.1}$$

Example 3.1. It is clear that (3.1.1) as a model describing a population of rabbits is oversimplified: rabbits do not die, they are always fertile as soon as they are mature, etc. However, there are biological phenomena for which (3.1.1) 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. 3.1 we present a family tree of a drone. It is clear that the number of ancestors kth generations earlier exactly satisfies (3.1.1).

Our aim here is not to reheat the classical Fibonacci equation but rather use it to explain a more general structure.

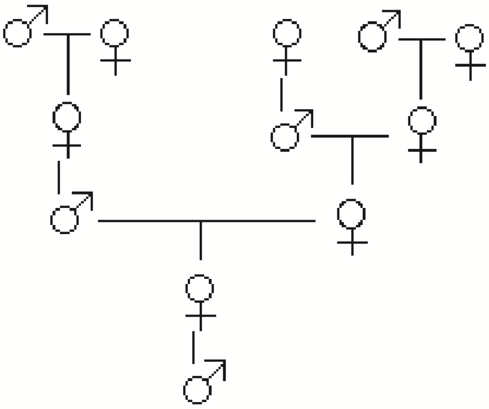

Fig. 3.1. The family tree of a drone

1.1 Models in discrete time

Fibonacci model is an example of an *age-structured population model*: in this particular case each month the population is represented by two classes of rabbits, adults $v_1(n)$ and juveniles $v_0(n)$. Thus the state of the population is described by the vector

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

Since the number of juvenile (one-month old) pairs in month $n+1$ is equal to the number of adults in the month n and the number of adults is the number of adults from the previous month and the number of juveniles from the previous month (who became adults). In other words

$$\begin{aligned} v_0(n+1) &= v_1(n), \\ v_1(n+1) &= v_1(n) + v_0(n) \end{aligned} \tag{3.1.2}$$

or, in a more compact form,

$$\mathbf{v}(n+1) = \mathcal{L}\mathbf{v}(n) := \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \mathbf{v}(n). \tag{3.1.3}$$

The solution can be found by iterations

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

that is,

$$\mathbf{v}(n) = \mathcal{L}^n \mathbf{v}(0), \tag{3.1.4}$$

Leslie matrices

How do we generalize this? Assume that we are tracking only females and not pairs and that census is taken immediately after the reproductive period (the length of which is negligible). 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 is **not** the case for Fibonacci rabbits, where we allowed adults to stay adults forever). We introduce the year-to-year survival probability s_i and the age dependent maternity function m_i; that is, s_i is probability of survival from age i to age $i+1$ and each individual of age i produces m_i offspring on average. Thus, say in the kth breeding season, we have $v_i(k)$ individuals of age i, s_i of them survives to the $k+1$th breeding season, that is, to age $i+1$, producing on average $f_i v_i(k) := m_{i+1} s_i v_i(k)$ offspring. Thus, the surviving individuals of all ages will produce

$$v_0(k+1) = \sum_{i=0}^{n-1} f_i v_i(k) = \sum_{i=0}^{n-1} m_{i+1} s_i v_i(k)$$

offspring. In this case, the evolution of the population can be described by the difference system

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

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

$$\mathcal{L} := \begin{pmatrix} f_0 & f_1 & \cdots & f_{n-2} & f_{n-1} \\ s_0 & 0 & \cdots & 0 & 0 \\ 0 & s_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & s_{n-2} & 0 \end{pmatrix}. \tag{3.1.5}$$

The matrix of the form (3.1.5) is referred to as a *Leslie matrix*.

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

$$\mathcal{L} := \begin{pmatrix} f_0+\tau_0 & f_1 & \cdots & f_{n-2} & f_{n-1} \\ s_0 & \tau_1 & \cdots & 0 & 0 \\ 0 & s_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & s_{n-2} & \tau_{n-1} \end{pmatrix}, \tag{3.1.6}$$

Such matrices are called *Usher matrices*. We note that the matrix of the Fibonacci process in an Usher matrix.

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

$$\mathcal{L} := \begin{pmatrix} \tau_0 & f_1 & 0 \\ s_0 & \tau_1 & 0 \\ 0 & s_1 & \tau_2 \end{pmatrix}, \tag{3.1.7}$$

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 $i = n$.

Projection matrices

Leslie matrices fit into a more general mathematical structure describing evolution of populations divided in states, or subpopulations, not necessarily related to age. Matrices resulting from such a modelling, that is, describing changes of a structured populations from one generation to another due to migrations between structural states and (generalized) birth processes are called *projection matrices*. 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 to, say, a_{ij} individuals in state j. Typically, this is due to a 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};
- dying with probability d_j, that is, surviving with probability $1 - d_j$.

It must be borne in mind that the exact interpretation of the coefficients can vary depending on the way of taking census and when the death and migrations occur in the process, see Remark 2.1.

Other choices and interpretations are, however, also possible. For instance, if we consider size structured population of clusters of cells divided into subpopulations according to their size i, an n-cluster can split into several smaller clusters, contributing thus to 'births' of clusters in subpopulations indexed by $i < n$. Hence, a_{ij} are non-negative but otherwise arbitrary numbers. Denoting, as before, by $v_i(k)$ the number of individuals at time k in state i, with $\mathbf{v}(k) = (v_1(k), \ldots, v_n(k))$, we have

$$\mathbf{v}(k+1) = \mathcal{A}\mathbf{v}(k), \tag{3.1.8}$$

where

$$\mathcal{A} := \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1\,n-1} & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2\,n-1} & a_{2n} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{n\,n-1} & a_{nn} \end{pmatrix}. \tag{3.1.9}$$

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 3.2. Any chromosome ends with a *telomer* which protects it agains 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 an 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 daughters with length $i-1$ telomer is $2(1-a_i)$. However, to have offspring, the cell must have 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

$$\mathcal{A} := \begin{pmatrix} 0 & d_1 & 0 & \cdots & 0 \\ 0 & r_1 & d_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & r_n \end{pmatrix}. \tag{3.1.10}$$

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

Markov matrices

A particular version of (3.1.9) 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 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

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

we obtain

$$\sum_{1\leq i\leq n} p_{ij} = 1. \tag{3.1.11}$$

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 $N(k) = v_1(k) + \ldots + v_n(k)$ so that

$$\begin{aligned} N(k+1) = \sum_{1\leq i\leq n} v_i(k+1) = \sum_{1\leq i\leq n} \left(\sum_{1\leq j\leq n} p_{ij} v_j(k) \right) \\ = \sum_{1\leq j\leq n} v_j(k) \left(\sum_{1\leq i\leq n} p_{ij} \right) = \sum_{1\leq j\leq n} v_j(k) = N(k). \end{aligned} \tag{3.1.12}$$

1.2 Transition matrices for continuous time processes

Let us consider a model with population divided into n subpopulation but with transitions between them happening in a 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 slightly different approach and will be considered later.

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 due 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,

$$\begin{aligned} v_i(t+\Delta t) = v_i(t) - d_i \Delta t v_i(t) + \sum_{j=1}^{n} b_{ij} \Delta t v_j(t) \\ + \sum_{\substack{j=1 \\ j\neq i}}^{n} \left(p_{ij} \Delta t v_j(t) - p_{ji} \Delta t v_i(t) \right), \quad i = 1, \ldots, n. \end{aligned} \tag{3.1.13}$$

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 (3.1.13), dividing by Δt and passing to the limit with $\Delta t \to 0$ we obtain

$$v_i'(t) = \sum_{j=1}^{n} q_{ij} v_j(t), \quad , i = 1, \ldots, n, \tag{3.1.14}$$

or

$$\mathbf{v}' = \mathcal{Q}\mathbf{v}, \tag{3.1.15}$$

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

$$\begin{aligned}
\frac{dN}{dt} &= \sum_{1\leq i\leq n} \frac{dv_i(t)}{dt} = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} q_{ij} v_j(t) \right) \\
&= \sum_{i=1}^{n} q_{ii} v_i(t) + \sum_{i=1}^{n} \left(\sum_{\substack{j=1\\j\neq i}}^{n} q_{ij} v_j(t) \right) \\
&= -\sum_{i=1}^{n} v_i(t) \left(\sum_{\substack{j=1\\j\neq i}}^{n} p_{ji} \right) + \sum_{i=1}^{n} \left(\sum_{\substack{j=1\\j\neq i}}^{n} p_{ij} v_j(t) \right) \\
&= -\sum_{i=1}^{n} v_i(t) \left(\sum_{\substack{j=1\\j\neq i}}^{n} p_{ji} \right) + \sum_{j=1}^{n} v_j(t) \left(\sum_{\substack{i=1\\i\neq j}}^{n} p_{ij} \right) \\
&= -\sum_{i=1}^{n} v_i(t) \left(\sum_{\substack{j=1\\j\neq i}}^{n} p_{ji} \right) + \sum_{i=1}^{n} v_i(t) \left(\sum_{\substack{j=1\\j\neq i}}^{n} p_{ji} \right) = 0,
\end{aligned} \tag{3.1.16}$$

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

Remark 3.3. 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. To certain extent we can compare the increments in the discrete time process

$$\begin{aligned}\mathbf{v}(k+1)-\mathbf{v}(k) &= (-I+\mathcal{P})\mathbf{v}(k) \\ &= \begin{pmatrix} -1+p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & -1+p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ p_{n1} & p_{n2} & \cdots & -1+p_{nn} \end{pmatrix}\mathbf{v}(k),\end{aligned} \tag{3.1.17}$$

so that the ‘increment’ matrix has the property that each row adds up to zero due to (3.1.11). 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 1. In fact, they can be arbitrary numbers and represent probability rates with $p_{ij}\Delta t$ being approximate interstate transition probabilities.

2 Long time behaviour of structured population models

The main interest in population theory is to determine the long time structure of the population.

2.1 Numerical example of the AEG property

Before we embark on mathematical analysis, let us consider two numerical examples which indicate what we should expect from the models.

Example 3.4. Let us consider a population divided into three classes, evolution of which is modelled by the Leslie matrix

$$\mathcal{L} = \begin{pmatrix} 2 & 1 & 1 \\ 0.5 & 0 & 0 \\ 0 & 0.4 & 0 \end{pmatrix},$$

so that the population $\mathbf{v} = (v_1, v_2, v_3)$ evolves according to

$$\mathbf{v}(k+1) = \mathcal{L}\mathbf{v}(k), \quad k = 0, 1, 2 \ldots,$$

or

$$\mathbf{v}(k) = \mathcal{L}^k \overset{\circ}{\mathbf{v}},$$

where $\overset{\circ}{\mathbf{v}}$ is an initial distribution of the population. In Fig. 3.2 we observe that each component grows very fast with k. However, if we compare growth of $v_1(k)$ with $v_2(k)$ and of $v_2(k)$ with $v_3(k)$ (see Fig. 3.3) we see that the ratios stabilize quickly around 4.5 in the first case and around 5.62 in the second case. This suggests that there is a scalar function $f(k)$ and a vector $\mathbf{e} = (e_1, e_2, e_3) = (25.29, 5.62, 1)$ such that for large k

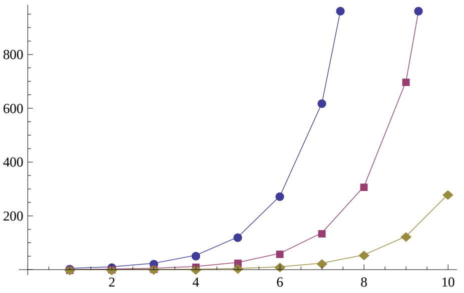

Fig. 3.2. Evolution of $v_1(k)$ (circles), $v_2(k)$ (squares) and $v_3(k)$ (rhombuses) for the initial distribution $\overset{\circ}{\mathbf{v}} = (1, 0, 3)$ and $k = 1, \ldots, 20$.

$$\mathbf{v}(k) \approx f(k)\mathbf{e}. \tag{3.2.18}$$

Let us consider another initial condition, say, $\overset{\circ}{v}= (2, 1, 4)$ and do the same comparison. It turns out that

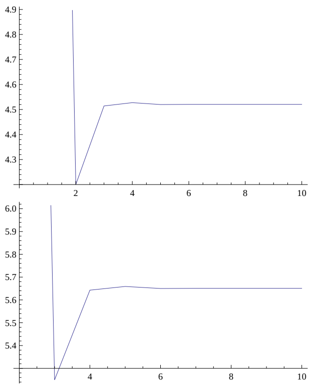

Fig. 3.3. Evolution of $v_1(k)/v_2(k)$ (top) and $v_2(k)/v_3(k)$ (bottom) for the initial distribution $\overset{\circ}{\mathbf{v}}= (1, 0, 3)$ and $k = 1, \ldots, 20$.

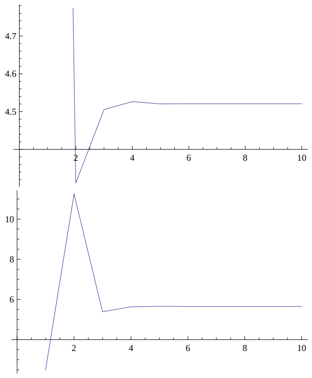

Fig. 3.4. Evolution of $v_1(k)/v_2(k)$ (top) and $v_2(k)/v_3(k)$ (bottom) for the initial distribution $\overset{\circ}{\mathbf{v}}= (2, 1, 4)$ and $k = 1, \ldots, 20$.

the ratios stabilize at the same level which further suggest that $\mathbf{e}$ does not depend on the initial condition so that (5.4.34) can be refined to

$$\mathbf{v}(k) \approx f_1(k)g(\overset{\circ}{\mathbf{v}})\mathbf{e}, \quad k \to \infty \tag{3.2.19}$$

where g is a linear function. Anticipating the development of the theory, it can be proved that $f_1(k) = \lambda^k$ where λ is the largest eigenvalue of $\mathcal{L}$, $\mathbf{e}$ is the eigenvector corresponding to λ and $g(\mathbf{x}) = \mathbf{g} \cdot \mathbf{x}$ with $\mathbf{g}$ being the eigenvector of the transpose matrix corresponding to λ. In our case, $\lambda \approx 2.26035$ and the ratios $v_i(k)/\lambda^k$ stabilize as seen in Fig. 3.5.

The situation in which the structure of the population after long time does not depend on the initial condition but only on the intrinsic properties of the model (here the leading eigenvalue) is called the *asynchronous exponential growth (AEG)* property.

Unfortunately, not all Leslie matrices enjoy this property.

Example 3.5. Consider a Leslie matrix given by

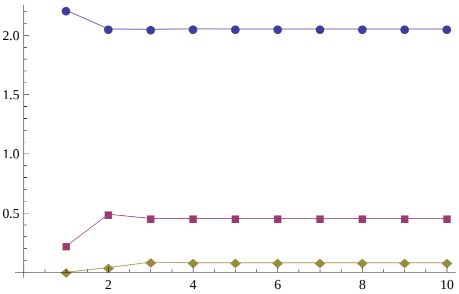

Fig. 3.5. Evolution of $v_1(k)/\lambda^k)$ (circles), $v_2(k)/\lambda^k$ (squares)and $v_3(k)/\lambda^k$ (rhombuses) for the initial distribution $\overset{\circ}{\mathbf{v}}= (1, 0, 3)$ and $k = 1, \ldots, 20$.

$$\mathcal{L} = \begin{pmatrix} 0 & 0 & 3 \\ 0.5 & 0 & 0 \\ 0 & 0.4 & 0 \end{pmatrix}$$

and a population evolving according to

$$\mathbf{y}(k) = \mathcal{L}^k \overset{\circ}{\mathbf{y}}$$

with $\overset{\circ}{\mathbf{y}}= (2, 3, 4)$. The solution is given in Fig. 3.6. The picture is completely different from that obtained

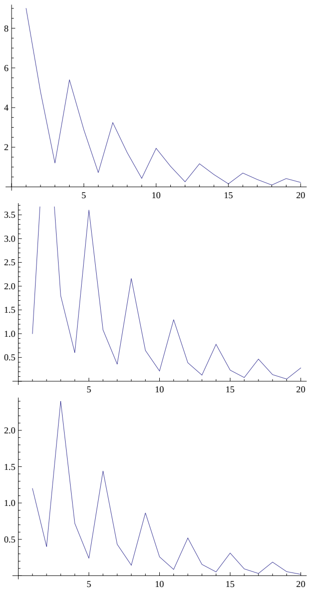

Fig. 3.6. Evolution of $y_1(k)$ (top) and $y_2(k)$ (middle) and $y_3(k)$ (bottom) for the initial distribution $\overset{\circ}{\mathbf{v}}= (2, 3, 4)$ and $k = 1, \ldots, 10$.

in Example 3.4. We observe some pattern but the ratios do not tend to a fixed limit but oscillate, as shown in Fig. 3.7. This can be explained using the spectral decomposition: indeed, the eigenvalues are given by $\lambda_1 = 0.843433, \lambda_2 = -0.421716 + 0.730434i, \lambda_2 = -0.421716 - 0.730434i$ and we can check that $|\lambda_1| = |\lambda_2| = |\lambda_3| = 0.843433$ and thus we do not have the dominant eigenvalue. The question we will try to answer in the next chapter is what features of the population are responsible for such behaviour.

The next example shows that structured population models in continuous time have the same property.

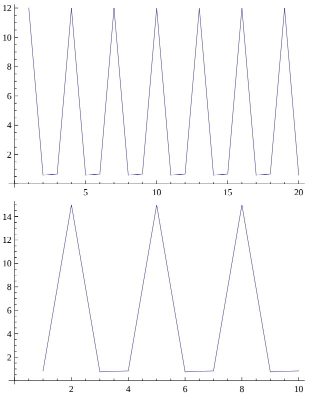

Fig. 3.7. Evolution of $y_1(k)/y_2(k)$ (top) and $y_2(k)/y_3(k)$ (bottom) for the initial distribution $\overset{\circ}{\mathbf{v}}= (2,3,4)$ and $k = 1,\dots,20$.

Example 3.6. **Population in continuous time.**
Consider the following problem

$$\frac{d\mathbf{v}}{dt} = \mathcal{A}\mathbf{v}, \tag{3.2.20}$$

where

$$\mathcal{A} = \begin{pmatrix} -1 & 1 & 1 \\ 0.5 & -0.5 & 0 \\ 0 & 0.4 & -1 \end{pmatrix}.$$

We consider this equation with the initial conditions $\overset{\circ}{\mathbf{v}}= (1,0,3)$ and $\overset{\circ}{\mathbf{v}}= (2,1,4)$.

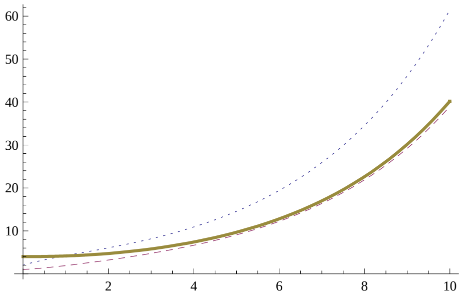

Fig. 3.8. Solutions $v_1(t)$ (dotted), $v_2(t)$ (dashed) and $v_3(t)$ (continuous) for the initial condition $\overset{\circ}{\mathbf{v}}= (2,1,4)$

As before we see that the components grow fast but $v_1(t)/v_2(t)$ and $v_2(t)/v_3(t)$ stabilize quickly around 1.57631 and 0.970382, respectively, see Fig. 3.9, and these ratios are independent of the initial conditions. Thus,

$$\mathbf{v}(t) \approx f(t)g(\overset{\circ}{\mathbf{v}})\mathbf{e}$$

for large t, where $\mathbf{e} = (1.5296, 0.970382, 1)$ and g is a scalar linear function of $\overset{\circ}{\mathbf{v}}$. As illustrated in Fig. 3.10, $f(t) = e^{0.288153t}$ and the number 0.288153 is the largest eigenvalue of $\mathcal{A}$.

It is interesting that the behaviour observed in Example 3.5 cannot occur in continuous time.

The question whether any matrix with nonnegative entries in discrete time and with positive off-diagonal entries gives rise to such a behaviour and, if not, what models lead to AEG, is quite delicate and requires invoking the Frobenius-Perron theorem which will be discussed in the next chapter. First, however, we have to review general properties of solution to systems of difference and differential equations.

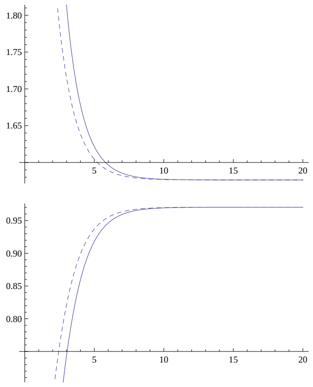

Fig. 3.9. Evolution of $v_1(t)/v_2(t)$ (top) and $v_2(t)/v_3(t)$ (bottom) for the initial distributions $\overset{\circ}{\mathbf{v}}= (1,0,3)$ (continuous line) and $\overset{\circ}{\mathbf{v}}= (2,1,4)$ (dashed line).

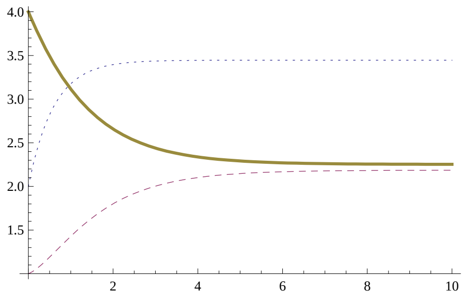

Fig. 3.10. Evolution $v_1(t)/e^{0.288153t}$ (dotted), $v_2(t)/e^{0.288153t}$ (dashed) and $v_3(t)/e^{0.288153t}$ (continuous) for the initial condition $\overset{\circ}{\mathbf{v}}= (2,1,4)$.

3 Systems of linear differential and difference equations.

To explain and be able to predict similar behaviour in population models, first we discuss basic facts solvability of initial value problems systems of difference

$$\mathbf{y}(k+1) = \mathcal{A}\mathbf{y}(k), \quad \mathbf{y}(0) =\overset{\circ}{\mathbf{y}} \tag{3.3.21}$$

and differential,

$$\frac{d\mathbf{y}}{dt} = \mathcal{A}\mathbf{y}, \tag{3.3.22}$$

equations. Here $\mathbf{y} = (y_1,\ldots,y_n) \in \mathbb{R}^n$ and $\mathcal{A} = (a_{ij})_{1\leq i,j\leq n}$ is an $n\times n$ matrix.

We begin with introducing relevant notation and mathematical concepts.

3.1 Basic mathematical notions

To make further progress, we have to formalize a number of statements made in the previous sections and, in particular, the meaning of the approximate equality (3.3.76). For this, we have to set the problem in an appropriate mathematical framework.

First, we note that, in the context of population theory, if a given difference equation/system of equations is to describe evolution of a population; that is, if the solution is the population size or density, then clearly solutions emanating from non-negative data must stay non-negative. Thus we have to extend the notion of positivity to vectors. We say that a vector $\mathbf{x} = (x_1, \ldots, x_n)$ is non-negative, (resp. positive), if for all $i = 1, \ldots, n$, $x_i \geq 0$, (resp. $x_i > 0$). We denote these as $\mathbf{x} \geq 0$, (resp. $\mathbf{x} > 0$) and put

$$X_+ = \{\mathbf{x} \in \mathbb{R}^n;\ \mathbf{x} \geq 0\}.$$

Similarly, we say that a matrix $\mathcal{A} = (a_{ij})_{1 \leq i,j \leq n}$ is non-negative (resp. positive) and write $\mathcal{A} \geq 0$ (resp. $\mathcal{A} > 0$) if $a_{ij} \geq 0$ (resp. $a_{ij} > 0$) for all $i, j = 1, \ldots, n$.

As we noted earlier, writing (3.1.4), the solution to (3.3.21) is given by the sequence $(\mathcal{A}^k)_{k \geq 1}$ of iterates operating in the state space $X = \mathbb{R}^n$. Then, in fact, (3.3.76) is a statement about the limit of $\mathcal{A}^k \overset{\circ}{\mathbf{y}}$ as $k \to \infty$ so we must introduce a way of measuring distances in X. Usually we want to make the metric consistent with the linear structure of $\mathbb{R}^n$; that is, with addition of vectors and multiplication of vectors by scalars. Such metrics are defined by the so-called norms. A norm is a function $\| \cdot \| : X \to \mathbb{R}_+$ satisfying, for any $\mathbf{x}, \mathbf{y} \in X, \alpha \in \mathbb{R}$,

$$\|\mathbf{x}\| = 0 \quad \text{iff} \quad \mathbf{x} = 0, \qquad \|\alpha \mathbf{x}\| = |\alpha| \|\mathbf{x}\|, \qquad \|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|.$$

Then, the distance between two points $\mathbf{x}$ and $\mathbf{y}$ is given by $\|\mathbf{x} - \mathbf{y}\|$.

In what follows, we say that the sequence $(\mathbf{x}(k))_{k \geq 0}$ converges to $\mathbf{x}$, and write

$$\lim_{k \to \infty} \mathbf{x}(k) = \mathbf{x}$$

if the numerical sequence $(\|\mathbf{x}(k) - \mathbf{x}\|)_{k \geq 0}$ converges to 0. There is a variety norms in $\mathbb{R}^n$ (which, however, define the same convergence of sequences, see e.g. [14, Chapter 5]), the most common being the Euclidean norm

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}.$$

In a plain language, a norm is a parameter which we use to describe the state of a system. In many applications discussed in this book, we consider populations and one of the most natural parameters with which a population is described, is its size, see e.g. Eqns (3.1.12) and (3.1.16). To accommodate this, we consider the norm

$$\|\mathbf{y}\| = \sum_{i=1}^{n} |y_i| \tag{3.3.23}$$

which, for $\mathbf{y} \geq 0$ simplifies to

$$\|\mathbf{y}\| = \sum_{i=1}^{n} y_i \tag{3.3.24}$$

which is the total population of the ensemble. This norm is often called l^1-norm.

Since we want $\mathcal{A}$ to act from X to X with the same way of measuring distances, we should have

$$\|\mathcal{A}\mathbf{x}\| = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leq \sum_{j=1}^{n} |x_j| \sum_{i=1}^{n} |a_{ij}| \leq \|\mathbf{x}\| \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}| =: \|\mathcal{A}\| \|\mathbf{x}\|$$

where

$$\|\mathcal{A}\| = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}|$$

is called the norm of the matrix/operator $\mathcal{A}$. It is a norm of an operator in the functional analytic sense, see [14, Chapter 5], and though one can define other norms, the above one is consistent with the interpretation of the problem and will be used in the notes.

3.2 Systems of difference equations I.

We are interested in solving

$$\mathbf{y}(k+1)=\mathcal{A}\mathbf{y}(k),$$

where $\mathcal{A}$ is an $n\times n$ matrix $\mathcal{A}=(a_{ij})_{1\leq i,j\leq n}$; that is

$$\mathcal{A}=\begin{pmatrix} a_{11} & \ldots & a_{1n}\\ \vdots & & \vdots\\ a_{n1} & \ldots & a_{nn}\end{pmatrix},$$

and $\mathbf{y}(k)=(y_1(k),\ldots,y_n(k))$.

Eq. (3.3.21) is usually supplemented by the initial condition $\mathbf{y}(0)=\mathbf{y^0}$. As noted earlier, by induction, to the solution to (3.3.21) is given by

$$\mathbf{y}(k)=\mathcal{A}^k\mathbf{y^0}, k=1,2,\ldots. \tag{3.3.25}$$

The problem with (3.3.25) is that it is rather difficult to give an explicit form of $\mathcal{A}^k$.

To proceed, we assume that the matrix $\mathcal{A}$ is nonsingular (this is not serious restriction as then one can consider action of the matrix in a subspace). This means, in particular, that if $\mathbf{v^1},\ldots,\mathbf{v^n}$ are linearly independent vectors, then also $\mathcal{A}\mathbf{v^1},\ldots,\mathcal{A}\mathbf{v^n}$ are linearly independent. Since $\mathbb{R}^n$ is n-dimensional, it is enough to find n linearly independent vectors $\mathbf{v^i}$, $i=1,\ldots,n$ for which $\mathcal{A}^k\mathbf{v^i}$ can be easily evaluated. Assume for a moment that such vectors have been found. Then, for arbitrary $\mathbf{x^0}\in\mathbb{R}^n$ we can find constants $c_1,\ldots,c_n$ such that

$$\mathbf{x^0}=c_1\mathbf{v^1}+\ldots+c_n\mathbf{v^n}.$$

Precisely, let $\mathcal{V}$ be the matrix having vectors $\mathbf{v^i}$ as its columns

$$\mathcal{V}=\begin{pmatrix} | & \ldots & |\\ \mathbf{v^1} & \ldots & \mathbf{v^n}\\ | & \ldots & |\end{pmatrix}. \tag{3.3.26}$$

Note, that $\mathcal{V}$ is invertible as the vectors $\mathbf{v^i}$ are linearly independent. Denoting $\mathbf{c}=(c_1,\ldots,c_n)$, we obtain

$$\mathbf{c}=\mathcal{V}^{-1}\mathbf{x^0}. \tag{3.3.27}$$

Thus, for an arbitrary $\mathbf{x^0}$ we have

$$\mathcal{A}^k\mathbf{x^0}=\mathcal{A}^k(c_1\mathbf{v^1}+\ldots+c_n\mathbf{v^n})=c_1\mathcal{A}^n\mathbf{v^1}+\ldots+c_k\mathcal{A}^n\mathbf{v^n}. \tag{3.3.28}$$

Now, if we denote by $\mathcal{A}_k$ the matrix whose columns are vectors $\mathcal{A}^k\mathbf{v^1},\ldots,\mathcal{A}^k\mathbf{v^n}$, then we can write

$$\mathcal{A}^k\mathbf{x^0}=\mathcal{A}_k\mathbf{c}=\mathcal{A}_k\mathcal{V}^{-1}\mathbf{x^0}. \tag{3.3.29}$$

Hence, the problem is to find linearly independent vectors $\mathbf{v^i}$, $i=1,\ldots,k$, on which powers of $\mathcal{A}$ can be easily evaluated. We shall 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

$$\mathcal{A}^k\mathbf{v^1}=\lambda_1^k\mathbf{v^1}.$$

Therefore, if we have n linearly independent eigenvectors $\mathbf{v^1},\ldots,\mathbf{v^n}$ corresponding to eigenvalues $\lambda_1,\ldots,\lambda_n$ (not necessarily distinct), then from (3.3.28) we obtain

$$\mathcal{A}^k\mathbf{x^0}=c_1\lambda_1^k\mathbf{v^1}+\ldots+c_n\lambda_n^k\mathbf{v^n}.$$

with $c_1,\ldots,c_n$ given by (3.3.27), or

$$\mathcal{A}^k\mathbf{x^0}=\begin{pmatrix} | & \ldots & |\\ \lambda_1^k\mathbf{v^1} & \ldots & \lambda_n^k\mathbf{v^n}\\ | & \ldots & |\end{pmatrix}\mathcal{V}^{-1}\mathbf{x_0} \tag{3.3.30}$$

Systems of differential equations I.

Considerations of the previous paragraph to some extent can be repeated for systems of differential equations

$$\mathbf{y}' = \mathcal{A}\mathbf{y},$$

where $\mathbf{y}(t) = (y_1(t), \ldots, y_n(t))$ and, as before, $\mathcal{A} = \{a_{ij}\}_{1\leq i,j\leq n}$ is an $n \times n$ matrix. The system (3.3.22) is considered together with the following initial conditions

$$\mathbf{y}(t_0) = \overset{\circ}{\mathbf{y}} . \tag{3.3.31}$$

The question of solvability and uniqueness of (3.3.22), (3.3.31) is not as easy as for difference equations but it follows from the Picard theorem. We summarize the relevant properties of solutions in the following theorem Thus, we can state

Theorem 3.7. *[5, 10]*

i. *There exists one and only one solution of the initial value problem (3.3.22), (3.3.31), which is defined for all $t \in \mathbb{R}$.*

ii. *The set $\mathbf{X}$ of all solutions to (3.3.22) is a linear space of dimension n.*

iii. *If $\mathbf{y_1}(t), \ldots, \mathbf{y_r}(t)$ are linearly independent solutions of (3.3.22) and let $t_0 \in \mathbb{R}$ be an arbitrary number. Then, $\{\mathbf{y_1}(t), \ldots, \mathbf{y_r}(t)\}$ form a linearly independent set of functions if and only if $\{\mathbf{y_1}(t_0), \ldots, \mathbf{y_r}(t_0)\}$ is a linearly independent set of vectors in $\mathbb{R}^n$.*

An important consequence of iii. is that solutions starting from linearly independent initial conditions remain linearly independent. Note that this is not necessarily the case in systems of difference equations – to have this property we required $\mathcal{A}$ to be nonsingular.

Theorem 3.7 implies that there is matrix $\mathcal{E}(t)$ such that the solution $\mathbf{y}(t)$ can be represented as

$$\mathbf{y}(t) = \mathcal{E}(t) \overset{\circ}{\mathbf{y}} \tag{3.3.32}$$

which satisfies $\mathcal{E}(0) = \mathcal{I}$ (the identity matrix). Then we proceed as in the discrete case by assuming that we can find n linearly independent vectors $\mathbf{v^i}$, $i = 1, \ldots, n$ for which $\mathcal{E}(t)\mathbf{v^i}$ can be easily evaluated. Then, for arbitrary $\overset{\circ}{\mathbf{y}} \in \mathbb{R}^n$ we can find constants $c_1, \ldots, c_n$ such that

$$\overset{\circ}{\mathbf{y}} = c_1\mathbf{v^1} + \ldots + c_n\mathbf{v^n};$$

that is, denoting $\mathbf{c} = (c_1, \ldots, c_n)$,

$$\mathbf{c} = \mathcal{V}^{-1} \overset{\circ}{\mathbf{x}}, \tag{3.3.33}$$

where $\mathcal{V}$ was defined in (3.3.26). Thus, for an arbitrary $\overset{\circ}{\mathbf{y}}$ we have

$$\mathcal{E}(t) \overset{\circ}{\mathbf{y}} = \mathcal{E}(t)(c_1\mathbf{v^1} + \ldots + c_2\mathbf{v^n}) = c_1\mathcal{E}(t)\mathbf{v^1} + \ldots + c_k\mathcal{E}(t)\mathbf{v^n}. \tag{3.3.34}$$

Now, if we denote by $\mathcal{E}_{\mathbf{v}}(t)$ the matrix whose columns are vectors $\mathcal{E}(t)\mathbf{v^1}, \ldots, \mathcal{E}(t)\mathbf{v^n}$, then we can write

$$\mathcal{E}(t) \overset{\circ}{\mathbf{y}} = \mathcal{E}_{\mathbf{v}}(t)\mathbf{c} = \mathcal{E}_{\mathbf{v}}(t)\mathcal{V}^{-1} \overset{\circ}{\mathbf{y}} . \tag{3.3.35}$$

Hence, again, the problem lies in finding linearly independent vectors $\mathbf{v^i}$, $i = 1, \ldots, k$, on which $\mathcal{E}$ can be easily evaluated. Mimicking the scalar case, 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 (3.3.22) if and only if

$$\mathcal{A}\mathbf{v} = \lambda\mathbf{v}, \tag{3.3.36}$$

or in other words, $\mathbf{y}(t) = e^{\lambda t}\mathbf{v}$ 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 3.7, 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 $\lambda_1, \ldots, \lambda_n$ (not necessarily distinct), then the general solution of (3.3.44) is of the form

$$\mathbf{y}(t) = c_1 e^{\lambda_1 t}\mathbf{v^1} + \ldots + c_n e^{\lambda_n t}\mathbf{v^n}. \tag{3.3.37}$$

with $c_1, \ldots, c_n$ given by (3.3.27), or

$$\mathcal{E}(t) \overset{\circ}{\mathbf{y}} = \begin{pmatrix} | & \ldots & | \\ e^{\lambda_1 t}\mathbf{v^1} & \ldots & e^{\lambda_n t}\mathbf{v^n} \\ | & \ldots & | \end{pmatrix} \mathcal{V}^{-1} \overset{\circ}{\mathbf{y}} . \tag{3.3.38}$$

Unfortunately, in many cases there are insufficiently many eigenvectors to generate all solutions.

Eigenvalues, eigenvectors and associated eigenvectors.

Let $\mathcal{A}$ be an $n \times n$ matrix. We say that a number λ (real or complex) is an *eigenvalue* of $\mathcal{A}$ is there exist a non-zero solution of the equation

$$\mathcal{A}\mathbf{v} = \lambda\mathbf{v}. \tag{3.3.39}$$

Such a solution is called an *eigenvector* of $\mathcal{A}$. The set of eigenvectors corresponding to a given eigenvalue is a vector subspace. Eq. (3.3.39) 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

$$det(\mathcal{A} - \lambda\mathcal{I}) = \begin{vmatrix} a_{11} - \lambda \ldots & a_{1n} \\ \vdots & \vdots \\ a_{n1} & \ldots a_{nn} - \lambda \end{vmatrix} = 0. \tag{3.3.40}$$

Evaluating the determinant we obtain a polynomial in λ of degree n. This polynomial is also called the *characteristic polynomial* of the system (3.3.22). We shall denote this polynomial by $p(\lambda)$. From algebra we know that there are exactly n, possibly complex, roots of $p(\lambda)$. The set of eigenvalues eigenvalues is called the *spectrum* of $\mathcal{A}$ and denoted by $\sigma(\mathcal{A})$. An important role in applications is played by the spectral radius of $\mathcal{A}$, defined as

$$r(A) = \sup_{\lambda \in \sigma(\mathcal{A})} |\lambda|. \tag{3.3.41}$$

Some of the roots of the characteristic polynomial can be multiple, so that in general $p(\lambda)$ factorizes into

$$p(\lambda) = (\lambda_1 - \lambda)^{n_1} \cdot \ldots \cdot (\lambda_k - \lambda)^{n_k}, \tag{3.3.42}$$

with $n_1 + \ldots + n_k = n$. 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 the roots of the characteristic polynomial of $\mathcal{A}$. The exponent n_i appearing in the factorization (3.3.42) 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 $\tilde{E}_{\lambda_i}$. The dimension of $\tilde{E}_{\lambda_i}$ 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 particular, 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 difference and 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 equations

$$(\mathcal{A}-\lambda_i\mathcal{I})^j\mathbf{v}=0$$

for $j\geq 1$. Using Theorem 7.17 we see that the solutions of these equations form an increasing set terminating at worst at

$$(\mathcal{A}-\lambda_i\mathcal{I})^{n_i}\mathbf{v}=0$$

with n_i linearly independent solutions. In practice, we first find linearly independent eigenvectors and, if there are fewer than n_i of them, we look for solutions to $(\mathcal{A}-\lambda_i\mathcal{I})^2\mathbf{v}=0$ selecting the ones that are not eigenvectors (then they must be linearly independent of eigenvectors) and continue with higher powers until we find n_i linearly independent solutions. In each the step, say j, we select solutions that are independent of the solutions obtained in step $j-1$; for this it is enough to find solutions to $(\mathcal{A}-\lambda_i\mathcal{I})^j\mathbf{v}=0$ that satisfy $(\mathcal{A}-\lambda_i\mathcal{I})^{j-1}\mathbf{v}\neq 0$.

Vectors $\mathbf{v}$ obtained in this way for a given λ_i are called *generalized* or *associated eigenvectors* corresponding to λ_i and they span an n_i dimensional subspace called a *generalized* or *associated eigenspace* corresponding to λ_i, denoted hereafter by E_{λ_i}.

Now we show how to apply the concepts discussed above to solve systems of difference and differential equations.

Systems of difference equations II.

Let us return to the system

$$\mathbf{y}(k+1)=\mathcal{A}\mathbf{y}(k),\quad \mathbf{y}(0)=\overset{\circ}{\mathbf{y}}\,.$$

As discussed, we need to find formulae for $\mathcal{A}^k\mathbf{v}$ for a selected n linearly independent vectors $\mathbf{v}$. Let us take as $\mathbf{v}$ the collection of all eigenvectors and associated eigenvectors of $\mathcal{A}$. We know that if $\mathbf{v^i}$ is an eigenvector associated to an eigenvalue λ^i, then $\mathcal{A}^k\mathbf{v^i}=\lambda_i^k\mathbf{v^i}$. Thus, the question is whether $\mathcal{A}^k$ can be effectively evaluated on associated eigenvectors.

Let $\mathbf{v^j}$ be an associated eigenvector found as a solution to $(\mathcal{A}-\lambda_i\mathcal{I})^j\mathbf{v^j}=\mathbf{0}$ with $j\leq n_i$. Then, using the binomial expansion, we find

$$\begin{aligned}\mathcal{A}^k\mathbf{v^j}&=(\lambda_i\mathcal{I}+\mathcal{A}-\lambda_i\mathcal{I})^k\mathbf{v^j}=\sum_{r=0}^{k}\lambda_i^{k-r}\binom{k}{r}(\mathcal{A}-\lambda_i\mathcal{I})^r\mathbf{v^j}\\&=\Big(\lambda_i^k\mathcal{I}+k\lambda_i^{k-1}(\mathcal{A}-\lambda_i\mathcal{I})+\ldots\\&\qquad+\frac{k!}{(j-1)!(k-j+1)!}\lambda_i^{k-j+1}(\mathcal{A}-\lambda_i\mathcal{I})^{j-1}\Big)\mathbf{v^j},\end{aligned}\tag{3.3.43}$$

where

$$\binom{k}{r} = \frac{k!}{r!(k-r)!}$$

is the Newton symbol. It is important to note that (3.3.43) is a finite sum for any k; it always terminates at most at the term $(\mathcal{A} - \lambda_1\mathcal{I})^{n_i-1}$ where n_i is the algebraic multiplicity of λ_i.

We shall illustrate these considerations by several examples.

Example 3.8. Find $\mathcal{A}^k$ 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 $\lambda = 4$, 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

$$\mathbf{v^1} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{v^2} = \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix}.$$

Therefore

$$\mathcal{A}^k\mathbf{v^1} = 4^k \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \mathcal{A}^k\mathbf{v^2} = 4^k \begin{pmatrix} 0 \\ -2 \\ 1 \end{pmatrix}.$$

To find the associated eigenvector 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 (3.3.43)

$$\begin{aligned} \mathcal{A}^k\mathbf{v^3} &= \left(4^k\mathcal{I} + k4^{k-1}(\mathcal{A} - 4\mathcal{I})\right)\mathbf{v^3} \\ &= \left(\begin{pmatrix} 4^k & 0 & 0 \\ 0 & 4^k & 0 \\ 0 & 0 & 4^k \end{pmatrix} + k4^{k-1} \begin{pmatrix} 0 & 1 & 2 \\ 0 & -2 & -4 \\ 0 & 1 & 2 \end{pmatrix}\right) \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} 2k4^{k-1} \\ -k4^k \\ 4^k + 2k4^{-1} \end{pmatrix}. \end{aligned}$$

To find explicit expression for $\mathcal{A}^k$ we use (3.3.29). In our case

$$\mathcal{A}_k = \begin{pmatrix} 4^k & 0 & 2k4^{k-1} \\ 0 & -2\cdot 4^k & -k4^k \\ 0 & 4^k & 4^k+2k4^{k-1} \end{pmatrix},$$

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

$$\mathcal{A}^k = \mathcal{A}_k\mathcal{V}^{-1} = \begin{pmatrix} 4^k & k4^{k-1} & 2k4^{k-1} \\ 0 & 4^k-2k4^{k-1} & -k4^k \\ 0 & k4^{k-1} & 4^k+2k4^{k-1} \end{pmatrix}.$$

The next example shows how to deal with complex eigenvalues. We recall that 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 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 the eigenvector $\mathbf{v}$ 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 $\lambda^k\mathbf{v}$ and $\bar{\lambda}^k\bar{\mathbf{v}}$. Since taking real and imaginary parts is a linear operations:

$$\Re(\lambda^k\mathbf{v}) = \frac{\lambda^k\mathbf{v} + \bar{\lambda}^k\bar{\mathbf{v}}}{2}, \qquad \Im(\lambda^k\mathbf{v}) = \frac{\lambda^k\mathbf{v} - \bar{\lambda}^k\bar{\mathbf{v}}}{2i},$$

both $\Re(\lambda^k\mathbf{v})$ and $\Im(\lambda^k\mathbf{v})$ are real valued solutions. To find explicit expressions for them we write $\lambda = re^{i\phi}$ where $r = |\lambda|$ and $\phi = Arg\lambda$. Then

$$\lambda^n = r^n e^{in\phi} = r^n(\cos n\phi + i\sin n\phi)$$

and

$$\Re(\lambda^n\mathbf{v}) = r^n(\cos n\phi\Re\mathbf{v} - \sin n\phi\Im\mathbf{v}),$$
$$\Im(\lambda^n\mathbf{v}) = r^n(\sin n\phi\Re\mathbf{v} + \cos n\phi\Im\mathbf{v}).$$

Example 3.9. Find $\mathcal{A}^k$ 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

$$\begin{pmatrix} 1-2i & -5 \\ 1 & -1-2i \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix};$$

thus

$$\mathbf{v^1} = \begin{pmatrix} 1+2i \\ 1 \end{pmatrix}$$

and

$$\mathbf{x}(k) = \mathcal{A}^n \mathbf{v^1} = (2i)^k \begin{pmatrix} 1+2i \\ 1 \end{pmatrix}.$$

To find real valued solutions, we have to take real and imaginary parts of $\mathbf{x}(k)$. Since $i = \cos\frac{\pi}{2} + i\sin\frac{\pi}{2}$ we have, by de Moivre's formula,

$$(2i)^k = 2^k \left(\cos\frac{\pi}{2} + i\sin\frac{\pi}{2}\right)^k = 2^k \left(\cos\frac{k\pi}{2} + i\sin\frac{k\pi}{2}\right).$$

Therefore

$$\begin{aligned} \Re\mathbf{x}(k) &= 2^k \left(\cos\frac{k\pi}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \sin\frac{k\pi}{2}\begin{pmatrix} 2 \\ 0 \end{pmatrix}\right) \\ \Im\mathbf{x}(k) &= 2^k \left(\cos\frac{k\pi}{2}\begin{pmatrix} 2 \\ 0 \end{pmatrix} + \sin\frac{k\pi}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right). \end{aligned}$$

The initial values for $\Re\mathbf{x}(k)$ and $\Im\mathbf{x}(k)$ are, respectively, $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 2 \\ 0 \end{pmatrix}$. Since $\mathcal{A}^k$ is a real matrix, we have $\Re\mathcal{A}^k\mathbf{v^1} = \mathcal{A}^k\Re\mathbf{v^1}$ and $\Im\mathcal{A}^k\mathbf{v^1} = \mathcal{A}^k\Im\mathbf{v^1}$, thus

$$\mathcal{A}^k\begin{pmatrix} 1 \\ 1 \end{pmatrix} = 2^k \left(\cos\frac{k\pi}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \sin\frac{k\pi}{2}\begin{pmatrix} 2 \\ 0 \end{pmatrix}\right) = 2^k \begin{pmatrix} \cos\frac{k\pi}{2} - 2\sin\frac{k\pi}{2} \\ \cos\frac{k\pi}{2} \end{pmatrix}$$

and

$$\mathcal{A}^k\begin{pmatrix} 2 \\ 0 \end{pmatrix} = 2^k \left(\cos\frac{k\pi}{2}\begin{pmatrix} 2 \\ 0 \end{pmatrix} + \sin\frac{k\pi}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) = 2^k \begin{pmatrix} 2\cos\frac{k\pi}{2} + \sin\frac{k\pi}{2} \\ \sin\frac{k\pi}{2} \end{pmatrix}.$$

To find $\mathcal{A}^k$ we use again (3.3.29). In our case

$$\mathcal{A}_k = 2^k \begin{pmatrix} \cos\frac{k\pi}{2} - 2\sin\frac{k\pi}{2} & 2\cos\frac{k\pi}{2} + \sin\frac{k\pi}{2} \\ \cos\frac{k\pi}{2} & \sin\frac{k\pi}{2} \end{pmatrix},$$

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

$$\mathcal{A}^k = \mathcal{A}_k\mathcal{V}^{-1} = -2^{k-1}\begin{pmatrix} -2\cos\frac{k\pi}{2} - \sin\frac{k\pi}{2} & 5\sin\frac{k\pi}{2} \\ -\sin\frac{k\pi}{2} & -2\cos\frac{k\pi}{2} + \sin\frac{k\pi}{2} \end{pmatrix}.$$

Systems of differential equations II.

Let us return to the system

$$\mathbf{y}' = \mathcal{A}\mathbf{y}. \tag{3.3.44}$$

As before, our goal is to find n linearly independent solutions of (3.3.44). For the solution matrix $\mathcal{E}(t)$ we do not have a natural expression as was the case for the difference system. If all eigenvalues are simple, then we have a sufficient number of eigenvector to define $\mathcal{E}(t)$ by (3.3.38). The same formula is valid if there are multiple eigenvalues but algebraic and geometric multiplicities of each eigenvalue are the same. However, it still remains to find a formula for $\mathcal{E}(t)$ when $\mathcal{A}$ has less than n linearly independent eigenvectors.

Recall that for a single equation $y' = ay$, where a is a constant, the general solution is given by $y(t) = e^{at}C$, where C is a constant. In a similar way, we would like to say that the general solution to (3.3.44) 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

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \ldots + \frac{x^k}{k!} + \ldots,$$

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

$$e^{\mathcal{A}} = \mathcal{I} + \mathcal{A} + \frac{1}{2}\mathcal{A}^2 + \frac{1}{3!}\mathcal{A}^3 + \ldots + \frac{1}{k!}\mathcal{A}^k + \ldots. \tag{3.3.45}$$

The problem is that the sum is infinite and we have to define what it means for a series of matrices to converge. This can be done but here we will avoid this problem by showing that, in fact, the sum in (3.3.45) can be always replaced by a finite sum. We note, however, that in some simple cases we can evaluate the infinite sum. 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

$$\begin{aligned} e^{\lambda\mathcal{I}} &= \mathcal{I} + \lambda\mathcal{I} + \frac{\lambda^2}{2}\mathcal{I} + \frac{\lambda^3}{3!}\mathcal{I} + \ldots + \frac{\lambda^k}{k!} + \ldots \\ &= \left(1 + \lambda + \frac{\lambda^2}{2} + \frac{\lambda^3}{3!} + \ldots + \frac{\lambda^k}{k!} + \ldots\right)\mathcal{I} \\ &= e^{\lambda}\mathcal{I}. \end{aligned} \tag{3.3.46}$$

Unfortunately, in most cases finding the explicit form for $e^{\mathcal{A}}$ directly is very difficult.

To justify algebraic manipulations below, we note that, in general, matrix exponentials have the following algebraic properties

$$\left(e^{\mathcal{A}}\right)^{-1} = e^{-\mathcal{A}}$$

and

$$e^{\mathcal{A}+\mathcal{B}} = e^{\mathcal{A}}e^{\mathcal{B}} \tag{3.3.47}$$

provided the matrices $\mathcal{A}$ and $\mathcal{B}$ commute: $\mathcal{AB} = \mathcal{BA}$. Furthermore, defining a function of t by

$$e^{t\mathcal{A}} = \mathcal{I} + t\mathcal{A} + \frac{t^2}{2}\mathcal{A}^2 + \frac{t^3}{3!}\mathcal{A}^3 + \ldots + \frac{t^k}{k!}\mathcal{A}^k + \ldots\ , \tag{3.3.48}$$

and formally differentiating it with respect to t we find, as in the scalar case, that

$$\begin{aligned}\frac{d}{dt}e^{t\mathcal{A}} &= \mathcal{A} + t\mathcal{A}^2 + \frac{t^2}{2!}\mathcal{A}^3 + \ldots + \frac{t^{k-1}}{(k-1)!}\mathcal{A}^k + \ldots \\ &= \mathcal{A}\left(\mathcal{I} + t\mathcal{A} + \frac{t^2}{2!}\mathcal{A}^2 + \ldots + \frac{t^{k-1}}{(k-1)!}\mathcal{A}^{k-1} + \ldots\right) \\ &= \mathcal{A}e^{t\mathcal{A}} = e^{t\mathcal{A}}\mathcal{A},\end{aligned}$$

proving thus that $y(t) = e^{t\mathcal{A}}\mathbf{v}$ is a solution to our system of equations for any constant vector $\mathbf{v}$ (provided, of course, that we can justify all the above operations in a rigorous way).

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 $e^{t\mathcal{A}}\mathbf{v}$ is finite. This is based on the following two observations. Firstly, since $\lambda\mathcal{I}$ and $\mathcal{A} - \lambda\mathcal{I}$ commute, we have by (3.3.46) and (3.3.47)

$$e^{t\mathcal{A}}\mathbf{v} = e^{t(\mathcal{A}-\lambda\mathcal{I})}e^{t\lambda\mathcal{I}}\mathbf{v} = e^{\lambda t}e^{t(\mathcal{A}-\lambda\mathcal{I})}\mathbf{v}.$$

Secondly, if $(\mathcal{A} - \lambda\mathcal{I})^m\mathbf{v} = \mathbf{0}$ for some m, then

$$(\mathcal{A} - \lambda\mathcal{I})^r\mathbf{v} = \mathbf{0}, \tag{3.3.49}$$

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 $\mathbf{v}$

$$e^{t(\mathcal{A}-\lambda\mathcal{I})}\mathbf{v} = \mathbf{v} + t(\mathcal{A} - \lambda\mathcal{I})\mathbf{v} + \ldots + \frac{t^{m-1}}{(m-1)!}(\mathcal{A} - \lambda\mathcal{I})^{m-1}\mathbf{v}.$$

and

$$e^{t\mathcal{A}}\mathbf{v} = e^{\lambda t}e^{t(\mathcal{A}-\lambda\mathcal{I})}\mathbf{v} = e^{\lambda t}\left(\mathbf{v} + t(\mathcal{A} - \lambda\mathcal{I})\mathbf{v} + \ldots + \frac{t^{m-1}}{(m-1)!}(\mathcal{A} - \lambda\mathcal{I})^{m-1}\mathbf{v}\right). \tag{3.3.50}$$

Thus, to find all solutions to $\mathbf{y}' = \mathcal{A}\mathbf{y}$ it is sufficient to find n independent vectors $\mathbf{v}$ satisfying (3.3.49) 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 $(\mathcal{A} - \lambda_1\mathcal{I})\mathbf{v^1} = 0$, thus m of (3.3.49) is equal to 1. Consequently, the sum in (3.3.50) terminates after the first term and we obtain

$$\mathbf{y_1}(t) = e^{\lambda_1 t}\mathbf{v^1}$$

in accordance with (3.3.37). 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 ν_i is less then 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_1.

Remark 3.10. Let us mention here that the exponential function $e^{t\mathcal{A}}$ has been introduced just as a guideline, to explain how the formula (3.3.50) was arrived at. Once we have this formula, we can directly check that it gives a solution to (3.3.44). Indeed,

$$\begin{aligned}
&\frac{d}{dt}e^{\lambda t}\left(\mathbf{v}+t(\mathcal{A}-\lambda\mathcal{I})\mathbf{v}+\ldots+\frac{t^{m-1}}{(m-1)!}(\mathcal{A}-\lambda\mathcal{I})^{m-1}\mathbf{v}\right)\\
&=\lambda e^{\lambda t}\left(\mathbf{v}+t(\mathcal{A}-\lambda\mathcal{I})\mathbf{v}+\ldots+\frac{t^{m-1}}{(m-1)!}(\mathcal{A}-\lambda\mathcal{I})^{m-1}\mathbf{v}\right)\\
&\quad+e^{\lambda t}\left((\mathcal{A}-\lambda\mathcal{I})\mathbf{v}+t(\mathcal{A}-\lambda\mathcal{I})^2\mathbf{v}\ldots+\frac{t^{m-2}}{(m-2)!}(\mathcal{A}-\lambda\mathcal{I})^{m-1}\mathbf{v}\right)\\
&=\lambda e^{\lambda t}\left(\mathbf{v}+t(\mathcal{A}-\lambda\mathcal{I})\mathbf{v}+\ldots+\frac{t^{m-1}}{(m-1)!}(\mathcal{A}-\lambda\mathcal{I})^{m-1}\mathbf{v}\right)\\
&\quad+e^{\lambda t}(\mathcal{A}-\lambda\mathcal{I})\left(\mathbf{v}+t(\mathcal{A}-\lambda\mathcal{I})^2\mathbf{v}\ldots+\frac{t^{m-2}}{(m-2)!}(\mathcal{A}-\lambda\mathcal{I})^{m-2}\mathbf{v}\right)\\
&=\lambda e^{\lambda t}\frac{t^{m-1}}{(m-1)!}(\mathcal{A}-\lambda\mathcal{I})^{m-1}\mathbf{v}\\
&\quad+e^{\lambda t}\mathcal{A}\left(\mathbf{v}+t(\mathcal{A}-\lambda\mathcal{I})^2\mathbf{v}\ldots+\frac{t^{m-2}}{(m-2)!}(\mathcal{A}-\lambda\mathcal{I})^{m-2}\mathbf{v}\right)\\
&=\mathcal{A}e^{\lambda t}\frac{t^{m-1}}{(m-1)!}(\mathcal{A}-\lambda\mathcal{I})^{m-1}\mathbf{v}-(\mathcal{A}-\lambda\mathcal{I})e^{\lambda t}\frac{t^{m-1}}{(m-1)!}(\mathcal{A}-\lambda\mathcal{I})^{m-1}\mathbf{v}\\
&\quad+e^{\lambda t}\mathcal{A}\left(\mathbf{v}+t(\mathcal{A}-\lambda\mathcal{I})^2\mathbf{v}\ldots+\frac{t^{m-2}}{(m-2)!}(\mathcal{A}-\lambda\mathcal{I})^{m-2}\mathbf{v}\right)\\
&\quad-e^{\lambda t}\frac{t^{m-1}}{(m-1)!}(\mathcal{A}-\lambda\mathcal{I})^{m}\mathbf{v}\\
&\quad+e^{\lambda t}\mathcal{A}\left(\mathbf{v}+t(\mathcal{A}-\lambda\mathcal{I})^2\mathbf{v}\ldots+\frac{t^{m-1}}{(m-1)!}(\mathcal{A}-\lambda\mathcal{I})^{m-1}\mathbf{v}\right)\\
&=e^{\lambda t}\mathcal{A}\left(\mathbf{v}+t(\mathcal{A}-\lambda\mathcal{I})^2\mathbf{v}\ldots+\frac{t^{m-1}}{(m-1)!}(\mathcal{A}-\lambda\mathcal{I})^{m-1}\mathbf{v}\right)
\end{aligned}$$

where we used $(A-\lambda\mathcal{I})^m\mathbf{v}=0$.

We illustrate the theory on a few examples.

Example 3.11. 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

$$-v_2 + 4v_3 = 0, \qquad 3v_1 + v_2 - v_3 = 0, \qquad 2v_1 + v_2 - 2v_3 = 0$$

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

$$-2v_1 - v_2 + 4v_3 = 0, \qquad 3v_1 - v_2 - v_3 = 0, \qquad 2v_1 + v_2 - 4v_3 = 0.$$

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

$$\mathbf{v^2} = C_2 \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$$

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

$$3v_1 - v_2 + 4v_3 = 0, \qquad 3v_1 + 4v_2 - v_3 = 0, \qquad 2v_1 + v_2 + v_3 = 0.$$

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

$$\mathbf{y}(t) = C_1 e^t \begin{pmatrix} -1 \\ 4 \\ 1 \end{pmatrix} + C_2 e^{3t} \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + C_3 e^{-2t} \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}.$$

If we single complex eigenvalue λ with eigenvector $\mathbf{v}$ then, as explained before Example 3.9, $\bar{\lambda}$ is also an eigenvalue with corresponding eigenvector $\bar{\mathbf{v}}$. Thus, we have two linearly independent (complex) solutions

$$\mathbf{z^1}(t) = e^{\lambda t}\mathbf{v}, \qquad \mathbf{z^2}(t) = e^{\bar{\lambda} t}\bar{\mathbf{v}} = \overline{\mathbf{z^1}}(t).$$

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

$$\begin{aligned} \mathbf{z^1}(t) &= e^{\lambda t}\mathbf{v} = e^{\xi t}(\cos\omega t + i \sin\omega t)(\Re\mathbf{v} + i\Im\mathbf{v}) \\ &= e^{\xi t}(\cos\omega t\, \Re\mathbf{v} - \sin\omega t\, \Im\mathbf{v}) + ie^{\xi t}(\cos\omega t\, \Im\mathbf{v} + \sin\omega t\, \Re\mathbf{v}) \\ &= \mathbf{y^1}(t) + i\mathbf{y^2}(t) \end{aligned}$$

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 the we can use two real linearly independent solutions

$$\begin{aligned} \mathbf{y^1}(t) &= e^{\xi t}(\cos\omega t\, \Re\mathbf{v} - \sin\omega t\, \Im\mathbf{v}) \\ \mathbf{y^2}(t) &= e^{\xi t}(\cos\omega t\, \Im\mathbf{v} + \sin\omega t\, \Re\mathbf{v}) \end{aligned} \tag{3.3.51}$$

Example 3.12. Solve the initial value problem

$$\mathbf{y}' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & 1 \end{pmatrix} \mathbf{y}, \qquad \mathbf{y}(0) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

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

$$-iv_1 = 0, \qquad -iv_2 - v_3 = 0, \qquad v_2 - iv_3 = 0.$$

The first equation gives $v_1 = 0$ and the other two yield $v_2 = iv_3$ so that each vector

$$\mathbf{v^2} = C_2 \begin{pmatrix} 0 \\ i \\ 1 \end{pmatrix}$$

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:

$$\begin{aligned} e^{(1+i)t} \begin{pmatrix} 0 \\ i \\ 1 \end{pmatrix} &= e^t(\cos t + i\sin t)\left(\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + i\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}\right) \\ = e^t &\left(\cos t \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} - \sin t \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + i\sin t \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + i\cos t \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}\right) \\ = e^t &\begin{pmatrix} 0 \\ -\sin t \\ \cos t \end{pmatrix} + ie^t \begin{pmatrix} 0 \\ \cos t \\ \sin t \end{pmatrix}. \end{aligned}$$

Thus, we obtain two real solutions

$$\mathbf{y^1}(t) = e^t \begin{pmatrix} 0 \\ -\sin t \\ \cos t \end{pmatrix}$$
$$\mathbf{y^2}(t) = e^t \begin{pmatrix} 0 \\ \cos t \\ \sin t \end{pmatrix}$$

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

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \qquad \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix},$$

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

$$\mathbf{y}(t) = e^t\begin{pmatrix}1\\0\\0\end{pmatrix} + e^t\begin{pmatrix}0\\-\sin t\\\cos t\end{pmatrix} + e^t\begin{pmatrix}0\\\cos t\\\sin t\end{pmatrix} = e^t\begin{pmatrix}1\\\cos t-\sin t\\\cos t+\sin t\end{pmatrix}.$$

The last example deals with multiple eigenvalues.

Example 3.13. 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

$$\mathbf{y^1}(t) = C_1e^t\begin{pmatrix}1\\0\\0\end{pmatrix}.$$

However, this is only one solution and $\lambda_1 = 1$ has algebraic multiplicity 2, so we have to look for one more solution. To this end we consider

$$\begin{aligned}(\mathcal{A}-1\mathcal{I})^2\mathbf{v} &= \begin{pmatrix}0\,1\,0\\0\,0\,0\\0\,0\,1\end{pmatrix}\begin{pmatrix}0\,1\,0\\0\,0\,0\\0\,0\,1\end{pmatrix}\begin{pmatrix}v_1\\v_2\\v_3\end{pmatrix}\\&= \begin{pmatrix}0\,0\,0\\0\,0\,0\\0\,0\,1\end{pmatrix}\begin{pmatrix}v_1\\v_2\\v_3\end{pmatrix} = \begin{pmatrix}0\\0\\0\end{pmatrix}\end{aligned}$$

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.

3.3 Higher order difference and differential equations

Once we know how to solve systems of difference and differential equations, it is easy to adopt the theory to cater for higher order scalar equations.

First consider the linear difference equation of order n:

$$y(k+n) + a_1 y(k+n-1) + \ldots + a_n y(k) = 0, \qquad n \geq 0 \tag{3.3.52}$$

where $a_1, \ldots, a_n$ are known numbers. This equation determines the values of $y(N)$, $N > n$ by n preceding values of $y(k)$. Thus, it is clear that to be able to solve this equation, that is, to start the recurrence procedure, we need n initial values $y(0), y(1), \ldots, y(n-1)$. Equation (3.3.52) can be written as a system of first order equations of dimension n. We let

$$\begin{aligned} z_1(k) &= y(k),\\ z_2(k) &= y(k+1) = z_1(k+1),\\ z_3(k) &= y(k+2) = z_2(k+1),\\ &\vdots\ \vdots\ \vdots,\\ z_n(k) &= y(k+n-1) = z_{n-1}(k-1), \end{aligned} \tag{3.3.53}$$

hence we obtain the system

$$\begin{aligned} z_1(k+1) &= z_2(k),\\ z_2(k+1) &= z_3(k),\\ &\vdots\ \vdots\ \vdots,\\ z_{n-1}(k+1) &= z_n(k),\\ z_n(k+1) &= -a_n z_1(k) - a_2 z_2(k) \ldots - a_1 z_n(k), \end{aligned}$$

or, in matrix notation,

$$\mathbf{z}(k+1) = \mathcal{A}\mathbf{z}(k)$$

where $\mathbf{z} = (z_1, \ldots, z_n)$, and

$$\mathcal{A} = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -a_n & -a_{n-1} & -a_{n-2} & \ldots & -a_1 \end{pmatrix}.$$

The matrix $\mathcal{A}$ often is called the companion matrix of the equation (3.3.52). It is clear that the initial values $y(0), \ldots, y(n-1)$ give the initial vector $\mathbf{z}^0 = (y(0), \ldots, y(n-1))$. Next we observe that the eigenvalues of $\mathcal{A}$ can be obtained by solving the equation

$$p_n(\lambda) = \begin{vmatrix} -\lambda & 1 & 0 & \ldots & 0 \\ 0 & -\lambda & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -a_n & -a_{n-1} & -a_{n-2} & \ldots & -a_1 - \lambda \end{vmatrix} = 0.$$

It is not obvious how to directly calculate the determinant. Here we present one method which is related to the difference equation. Later, is Remark 3.28, we present another way which is more related to the interpretation of the problem. Expanding first the determinant along the first row, we find

$$\begin{aligned} p_n(\lambda) &= -\lambda \begin{vmatrix} -\lambda & 1 & 0 & \ldots & 0 \\ 0 & -\lambda & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -a_{n-1} & -a_{n-2} & -a_{n-3} & \ldots & -a_1 - \lambda \end{vmatrix} - \begin{vmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -a_n & -a_{n-2} & -a_{n-3} & \ldots & -a_1 - \lambda \end{vmatrix} \\ &= -\lambda p_{n-1}(\lambda) + (-1)^n a_n. \end{aligned}$$

This is a linear difference equation and we can solve it using (2.3.49). Taking into account that $p_1(\lambda) = -a - 1 - \lambda$ and adjusting (2.3.49) to cater for the fact that we start with $n = 1$, we obtain

$$p_n(\lambda) = (-\lambda)^{n-1}(a_1 + \lambda) + \sum_{k=1}^{n-1} (-\lambda)^{n-k-1}(-1)^{k+1} a_{k+1} = (-1)^n(\lambda^n + \lambda^{n-1}a_1 + \cdots + a_n). \quad (3.3.54)$$

We note that the characteristic polynomial of the companion matrix can be obtained by just replacing $y(k+n-i)$ in (3.3.52) by λ^{n-i}, $i = 0, \ldots, n$. 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 an $n \times n$ system we have to construct n linearly independent vectors $\mathbf{v^1}, \ldots, \mathbf{v^n}$ so that the solution is given by $\mathbf{z^1}(k) = \mathcal{A}^k\mathbf{v^1}, \ldots \mathbf{z^n}(k) = \mathcal{A}^k\mathbf{v^n}$ and coordinates of each $\mathbf{z^i}$ are products of λ_i and polynomials in k of degree strictly smaller than the algebraic multiplicity of λ_i. Thus, to obtain n_i solutions of the higher order equation corresponding to the eigenvalue λ_i, by (3.3.53), we take only the first coordinates of all $\mathbf{z^i}(k)$ 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 $\{\lambda_i^k, k\lambda_i^k, \ldots, k^{n_i-1}\lambda_i^k\}$ 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 3.14. Consider the Fibonacci equation (3.1.1):

$$y(k+2) = y(k+1) + y(k) \quad (3.3.55)$$

to be consistent with the notation of the present chapter. Introducing new variables $z_1(k) = y(k), z_2(k) = y(k+1) = z_1(k+1)$ so that $y(k+2) = z_2(k+1)$, we re-write the equation as the system

$$\begin{aligned} z_1(k+1) &= z_2(k), \\ z_2(k+1) &= z_1(k) + z_2(k); \end{aligned}$$

note that it is not the same form as (3.1.3). 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 (3.3.55) is given by

$$y(n) = c_1 \left(\frac{1+\sqrt{5}}{2}\right)^n + c_2 \left(\frac{1-\sqrt{5}}{2}\right)^n. \tag{3.3.56}$$

Let us find the particular solution satisfying the initial conditions $y(0) = 1$, $y(1) = 2$ (corresponding to one pair of adult 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 $c_2 = -3\sqrt{5}/5$.

3.4 Spectral Decomposition.

If $\mathbf{v}$ is an eigenvector of a matrix $\mathcal{A}$ corresponding to an eigenvalue λ, then the one dimensional eigenspace space $\tilde{E}_\lambda$ has an important property of being *invariant* under $\mathcal{A}$ as well as under $\mathcal{A}^k$ and $e^{t\mathcal{A}}$; that is, if $\mathbf{y} \in \tilde{E}_\lambda$, then $\mathcal{A}\mathbf{y} \in \tilde{E}_\lambda$ (and $\mathcal{A}^k\mathbf{y}, e^{t\mathcal{A}}\mathbf{y} \in \tilde{E}_\lambda$ for all $k = 1, 2, \ldots$ and $t > 0$). In fact, in this case, $\mathbf{y} = \alpha\mathbf{v}$ for some $\alpha \in \mathbf{R}$ and

$$\mathcal{A}\mathbf{y} = \alpha\mathcal{A}\mathbf{v} = \alpha\lambda\mathbf{v} \in \tilde{E}_\lambda.$$

Similarly, $\mathcal{A}^k\mathbf{y} = \lambda^k\alpha\mathbf{v} \in \tilde{E}_\lambda$ and $e^{t\mathcal{A}}\mathbf{y} = e^{\lambda t}\alpha\mathbf{v} \in \tilde{E}_\lambda$. Thus, if $\mathcal{A}$ is diagonalizable, then the evolution governed by $\mathbf{A}$ can be decomposed into n independent scalar evolutions occurring in eigenspaces of $\mathcal{A}$. The situation is more complicated when we have multiple eigenvalues as the one dimensional spaces spanned by generalized eigenvectors are not invariant under $\mathcal{A}$. However, we can show that the each generalized eigenspace spanned by all eigenvectors and generalized eigenvectors corresponding to the same eigenvalue is invariant under $\mathcal{A}$.

We start with the following property of E_{λ_i} which is important in this context.

Lemma 3.15. *Let $E_{\lambda_i} = Span\{\mathbf{v}^1, \ldots, \mathbf{v}^{n_i}\}$ be the generalized eigenspace corresponding to an eigenvalue λ_i and $\mathbf{v}^r$ satisfies*

$$(\mathcal{A} - \lambda_i\mathcal{I})^k\mathbf{v}^r = 0,$$

for some $1 < k < n_i$, while $(\mathcal{A} - \lambda_i\mathcal{I})^{k-1}\mathbf{v}^r = 0$. Then $\mathbf{v}^r$ satisfies

$$(\mathcal{A} - \lambda_i\mathcal{I})\mathbf{v}^r = \mathbf{v}^{r'}, \tag{3.3.57}$$

where $(\mathcal{A} - \lambda_i\mathcal{I})^{k-1}\mathbf{v}^{r'} = 0$ and

$$(\mathcal{A} - \lambda_i\mathcal{I})^{k-1}\mathbf{v}^r = \mathbf{v}^{r'}, \tag{3.3.58}$$

where $\mathbf{v}^{r'}$ is an eigenvector.

Proof. Let $E_{\lambda_i} = Span\{\mathbf{v}^1, \ldots, \mathbf{v}^{n_j}\}$ be grouped so that the first ν_i elements: $\{\mathbf{v}^1, \ldots, \mathbf{v}^{\nu_i}\}$ are the eigenvectors, $\{\mathbf{v}^\rho\}_{\nu_i+1\leq\rho\leq r'}$ satisfy $(\mathcal{A}-\lambda\mathcal{I})^2\mathbf{v}^\rho = 0$, etc. Then $\mathbf{v}^\rho$, $\nu_i+1\leq\rho\leq r'$ satisfies

$$0 = (\mathcal{A}-\lambda\mathcal{I})^2\mathbf{v}^\rho = (\mathcal{A}-\lambda\mathcal{I})((\mathcal{A}-\lambda\mathcal{I})\mathbf{v}^\rho).$$

Since $\mathbf{v}^\rho$ is not an eigenvector, $0 \neq (\mathcal{A}-\lambda\mathcal{I})\mathbf{v}^\rho$ must be an eigenvector so that any $\mathbf{v}^\rho$ with $\nu_i+1\leq\rho\leq r'$ satisfies (after possibly multiplication by a scalar)

$$(\mathcal{A}-\lambda\mathcal{I})\mathbf{v}^\rho = \mathbf{v}^j$$

for some eigenvector $\mathbf{v}^j$, $j\leq\nu_i$. If $r' < n_i$, then the elements from the next group, $\{\mathbf{v}^\rho\}_{r'+1\leq\rho\leq r''}$ satisfy

$$0 = (\mathcal{A}-\lambda\mathcal{I})^3\mathbf{v}^\rho = (\mathcal{A}-\lambda\mathcal{I})(\mathcal{A}-\lambda\mathcal{I})^2\mathbf{v}^\rho \tag{3.3.59}$$

and since $\mathbf{v}^\rho$ in this range does not satisfy $(\mathcal{A}-\lambda\mathcal{I})^2\mathbf{v}^\rho = 0$, we may put

$$(\mathcal{A}-\lambda\mathcal{I})^2\mathbf{v}^\rho = \mathbf{v}^j \tag{3.3.60}$$

for some $1\leq j\leq\nu_i$; that is, for some eigenvector $\mathbf{v}^j$. Alternatively, we can write (3.3.59) as

$$(\mathcal{A}-\lambda\mathcal{I})^2(\mathcal{A}-\lambda\mathcal{I})\mathbf{v}^\rho = 0$$

and since $\mathbf{v}^\rho$ is not an eigenvector,

$$(\mathcal{A}-\lambda\mathcal{I})\mathbf{v}^\rho = v^{\rho'} \tag{3.3.61}$$

for some ρ' between ν_i+1 and r'. By induction, we obtain a basis of E_λ consisting of vectors satisfying (3.3.60) where on the right-hand side stands a vector of the basis constructed in the previous cycle. □

An important corollary of this lemma is

Corollary 3.16. *Each generalized eigenspace E_{λ_i} of $\mathcal{A}$ is invariant under $\mathcal{A}$; that is, for any $\mathbf{v}\in E_{\lambda_i}$ we have $\mathcal{A}\mathbf{v}\in E_{\lambda_i}$. It is also invariant under $\mathcal{A}^k, k=1,2,\ldots$ and $e^{t\mathcal{A}}, t>0$.*

Proof. We use the representation of E_{λ_i} obtained in the previous lemma. Indeed, let $\mathbf{x} = \sum_{j=1}^{n_i} a_j\mathbf{v}^j$ be an arbitrary element of E_{λ_i}. Then

$$(\mathcal{A}-\lambda_i\mathcal{I})\mathbf{x} = \sum_{j=1}^{n_i} a_j(\mathcal{A}-\lambda_i\mathcal{I})\mathbf{v}^j$$

and, by construction, $(\mathcal{A}-\lambda_i\mathcal{I})\mathbf{v}^j = \mathbf{v}^{j'}$ for some $j'<j$ (belonging to the previous 'cycle'). In particular, $(\mathcal{A}-\lambda_i\mathcal{I})\mathbf{v}^j = 0$ for $1\leq j\leq\nu_i$ (eigenvectors). Thus

$$\mathcal{A}\mathbf{x} = \lambda\mathbf{x} - \sum_{j'>\nu_i} a_{j'}\mathbf{v}^{j'} \in E_\lambda,$$

which ends the proof of the first part.

From the first part, by induction, we obtain that $(\mathcal{A}-\lambda_i\mathcal{I})^k E_{\lambda_i} \subset E_{\lambda_i}$. In fact, let $\mathbf{x}\in E_{\lambda_i}$ and assume $(\mathcal{A}-\lambda_i\mathcal{I})^{k-1}\mathbf{x}\in E_{\lambda_i}$. Then $(\mathcal{A}-\lambda_i\mathcal{I})^k\mathbf{x} = (\mathcal{A}-\lambda_i\mathcal{I})(\mathcal{A}-\lambda_i\mathcal{I})^{k-1}\mathbf{x}\in E_{\lambda_i}$ by the induction assumption and the first part.

For $\mathcal{A}^k$ we have

$$\begin{aligned}\mathcal{A}^k\mathbf{x} = (\mathcal{A}-\lambda_i\mathcal{I}+\lambda_i\mathcal{I})^k\mathbf{x} &= \sum_{j=1}^{n_i} a_j(\mathcal{A}-\lambda_i\mathcal{I}+\lambda_i\mathcal{I})^k\mathbf{v}^j\\ &= \sum_{j=1}^{n_i} a_j\sum_{r=0}^{k}\lambda_i^{k-r}\binom{k}{r}(\mathcal{A}-\lambda_i\mathcal{I})^r\mathbf{v^j}\end{aligned}$$

where the inner sum must terminate at at most $n_i - 1$ term since $\mathbf{v}^j$ are determined by solving $(\mathcal{A} - \lambda\mathcal{I})^\nu \mathbf{v} = 0$ with ν being at most equal to n_i. From the previous part of the proof we see that $(\mathcal{A} - \lambda_i\mathcal{I})^r \mathbf{v^j} \in E_{\lambda_i}$ and thus $\mathcal{A}^k\mathbf{x}$.

The same argument works for $e^{t\mathcal{A}}$. Indeed, for $\mathbf{x} \in E_{\lambda_i}$ and using (3.3.50) we obtain

$$e^{t\mathcal{A}}\mathbf{x} = e^{\lambda_i t}\sum_{j=1}^{n_i} a_j e^{t(\mathcal{A}-\lambda\mathcal{I})}\mathbf{v^j} = e^{\lambda_i t}\sum_{j=1}^{n_i} a_j \sum_{r=0}^{r_j} \frac{t^{r-1}}{(r-1)!}(\mathcal{A}-\lambda\mathcal{I})^{r-1}\mathbf{v^j}. \tag{3.3.62}$$

with $r_j \leq n_i$ and the conclusion follows as above. □

This result suggests that the the evolution governed by $\mathcal{A}$ in both discrete and continuous case can be broken into several simpler and independent pieces occurring in each generalized eigenspace. To write this in proper mathematical terms, we need to introduce some notation.

Let us recall that we have representations

$$\mathcal{A}^k \overset{\circ}{\mathbf{x}}= \begin{pmatrix} | & \dots & | \\ \mathcal{A}^k\mathbf{v^1} & \dots & \mathcal{A}^k\mathbf{v^n} \\ | & \dots & | \end{pmatrix} \mathcal{V}^{-1} \overset{\circ}{\mathbf{x}} \tag{3.3.63}$$

and

$$e^{t\mathcal{A}} \overset{\circ}{\mathbf{x}}= \begin{pmatrix} | & \dots & | \\ e^{t\mathcal{A}}\mathbf{v^1} & \dots & e^{t\mathcal{A}}\mathbf{v^n} \\ | & \dots & | \end{pmatrix} \mathcal{V}^{-1} \overset{\circ}{\mathbf{x}}, \tag{3.3.64}$$

where

$$\mathcal{V} = \begin{pmatrix} | & \dots & | \\ \mathbf{v^1} & \dots & \mathbf{v^n} \\ | & \dots & | \end{pmatrix}. \tag{3.3.65}$$

Following our considerations, we select the vectors $\mathbf{v^1}, \dots, \mathbf{v^n}$ to be eigenvectors and generalized eigenvectors of $\mathcal{A}$ as then the entries of the solution matrices can be evaluated explicitly with relative ease. We want to split these expressions into generalized eigenspaces.

Let us introduce the matrix

$$\mathcal{P}_i = \begin{pmatrix} 0 \dots & | & \dots 0 \\ 0 \dots & \mathbf{v^i} & \dots 0 \\ 0 \dots & | & \dots 0 \end{pmatrix} \begin{pmatrix} | & \dots & | \\ \mathbf{v^1} & \dots & \mathbf{v^n} \\ | & \dots & | \end{pmatrix}^{-1}. \tag{3.3.66}$$

and note that, for $\mathbf{x} = c_1\mathbf{v^1} + \dots + c_n\mathbf{v^n}$, $\mathcal{P}_i\mathbf{x} = c_i\mathbf{v^i}$; that is, $\mathcal{P}_i$ selects the part of $\mathbf{x}$ along $\mathbf{v^i}$. It is easy to see, that

$$\mathcal{P}_i^2 = \mathcal{P}_i, \qquad \mathcal{P}_i\mathcal{P}_j = 0, \tag{3.3.67}$$

Matrices with such properties are called *projections*; in particular $\mathcal{P}_i$ is a projection onto $\mathbf{v^i}$. Clearly,

$$\mathcal{I} = \sum_{i=1}^{n} \mathcal{P}_i,$$

however, $\mathcal{A}\mathcal{P}_i\mathbf{x} = c_i\mathcal{A}\mathbf{v^i}$ is in the span of $\mathbf{v^i}$ only if $\mathbf{v^i}$ is an eigenvector. Thus, as we said earlier, this decomposition is not useful unless all $\mathbf{v^i}$s are eigenvectors.

On the other hand, if we consider operators

$$\mathcal{P}_{\lambda_i} = \sum_{j;\ \mathbf{v^j}\in E_{\lambda_i}} \mathcal{P}_j, \tag{3.3.68}$$

where $\mathcal{P}_i$, then such operators again will be projections. This follows from (3.3.67) by termwise multiplication. They are called *spectral projections*. Let $\sigma(\mathcal{A})$ denotes the set of all eigenvalues of $\mathcal{A}$, called the *spectrum* of $\mathcal{A}$. The decomposition

$$\mathcal{I} = \sum_{\lambda\in\sigma(\mathcal{A})} \mathcal{P}_\lambda, \tag{3.3.69}$$

is called the *spectral resolution of identity.*

In particular, if all eigenvalues are simple (or semi-simple), we obtain the spectral decomposition of $\mathcal{A}$ in the form

$$\mathcal{A} = \sum_{\lambda\in\sigma(\mathcal{A})} \lambda\mathcal{P}_\lambda,$$

and, for $\mathcal{A}^k$ and $e^{t\mathcal{A}}$,

$$\mathcal{A}^k = \sum_{\lambda\in\sigma(\mathcal{A})} \lambda^k\mathcal{P}_\lambda, \tag{3.3.70}$$

and

$$e^{t\mathcal{A}} = \sum_{\lambda\in\sigma(\mathcal{A})} e^{\lambda t}\mathcal{P}_\lambda, \tag{3.3.71}$$

which is another way of writing (3.3.30) and (3.3.38), respectively.

In general case, we use (3.3.69) to write

$$\mathcal{A}\mathbf{x} = \sum_{\lambda\in\sigma(A)} \mathcal{A}\mathcal{P}_\lambda\mathbf{x}, \tag{3.3.72}$$

where, by Corollary 3.16, we have $\mathcal{A}\mathcal{P}_\lambda\mathbf{x} \in E_\lambda$. Thus, using (3.3.67), we get $\mathcal{P}_{\lambda_i}\mathcal{A}\mathcal{P}_{\lambda_j} = 0$ for $i \neq j$. Using (3.3.68) and we obtain

$$\mathcal{P}_\lambda\mathcal{A}\mathbf{x} = \mathcal{P}_\lambda\mathcal{A}\mathcal{P}_\lambda\mathbf{x} = \mathcal{A}\mathcal{P}_\lambda\mathbf{x}.$$

Thus, (3.3.72) defines a decomposition of the action of $\mathcal{A}$ into non-overlapping subspaces E_λ, $\lambda \in \sigma(\mathcal{A})$, which is called the *spectral decomposition* of $\mathcal{A}$.

To give spectral decomposition of $\mathcal{A}^k$ and $e^{t\mathcal{A}}$, generalizing (3.3.70) and (3.3.71), we observe that, by Corollary 3.16, also $\mathcal{A}^k\mathcal{P}_\lambda\mathbf{x} \in E_\lambda$ and $e^{t\mathcal{A}}\mathcal{P}_\lambda\mathbf{x} \in E_\lambda$. Therefore

$$\mathcal{A}^k\mathbf{x} = \sum_{\lambda\in\sigma(\mathcal{A})} \mathcal{A}^k\mathcal{P}_\lambda\mathbf{x} = \sum_{\lambda\in\sigma(\mathcal{A})} \lambda^k\mathbf{p}_\lambda(k)\mathbf{x}, \tag{3.3.73}$$

and

$$e^{t\mathcal{A}}\mathbf{x} = \sum_{\lambda\in\sigma(\mathcal{A})} e^{\lambda t}\mathcal{P}_\lambda\mathbf{x} = \sum_{\lambda\in\sigma(\mathcal{A})} e^{\lambda t}\mathbf{q}_\lambda(t)\mathbf{x}, \tag{3.3.74}$$

where $\mathbf{p}_\lambda$ and $\mathbf{q}_\lambda$ are polynomials in k and, respectively, in t, of degree strictly smaller than the algebraic multiplicity of λ, and with vector coefficients being linear combinations of eigenvectors and associated eigenvectors corresponding to λ.

Returning to our main problem, that is, to the long time behaviour of iterates $\mathcal{A}^k$ and the exponential function $e^{t\mathcal{A}}$, then, from (3.3.73) and (3.3.74), we see that on each eigenspace the long time behaviour of $\mathcal{A}^k$ (respectively, of $e^{t\mathcal{A}}$) is determined by λ^n (respectively, $e^{t\lambda}$), possibly multiplied by a polynomial of degree smaller than the algebraic multiplicity of λ.

The situation observed in Examples 3.4 and 3.6 corresponds to the situation when there is a real positive simple eigenvalue, say, λ_1 satisfying $\lambda_1 > |\lambda|$ in discrete time, or $\lambda_1 > \Re\lambda$ in continuous time, for any other λ. Such an eigenvalue is called the *principal* or *dominant* eigenvalue. In such a case, for any initial condition $\overset{\circ}{\mathbf{x}}$ for which $\mathcal{P}_{\lambda_1}\overset{\circ}{\mathbf{x}}\neq 0$, we have

$$\mathcal{A}^k \overset{\circ}{\mathbf{x}}\approx c_1\lambda_1^k\mathbf{v}^\mathbf{1}$$

for large k in discrete time or, in continuous time,

$$e^{t\mathcal{A}} \overset{\circ}{\mathbf{x}}\approx c_1e^{\lambda_1 t}\mathbf{v}^\mathbf{1},$$

for large t. In such a case the vector $\mathbf{v^1}$ is called a *stable age structure*. An important question is to determine c_1 (an possibly other coefficients of the spectral decomposition). Clearly, $c_1\mathbf{v^1} = \mathcal{P}_1$ but the definition of $\mathcal{P}_i$ involves knowing all eigenvectors and associated eigenvectors of $\mathcal{A}$ and thus is not particularly handy. Here we shall describe a simpler method.

Let us recall that the transposed matrix $\mathcal{A}^*$ satisfies

$$< \mathcal{A}^*\mathbf{x}^*, \mathbf{y} >=< \mathbf{x}^*, \mathcal{A}\mathbf{y} >$$

where $< \mathbf{x}^*, \mathbf{y} >= \mathbf{x}^* \cdot \mathbf{y} = \sum_{i=1}^n x_i^* y_i$ Matrices $\mathcal{A}$ and $\mathcal{A}^*$ have the same eigenvalues and, though eigenvectors and associated eigenvectors are different (unless $\mathcal{A}$ is symmetric), the structure of the generalized eigenspaces corresponding to the same eigenvalue is identical (that is, the geometric multiplicities of λ are equal and we have the same number of associated eigenvectors solving $(\mathcal{A} - \lambda\mathcal{I})^\nu\mathbf{v} = 0$ and $(\mathcal{A}^* - \lambda\mathcal{I})^\nu\mathbf{v}^* = 0$). This follows from the fact that determinant, nullity and rank of a matrix and its transpose are the same.

Theorem 3.17. *Let E_λ and $E^*_{\lambda^*}$ be generalized eigenspaces of, respectively, $\mathcal{A}$ and $\mathcal{A}^*$, corresponding to different eigenvalues: $\lambda \neq \lambda^*$. If $\mathbf{v}^* \in E^*_{\lambda^*}$ and $\mathbf{v} \in E_\lambda$, then*

$$< \mathbf{v}^*, \mathbf{v} >= 0 \tag{3.3.75}$$

Proof. We can assume that $\lambda^* \neq 0$ since, if $\lambda^* = 0$, then $\lambda \neq 0$ and we can repeat the calculations below starting with λ instead of λ^*. We begin with $\mathbf{v} \in E_\lambda$ and $\mathbf{v}^* \in E^*_{\lambda^*}$ being eigenvectors. Then

$$< \mathbf{v}^*, \mathbf{v} >= \frac{1}{\lambda^*} < \mathcal{A}^*\mathbf{v}^*, \mathbf{v} >= \frac{1}{\lambda^*} < \mathbf{v}^*, \mathcal{A}\mathbf{v} >= \frac{\lambda}{\lambda^*} < \mathbf{v}^*, \mathbf{v} > .$$

Thus, $\left(\frac{\lambda}{\lambda^*} - 1\right) < \mathbf{v}^*, \mathbf{v} >= 0$ and, since $\lambda \neq \lambda^*$, we must have $< \mathbf{v}^*, \mathbf{v} >= 0$. Next we assume, that $\mathbf{v}^*$ is an eigenvector and $\mathbf{v}$ is an associated eigenvector which solves $(\mathcal{A} - \lambda\mathcal{I})^k\mathbf{v} = 0$ with $k > 1$. Then, by Lemma 3.15, $(\mathcal{A} - \lambda\mathcal{I})\mathbf{v} = \mathbf{v}'$, where $(\mathcal{A} - \lambda\mathcal{I})^{k-1}\mathbf{v}' = 0$. We adopt induction assumption that $< \mathbf{v}^*, \mathbf{v}' >= 0$ for any $\mathbf{v}'$ which satisfy $(\mathcal{A} - \lambda\mathcal{I})^{k-1}\mathbf{v}' = 0$. Then, as above

$$< \mathbf{v}^*, \mathbf{v} >= \frac{1}{\lambda^*} < \mathbf{v}^*, \mathcal{A}\mathbf{v} >= \frac{1}{\lambda^*} < \mathbf{v}^*, \lambda\mathbf{v} + \mathbf{v}' >= \frac{\lambda}{\lambda^*} < \mathbf{v}^*, \mathbf{v} > .$$

and the proof follows as before. Finally, let $(\mathcal{A}^* - \lambda^*\mathcal{I})^k\mathbf{v}^* = 0$ with $k > 1$. Then $\lambda^*\mathbf{v}^* = \mathcal{A}^*\mathbf{v}^* - \tilde{\mathbf{v}}^*$, where $(\mathcal{A}^* - \lambda^*\mathcal{I})^{k-1}\tilde{\mathbf{v}}^* = 0$. We can adopt the induction assumption that $< \tilde{\mathbf{v}}^*, \mathbf{v} >= 0$ for any $\tilde{\mathbf{v}}^*$ satisfying $(\mathcal{A}^* - \lambda^*\mathcal{I})^{k-1}\tilde{\mathbf{v}}^* = 0$. Then

$$\begin{aligned} < \mathbf{v}^*, \mathbf{v} > &= \frac{1}{\lambda^*} < \lambda^*\mathbf{v}^*, \mathbf{v} >= \frac{1}{\lambda^*} < \mathcal{A}^*\mathbf{v}^* - \tilde{\mathbf{v}}^*, \mathbf{v} >= \frac{1}{\lambda^*} < \mathcal{A}^*\mathbf{v}^*, \mathbf{v} > \\ &= \frac{\lambda}{\lambda^*} < \mathbf{v}^*, \mathcal{A}\mathbf{v} > . \end{aligned}$$

□

Summarizing, to determine a long time behaviour of a population described by either discrete $\mathbf{y}(k+1) = \mathcal{A}\mathbf{y}$ or continuous system $\mathbf{y}' = \mathcal{A}\mathbf{y}$ we have to

1. Find eigenvalues of $\mathcal{A}$ and determine whether there is the dominant eigenvalue, that is, a simple real eigenvalue, say, λ_1 satisfying $\lambda_1 > |\lambda|$ in discrete time, or $\lambda_1 > \Re\lambda$ in continuous time, for any other λ.
2. If this is the case, we find the eigenvector $\mathbf{v}$ of $\mathcal{A}$ and $\mathbf{v}^*$ of $\mathcal{A}^*$ corresponding to λ_1.
3. The long time behaviour of the population is then described by

   $$\mathcal{A}^k\mathbf{x} \approx \lambda_1^k < \mathbf{v}^*, \mathbf{x} > \mathbf{v} \tag{3.3.76}$$

   for large k in discrete time or, in continuous time, by

   $$e^{t\mathcal{A}}\mathbf{x} \approx e^{\lambda_1 t} < \mathbf{v}^*, \mathbf{x} > \mathbf{v} \tag{3.3.77}$$

   for large time, for any initial distribution of the population satisfying $< \mathbf{v}^*, \mathbf{x} >\neq 0$.

We illustrate this result by finding the long time behaviour of solutions to the system discussed in Example 3.11.

Example 3.18. Consider

$$\mathbf{y}' = \begin{pmatrix} 1 & -1 & 4 \\ 3 & 2 & -1 \\ 2 & 1 & -1 \end{pmatrix} \mathbf{y}.$$

The eigenvalues of $\mathcal{A}$ are $\lambda_1 = 1$, $\lambda_2 = 3$ and $\lambda_3 = -2$. We found eigenvectors corresponding to this eigenvalues to be

$$\mathbf{v^1} = \begin{pmatrix} -1 \\ 4 \\ 1 \end{pmatrix}, \quad \mathbf{v^2} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \quad \mathbf{v^3} = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix},$$

and the general solution

$$\mathbf{y}(t) = C_1 e^t \begin{pmatrix} -1 \\ 4 \\ 1 \end{pmatrix} + C_2 e^{3t} \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + C_3 e^{-2t} \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}.$$

Clearly, writing

$$\mathbf{y}(t) = e^{3t} \left(C_2 \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + C_1 e^{-2t} \begin{pmatrix} -1 \\ 4 \\ 1 \end{pmatrix} + C_3 e^{-5t} \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} \right). \tag{3.3.78}$$

we see that the dominant eigenvalue is $\lambda_2 = 3$ and for large time

$$\mathbf{y}(t) \approx e^{3t} C_2 \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \tag{3.3.79}$$

where C_2 depends on the initial condition.

The transposed matrix is given by

$$\mathcal{A}^* = \begin{pmatrix} 1 & 3 & 2 \\ -1 & 2 & 1 \\ 4 & -1 & -1 \end{pmatrix}$$

and the eigenvector $\mathbf{v}^*$ corresponding to $\lambda = 3$ can be calculated by

$$(\mathcal{A}^* - 3\mathcal{I})\mathbf{v} = \begin{pmatrix} -2 & 3 & 2 \\ -1 & -1 & 1 \\ 4 & -1 & -4 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

and we get $v_2 = 0$ and $v_1 = v_3$. Thus, $\mathbf{v_2}^* = (1, 0, 1)$ and we can check that, indeed, $< \mathbf{v_2}^*, \mathbf{v^1} >=< \mathbf{v_2}^*, \mathbf{v^2} >= 0$. Then, multiplying (3.3.78) by $\mathbf{v_2}^*$ we obtain

$$< \mathbf{v_2}^*, \mathbf{y}(t) >= C_2 e^{\lambda_2 t} < \mathbf{v_2}^*, \mathbf{v^2} >$$

and, taking $t = 0$ we have

$$< \mathbf{v_2}^*, \overset{\circ}{\mathbf{x}} >= C_2 < \mathbf{v_2}^*, \mathbf{v^2} >$$

and $C_2 = \frac{1}{2}(\overset{\circ}{x}_1 + \overset{\circ}{x}_3)$. Clearly, long time picture of evolution given by (3.3.79) will not be realized if $\overset{\circ}{\mathbf{x}}$ is orthogonal to $\mathbf{v_2}^*$.

Example 3.19. Returning to Fibonacci rabbits, we see that the eigenvalues of $\mathcal{L}$ are exactly numbers

$$\lambda_{1,2} = r_\pm = \frac{1 \pm \sqrt{5}}{2}$$

and clearly, $\lambda_1 = (1+\sqrt{5})/2$ is the dominant eigenvalue. The eigenvector associated with this eigenvalue is $\mathbf{v^1} = (\lambda_1, 1) = ((\sqrt{5}+1)/2, 1)$ and this gives the stable age structure. Moreover, the matrix $\mathcal{L}$ is symmetric and thus the eigenvectors of $\mathcal{L}^*$ are the same as of $\mathcal{L}$. Thus

$$\mathbf{v}(k) = \begin{pmatrix} v_1(k) \\ v_0(k) \end{pmatrix} \approx C_1 r_+^k \begin{pmatrix} \frac{\sqrt{5}+1}{2} \\ 1 \end{pmatrix}$$

where

$$C_1 = \frac{2\left(v_1(0)\frac{\sqrt{5}+1}{2} + v_0(0)\right)}{5+\sqrt{5}}$$

as $<\mathbf{v^1}, \mathbf{v^1}>= (5+\sqrt{5})/2$.

Taking, for instance, the initial condition discussed in Section 1: $v_1(0) = 0, v_0(0) = 1$, we find $C_1 = 2/(5+\sqrt{5})$ and if we like to estimate the growth of the whole population, we have

$$y(k) = v_1(k) + v_0(k) \approx \frac{2}{5+\sqrt{5}}\left(\frac{\sqrt{5}+1}{2} + 1\right) r_+^k = \frac{3+\sqrt{5}}{5+\sqrt{5}} r_+^k = \frac{1+\sqrt{5}}{2\sqrt{5}} r_+^k,$$

in accordance with (3.3.56).

Example 3.20. Consider a population with individuals living up to two years in which both juveniles and adults can reproduce with effective maternity rate of juveniles and adults per capita is, respectively, 1 and 4. Assume further that the survival rate of juveniles is 1/2. Find the formula for the evolution of this population.

The question amounts to finding iterates $\mathcal{L}^k$ of the Leslie matrix

$$\mathcal{L} = \begin{pmatrix} 1 & 4 \\ \frac{1}{2} & 0 \end{pmatrix}.$$

We start with finding eigenvalues of $\mathcal{L}$:

$$p(\lambda) = \begin{vmatrix} 1-\lambda & 4 \\ \frac{1}{2} & -\lambda \end{vmatrix} = \lambda^2 - \lambda - 2 = 0$$

gives the eigenvalues $\lambda = 2$ and -1 of algebraic multiplicity 1. Thus, both eigenvalues are simple. To find eigenvectors corresponding to $\lambda_1 = 2$, we solve

$$(\mathcal{L} - 2\mathcal{I})\mathbf{v} = \begin{pmatrix} -1 & 4 \\ \frac{1}{2} & -2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Thus, v_2 is arbitrary, say $v_2 = 1$ and therefore $v_1 = 4$ so that

$$\mathbf{v^1} = \begin{pmatrix} 4 \\ 1 \end{pmatrix}.$$

Repeating this for $\lambda_2 = -1$, we obtain

$$(\mathcal{L} + 1\mathcal{I})\mathbf{v} = \begin{pmatrix} 2 & 4 \\ \frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Again, we take $v_2 = 1$ and therefore $v_1 = -2$ so that

$$\mathbf{v^2} = \begin{pmatrix} -2 \\ 1 \end{pmatrix}.$$

Therefore

$$\mathcal{L}^k\mathbf{v^1} = 2^k\begin{pmatrix}4\\1\end{pmatrix}, \qquad \mathcal{L}^k\mathbf{v^2} = (-1)^k\begin{pmatrix}-2\\1\end{pmatrix}.$$

To find explicit expression for $\mathcal{L}^k$ we use (3.3.29) we recall that the solution originating from $\overset{\circ}{\mathbf{x}}= (\overset{\circ}{x}_1, \overset{\circ}{x}_2)$ is given by

$$\mathcal{L}^k\overset{\circ}{\mathbf{x}}= c_1 2^k\begin{pmatrix}4\\1\end{pmatrix} + c_2(-1)^k\begin{pmatrix}-2\\1\end{pmatrix} = \begin{pmatrix}4\cdot 2^k & -2\cdot(-1)^k\\ 2^k & (-1)^k\end{pmatrix}\begin{pmatrix}c_1\\c_2\end{pmatrix}$$

where c_1 and c_2 are determined from $\overset{\circ}{\mathbf{x}}= c_1\mathbf{v^1} + c_2\mathbf{v^2}$ so that

$$\begin{pmatrix}c_1\\c_2\end{pmatrix} = \mathcal{V}^{-1}\overset{\circ}{\mathbf{x}}= \begin{pmatrix}\frac{1}{6} & \frac{1}{3}\\ -\frac{1}{6} & \frac{2}{3}\end{pmatrix}\overset{\circ}{\mathbf{x}}\,.$$

where

$$\mathcal{V} = \begin{pmatrix}| & |\\ \mathbf{v^1} & \mathbf{v^2}\\ | & |\end{pmatrix} = \begin{pmatrix}4 & -2\\ 1 & 1\end{pmatrix}.$$

Finally

$$\mathcal{L}^k\overset{\circ}{\mathbf{x}}= \begin{pmatrix}4\cdot 2^k & -2\cdot(-1)^k\\ 2^k & (-1)^k\end{pmatrix}\begin{pmatrix}\frac{1}{6} & \frac{1}{3}\\ -\frac{1}{6} & \frac{2}{3}\end{pmatrix}\overset{\circ}{\mathbf{x}},$$

where the matrix on the left hand side is the matrix $\mathcal{A}_k$, introduced in (3.3.29). Now, we see that for large k we have

$$\mathcal{L}^k\overset{\circ}{\mathbf{x}}= \begin{pmatrix}x_1(k)\\x_2(k)\end{pmatrix} \approx c_1 2^k\begin{pmatrix}4\\1\end{pmatrix} = 2^k\begin{pmatrix}4\\1\end{pmatrix}\left(\frac{1}{6}\overset{\circ}{x}_1 + \frac{1}{3}\overset{\circ}{x}_2\right).$$

To provide a better interpretation of the above formula, we note that the total population approximately evolves as

$$P(k) = x_1(k) + x_2(k) \approx 2^k\left(\frac{1}{6}\overset{\circ}{x}_1 + \frac{1}{3}\overset{\circ}{x}_2\right)(4+1)$$

and therefore the fractions of juveniles and adults in the total population approximately are given by, respectively,

$$J = \frac{x_1(k)}{P(k)} \approx \frac{4}{5}, \qquad A = \frac{x_2(k)}{P(k)} = \frac{1}{5}.$$

Thus, the eigenvector of the dominant eigenvalue normalized with respect to the total population:

$$(J, A) = \frac{1}{1+4}(4, 1).$$

gives the so called stable age profile of the population.

The next example shows how to deal with complex eigenvalues.

Example 3.21. Consider a population with individuals living up to 3 years in which only the eldest generation reproduces with effective birth rate 32 juveniles per individual. Assume further that the survival rate of juveniles is 3/4 and yearlings to the 2 years old is 1/3. Find the formula for the evolution of this population. Here we have

$$\mathcal{L} = \begin{pmatrix}0 & 0 & 32\\ \frac{3}{4} & 0 & 0\\ 0 & \frac{1}{3} & 0\end{pmatrix}.$$

The eigenvalues are determined from

$$\begin{vmatrix}-\lambda & 0 & 32\\ \frac{3}{4} & -\lambda & 0\\ 0 & \frac{1}{3} & -\lambda\end{vmatrix} = -\lambda^3 + 8 = 0$$

so that

$$\lambda_1 = 2, \qquad \lambda_2 = 2e^{2\pi i/3} = 2\left(\frac{-1}{2} + i\frac{\sqrt{3}}{2}\right), \qquad \lambda_3 = \bar{\lambda}_2 = 2\left(\frac{-1}{2} - i\frac{\sqrt{3}}{2}\right).$$

We find general form of eigenvectors $\mathbf{v^i}$ corresponding to eigenvalues λ_i, $i = 1, 2, 3$, by solving

$$\begin{pmatrix} -\lambda_i & 0 & 32 \\ \frac{3}{4} & -\lambda_i & 0 \\ 0 & \frac{1}{3} & -\lambda_i \end{pmatrix} \begin{pmatrix} v_1^i \\ v_2^i \\ v_3^i \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Taking $v_3^i = 1$, we find $v_2^i = 3\lambda_i$ and $v_1^i = 4\lambda_i^2$ (note that the first equation of the system gives the eigenvalue equation). Thus we can write

$$\mathbf{v^1} = \begin{pmatrix} 16 \\ 6 \\ 1 \end{pmatrix},$$

then

$$\mathbf{v^2} = \begin{pmatrix} 16e^{4\pi i/3} \\ 6e^{2\pi i/3} \\ 1 \end{pmatrix} = \begin{pmatrix} 16\left(\frac{-1}{2} - i\frac{\sqrt{3}}{2}\right) \\ 6\left(\frac{-1}{2} + i\frac{\sqrt{3}}{2}\right) \\ 1 \end{pmatrix}$$

and

$$\mathbf{v^3} = \overline{\mathbf{v^2}} = \begin{pmatrix} 16e^{-4\pi i/3} \\ 6e^{-2\pi i/3} \\ 1 \end{pmatrix} = \begin{pmatrix} 16\left(\frac{-1}{2} + i\frac{\sqrt{3}}{2}\right) \\ 6\left(\frac{-1}{2} - i\frac{\sqrt{3}}{2}\right) \\ 1 \end{pmatrix}$$

One could write the general solution of the problem as

$$\mathcal{L}^k \overset{\circ}{\mathbf{x}} = \begin{pmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \end{pmatrix} = c_1 2^k \mathbf{v^1} + c_2 2^k e^{2k\pi i/3}\mathbf{v^2} + c_3 2^k e^{-2k\pi i/3}\overline{\mathbf{v^2}} \tag{3.3.80}$$

where $\overset{\circ}{\mathbf{x}} = c_1\mathbf{v^1} + c_2\mathbf{v^2} + c_3\mathbf{v^3}$ but the expressions above are complex and we require real solutions. We take advantage of the fact that complex solutions appear as a complex conjugate pair.

To find explicit real expressions for them, we write $\lambda = re^{i\phi}$ where $r = |\lambda|$ and $\phi = Arg\lambda$. Then

$$\lambda^k = r^n e^{in\phi} = r^k(\cos k\phi + i\sin k\phi)$$

and

$$\Re(\lambda^k\mathbf{v}) = r^k(\cos k\phi\Re\mathbf{v} - \sin k\phi\Im\mathbf{v}),$$
$$\Im(\lambda^k\mathbf{v}) = r^k(\sin k\phi\Re\mathbf{v} + \cos k\phi\Im\mathbf{v}).$$

In our case

$$\Re\lambda_2^k\mathbf{v^2} = 2^k\left(\cos 2k\pi/3\begin{pmatrix} -8 \\ -3 \\ 1 \end{pmatrix} - \sin 2k\pi/3\begin{pmatrix} -8\sqrt{3} \\ 3\sqrt{3} \\ 0 \end{pmatrix}\right)$$

and

$$\Im\lambda_2^k\mathbf{v^2} = 2^k\left(\sin 2k\pi/3\begin{pmatrix} -8 \\ -3 \\ 1 \end{pmatrix} + \cos 2k\pi/3\begin{pmatrix} -8\sqrt{3} \\ 3\sqrt{3} \\ 0 \end{pmatrix}\right).$$

The initial values for $\Re\lambda^k\mathbf{v^2}$ and $\Im\lambda^k\mathbf{v^2}$ are, respectively, $\begin{pmatrix} -8 \\ -3 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} -8\sqrt{3} \\ 3\sqrt{3} \\ 0 \end{pmatrix}$. Since $\mathcal{L}^k$ is a real matrix, we have $\Re\mathcal{L}^k\mathbf{v^2} = \mathcal{L}^k\Re\mathbf{v^2}$ and $\Im\mathcal{L}^k\mathbf{v^2} = \mathcal{L}^k\Im\mathbf{v^2}$, thus

$$\begin{aligned}\mathcal{L}^k\begin{pmatrix}-8\\-3\\1\end{pmatrix} &= 2^k\left(\cos 2k\pi/3\begin{pmatrix}-8\\-3\\1\end{pmatrix} - \sin 2k\pi/3\begin{pmatrix}-8\sqrt{3}\\3\sqrt{3}\\0\end{pmatrix}\right)\\ &= 2^k\begin{pmatrix}-8(\cos 2k\pi/3 - \sqrt{3}\sin 2k\pi/3)\\-3(\cos 2k\pi/3 + \sqrt{3}\sin 2k\pi/3)\\\cos 2k\pi/3\end{pmatrix}\end{aligned}$$

and

$$\begin{aligned}\mathcal{L}^k\begin{pmatrix}-8\sqrt{3}\\3\sqrt{3}\\0\end{pmatrix} &= 2^k\left(\sin 2k\pi/3\begin{pmatrix}-8\\-3\\1\end{pmatrix} + \cos 2k\pi/3\begin{pmatrix}-8\sqrt{3}\\3\sqrt{3}\\0\end{pmatrix}\right)\\ &= 2^k\begin{pmatrix}-8(\sin 2k\pi/3 + \sqrt{3}\cos 2k\pi/3)\\-3(\sin 2k\pi/3 - \sqrt{3}\cos 2k\pi/3)\\\sin 2k\pi/3\end{pmatrix}\end{aligned}$$

Thus we see that (3.3.80) can be written using only real expressions as

$$\begin{aligned}\mathcal{L}^k\,\overset{\circ}{\mathbf{x}} = c_1 2^k\begin{pmatrix}16\\6\\1\end{pmatrix} &+ c_2 2^k\begin{pmatrix}-8(\cos 2k\pi/3 - \sqrt{3}\sin 2k\pi/3)\\-3(\cos 2k\pi/3 + \sqrt{3}\sin 2k\pi/3)\\\cos 2k\pi/3\end{pmatrix}\\ &+ c_3 2^k\begin{pmatrix}-8(\sin 2k\pi/3 + \sqrt{3}\cos 2k\pi/3)\\-3(\sin 2k\pi/3 - \sqrt{3}\cos 2k\pi/3)\\\sin 2k\pi/3\end{pmatrix}\end{aligned}$$

where $\mathbf{c} = (c_1, c_2, c_3)$ are solutions to

$$\overset{\circ}{\mathbf{x}} = \mathcal{V}\mathbf{c}$$

with

$$\mathcal{V} = \begin{pmatrix}16 & -8 & -8\sqrt{3}\\6 & -3 & 3\sqrt{3}\\1 & 1 & 0\end{pmatrix}$$

The final example is to illustrate the case when we do not have sufficiently many eigenvectors.

Example 3.22. Consider a population divided into two classes, reproductive and post reproductive. Assume that the reproductive season lasts exactly 1 year and the effective fertility is 0.5. Assume further that 20% percent of this class survives the first year and moves to the postreproductive class. The individuals can live indefinitely but every year 50% of individuals in the postreproductive class dies. Write down the Usher matrix for this population and find its eigenvalues and eigenvectors.

Here we have

$$\mathcal{U} = \begin{pmatrix}0.5 & 0\\0.2 & 0.5\end{pmatrix}.$$

The characteristic equation is

$$p(\lambda) = \begin{vmatrix}0.5-\lambda & 0\\0.2 & 0.5-\lambda\end{vmatrix} = \lambda^2 - \lambda + 0.25 = 0$$

gives the double eigenvalue $\lambda = 0.5$ so that the algebraic multiplicity of λ is 2. Let us find eigenvectors. We solve find eigenvectors corresponding to $\lambda_1 = 2$, we solve

$$(\mathcal{U} - 0.5\mathcal{I})\mathbf{v} = \begin{pmatrix}0 & 0\\0.2 & 0\end{pmatrix}\begin{pmatrix}v_1\\v_2\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix}$$

which gives $\mathbf{v}^2 = (0, 1)$ and this is the only eigenvector so that the eigenvalue is not semisimple. Hence, we have to find the associated eigenvector by solving

$$(\mathcal{U}-0.5\mathcal{I})^2\mathbf{v} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Any vector satisfies the above equation. We select the simplest vector linearly independent of $\mathbf{v}^2 = (0,1)$ which is $\mathbf{v}^1 = (1,0)$. In this way, we obtain $\mathcal{V} = \mathcal{I}$.

Thus we can write, for $\mathbf{x} = (x_1, x_2)$,

$$\begin{aligned} \mathcal{U}^k\mathbf{x} &= x_1\mathcal{U}^k\mathbf{v}^1 + x_2\mathcal{U}^k\mathbf{v}^2 = x_1(0.5\mathcal{I} + (\mathcal{U}-0.5\mathcal{I}))^k\mathbf{v}^1 + x_2(0.5)^k\mathbf{v}^2 \\ &= x_1\left((0.5)^k\mathbf{v}^1 + k(0.5)^{k-1}\begin{pmatrix} 0 & 0 \\ 0.2 & 0 \end{pmatrix}\mathbf{v}^1\right) + x_2(0.5)^k\mathbf{v}^2 \\ &= x_1\left((0.5)^k\begin{pmatrix} 1 \\ 0 \end{pmatrix} + k(0.5)^{k-1}\begin{pmatrix} 0 \\ 0.2 \end{pmatrix}\right) + x_2(0.5)^k\begin{pmatrix} 0 \\ 1 \end{pmatrix}. \end{aligned}$$

The question whether any matrix with nonnegative entries gives rise to such a behaviour and, if not, what models exhibit AEG, is much more delicate and requires invoking the Frobenius-Perron theorem which will be discussed in the next section.

4 Positive linear dynamical systems – Frobenius-Perron theorem

4.1 Problems with general linear models – the Natchez structure

Many societies are divided into classes membership of which is largely hereditary. Such a class structure is essential in maintaining power and control the distribution of resources by an appropriately defined elite. One of the ways to prevent watering down of the elite is the practice of endogamy; that is, marrying within one's own class. Some societies, however, practice an open class system to prevent stagnation of the structure. We describe and analyse on of the most famous societies of such a type – the civilization of Natchez, see [17, 18, 23].

Natchez were Native Americans who lived in the lower Mississipi in North America. The early accounts about them came from the Spanish explorer Hernando de Soto in 1542. Most of the information about Natchez is due to their contacts with French colonisers who first met them in 1682. The civilisation ceased to exist after the so-called Natchez massacre in 1731 when they were defeated and then dispersed or were enslaved.

Natchez created a complex system of open class structure based on the exogamous marriages so that the power is passed between people born to different social classes. The society was divided into two main classes – nobility and commoners (so-called Stinkards). The nobility was further divided ito subclasses (casts): Suns, Nobles and Honoured People. A member of nobility only could marry a Stinkard. According to [23], a person's social status and class were determined matrilineally; that is, the children of female Suns, Nobles, or Honoureds kept the status of their mothers. However, the children of male Suns and Nobles did not become commoners, as noble exogamy and matrilineal descent would appear to dictate, but rather moved one class below the class of their fathers. In other words, children of male Suns became Nobles, while children of male Nobles became Honoured.

We summarize the permissible marriages and inherited statuses in the table below. An empty place intersection of the row and a column in the table means that such a marriage is not permitted; if it is permitted, then the entry at the intersection indicates the status of the offspring. To analyse the evolution of this population, we shall create a mathematical model based on Table 3.1. To make it feasible, we adopt the following simplifying assumptions.

1. There is the same number of males and females in each class in each generation.
2. Each person marries only once and the spouse is from the same generation.
3. Each pair has exactly one son and one daughter.

Mother/Father	**Sun**	**Noble**	**Honoured**	**Stinkard**
Sun				Sun
Noble				Noble
Honoured				Honoured
Stinkard	Noble	Honoured	Stinkard	Stinkard

Table 3.1. Possible marriages in the Natchez population and the status of their offspring

Since we have the same number of males and females in each generation, in our model we will only track the males. Hence, let $x_i(k)$ denote the number of males in the class i and in the generation k, where the classes are numbered as follows: 1 -Sun, 2 -Noble, 3 -Honoured, 4 -Stinkard. Let us consider the class distribution in the generation $k+1$. Since a Sun son only can be born to a Sun mother and there is no other way to become a Sun, using the fact that the number of female Suns equals the number of male Suns we can write

$$x_1(k+1) = x_1(k).$$

Next, a Noble son only can be born to Sun father or to Noble mother, using the parity of males and females in the Noble class we get

$$x_2(k+1) = x_1(k) + x_2(k).$$

A Honoured son is either a descendent of Noble father or Honoured mother hence, as before,

$$x_3(k+1) = x_2(k) + x_3(k).$$

Finally, the number of male offspring in the Stinkard class is equal to the number of Stinkard males who are not married to females from the nobility plus the number of sons of Stinkard mothers and Honoured fathers (remember that the son of a Stinkard father and the Honoured mother is Honoured but then the son of a Stinkard mother and Honoured father is a Stinkard). Hence, using again the fact that the numbers of males and females in each class are equal, we arrive at

$$x_4(k+1) = -x_1(k) - x_2(k) + x_4(k).$$

Writing these equations in a matrix form we obtain

$$\begin{pmatrix} x_1(k+1) \\ x_2(k+1) \\ x_3(k+1) \\ x_4(k+1) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ -1 & -1 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \\ x_4(k) \end{pmatrix} \qquad (3.4.81)$$

Let us denote the coefficient matrix by $\mathcal{A}$. The form of the matrix allows for expressing $\mathcal{A}^k$ in an explicit form. We write $\mathcal{A} = \mathcal{I} + \mathcal{B}$, where

$$\mathcal{B} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & -1 & 0 & 0 \end{pmatrix}.$$

It is easy to see that $\mathcal{B}$ is idempotent. Indeed,

$$\mathcal{B}^2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix}, \qquad \mathcal{B}^3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Hence, using the Newton formula, we get

$$\mathcal{A}^k = (I+B)^k = \mathcal{I} + k\mathcal{B} + \frac{k(k-1)}{2}\mathcal{B}^2$$

and, using the formulae for the powers of $\mathcal{B}$

$$\mathcal{A}^k = \begin{pmatrix} 1 & 0 & 0 & 0 \\ k & 1 & 0 & 0 \\ \frac{k(k-1)}{2} & k & 1 & 0 \\ \frac{-k(k+1)}{2} & -k & 0 & 1 \end{pmatrix}.$$

Hence we obtain

$$\mathbf{x}(k) = (x_1(0), kx_1(0) + x_2(0), \frac{1}{2}k(k-1)x_1(0) + kx_2(0) + x_3(0), -\frac{1}{2}k(k+1)x_1(0) - kx_2(0) + x_4(0)).$$

In particular, if $x_1(0) = x_2(0) = 0$, then the class structure of the population does not change as we have

$$\mathbf{x}(k) = (0, 0, x_3(0), x_4(0)).$$

If, however, originally we have some members of either Sun or Noble class, then the size of the Stinkard shrinks and becomes negative in finite time. Indeed, e.g. solving for k

$$-\frac{1}{2}k(k+1)x_1(0) - kx_2(0) + x_4(0) = 0,$$

we obtain the positive solution

$$k_0 = \sqrt{(1+2p)^2 + 8q} - (1+2p),$$

where $p = x_2(0)/x_1(0)$ and $q = x_4(0)/x_1(0)$. For instance, if the number of Suns and Nobles is equal ($p = 1$) and the number of Stinkards is 5 times the number of Suns ($q = 5$), we obtain $k_0 = 4$, the the number of Stinkards will become 0 after four generations. Thus, nobody from the nobility will be able to marry and have offspring and thus the population will become extinct. Since, according to the historical records, the Natchez civilisation survived for several hundred years, the model in the presented form cannot be correct.

We recognize, however, that we have made a number of assumptions in order to simplify the model. It is possible that we can modify some of them to get a model which gives a more realistic evolution of the population by avoiding its collapse in such a short time. Let us try to adjust the birth coefficients. Indeed, in many societies it is observed that the birth rate depends on the status of the parents. Thus, let us address the following question: Can we find a birth rate for each combination of parents in the Natchez community which will result in its a stable class distribution?

To simplify the analysis without compromising the general picture, we restrict ourselves to three classes which are labelled **A**, **B** and **C** and, following Table 3.1, we construct the intermarriage scheme summarized in Table 3.2. Interpretacja of the table is the same as before. The Roman letter at the intersection of a column and a row indicates the class of the offspring coming from the mother in the row and father in the column, while the Greek letter α_i gives the average number of male offspring from in such a marriage. Further, as before, we assume that

Mother/Father	**A**	**B**	**C**
A			**A** α_1
B			**B** α_3
C	**B** α_2	**C** α_4	**C** α_5

Table 3.2. Intermarriage scheme in a simplified Natchez-like community

1. The numbers of males and females are equal in each class and each generation;
2. there are no inter-generation marriages,

and by $x_1(k)$, $x_2(k)$, $x_3(k)$ we denote the number of male members of the population in the class, **A**, **B** and **C** in the kth generation. Then

$$x_1(k+1) = \alpha_1 x_1(k). \tag{3.4.82}$$

Similarly, since a child in **B** may be born either to a father from class **A** (and mother from class **C**) or to mother from class **B** (and father from **C**), we have

$$x_2(k+1) = \alpha_2 x_1(k) + \alpha_3 x_2(k).$$

Finally, a child of class **C** may be born to a father of class **B** or father of class **C** except those who married females of class **A** or **B**. Hence

$$x_3(k+1) = \alpha_4 x_2(k) + \alpha_5 \left(x_3(k) - x_1(k) - x_2(k)\right).$$

In matrix form

$$\begin{pmatrix} x_1(k+1) \\ x_2(k+1) \\ x_3(k+1) \end{pmatrix} = \begin{pmatrix} \alpha_1 & 0 & 0 \\ \alpha_2 & \alpha_3 & 0 \\ -\alpha_5 & (\alpha_4 - \alpha_5) & \alpha_5 \end{pmatrix} \begin{pmatrix} x_1(k) \\ x_2(k) \\ x_3(k) \end{pmatrix}, \tag{3.4.83}$$

or, in compact form

$$\mathbf{x}(k+1) = \mathcal{A}\mathbf{x}(k).$$

Before we analyze the model, let us check that in the case $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 1$ the model reproduces the behaviour of the full, four class, Natchez model. In such a case we have

$$\mathcal{A} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

As in the previous case, we write $\mathcal{A} = \mathcal{I} + \mathcal{B}$ and see that $\mathcal{B}^2 = 0$. Hence, $\mathcal{A}^k = \mathcal{I} + k\mathcal{B}$, or

$$\mathcal{A}^k = \begin{pmatrix} 1 & 0 & 0 \\ k & 1 & 0 \\ -k & 0 & 1 \end{pmatrix}.$$

Therefore, $x_3(k) = -kx_1(0) + x_3(0)$ and thus $x_1(0) > 0$ yields $x_3(k) < 0$ for $k > x_3(0)/x_1(0)$. We see that the simplified system displays the same collapse as the original Natchez system.

Let us return to the model (3.4.83). Following the analysis of Subsection 3.4, we will try to find a positive eigenvector of $\mathcal{A}$ associated with the largest positive eigenvalue. Such a vector would correspond to a stable population structure which, once attained, would not change in time. The total population would change as powers of the eigenvalue. In this way we see that if the structure is perturbed, then the model will return to the stable distribution. Moreover, we must ensure that in our model we always have sufficiently many members of the class **C** to provide spouses to the higher classes. Since the matrix $\mathcal{A}$ is triangular, its eigenvalues are given by the diagonal elements; that is, eigenvalues are α_1, α_3 and α_5. First, let $\alpha_5 > \alpha_1$ and $\alpha_5 > \alpha_3$; that is, α_5 is the dominant eigenvalue of $\mathcal{A}$. This assumption corresponds to an increasing fertility of the lowest class. To find the corresponding eigenvector, we solve

$$\begin{pmatrix} (\alpha_1 - \alpha_5) & 0 & 0 \\ \alpha_2 & (\alpha_3 - \alpha_5) & 0 \\ \alpha_5 & (\alpha_4 - \alpha_5) & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Hence $x_1 = 0$ hence there is no positive eigenvector. Moreover, this leads to $x_2 = 0$ and arbitrary x_3, so that the only possible long term structure is given by $(0, 0, 1)$; that is, the population will only consists of commoners. While it is certainly possible, we are interested in survival of the community as a whole which is not possible if α_5 is the dominant eigenvalue. A similar outcome is obtained if we assume that α_3 is the dominant eigenvalue. Here also $x_1 = 0$ and though under additional assumption $\alpha_4 > \alpha_5$ we obtain a nonnegative eigenvector $(0, 1, (\alpha_4 - \alpha_5)/(\alpha_3 - \alpha_5)$, but the structure is destroyed.

The last option is to assume that α_1 is a dominant eigenvalue. Thus, let $\alpha_1 > \alpha_3$ and $\alpha_1 > \alpha_5$. The eigenvector corresponding to α_1 satisfies

$$\begin{pmatrix} 0 & 0 & 0 \\ \alpha_2 & (\alpha_3 - \alpha_1) & 0 \\ -\alpha_5 & (\alpha_4 - \alpha_5) & (\alpha_5 - \alpha_1) \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The second equation yields

$$x_2 = \frac{\alpha_2}{(\alpha_1 - \alpha_3)} x_1.$$

Since $\alpha_1 > \alpha_3$, we see that $x_2 > 0$ provided $x_1 > 0$. Then from the last equation we get

$$x_3 = \frac{1}{\alpha_1 - \alpha_5} \left(-\alpha_5 + \alpha_2 \frac{\alpha_4 - \alpha_5}{\alpha_1 - \alpha_3} \right) x_1.$$

Since $\alpha_1 > \alpha_5$ and $\alpha_1 > \alpha_3$, we have $x_3 > 0$ if and only if

$$-\alpha_5(\alpha_1 - \alpha_3) + \alpha_2(\alpha_4 - \alpha_5) > 0.$$

In other words, the eigenvector corresponding to the dominant eigenvalue α_1 is positive provided

$$\begin{cases} \alpha_1 > \alpha_3 \\ \alpha_1 > \alpha_5 \\ \alpha_2(\alpha_4 - \alpha_5) > \alpha_5(\alpha_1 - \alpha_3). \end{cases} \tag{3.4.84}$$

We observe that a necessary condition for the last inequality to be satisfied, we must have $\alpha_4 > \alpha_5$. In particular, the inequality will be satisfied if α_5 is sufficiently small. Thus, we have a positive stable distribution if the birth rate in the lowest class is small.

To complete our considerations, we must show that in each generation we have sufficiently many members of class **C** for the members of classes **B** and **A** to marry. Thus we require that the components of the stable distribution vector additionally satisfy

$$x_3 \geq x_1 + x_2. \tag{3.4.85}$$

Substituting here the formulae for x_2 and x_3, which were derived earlier, we obtain the inequality

$$\frac{(\alpha_4 - \alpha_5)}{(\alpha_1 - \alpha_5)} - \frac{\alpha_5(\alpha_1 - \alpha_3)}{\alpha_2(\alpha_1 - \alpha_5)} - \frac{(\alpha_1 - \alpha_3)}{\alpha_2} - 1 \geq 0. \tag{3.4.86}$$

We observe here that we can satisfy this inequality by taking sufficiently large α_4. In particular, we must have

$$\alpha_4 > \alpha_5. \tag{3.4.87}$$

Thus, to ensure that in the stable population distribution we have sufficient number of members of class **C**, we should ensure that α_4; that is, the birth rate of in marriages of mothers of class **C** and fathers of class **B**, is sufficiently large. Of course, this is not a necessary condition. Any set of birth rates satisfying (3.4.85), in addition to (3.4.84), will ensure (3.4.85.

The birth coefficients $\alpha_1 = 1.2$, $\alpha_2 = 0.9$, $\alpha_3 = 1.1$, $\alpha_4 = 1.4$ and $\alpha_5 = 0.8$ satisfy (3.4.86). Then the stable population structure is given by

$$\mathbf{x} = (1, 9, 11.5).$$

We observe that the provided analysis is not complete. It shows that there is a structure of the society which can persist in a stable way. However, we do not know whether any initial population will eventually stabilize at the structure determined by the eigenvector found above. Fortunately, the problem allows for a more detailed solution which confirm the results obtained by the analysis of the dominant eigenvalue and the corresponding eigenvector. First, we observe that the first equation in (3.4.83) is independent of the other two and the second one does not depend on the third one. Then we obtain

$$x_1(k) = \alpha_1^k \overset{\circ}{x}_1 \tag{3.4.88}$$

and, using (2.3.49),

$$x_2(k) = \alpha_3^k \overset{\circ}{x}_2 + \alpha_2 \overset{\circ}{x}_1 \frac{\alpha_1^k - \alpha_3^k}{\alpha_1 - \alpha_3}. \tag{3.4.89}$$

Finally, using (3.4.88) and (3.4.89) and, again, (2.3.49), we have

$$\begin{aligned} x_3(k) &= \alpha_5^k \overset{\circ}{x}_3 + \overset{\circ}{x}_2 (\alpha_4 - \alpha_5) \sum_{i=0}^{k-1} \alpha_5^{k-i-1} \alpha_3^i + \overset{\circ}{x}_1 \frac{\alpha_2(\alpha_4 - \alpha_5}{\alpha_1 - \alpha_3} \left(\sum_{i=0}^{k-1} \alpha_5^{k-i-1} \alpha_1^i - \sum_{i=0}^{k-1} \alpha_5^{k-i-1} \alpha_3^i \right) \\ &\quad - \alpha_5 \overset{\circ}{x}_1 \sum_{i=0}^{k-1} \alpha_5^{k-i-1} \alpha_1^i \\ &= \alpha_5^k \overset{\circ}{x}_3 + \frac{\alpha_3^k - \alpha_5^k}{\alpha_3 - \alpha_5} (\alpha_4 - \alpha_5) \left(\overset{\circ}{x}_2 - \overset{\circ}{x}_1 \frac{\alpha_2}{\alpha_1 - \alpha_3} \right) + \frac{\alpha_1^k - \alpha_5^k}{\alpha_1 - \alpha_5} \left(\frac{\alpha_2(\alpha_4 - \alpha_5)}{\alpha_1 - \alpha_3} - \alpha_5 \right) \overset{\circ}{x}_1 . \end{aligned} \tag{3.4.90}$$

Assuming that (3.4.87) is satisfied, we see that for $x_3(k)$ to be positive, we need (3.4.84) to be satisfied as well as

$$\frac{\overset{\circ}{x}_1}{\overset{\circ}{x}_2} < \frac{\alpha_1 - \alpha_3}{\alpha_2}. \tag{3.4.91}$$

Hence, though in general some non-negative initial conditions may lead negative solution; that is, to the collapse of the society, positive initial conditions satisfying (3.4.91) ensure that the population will stay positive for all times and distribution among classes will eventually stabilize at the vector

$$\left(1, \frac{\alpha_2}{(\alpha_1 - \alpha_3)}, \frac{1}{\alpha_1 - \alpha_5} \left(-\alpha_5 + \alpha_2 \frac{\alpha_4 - \alpha_5}{\alpha_1 - \alpha_3} \right) \right).$$

Note the relation with condition (3.4.91). The latter simply states that for the solution to remain positive, the ratio of the initial populations in classes **A** and **B** must be smaller that the ratio of these populations in the stable distribution vector. The analysis is, however, still not complete. The reason is that though the model (3.4.83) is a proper reflection of the assumptions, we cannot claim that the model determines the assumptions in a unique way. For instance, the equation (3.4.82) only specifies how many offspring the class **A** produces in each cycle but it does not tell that the offspring must result from the marriage of a class **A** mother and class **C** father. In other words, we do not know whether

$$x_3(k) \geq x_1(k) + x_2(k) \tag{3.4.92}$$

is satisfied for any k; that is, we do not know that in each cycle the society can obey its marriage scheme. For this, using (3.4.88)–(3.4.90), we have to prove that

$$\begin{aligned} &\alpha_5^k \overset{\circ}{x}_3 + \frac{\alpha_3^k - \alpha_5^k}{\alpha_3 - \alpha_5} (\alpha_4 - \alpha_5) \left(\overset{\circ}{x}_2 - \overset{\circ}{x}_1 \frac{\alpha_2}{\alpha_1 - \alpha_3} \right) + \frac{\alpha_1^k - \alpha_5^k}{\alpha_1 - \alpha_5} \left(\frac{\alpha_2(\alpha_4 - \alpha_5)}{\alpha_1 - \alpha_3} - \alpha_5 \right) \overset{\circ}{x}_1 \\ &\qquad \geq \alpha_3^k \overset{\circ}{x}_2 + \alpha_2 \overset{\circ}{x}_1 \frac{\alpha_1^k - \alpha_3^k}{\alpha_1 - \alpha_3} + \alpha_1^k \overset{\circ}{x}_1 . \end{aligned} \tag{3.4.93}$$

It is natural to assume that the initial conditions satisfy $\overset{\circ}{x}_3 \geq \overset{\circ}{x}_1 + \overset{\circ}{x}_2$. We see that if we prove (3.4.93) for $\overset{\circ}{x}_3 = \overset{\circ}{x}_1 + \overset{\circ}{x}_2$, then it will be valid for any $\overset{\circ}{x}_3 \geq \overset{\circ}{x}_1 + \overset{\circ}{x}_2$. So, let $\overset{\circ}{x}_3 = \overset{\circ}{x}_1 + \overset{\circ}{x}_2$ and, denoting $q = \overset{\circ}{x}_2 / \overset{\circ}{x}_1$, we must find conditions for

$$\begin{aligned} \Psi_k(q) &:= (\alpha_3^k - \alpha_5^k) \frac{\alpha_5 - \alpha_3 + 1}{\alpha_3 - \alpha_5} q + \alpha_5^k - \alpha_1^k - \frac{\alpha_2(\alpha_3^k - \alpha_5^k)}{(\alpha_3 - \alpha_5)(\alpha_1 - \alpha_3)} + \frac{\alpha_1^k - \alpha_5^k}{\alpha_1 - \alpha_5} \left(\frac{\alpha_2(\alpha_4 - \alpha_5)}{\alpha_1 - \alpha_3} - \alpha_5 \right) \\ &\quad - \alpha_2 \frac{\alpha_1^k - \alpha_3^k}{\alpha_1 - \alpha_3} \geq 0 \end{aligned} \tag{3.4.94}$$

for all $q \geq \alpha_2/(\alpha_1 - \alpha_3)$ and $k \geq 1$.

We re-write the previous equation as

$$\begin{aligned}\Psi_k(q) = & \frac{\alpha_3^k-\alpha_5^k}{\alpha_3-\alpha_5}\left((\alpha_5-\alpha_3+1)q - \frac{\alpha_2}{\alpha_1-\alpha_3}\right) + \frac{\alpha_2(\alpha_1^k-\alpha_5^k)}{\alpha_1-\alpha_3}\left(\frac{\alpha_4-\alpha_5}{\alpha_1-\alpha_5} - \frac{\alpha_5(\alpha_1-\alpha_3)}{\alpha_2(\alpha_1-\alpha_5)} - \frac{\alpha_1-\alpha_3}{\alpha_2}\right)\\ & -\alpha_2\frac{\alpha_1^k-\alpha_3^k}{\alpha_1-\alpha_3} \end{aligned}\qquad (3.4.95)$$

and, using (3.4.86),

$$\begin{aligned}\Psi_k(q) &\geq \frac{\alpha_3^k-\alpha_5^k}{\alpha_3-\alpha_5}\left((\alpha_5-\alpha_3+1)q - \frac{\alpha_2}{\alpha_1-\alpha_3}\right) + \frac{\alpha_2}{\alpha_1-\alpha_3}(\alpha_3^k-\alpha_5^k)\\ &\geq \frac{\alpha_3^k-\alpha_5^k}{\alpha_3-\alpha_5}\left((\alpha_5-\alpha_3+1)\left(q-\frac{\alpha_2}{\alpha_1-\alpha_3}\right)\right)\end{aligned}$$

and the last line is nonnegative whenever $\alpha_5-\alpha_3+1\geq 0$ (since $q\geq \alpha_2/(\alpha_1-\alpha_3)$).

4.2 Positive dynamical systems

If a given difference or differential equation/system of equations is to describe evolution of a population; that is, if the solution is the population size or density, then clearly solutions emanating from non-negative data must stay non-negative. If we deal with systems of equations, then non-negativity must be understood in the sense defined above. We note that there are models admitting negative solutions such as the discrete logistic equation (see discussion preceding (2.2.23) or the Natchez population in the previous section), but then the moment the solution becomes negative is interpreted as the extinction of the population and the model ceases to be applicable for later times.

Let us first consider processes occurring in discrete time.

Proposition 3.23. *The solution* $\mathbf{y}(k)$ *of*

$$\mathbf{y}(k+1)=\mathcal{A}\mathbf{y}(k),\quad \mathbf{y}(0)=\overset{\circ}{\mathbf{y}}$$

satisfies $\mathbf{y}(k)\geq 0$ *for any* $k=1,\ldots,$ *for arbitrary* $\overset{\circ}{\mathbf{y}}\geq 0$ *if and only if* $\mathcal{A}\geq 0$.

Proof. The ‘if’ part is easy. We have $y_i(k)=\sum\limits_{j=1}^{n} a_{ij}y_j(k-1)$ for $k\geq 1$ so if $a_{ij}\geq 0$ and $\overset{\circ}{y}_j\geq$ for $i,j=1,\ldots,n$, then $y_i(1)\geq 0$ for all $i=1,\ldots,n$ and the extension for $k>1$ follows by induction.

On the other hand, assume that $a_{ij}<0$ for some i,j and consider $\overset{\circ}{\mathbf{y}}=\mathbf{e}_j=(0,\ldots,0,1,0,\ldots,0)$, where 1 is on the jth place. Then $\mathcal{A}\overset{\circ}{\mathbf{y}}=(a_{1j},\ldots,a_{ij},\ldots,a_{nj})$ so the output is not non-negative. Thus, the condition $\mathcal{A}\geq 0$ is also necessary. □

The proof of analogous result in continuous time is slightly more involved.

Proposition 3.24. *The solution* $\mathbf{y}(t)$ *of*

$$\mathbf{y}'=\mathcal{A}\mathbf{y},\quad \mathbf{y}(0)=\overset{\circ}{\mathbf{y}}$$

satisfies $\mathbf{y}(t)\geq 0$ *for any* $t>0$ *for arbitrary* $\overset{\circ}{\mathbf{y}}\geq 0$ *if and only if* $\mathcal{A}$ *has non-negative off-diagonal entries.*

Proof. First let us consider $\mathcal{A}\geq 0$. Then, using the representation (3.3.48)

$$e^{t\mathcal{A}}=\mathcal{I}+t\mathcal{A}+\frac{t^2}{2}\mathcal{A}^2+\frac{t^3}{3!}\mathcal{A}^3+\ldots+\frac{t^k}{k!}\mathcal{A}^k+\ldots,$$

and the results of the previous proposition we see that $e^{t\tilde{\mathcal{A}}} \geq 0$. Next, we observe that for any real a and $\overset{\circ}{\mathbf{y}}$ the function $\mathbf{y}(t) = e^{at}e^{t\mathcal{A}}\, \overset{\circ}{\mathbf{y}} \geq 0$ and satisfies the equation

$$\mathbf{y}' = a\mathbf{y} + \mathcal{A}\mathbf{y} = (a\mathcal{I} + \mathcal{A})\mathbf{y}.$$

Hence if the diagonal entries of $\mathcal{A}$, a_{ii}, are negative, then denoting $r = \max_{1\leq i\leq n}\{-a_{ii}\}$ we find that $\tilde{\mathcal{A}} = r\mathcal{I} + \mathcal{A} \geq 0$. Using the first part of the proof, we see that

$$e^{t\mathcal{A}} = e^{-rt}e^{t\tilde{\mathcal{A}}} \geq 0. \tag{3.4.96}$$

Let us write

$$e^{t\mathcal{A}} = \mathcal{E}(t) = \begin{pmatrix} \epsilon_{11}(t) & \ldots & \epsilon_{1n}(t) \\ \vdots & & \vdots \\ \epsilon_{n1}(t) & \ldots & \epsilon_{nn}(t) \end{pmatrix},$$

so $\epsilon_{ij}(t) \geq 0$ for all $i, j = 1, \ldots, n$, and consider $\mathcal{E}(t)\mathbf{e}_i = (\epsilon_{1i}(t), \ldots, \epsilon_{ii}(t), \ldots, \epsilon_{ni}(t))$. Then

$$\begin{aligned}(a_{1i}, \ldots, a_{ii}, \ldots, a_{ni}) &= \mathcal{A}\mathcal{E}(t)\mathbf{e}_i|_{t=0} = \left.\frac{d}{dt}\mathcal{E}(t)\mathbf{e}_i\right|_{t=0} \\ &= \lim_{h\to 0^+}\left(\frac{\epsilon_{1i}(h)}{h}, \ldots, \frac{\epsilon_{ii}(h) - 1}{h}, \ldots, \frac{\epsilon_{ni}(h)}{h}\right),\end{aligned}$$

so that $a_{ji} \geq 0$ for $j \neq i$. □

4.3 Classification of projection matrices

The long time behaviour of $(\mathcal{A}^k)_{k\geq 1}$ is fully determined by whether $\mathcal{A}$ is a primitive irreducible, imprimitive irreducible or a reducible matrix. These concepts are also easily interpreted in the context of population dynamics.

For a matrix $\mathcal{A} = (a_{ij})_{1\leq i,j\leq n}$, we say that there is an *arc* from i to j if $a_{ij} > 0$; a *path* from i to j is a sequence of arcs starting from i and ending in j in which the endpoint of each arc (apart from the last) is the beginning of the subsequent arc. A *loop* is a path from i to itself.

We say that a non-negative matrix is *irreducible* if, for each i and j, there is a path from i to j. Otherwise, we say that it is reducible.

This definition easily can be interpreted in terms of graphs. A *graph* is a nonempty finite set of vertices and (possibly empty) set of edges (edge can be interpreted as a unordered pair of vertices). An directed graph or a *digraph* is a graph with directed edges (a directed edge is then an ordered pair of vertices). As with matrices, a *path* in a graph is a finite sequence of directed edges $((i_1, i_2), (i_2, i_3), \ldots, (i_{k-1}, i_k))$ in which no vertex is repeated apart from possibly $i_1 = i_k$; in the latter case the path is called a loop. We say that a digraph is strongly connected if for any pair of vertices i and j there is a path which connects them.

By the *incidence matrix* associated to a nonnegative matrix $\mathcal{A}$ we understand the matrix $\mathcal{D} = (d_{ij})_{1\leq i,j\leq n}$ where $d_{i,j} = 1$ if $a_{i,j} > 0$ and $d_{i,j} = 0$ otherwise. There is a one to one correspondence between diagraphs and incidence matrices (up to a permutation). For a given $\mathcal{D}$ we take $\{1, \ldots, n\}$ as the set of vertices and we draw a directed from j to i whenever $d_{i,j} > 0$. Conversely, given a diagraph with n vertices, we number them $\{1, \ldots, n\}$ and set $d_{i,j} = 1$ whenever there is an edge from j to i.

Then it is relatively easy to prove that a matrix $\mathcal{A}$ is irreducible if and only if its incidence matrix $\mathcal{D}$ is irreducible. Then one can prove that $\mathcal{A}$ is irreducible if and only if the digraph associated with the incidence matrix of $\mathcal{A}$ is strongly connected.

An equivalent, but more algebraic, condition must be preceded by some notation. We write

$$\mathcal{A}^k = (a^{(k)}_{i,j})_{1\leq i,j\leq n}.$$

It is easy to see that

$$a^{(k)}_{i,j} = \sum_{1\leq i_r\leq n, r=1,\ldots,k-1} a_{i,i_1}a_{i_1,i_2}\cdot\ldots\cdot a_{i_{k-1},j}$$

If some $a_{i,i_1}a_{i_1,i_2}\cdot\ldots\cdot a_{i_{k-1},j}\neq 0$ then there is a path starting from j and passing through $i_{k-1},\ldots,i_1$ to i. Since the matrix elements are nonnegative, for $a^{(k)}_{i,j}$ to be non-zero it is enough that there exists at least one such path. Thus, $\mathcal{A}$ is irreducible if for each pair (i,j) there is k such that $a^{(k)}_{i,j} > 0$.

If the matrix $\mathcal{A}$ is not irreducible, then we say that is is *reducible.* Thus, a matrix is reducible if the associated graph is not strongly connected, that is, if there are vertices i and j such that i is not accessible from j. An equivalent definition is that $\mathcal{A}$ is reducible if, by simultaneous permutation of rows and columns, it can be brought to the form

$$\begin{pmatrix} A & \mathbf{0} \\ B & C \end{pmatrix}$$

where A and B are square matrices.

In terms of age-structured population dynamics, a matrix is irreducible if each stage i can contribute to any other stage j. E.g., the Usher matrix

$$\begin{pmatrix} 0&0&1&0\\ 1&0&0&0\\ 0&1&0&0\\ 0&0&1&1 \end{pmatrix}$$

is reducible as the last state cannot contribute to any other state and fertility is only concentrated in one state.

Irreducible matrices are subdivided into two further classes. An irreducible matrix $\mathcal{A}$ is called *primitive* if

$$\mathcal{A}^k > 0,$$

otherwise it is called *imprimitive.*

Note the difference between irreducibility and primitivity. For irreducibility we require that for each (i,j) there is k such that $a^{(k)}_{i,j} > 0$ but for primitivity there must be k such that $a^{(k)}_{i,j} > 0$ for all (i,j).

In population dynamics, if the population has a single reproductive stage, then its projection matrix is imprimitive. E.g., the matrix

$$\begin{pmatrix} 0&0&1\\ 1&0&0\\ 0&1&0 \end{pmatrix}$$

describing a semelparous population is imprimitive.

The Perron-Frobenius theorem can be summarized as follows.

Theorem 3.25. *Let $\mathcal{A}$ be a nonnegative matrix.*

1. *There exists a real nonnegative eigenvalue $\lambda_{max} = r(\mathcal{A})$ such that $\lambda_{max} \geq |\lambda|$ for any $\lambda \in \sigma(\mathcal{A})$. There is an eigenvector (called the Perron eigenvector) corresponding to λ_{max} which is real and nonnegative.*
2. *If, in addition, $\mathcal{A}$ is irreducible, then λ_{max} is simple and strictly positive, $\lambda_{max} \geq |\lambda|$ for $\lambda \in \sigma(\mathcal{A})$. The eigenvector corresponding to λ_{max} may be chosen to be strictly positive.*

   (i)If $\mathcal{A}$ is additionally primitive, then $\lambda_{max} > |\lambda|$;

   (ii) If $\mathcal{A}$ is imprimitive, then there are $d-1$ (d is called the imprimitivity index) eigenvalues $\lambda_j = \lambda_{max}e^{2\pi i\frac{j}{d}}, j = 1,\ldots,j-1$, with $\lambda_{max} = |\lambda_j|$.

Let us apply the Perron-Frobenius theorem in the population context. Suppose that our population is divided into n age (or other) classes and the state of the population is given by the vector $\mathbf{x} = (x_1, \ldots, x_n)$ giving the number of individuals (or density) in each class. Let $\mathbf{x}_0 \geq 0$ denote the initial distribution of the population among the classes. Then

$$\mathbf{x}(k) = \mathcal{A}^k \mathbf{x}_0$$

is the distribution after k periods and

$$P(k, \mathbf{x}) = \|\mathcal{A}^k \mathbf{x}_0\| = \sum_{i=1}^{n} (\mathcal{A}^k \mathbf{x}_0)_i = \sum_{i=1}^{n} x_i(k) = \|\mathbf{x}(k)\|$$

is the total population at time k evolving from the initial distribution $\mathbf{x}_0$.

If $\mathcal{A}$ is nonnegative, irreducible or primitive, then the transpose $\mathcal{A}^T$ has the same property. Let $r := \lambda_{max}$ be the dominant eigenvalue of both matrices and $\mathbf{v}$ and $\mathbf{v}^*$ be the corresponding strictly positive eigenvectors of, respectively, $\mathcal{A}$ and $\mathcal{A}^T$, corresponding to λ_{max}. We normalize $\mathbf{v}$ so that $\|\mathbf{v}\| = 1$ and $\mathbf{v}^*$ so that $< \mathbf{v}^*, \mathbf{v} >= 1$.

Combining the Perron-Frobenius theorem with the spectral decomposition we arrive at the following result.

Theorem 3.26 (Fundamental Theorem of Demography). *Suppose that the projection matrix $\mathcal{A}$ is irreducible and primitive and let r be the strictly positive dominant eigenvalue of $\mathcal{A}$, $\mathbf{v}$ the strictly positive eigenvector of $\mathcal{A}$ and $\mathbf{v}^*$ strictly positive eigenvector of $\mathcal{A}^T$ corresponding to r. Then, for any $\mathbf{x}_0 \geq 0$,*

(a) $\mathcal{A}$ has the AEG property

$$\lim_{k\to\infty} r^{-k} \mathcal{A}^k \mathbf{x}_0 =< \mathbf{v}^*, \mathbf{x}_0 > \mathbf{v}. \tag{3.4.97}$$

(b)

$$\lim_{k\to\infty} \frac{\mathbf{x}(k)}{P(k, \mathbf{x}_0)} = \frac{\mathcal{A}^k \mathbf{x}_0}{P(k, \mathbf{x}_0)} = \mathbf{v}. \tag{3.4.98}$$

(c) If $r < 1$

$$\lim_{k\to\infty} P(k, \mathbf{x}_0) = 0,$$

and

$$\lim_{k\to\infty} P(k, \mathbf{x}_0) = \infty$$

if $r > 1$.

Proof. (a) We use (3.3.73), (3.3.76) and Theorem 3.25 1.(i)

$$\mathcal{A}^k \mathbf{x}_0 = \sum_{\lambda\in\sigma(\mathcal{A})} \lambda^k \mathbf{p}_\lambda(k, \mathcal{P}_\lambda \mathbf{x}_0) = r^k < \mathbf{v}^*, \mathbf{x}_0 > \mathbf{v} + \sum_{\lambda\in\sigma(\mathcal{A})\setminus\{r\}} \lambda^k \mathbf{p}_\lambda(k, \mathcal{P}_\lambda \mathbf{x}_0). \tag{3.4.99}$$

By primitivity of $\mathcal{A}$, $r > |\lambda|$ for any $\lambda \in \sigma(\mathcal{A} \setminus \{r\})$ and since $p_\lambda(k, \mathcal{P}_\lambda, \mathbf{x}_0)$ are polynomials in k, we have

$$\left\| \left(\frac{\lambda}{r}\right)^k \mathbf{p}_\lambda(k, \mathcal{P}_\lambda \mathbf{x}_0) \right\| \to 0, \qquad k \to \infty,$$

and (a) is proved.

For (b) we see that, by (a),

$$\begin{aligned} \lim_{k\to\infty} \frac{P(k, \mathbf{x}_0)}{r^k} &= \lim_{k\to\infty} \frac{\|\mathcal{A}^k \mathbf{x}_0\|}{r^k} = \lim_{k\to\infty} \left\| \frac{\mathcal{A}^k \mathbf{x}_0}{r^k} \right\| \\ &= | < \mathbf{v}^*, \mathbf{x}_0 > | > 0. \end{aligned} \tag{3.4.100}$$

Hence, by (a) and (3.4.100),

$$\lim_{k\to\infty}\frac{\mathcal{A}^k\mathbf{x}_0}{P(k,\mathbf{x}_0)}=\lim_{k\to\infty}\frac{r^{-k}\mathcal{A}^k\mathbf{x}_0}{r^{-k}P(k,\mathbf{x}_0)}=\mathbf{v},$$

which gives (3.4.98).

To prove (c) we observe that

$$P(k,\mathbf{x}_0)=\left\|\mathcal{A}^k\left(\frac{\mathbf{x}_0}{P(k-1,\mathbf{x}_0)}\right)\right\|P(k-1,\mathbf{x}_0)=r_{k-1}P(k-1,\mathbf{x}_0)$$

where

$$r_{k-1}=\left\|\mathcal{A}^k\left(\frac{\mathbf{x}_0}{P(k-1,\mathbf{x}_0)}\right)\right\|=r\left\|\frac{r^{-k}\mathcal{A}^k\mathbf{x}_0}{r^{-(k-1)}P(k-1,\mathbf{x}_0)}\right\|\to r\frac{|<\mathbf{v}^*,\mathbf{x}_0>|\|\mathbf{v}\|}{|<\mathbf{v}^*,\mathbf{x}_0>|}=r,\quad \text{as}\quad k\to\infty$$

by (a), (b) and the normalization of $\mathbf{v},\mathbf{v}^*$. Thus, if $r<1$, then we can pick $\bar{r}<1$ such that $r_{k-1}\leq\bar{r}$ for all k larger than some k_0 and

$$P(k_0+i,\mathbf{x}_0)\leq\bar{r}^iP(k_0,\mathbf{x}_0)\to 0\quad\text{as}\quad i\to\infty.$$

Similarly, if $r>1$, then we can pick $\tilde{r}>1$ such that $r_{k-1}\geq\tilde{r}$ for all k larger than some k_0 and

$$P(k_0+i,\mathbf{x}_0)\geq\tilde{r}^iP(k_0,\mathbf{x}_0)\to\infty\quad\text{as}\quad i\to\infty,$$

as $P(k_0,\mathbf{x}_0)\neq 0$ for any finite k_0. Indeed, otherwise from nonnegativity we would have $\mathbf{x}(k_0)=0$ and thus $\mathbf{x}(k)=0$ for $k\geq k_0$, contradicting (a). □

If we note that for each $1\leq i\leq n$

$$\frac{(\mathcal{A}^k\mathbf{x})_i}{P(k,\mathbf{x})}$$

is the fraction of the population in the state i at time k, then the result above states that for large times the fraction of the population in the state i approximately is given by the i coordinate of the Perron eigenvector and is independent of the initial distribution $\mathbf{x}$. Moreover $\mathbf{v}$ is approached (or departed from) at an exponential rate, hence the name *asynchronous exponential growth.*

4.4 Example–irreducible case

Let us consider the Leslie matrix

$$\mathcal{L}:=\begin{pmatrix} f_0 & f_1 & \cdots & f_{n-2} & f_{n-1}\\ s_0 & 0 & \cdots & 0 & 0\\ 0 & s_1 & \cdots & 0 & 0\\ \vdots & \vdots & \cdots & \vdots & \vdots\\ 0 & 0 & \cdots & s_{n-2} & 0\end{pmatrix}, \tag{3.4.101}$$

and find under what conditions the population described by $\mathcal{L}$ exhibits asynchronous exponential growth.

First we observe that for irreducibility we need all $s_i\neq 0, 0\leq i\leq n-2$. Indeed, if for some i the coefficient $s_i=0$, then there would be no path from $k\leq i$ to $k>i$. In other words, there would be no way of reaching the age $k>i$. Assuming this, $\mathcal{L}$ is irreducible if and only if $f_{n-1}>0$. Clearly, if $f_{n-1}=0$, then there is no communication from class $n-1$ to any other class and thus $\mathcal{L}$ is reducible. Now, let $f_{n-1}>0$ and pick a (i,j). If $j<i$, then there is a path $(j,j+1)...(i-1,i)$ ensured by the survival coefficients $s_j,s_{j+1},\dots s_{i-1}$. If $j\geq i$, then the survival coefficients ensure that we reach the last class $n-1$, then since $f_{n-1}>0$ we reach the class 0 and then we arrive at i by aging, that is $(j,j+1)\dots(n-2,n-1)(n-1,0)(0,1),\dots(i-1,i)$.

The question of primitivity is more complicated. Let us first assume that $f_j>0$ for $j=0,\dots,n-1$, that is, that any age group is capable of reproduction. Let us consider arbitrary initial state j. Then there is an arc between j and 0 ($a_{0j}=f_j>0$) and then from state 0 one can reach any state i in exactly i steps ($s_0s_1\cdot\dots\cdot s_i$). Thus, there is a path joining j and i of length $i+1$ which still depends on the target state.

However, there is an arc from 0 to itself, so we can wait at 0 for any number of steps. In particular we can wait for $n-i$ steps so that j can be connected with i is $n+1$ steps. In other words

$$s_{i-1}\cdot\dots\cdot s_1 s_0 f_0\cdot\ldots\cdot f_0 f_j > 0$$

where f_1 occurs $n-i$ times. Hence $\mathcal{L}^n > 0$.

Remark 3.27. The above argument shows that any irreducible matrix in which at least one diagonal entry is not equal to zero is primitive.

This result assumes too much - typically young individuals cannot reproduce. We will strengthen this result. Let

$$0 = \det(\lambda I - \mathcal{L}) = \lambda^n + a_{n_1}\lambda^{n_1} + \ldots a_{n_i}\lambda^{n_i} \tag{3.4.102}$$

with $n > n_1 > \ldots > n_i$, $a_{n_k} \neq 0, k = 1,\ldots,i$ be the characteristic equation of $\mathcal{L}$. It follows that if $\mathcal{L}$ is imprimitive of index d, then d is the greatest common divisor of $n-n_1, n_1-n_2,\ldots,n_{i-1}-n_i$. This is related to the fact that $\lambda^d - r^d$ is a factor of the characteristic polynomial but full proof requires more subtle characterization of the spectrum of imprimitive matrices. The characteristic equation of a Leslie matrix can be calculated explicitly and it has its own biological interpretation. Let us start with $n=2$. Then the determinant D_2 equals

$$D_2 = \lambda^2 - \lambda f_0$$

and for $n=3$

$$D_3(\lambda) = \det\begin{pmatrix} f_0-\lambda & f_1 & f_2 \\ s_0 & -\lambda & 0 \\ 0 & s_1 & -\lambda \end{pmatrix} = -\lambda^3 + \lambda^2 f_0 + \lambda f_1 s_0 + f_2 s_0 s_1, \tag{3.4.103}$$

We make the induction assumption

$$D_{n-1}(\lambda) = (-1)^{n-1}\left(\lambda^{n-1} - \sum_{k=0}^{n-2}\lambda^{n-2-k} f_k \prod_{i=0}^{k-1} s_i\right),$$

with $\prod\limits_{i=0}^{-1} = 1$. Then

$$\begin{aligned} D_n(\lambda) &= \det\begin{pmatrix} f_0-\lambda & f_1 & \cdots & f_{n-2} & f_{n-1} \\ s_0 & -\lambda & \cdots & 0 & 0 \\ 0 & s_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & s_{n-2} & -\lambda \end{pmatrix} \\ &= -\lambda D_{n-1} + (-1)^{2n-1} s_{n-2}(-1)^n f_{n-1}\prod_{i=0}^{n-1} s_i = (-1)^n\left(\lambda^n - \sum_{k=0}^{n-1}\lambda^{n-1-k} f_k \prod_{i=0}^{k-1} s_i\right). \end{aligned}$$

Now, this equation can be simplified if we remember that $s_i = l_{i+1}/l_i$ with $l_0 = 1$, where l_i is the probability of surviving from birth to age i and $f_i = m_{i+1}s_i$. Thus

$$f_k \prod_{i=0}^{k-1} s_i = m_{k+1}s_k s_{k-1}\cdot\ldots\cdot s_0 = m_{k+1}l_{k+1}$$

and thus the characteristic equation for a Leslie matrix can be written as

$$\sum_{k=1}^{n}\lambda^{n-k} m_k l_k = \lambda^n. \tag{3.4.104}$$

Using the criterion mentioned above, a Leslie matrix is irreducible and primitive if e.g. the fertility of the oldest generation m_n is not zero and two subsequent generations have nonzero fertility.

Remark 3.28. **Alternative derivation of the characteristic equation of a Leslie matrix and its eigenvectors.** Consider the eigenvalue-eigenvector equation for a Leslie matrix

$$\mathcal{L}\mathbf{v} = \begin{pmatrix} f_0 & f_1 & \cdots & f_{n-2} & f_{n-1} \\ s_0 & 0 & \cdots & 0 & 0 \\ 0 & s_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & \cdots & s_{n-2} & 0 \end{pmatrix} \begin{pmatrix} v_0 \\ v_1 \\ v_2 \\ \vdots \\ v_{n-1} \end{pmatrix} = \lambda \begin{pmatrix} v_0 \\ v_1 \\ v_2 \\ \vdots \\ v_{n-1} \end{pmatrix}.$$

The equations from the second row down read

$$s_0 v_0 = \lambda v_1, \quad s_1 v_1 = \lambda v_2, \quad \dots s_{n-2} v_{n-2} = \lambda v_{n-1}.$$

Taking $v_0 = 1$, we obtain

$$v_1 = \frac{s_0}{\lambda}, \quad v_2 = \frac{s_0 s_1}{\lambda^2}, \quad \dots v_{n-1} = \frac{s_0 s_1 \dots s_{n-2}}{\lambda^{n-1}}.$$

Now, the first row gives the equation

$$\lambda = \left(f_0 + \frac{f_1 s_0}{\lambda} + \frac{f_2 s_0 s_1}{\lambda^2} + \dots + \frac{f_{n-1} s_0 s_1 \dots s_{n-2}}{\lambda^{n-1}} \right).$$

Again we use $s_i = l_{i+1}/l_i$ where l_i is probability of survival till the $i+1$st reproductive cycle from birth (thus s_i is conditional probability of survival to the next reproductive cycle if one survived till i from birth) and $f_i = m_{i+1} s_i$ to rewrite the above as

$$1 = \left(\frac{m_1 l_1}{\lambda} + \frac{m_2 l_2}{\lambda^2} + \frac{m_3 l_3}{\lambda^3} + \dots + \frac{m_n l_n}{\lambda^n} \right),$$

where we used $l_0 = 1$. This equation is called the discrete Euler-Lotka equation.

Relative simplicity of the characteristic equation allows to strengthen this result even more. In fact, $\mathcal{L}$ is imprimtive if and only if the maternity function is periodic, that is, if the greatest common divisor of ages of positive reproduction, called the period, is greater than 1. For instance, the sequence $m_2, m_4, m_6 ...$ has period 2. In particular, the period is equal to the imprimitivity index.

The proof follows from the criterion given above but we shall provide a direct instructive proof. Indeed, suppose that

$$\lambda_j = re^{i\theta}, \quad \theta \neq 2\pi n,$$

is a negative or complex root to

$$\psi(\lambda) = \lambda^{-1} m_1 l_1 + \dots + \lambda^{n-1} m_{n-1} l_{n-1} + \lambda^{-n} m_n l_n = \sum_{k=1}^{n} \lambda^{-k} m_k l_k = 1. \tag{3.4.105}$$

Then

$$\sum_{k=1}^{n} r^{-k} e^{-ik\theta} l_k m_k = 1 \tag{3.4.106}$$

or, taking real and imaginary parts,

$$\sum_{k=1}^{n} r^{-k} \cos(k\theta) l_k m_k = 1, \tag{3.4.107}$$

$$\sum_{k=1}^{n} r^{-k} \sin(k\theta) l_k m_k = 0. \tag{3.4.108}$$

If m_k is periodic, then the only nonzero terms correspond to multiples of d, $m_{k_1d}, m_{k_2d}, m_{k_3d}, \ldots$. Taking $\theta_j = 2\pi j/d$, $j = 0, 1, \ldots, d-1$, we see $\cos k_l d\theta_j = 1$, $\sin k_l d\theta_j = 0$ and so, if the above equations are satisfied by r, they are also satisfied by any $\lambda_j = re^{i\theta_j}$.

If m_k is aperiodic, with $m_{k_i} \neq 0$, $i \in I \subset \{1, \ldots, n\}$, then there is no $\theta \neq 0$ for which $\cos k_i\theta = 1$ for all k_i. Indeed, otherwise there is $\theta \in (0, 2\pi)$ such that

$$\cos k_1\theta = 1.$$

This implies

$$\theta = 2\pi\frac{l}{k_1} = 2\pi\frac{j}{d},$$

where $l < k_1$ is an integer and j and d are relatively prime integers, so that $0 < j \leq d-1$ (note that if $j = 0$, then $\theta = 0$). But then

$$k_i\theta = k_i\frac{2\pi j}{d} = 2\pi l_i$$

for some integer l_i so that

$$k_i = l_i\frac{d}{j}.$$

However, k_i is an integer and j and d are relatively prime so that l_i must be divisible by j. Hence

$$k_i = r_i d$$

for some integer r_i, $i \in I$. Thus, whatever θ, for some k we must have $\cos k\theta < 1$. But then, if (3.4.107) was satisfied, we would have

$$\sum_{k=1}^{n} |\lambda_j|^{-k} l_k m_k > 1.$$

On the other hand, since

$$\sum_{k=1}^{n} r^{-k} l_k m_k = 1,$$

we obtain $|\lambda_j| < r$.

Remark 3.29. Eq. (3.4.105) is the characteristic equation of the so called Lotka renewal equation *and can be derived directly. Let $B(k)$ be the number of births at time k. These births can be divided into two classes: one class attributed to females born between time 0 and k and the other due to females which were alive at time $k = 0$. Females that are of age i at time k were born at time $k - i$. The number of females born at $k - i$ is given by $B(k - i)$. The number of them that survive till the age k is $l(i)B(k - i)\Delta t$ and thus the number of births at time k by females of age i is $l(i)B(k - i)m(i)$. Thus, this contribution is*

$$\sum_{i=1}^{k} B(k-i)l_i m_i.$$

To find the contribution of the females who were present at time $k = 0$ we begin with taking the number of females of age i present at $k = 0$, $x_i(0)$. These females must live till the age $k + i$, that is we must take survival rate till $k + i$, l_{k+i}, conditioned upon the female having survived till i. Hence we see that $x_i(0)l_{k+i}/li$ females survived till time k. These gave birth to $m_{k+i}x_i(0)l_{k+i}/l_i$ new females. To find the number of all births due to females older then k we again sum over all ages. However, no female survives beyond n so the summation terminates at $n - k$ (no female older that $n - k$ at time $k = 0$ will survive till k). Combining these two formulae we get

$$B(k) = \sum_{i=1}^{k} B(k-i)l_i m_i + G(k), \tag{3.4.109}$$

where

$$G_k = \sum_{i=1}^{n-k} x_i(0)\frac{l_{k+i}}{l_i}m_{k+i}.$$

For large k, $k > n$, the equation (3.4.109) reduces to

$$B(k) = \sum_{i=1}^{k} B(k-i)l_i m_i$$

which is a constant coefficient difference equation the characteristic equation of which is exactly (3.4.105).

4.5 Reducible case

Let us consider a more complicated case where the fertility is restricted to some interval $[n_1, n_2]$, that is, when $f_j > 0$ for $j \in [n_1, n_2]$. As we noted earlier, if $n_2 < n$, the matrix cannot be irreducible as there is no communication between postreproductive stages and the reproductive ones. Consequently, if we start only with individuals in postreproductive age, the population will die out in finite time. Nevertheless, if $n_1 < n_2$ then the population still displays asynchronous exponential growth, albeit with a slight modification, as explained below.

To analyse this model, we note that since we cannot move from stages with $j > n_2$ to earlier stages, the part of the population with $j \leq n_2$ evolves independently from postreproductive part (but feeds into it.) Assume that $n_1 < n_2$ and introduce the restricted matrix

$$\tilde{\mathcal{L}} = \begin{pmatrix} f_0 & f_1 & \cdots & f_{n_2-1} & f_{n_2} \\ s_0 & 0 & \cdots & 0 & 0 \\ 0 & s_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & s_{n_2-1} & 0 \end{pmatrix}$$

and the matrix providing (one-way) link from reproductive to postreproductive stages is given by

$$\mathcal{R} = \begin{pmatrix} 0 & \cdots & s_{n_2} & 0 & \cdots & 0 & 0 \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & s_{n-2} & 0 \end{pmatrix}$$

For the matrix $\tilde{\mathcal{L}}$, $f_{n_2} > 0$ and $f_{n_2-1} > 0$ and we can apply the considerations of the previous section and Theorem 3.26. Thus, there is $r > 0$ there are vectors $\mathbf{v} = (v_0, \ldots v_{n_2})$ and $\mathbf{v}^* = (v_0^*, \ldots v_{n_2}^*)$ such that $\tilde{\mathcal{L}}\mathbf{v} = r\mathbf{v}$ and

$$\lim_{k\to\infty} r^{-k}\mathbf{x}(k+1) = \lim_{k\to\infty} r^{-k}\tilde{\mathcal{L}}^k\mathbf{x}_0 = \mathbf{v} < \mathbf{v}^*, \mathbf{x}_0 >, \quad 0 \leq \mathbf{x}_0 \in \mathbb{R}^{n_2}. \tag{3.4.110}$$

For $n_2 \leq j < n, k \geq 0$, we have $x_{j+1}(k+1) = s_j x_j(k)$. Hence, starting from $x_{n_2}(k)$ we get $x_{n_2+i}(k+i) = c_i x_{n_2}(k)$, where $c_i = s_{n_2+i-1} \cdot \ldots \cdot s_{n_2}$, as long as $i \leq n - n_2 - 1$. So

$$\lim_{k\to\infty} r^{-k}x_{n_2+i}(k+i) = c_i v_{n_2} < \mathbf{v}^*, \mathbf{x}_0 >, \quad 0 \leq \mathbf{x}_0 \in \mathbb{R}^{n_2},$$

and hence, changing $k+i$ into k

$$\lim_{k\to\infty} r^{-k}x_{n_2+i}(k) = c_i r^{-i} v_{n_2} < \mathbf{v}^*, \mathbf{x}_0 >, \quad 0 \leq \mathbf{x}_0 \in \mathbb{R}^{n_2},$$

for any $i = 1, \ldots, n - n_2 - 1$.

Hence, we see that the formula (3.3.76) is satisfied if we take

$$\begin{aligned} \mathbf{v} &= (v_0, \ldots v_{n_2}, c_1 r^{-1} v_{n_2}, \ldots, c_{n-n_2-1} r^{-(n-n_2-1)} v_{n_2}) \\ \mathbf{v}^* &= (v_0^*, \ldots y_{n_2}^*, 0, \ldots, 0). \end{aligned}$$

Finally, we observe that if only one f_j is positive (semelparous population), then we do not have asynchronous exponential growth. Indeed, in this case starting from initial population in one class we will have a cohort of individuals in the same age group moving through the system. We have observed such a behaviour in Example 3.5.

5 Birth-and-death type problems

Consider a population consisting of $N(t)$ individuals at time t. We allow stochasticity to intervene in the process so that $N(t)$ becomes a random variable. Accordingly, we denote by

$$p_n(t) = P[N(t) = n], \quad n = 1, 2, \ldots \tag{3.5.111}$$

the probability that the population has n individuals at t.

5.1 Birth process

At first, we consider only births and we assume that each individual gives births to a new one independently of others. For a single individual, we assume that

$$P\{1 \text{ birth in } (t, t + \Delta t]|N(t) = 1\} = \beta\Delta t + o(\Delta), \tag{3.5.112}$$

$$P\{\text{more than 1 birth in } (t, t + \Delta t]|N(t) = 1\} = o(\Delta t), \tag{3.5.113}$$

$$P\{0 \text{ births in } (t, t + \Delta t]|N(t) = 1\} = 1 - \beta\Delta t + o(\Delta t). \tag{3.5.114}$$

If we have n individuals, than 1 births will occur if exactly one of them give birth to one offspring and the remaining $n - 1$ produce 0. This can happen in n ways. Thus

$$\begin{aligned} P\{1 \text{ birth in } (t, t + \Delta t]|N(t) = n\} &= n(\beta\Delta t + o(\Delta t))(1 - \beta\Delta t + o(\Delta t))^{n-1} \\ &= n\beta\Delta t + o(\Delta t). \end{aligned} \tag{3.5.115}$$

Similarly, more then one birth can occur if one individual give births to more than 1 offspring or at least two individuals give birth to one new one. Considering all possible combinations, we end up with finite sum each term of which is multiplied by Δt or its higher powers. Thus

$$P\{\text{more than 1 birth in } (t, t + \Delta t]|N(t) = n\} = o(\Delta t). \tag{3.5.116}$$

Finally, no birth occurs if none individual produces an offspring; that is

$$\begin{aligned} P\{0 \text{ births in } (t, t + \Delta t]|N(t) = n\} &= (1 - \beta\Delta t + o(\Delta t))^n \\ &= 1 - n\beta\Delta t + o(\Delta t). \end{aligned} \tag{3.5.117}$$

We can set up the equation describing evolution of $p_n(t)$. There can be n individuals at time $t + \Delta t$ if there were $n - 1$ individuals at time t and one births occurred or if there were n individuals and zero births occurred, or less than $n - 1$ individuals and more than 1 birth occurred. However, the last event occurs with probability $o(\Delta t)$ and will be omitted. Using the theorem of total probabilities

$$\begin{aligned} p_n(t + \Delta t) = p_{n-1}(t)P\{1 \text{ birth in } (t, t + \Delta t]|N(t) = n - 1\} \\ +p_n(t)P\{0 \text{ births in } (t, t + \Delta t]|N(t) = n\} \end{aligned} \tag{3.5.118}$$

that is, using the formulae

$$p_n(t) = (n - 1)\beta\Delta t p_{n-1} + (1 - n\beta\Delta t)p_n(t) + o(\Delta) + o(\Delta t). \tag{3.5.119}$$

After some algebra, we get

$$\frac{p_n(t+\Delta t)-p_n(t)}{\Delta t} = -n\beta p_n(t) + (n-1)\beta p_{n-1}(t) + \frac{o(\Delta t)}{\Delta t}$$

and, passing to the limit

$$\frac{dp_n(t)}{dt} = -n\beta p_n(t) + (n-1)\beta p_{n-1}(t). \tag{3.5.120}$$

This is an infinite chain of differential equations which must be supplemented by an initial condition. The population at $t = 0$ had to have some number of individuals, say, n_0. Hence,

$$p_n(0) = \begin{cases} 1 \text{ for } n = n_0, \\ 0 \text{ for } n \neq n_0. \end{cases} \tag{3.5.121}$$

Since this is purely birth process, $p_n(0) = 0$ for $t > 0$ and $n < n_0$.

Since the rate of change of p_n depends only on itself and on the preceding $p_{n-1}(t)$, we have

$$\frac{dp_{n_0}(t)}{dt} = -n_0\beta p_{n_0}(t), \tag{3.5.122}$$

so that

$$p_{n_0}(t) = e^{-\beta n_0 t}.$$

For $p_{n_0+1}(t)$ we obtain nonhomogeneous equation

$$\frac{dp_{n_0+1}(t)}{dt} = -(n_0+1)\beta p_{n_0+1}(t) + \beta n_0 e^{-\beta n_0 t}.$$

Using integrating factor $e^{\beta(n_0+1)t}$ we obtain

$$\left(p_{n_0+1}(t)e^{\beta(n_0+1)t}\right)' = \beta n_0 e^{\beta t}$$

or

$$p_{n_0+1}(t) = (n_0 e^{\beta t} + C)e^{-\beta(n_0+1)t}$$

so, using the initial condition $p_{n_0+1}(0) = 0$, we obtain

$$p_{n_0+1}(t) = n_0(1-e^{-\beta t})e^{-\beta n_0 t}$$

In general, it can be proved that

$$p_{n_0+m}(t) = \binom{n_0+m-1}{n_0-1} e^{-\beta n_0 t}(1-e^{-\beta t})^m.$$

Indeed, we proved the validity of the formula for $m = 1$. Next

$$\frac{dp_{n_0+m+1}(t)}{dt} = -(n_0+m+1)\beta p_{n_0+m+1}(t) + \beta \binom{n_0+m-1}{n_0-1} e^{-\beta n_0 t}(1-e^{-\beta t})^m.$$

and, as before

$$\left(p_{n_0+m+1}(t)e^{\beta(n_0+m+1)t}\right)' = \beta \binom{n_0+m-1}{n_0-1} e^{\beta t}(1-e^{-\beta t})^m.$$

and, integrating

$$\begin{aligned}
&p_{n_0+m+1}(t)e^{\beta(n_0+m+1)t} \\
&= C + \beta(n_0+m)\binom{n_0+m-1}{n_0-1}\int e^{\beta(m+1)t}(1-e^{-\beta t})^m dt \\
&= C + \beta(n_0+m)\binom{n_0+m-1}{n_0-1}\int e^{\beta t}(e^{\beta t}-1)^m dt \\
&= C + (n_0+m)\binom{n_0+m-1}{n_0-1}\int u^m du \\
&= C + \frac{n_0+m}{m+1}\binom{n_0+m-1}{n_0-1}(e^{\beta t}-1)^{m+1} \\
&= C + \binom{n_0+m}{n_0-1}(e^{\beta t}-1)^{m+1}
\end{aligned}$$

Using the initial condition $p_{n_0+m+1}(0) = 0$ we find $C = 0$ and so

$$p_{n_0+m+1}(t) = \binom{n_0+m}{n_0-1} e^{-\beta n_0 t}(1-e^{-\beta t})^{m+1}.$$

5.2 Birth-and-death system

The obvious drawback of the system discussed above is that individuals never die. We can easily remedy this by adding possibility of dying in the same way as we modelled births. Accordingly,

$$P\{1 \text{ birth in } (t, t+\Delta t]|N(t) = 1\} = \beta\Delta t + o(\Delta), \tag{3.5.123}$$
$$P\{1 \text{ death in } (t, t+\Delta t]|N(t) = 1\} = \delta\Delta t + o(\Delta), \tag{3.5.124}$$
$$P\{\text{no change in } (t, t+\Delta t]|N(t) = 1\} = 1 - (\beta+\delta)\Delta t + o(\Delta t). \tag{3.5.125}$$

Possibility of more then one births or death occurring in $(t, t+\Delta t]$ is assumed to be or order $o(\Delta t)$ and will be omitted in the discussion.

As before, we assume that in the population of n individuals births and deaths occur independently. The probability of 1 birth is given by

$$\begin{aligned} &P\{1 \text{ birth in } (t, t+\Delta t]|N(t) = n\} \\ &= n(\beta\Delta t + o(\Delta t))(1 - (\beta+\delta)\Delta t + o(\Delta t))^{n-1} \\ &= n\beta\Delta t + o(\Delta t). \end{aligned} \tag{3.5.126}$$

Similarly, probability of 1 (net) death in the population

$$\begin{aligned} &P\{1 \text{ birth in } (t, t+\Delta t]|N(t) = n\} \\ &= n(\delta\Delta t + o(\Delta t))(1 - (\beta+\delta)\Delta t + o(\Delta t))^{n-1} \\ &= n\delta\Delta t + o(\Delta t). \end{aligned} \tag{3.5.127}$$

and, finally,

$$\begin{aligned} P\{\text{no change in } (t, t+\Delta t]|N(t) = n\} &= (1 - (\beta+\delta)\Delta t + o(\Delta t))^n \\ &= 1 - n(\beta+\delta)\Delta t + o(\Delta t). \end{aligned} \tag{3.5.128}$$

We can set up the equation describing evolution of $p_n(t)$. Arguing as before

$$p_n(t) = (n-1)\beta\Delta t p_{n-1} + (n+1)\delta\Delta t p_{n+1} + (1 - n(\beta+\delta)\Delta t)p_n(t) + o(\Delta t) \tag{3.5.129}$$

and, finally

$$\frac{dp_n(t)}{dt} = -n(\beta+\delta)p_n(t) + (n-1)\beta p_{n-1}(t) + (n+1)\delta p_{n+1}(t). \tag{3.5.130}$$

This system has to be supplemented by the initial condition

$$p_n(0) = \begin{cases} 1 \text{ for } n = n_0, \\ 0 \text{ for } n \neq n_0. \end{cases} \tag{3.5.131}$$

Remark 3.30. Equations similar to (3.5.130) can occur in many other ways, not necessarily describing stochastic processes. In general, we can consider population consisting of individuals differentiated by a single feature, e.g., we can consider cells having n copies of a particular gen. Here, $u_n(t)$ will be the number of individuals having n copies of this gen. Due to mutations or other environmental influence, the number of genes can increase or decrease. We may assume that at sufficiently small period of time only one change may occur. Denoting by β_n and δ_n the rates of increasing (resp. decreasing) the number of genes if there are n of them, by the argument used above, we have

$$u_n'(t) = -(\beta_n + \delta_n)u_n(t) + \delta_{n+1}u_{n+1}(t) + \beta_{n-1}u_{n-1}(t), \quad n \geq 0.$$

Contrary to (3.5.120), the system (3.5.130) is much more difficult to solve. In fact, even proving that there is a solution to it is a highly nontrivial exercise. In what follows, we assume that $(p_0(t), p_1(t), \ldots,)$ exists and describes a probability; that is

$$\sum_{n=0}^{\infty} p_n(t) = 1, \quad t \geq 0. \tag{3.5.132}$$

Then, we will be able to find formulae for p_n by the generating function method. We define

$$F(t,x) = \sum_{n=0}^{\infty} p_n(t)x^n$$

Since $p_n \geq 0$, by (3.5.132), the generation function is defined in the closed circle $|x| \leq 1$ and analytic in $|x| < 1$. The generating function has the following properties:

(1) The probability of extinction at time t, $p_0(t)$, is given by

$$p_0(t) = F(t,0). \tag{3.5.133}$$

(2) The probabilities $p_n(t)$ are given by

$$p_n(t) = \left.\frac{1}{n!}\frac{\partial^n F}{\partial x^n}\right|_{x=0} \tag{3.5.134}$$

If $F(t,x)$ is analytic in a little larger circle, containing $x = 1$, we can use F to find other useful quantities. The expected value of $N(t)$ at time t is defined by

$$E(N(t)) = \sum_{n=0}^{\infty} np_n(t)$$

On the other hand,

$$\frac{\partial F}{\partial x}(t,x) = \sum_{n=0}^{\infty} np_n(t)x^{n-1}$$

so that

$$E[N(t)] = \sum_{n=0}^{\infty} np_n(t) = \left.\frac{\partial F}{\partial x}\right|_{x=1} \tag{3.5.135}$$

Similarly, the variance is defined by

$$Var[N(t)] = E[N^2(t)] - (E[N(t)])^2.$$

On the other hand,

$$\left.\frac{\partial^2 F}{\partial x^2}(t,x)\right|_{x=1} = \sum_{n=0}^{\infty} n(n-1)p_n(t) = E[N^2(t)] - E[N(t)].$$

Combining these formulae, we get

$$Var[N(t)] = \left.\left(\frac{\partial^2 F}{\partial x^2} + \frac{\partial F}{\partial x} - \left(\frac{\partial F}{\partial x}\right)^2\right)\right|_{x=1} \tag{3.5.136}$$

Let us find the equation satisfied by F. Using (3.5.130) and remembering that $p_{-1} = 0$, we have

$$\begin{aligned}\frac{\partial F}{\partial t}(t,x) &= \sum_{n=0}^{\infty} n\frac{dp_n}{dt}(t) = -(\beta+\delta)\sum_{n=0}^{\infty} np_n(t)x^n \\ &+\beta\sum_{n=0}^{\infty}(n-1)p_{n-1}(t)x^n + \delta\sum_{n=0}^{\infty}(n+1)p_{n+1}(t)x^n \\ &= -(\beta+\delta)x\frac{\partial F}{\partial x}(t,x) + \beta x^2\frac{\partial F}{\partial x}(t,x) + \delta\frac{\partial F}{\partial x}(t,x).\end{aligned}$$

That is, to find F we have to solve the equation

$$\frac{\partial F}{\partial t} = \left(\beta x^2 - (\beta+\delta)x + \delta\right)\frac{\partial F}{\partial x}. \tag{3.5.137}$$

supplemented by the initial condition

$$F(0,x) = x^{n_0}.$$

The equation can be solved by characteristics. This problem is slightly simpler than the McKendrick-van Foerster equation: F is constant along characteristics, which are given by

$$\frac{dx}{dt} = -(\beta x - \delta)(x-1)$$

that is

$$\begin{aligned} -t + C &= \int \frac{dt}{(\beta x - \delta)(x-1)} = \frac{1}{\beta - \delta}\left(-\int \frac{dx}{x - \frac{\delta}{\beta}} + \int \frac{dx}{x-1}\right) \\ &= \frac{1}{\beta-\delta}\ln\left|\frac{x-1}{x-\frac{\delta}{\beta}}\right| \end{aligned}$$

provided $\beta \neq \delta$ and $x \neq 1, \delta/\beta$. This gives

$$\left|\frac{\beta x - \delta}{x-1}\right| = Ce^{rt}$$

where $r = \beta - \delta$. Thus, we have the general solution

$$F(t,x) = G\left(e^{-rt}\left|\frac{\beta x-\delta}{x-1}\right|\right),$$

where G is an arbitrary function. Using the initial condition, we get

$$x^{n_0} = G\left(\left|\frac{\beta x - \delta}{x-1}\right|\right)$$

Assume $x < min\{1, \delta/\beta\}$ or $x > \max\{1, \delta/\beta\}$ so that we can drop absolute value bars. Solving

$$s = \frac{\beta x - \delta}{x-1}$$

we get

$$x = \frac{s-\delta}{s-\beta}$$

so that

$$G(s) = \left(\frac{s-\delta}{s-\beta}\right)^{n_0}.$$

Thus, the solution is given by

$$F(x,t) = \left(\frac{e^{-rt}\frac{\beta x - \delta}{x-1} - \delta}{e^{-rt}\frac{\beta x-\delta}{x-1} - \beta}\right)^{n_0} = \left(\frac{e^{rt}\delta(1-x) + (\beta x - \delta}{e^{rt}\beta(1-x) + (\beta x - \delta)}\right)^{n_0} \tag{3.5.138}$$

Consider zero of the denominator:

$$x = \frac{e^{rt} - \frac{\delta}{\beta}}{e^{rt}-1}$$

If $\delta/\beta < 1$, then $r > 0$ and we see that $x > 0$ and, as $t \to \infty$, x moves from $+\infty$ to 1 and thus F is analytical in the circle stretching from the origin to the first singularity, which is bigger than 1 for any finite t. If $\delta/\beta > 1$, then $r < 0$ and x above is again positive and moves from infinity to $\delta/\beta > 1$ so again F is analytic in a circle with radius bigger than 1. Since we know that the generating function (defined by the series, coincides with F defined above for $|x| < \min\{1, \delta/\beta\}$, by the principle of analytic continuation, the generation function coincides with F in the whole domain of its analyticity (note that this is not necessarily solution of the equation (3.5.137) outside this region as we have removed the absolute value bars).

Consider now the case $\beta = \delta$. Then the characteristic equation is

$$\frac{dx}{dt} = -\beta(x-1)^2$$

solving which we obtain

$$\frac{1}{x-1} = \beta t + \xi,$$

or

$$\xi = \frac{1 - x\beta t + \beta t}{x-1}.$$

Hence, the general solution is given by

$$F(t,x) = G\left(\frac{1 - x\beta t + \beta t}{x-1}\right).$$

Using the initial condition, we have

$$x^{n_0} = G\left(\frac{1}{x-1}\right).$$

Defining

$$s = \frac{1}{x-1}$$

or

$$x = 1 + \frac{1}{s}.$$

Hence

$$G(s) = \left(1 + \frac{1}{s}\right)^{n_0}.$$

Therefore

$$F(t,x) = \left(1 + \frac{x-1}{1 - x\beta t + \beta t}\right)^{n_0} = \left(\frac{\beta t + (1-\beta t)x}{1 - x\beta t + \beta t}\right)^{n_0}.$$

Summarizing,

$$F(t,x) = \begin{cases} \left(\frac{e^{rt}\delta(1-x) + (\beta x - \delta)}{e^{rt}\beta(1-x) + (\beta x - \delta)}\right)^{n_0} & \text{if } \beta \neq \delta \\ \left(\frac{\beta t + (1-\beta t)x}{1 - x\beta t + \beta t}\right)^{n_0} & \text{if } \beta = \delta \end{cases} \tag{3.5.139}$$

Let us complete this section by evaluating some essential parameters. The probability of extinction at time t is given by

$$p_0(t) = F(t,0) = \begin{cases} \left(\frac{\delta(e^{rt}-1)}{e^{rt}\beta - \delta}\right)^{n_0} & \text{if } \beta \neq \delta \\ \left(\frac{\beta t}{1+\beta t}\right)^{n_0} & \text{if } \beta = \delta. \end{cases} \tag{3.5.140}$$

Hence, the asymptotic probability of extinction is given by

$$\lim_{t\to\infty} p_0(t) = \begin{cases} \left(\frac{\delta}{\beta}\right)^{n_0} & \text{if } \beta > \delta \\ 1 & \text{if } \beta \leq \delta. \end{cases} \tag{3.5.141}$$

We note that even for positive net growth rates $\beta > \delta$ the probability of extinction is non-zero. Populations with small initial numbers are especially susceptible to extinction.

To derive the expected size of the population we use (3.5.135). We have

$$
\begin{aligned}
E[N(t)] &= \left.\frac{\partial F}{\partial x}\right|_{x=1} \\
&= n_0\left(\frac{e^{rt}\delta(1-x)+(\beta x-\delta)}{e^{rt}\beta(1-x)+(\beta x-\delta)}\right)^{n_0-1} \\
&\left.\frac{(-e^{rt}\delta+\beta)(e^{rt}\beta(1-x)+(\beta x-\delta))+\beta(e^{rt}-1)(e^{rt}\delta(1-x)+(\beta x-\delta))}{(e^{rt}\beta(1-x)+(\beta x-\delta))^2}\right|_{x=1} \\
&= n_0\frac{(-e^{rt}\delta+\beta)(\beta-\delta)+\beta(e^{rt}-1)(\beta-\delta)}{(\beta-\delta)^2} \\
&= n_0e^{rt}
\end{aligned}
$$

To get the variance, we have to find the second derivative. It is given by

$$
\begin{aligned}
&\frac{\partial^2 F}{\partial x^2} \\
&= n_0\left(\frac{x\beta-\delta+e^{rt}(1-x)\delta}{e^{rt}(1-x)\beta+x\beta-\delta}\right)^{-1+n_0} \\
&\left(-\frac{2(\beta-e^{rt}\beta)(\beta-e^{rt}\delta)}{(e^{rt}(1-x)\beta+x\beta-\delta)^2}+\frac{2(\beta-e^{rt}\beta)^2(x\beta-\delta+e^{rt}(1-x)\delta)}{(e^{rt}(1-x)\beta+x\beta-\delta)^3}\right)+ \\
&(-1+n_0)n_0\left(\frac{x\beta-\delta+e^{rt}(1-x)\delta}{e^{rt}(1-x)\beta+x\beta-\delta}\right)^{-2+n_0} \\
&\left(\frac{\beta-e^{rt}\delta}{e^{rt}(1-x)\beta+x\beta-\delta}-\frac{(\beta-e^{rt}\beta)(x\beta-\delta+e^{rt}(1-x)\delta)}{(e^{rt}(1-x)\beta+x\beta-\delta)^2}\right)^2
\end{aligned}
$$

Hence

$$
\begin{aligned}
Var[N(t)] &= \left.\left(\frac{\partial^2 F}{\partial x^2}+\frac{\partial F}{\partial x}-\left(\frac{\partial F}{\partial x}\right)^2\right)\right|_{x=1} \\
&= n_0\left(\frac{2(\beta-e^{rt}\beta)^2}{(\beta-\delta)^2}-\frac{2(\beta-e^{rt}\beta)(\beta-e^{rt}\delta)}{(\beta-\delta)^2}\right)+n_0\left(-\frac{\beta-e^{rt}\beta}{\beta-\delta}+\frac{\beta-e^{rt}\delta}{\beta-\delta}\right) \\
&+(-1+n_0)n_0\left(-\frac{\beta-e^{rt}\beta}{\beta-\delta}+\frac{\beta-e^{rt}\delta}{\beta-\delta}\right)^2-{n_0}^2\left(-\frac{\beta-e^{rt}\beta}{\beta-\delta}+\frac{\beta-e^{rt}\delta}{\beta-\delta}\right)^2 \\
&= \quad \frac{e^{rt}(-1+e^{rt})n_0(\beta+\delta)}{\beta-\delta}
\end{aligned}
$$

for $\beta \neq \delta$, while for $\beta = \delta$ we obtain

$$V(t) = 2n_0\beta t.$$

4

Discrete time non-linear models for interacting species and age structured populations

System of discrete equations occur when we have two, or more, interacting species. However, we also have seen systems in age structured one-species models. They were linear but can be easily generalized to non-linear by introducing density dependent coefficients (such as logistic growth). We have discuss two such systems, next we introduce tools for their analysis, and finally provide stability analysis of them.

1 Models

1.1 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 youngs. 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:

1. Hosts that have been parasitized will give rise to the next generation of parasitoids.
2. Hosts that have not been parasitized will give rise to their own prodigy.
3. 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_t – density (number) of host species in generation t,
- P_t – density (number) of parasitoid in generation t,
- $f = f(N_t, P_t)$ – 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:

$$\begin{aligned} N_{t+1} &= \lambda N_t f(N_t, P_t), \\ P_{t+1} &= c N_t (1 - f(N_t, P_t). \end{aligned} \tag{4.1.1}$$

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 to the list 1)-3).

4. Encounters occur randomly. The number of encounters N_e of the host with the parasitoid is therefore proportional to the product of their densisties (numbers):
$$N_e = \alpha N_t P_t,$$
where α is a constant, which represents the searching efficiency of the parasitoids. (This kind of assumption presupposing random encounters is is known as the *law of mass action.*)
5. 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.

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

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

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

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

that is, by 4.,

$$\mu = a P_t.$$

Hence, the probability of a host not having any encounter with parasitoid is

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

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

$$\begin{aligned} N_{t+1} &= \lambda N_t e^{-aP_t}, \\ P_{t+1} &= cN_\lambda(1 - e^{-aP_t}) \end{aligned} \tag{4.1.2}$$

1.2 Non-linear age structured model.

Consider a single species population with two age classes: juveniles and adults. Let X_t be the number of juveniles at time t and Y_t be the number of adults. We assume that the fertility rate for adults is b, c is the survival rate of juveniles; that is a fraction c of juveniles present at time t become adults at $t + 1$ and the rest dies. In each time period only the density dependent fraction $s - DY_t$ of the adult population survives. These assumptions lead to the system

$$\begin{aligned} X_{t+1} &= bY_t, \\ X_{t+1} &= cX_t + Y_t(s - DY_t). \end{aligned} \tag{4.1.3}$$

We re-write this equation in a form which is more convenient for analysis by introducing new unknowns $X_t = b\hat{X}_t/D$ and $Y_t = \hat{Y}_t/D$, which converts (4.1.3) into

$$\begin{aligned} \hat{X}_{t+1} &= \hat{Y}_t, \\ \hat{X}_{t+1} &= a\hat{X}_t + \hat{Y}_t(s - \hat{Y}_t), \end{aligned} \tag{4.1.4}$$

where $a = cb > 0$.

1.3 SIR model

Let us consider the population divided into three classes: susceptibles S, infectives I and removed (immune or dead) R. We do not consider any births in the process. Within one cycle from time k to time $k + 1$ the probability of an infective meeting someone is α' and thus meeting a susceptible is $\alpha' S/N$ where N is the size of the population at time k; further a fraction α'' of these encounters results in an infection. We denote $\alpha = \alpha'\alpha''$. 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. This results in the system

$$\begin{aligned} S(k+1) &= S(k) - \frac{\alpha}{N} I(k)S(k) + \beta(I(k) + R(k)) \\ I(k+1) &= I(k) + \frac{\alpha}{N} I(k)S(k) - \gamma I(k) - \beta I(k) \\ R(k+1) &= R(k) - \beta R(k) + \gamma I(k) \end{aligned} \tag{4.1.5}$$

We observe that

$$S(k+1) + I(k+1) + R(k+1) = S(k) + I(k) + R(k) = const = N$$

so that the total population does not change in time.

This can be used to reduce the (4.1.5) to a two dimensional system

$$\begin{aligned} S(k+1) &= S(k) - \frac{\alpha}{N} I(k)S(k) + \beta(N - S(k)) \\ I(k+1) &= I(k)(1 - \gamma - \beta) + \frac{\alpha}{N} I(k)S(k). \end{aligned} \tag{4.1.6}$$

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

2 Stability analysis

In both cases (that is, for the host-parasitoid models of for the age structured population model) our interest is in finding and determining stability of the equilibria. For this, however, we have to do some mathematics.

We shall be concerned with autonomous systems of difference equations

$$\mathbf{x}(n+1) = \mathbf{f}(\mathbf{x}(n)), \tag{4.2.7}$$

where $\mathbf{x}(0) = \mathbf{x}_0$ is given. Here, $\mathbf{x} = (x_1, \ldots, x_N)$ and $\mathbf{f}(t) = \{f_1(t), \ldots, f_N(t)\}$ is a continuous function from $\mathbb{R}^N$ into $\mathbb{R}^N$. In what follows, $\|\cdot\|$ is any norm on $\mathbb{R}^N$, unless specified otherwise.

As in the scalar case, $\mathbf{x}^* \in \mathbb{R}^N$ is called an equilibrium point of (4.2.7) if

$$\mathbf{x}^* = \mathbf{f}(\mathbf{x}^*) \tag{4.2.8}$$

The definition of stability is analogous to the scalar case.

Definition 4.1. *(a) The equilibrium $\mathbf{x}^*$ is stable if for given $\epsilon > 0$ there is $\delta > 0$ such that for any $\mathbf{x}$ and for any $n > 0$, $\|\mathbf{x} - \mathbf{x}^*\| < \delta$ implies $\|\mathbf{f}^n(\mathbf{x}) - \mathbf{x}^*| < \epsilon$ for all $n > 0$. If $\mathbf{x}^*$ is not stable, then it is called unstable (that is, $\mathbf{x}^*$ is unstable if there is $\epsilon >$ such that for any $\delta > 0$ there are $\mathbf{x}$ and n such that $\|\mathbf{x} - \mathbf{x}^*\| < \delta$ and $\|\mathbf{f}^n(\mathbf{x}) - \mathbf{x}^*\| \geq \epsilon$.)*

(b) The point $\mathbf{x}^*$ is called attracting if there is $\eta > 0$ such that

$$\|\mathbf{x}(0) - x^*\| < \eta \text{ implies } \lim_{n\to\infty} \mathbf{x}(n) = \mathbf{x}^*.$$

If $\eta = \infty$, then $\mathbf{x}^*$ is called a global attractor or globally attracting.

(c) The point $\mathbf{x}^*$ is called an asymptotically stable equilibrium if it is stable and attracting. If $\eta = \infty$, ten $\mathbf{x}^*$ is said to be globally asymptotically stable equilibrium.

It is worthwhile to note that in higher dimension we may have unstable and attracting equilibria.

2.1 Stability of linear systems

We consider the linear autonomous system

$$\mathbf{x}(n+1) = \mathcal{A}\mathbf{x}(n), \quad \mathbf{x}(0) = \overset{\circ}{\mathbf{x}}, \tag{4.2.9}$$

We assume that $\mathcal{A}$ is non-singular. The origin $\mathbf{0}$ is always an equilibrium point of (4.2.9). We have the following result:

Theorem 4.2. *The following statements hold:*

1. *The zero solution of (4.2.9) is stable if and only if the spectral radius of $\mathcal{A}$ satisfies $\rho(\mathcal{A}) \leq 1$ and the eigenvalues of unit modulus are semi-simple;*
2. *The zero solution is asymptotically stable if and only if $\rho(\mathcal{A}) < 1$.*

Proof. Let $\lambda_1, \ldots, \lambda_k$ be distinct eigenvalues of $\mathcal{A}$, each with algebraic multiplicity n_i so that $n_1 + \ldots + n_k = N$. We assume that $|\lambda_1| \geq |\lambda_2| \geq \ldots |\lambda_k| > 0$. For each $1 \leq r \leq k$, let $\mathbf{v}_r^1, \ldots, \mathbf{v}_r^{n_r}$ be the set of eigenvectors and associated eigenvectors belonging to λ_r. Each $\mathbf{v}_r^j$ is a solution to

$$(\mathcal{A} - \lambda_r \mathcal{I})^{m_j^r} \mathbf{v}_r^j = 0$$

for some $1 \leq m_j^r \leq n_r$ (some (even all) js may correspond to the same m_j^r. Then we can write the solution as

$$\mathcal{A}^n\mathbf{x}_0 = \sum_{r=1}^{k}\left(\sum_{j=1}^{n_r} c_r^j \mathcal{A}^n \mathbf{v}_r^j\right) \tag{4.2.10}$$

where c_r^j are coefficients of the expansion of $\mathbf{x}_0$ in the basis consisting of $\mathbf{v}_r^j$, $1 \leq r \leq k$, and $1 \leq j \leq n_r$ and

$$\begin{aligned}\mathcal{A}^n\mathbf{v}_r^j &= (\lambda_r\mathcal{I} + \mathcal{A} - \lambda_r\mathcal{I})^n\mathbf{v}_r^j = \sum_{l=0}^{n}\lambda_r^{n-l}\binom{n}{l}(\mathcal{A}-\lambda_r\mathcal{I})^l\mathbf{v}_r^j \\ &= \Big(\lambda_r^n\mathcal{I} + n\lambda_r^{n-1}(\mathcal{A}-\lambda_r\mathcal{I}) + \ldots \\ &\quad + \frac{n!}{(m_j^r-1)!(n-m_j^r+1)!}\lambda_r^{n-m_j^r+1}(\mathcal{A}-\lambda_r\mathcal{I})^{m_j^r-1}\Big)\mathbf{v}_r^j,\end{aligned} \tag{4.2.11}$$

It is important to note that (4.2.11) is a finite sum for any n as the term $(\mathcal{A}-\lambda_r\mathcal{I})^{m_j^r}\mathbf{v}_r^j$ and all subsequent ones are zero. Using the triangle inequality for norms, we obtain

$$\begin{aligned}&\|\mathcal{A}^n\mathbf{v}_r^j\| \\ &\leq |\lambda_r|^n\Big(1 + n|\lambda_r|^{-1}(\|\mathcal{A}\| + |\lambda_r|) + \ldots \\ &\quad + P_{m_j^r-1}(n)|\lambda_r|^{-m_j^r+1}(|\mathcal{A}| + |\lambda_r|)^{m_j^r-1}\Big)\|\mathbf{v}_r^j\| \\ &= |\lambda_r|^n n^{m_j^r-1}\Big(n^{-m_j^r+1} + n^{m_j^r-2}|\lambda_r|^{-1}(\|\mathcal{A}\| + |\lambda_r|) + \ldots \\ &\quad + \frac{P_{m_j^r-1}(n)}{n^{m_j^r-1}}|\lambda_r|^{-m_j^r+1}(\|\mathcal{A}\| + |\lambda_r|)^{m_j^r-1}\Big)\|\mathbf{v}_r^j\| \\ &\leq C_j^r|\lambda_r|^n n^{m_j^r-1} \leq C_j^r|\lambda_r|^n n^{n_r-1},\end{aligned} \tag{4.2.12}$$

where the constant C_j^r does not depend on n and we used $m_j^r \leq n_r$. Next we observe that the vector consisting of constants c_r^j is given by

$$\mathbf{c} = \begin{pmatrix} | & \ldots & | \\ \mathbf{v}_1^1 & \ldots & \mathbf{v}_k^{n_k} \\ | & \ldots & | \end{pmatrix}^{-1}\mathbf{x}_0$$

and thus, for some constant M

$$\|\mathbf{c}\| \leq M\|\mathbf{x}_0\|. \tag{4.2.13}$$

Assume now that $\rho(\mathcal{A}) < 1$; that is all eigenvalues have absolute values smaller than 1. Then

$$\|\mathcal{A}^n\mathbf{x_0}\| \leq \sum_{r=1}^{k}|\lambda_r|^n n^{n_r-1}\left(\sum_{j=1}^{n_r}|c_r^j|C_j^r\right) \leq M'\|\mathbf{x}_0\|\sum_{r=1}^{k}|\lambda_r|^n n^{n_r-1}$$

where

$$M' = M\max_{1\leq r\leq k}\sum_{j=1}^{n_r}C_j^r$$

and we used the fact that in (4.2.13) we can use $\|\mathbf{c}\| = \max\{|c_r^j|\}$. From $\rho(\mathcal{A}) < 1$ we infer that $1 > |\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_k|$ and hence there is $1 > \eta > |\lambda_1|$. With this η, we have $|\lambda_i|\eta^{-1} \leq \eta_0 < 1$ for any $i = 1, \ldots, k$ and

$$\|\mathcal{A}^n\mathbf{x_0}\| \leq M'k\|\mathbf{x}_0\|\eta^n\eta_0^n n^{N-1}$$

Now, for any $a < 1$ and $k > 0$ we have $\lim\limits_{n\to\infty} a^n n^k = 0$ so that $a^n n^k \leq L$ for some constant L. Thus, there is are constants $K > 0$ and $0 < \eta < 1$ such that

$$\|\mathcal{A}^n\mathbf{x_0}\| \leq K\|\mathbf{x}_0\|\eta^n \tag{4.2.14}$$

and the zero solution is asymptotically stable.

If there are eigenvalues of unit modulus but they are semi-simple, then the expansions (4.2.13) reduce to first term ($j_1 = 1$ in each case) so that in such a case

$$\|\mathcal{A}^n\mathbf{v}_r^j\| \leq C_j^r$$

and the solution is stable (but not asymptotically stable).

If an eigenvalue λ_r is not semi-simple, then some m_j^r is bigger then 1 and we have polynomial entries in (4.2.11). Consider an associated eigenvector $\mathbf{v}_r^j$ corresponding to this case. Then

$$\begin{aligned}
&\|\mathcal{A}^n\mathbf{v}_r^j\| \\
&= \|\mathbf{v}_r^j + n\lambda_r^{-1}(\mathcal{A}-\lambda_r)\mathbf{v}_r^j + \ldots + P_{m_j^r-1}(n)\lambda_r^{-m_j^r+1}(\mathcal{A}-\lambda_r)^{m_j^r-1}\mathbf{v}_r^j\| \\
&\geq n^{m_j^r-1}\Big| n^{-m_j^r+1}|P_{m_j^r-1}(n)|\|(\mathcal{A}-\lambda_r)^{m_j^r-1}\mathbf{v}_r^j\| \\
&\quad - n^{-1}\|n^{-m_j^r+2}\mathbf{v}_r^j + n^{-m_j^r+3}\lambda_r^{-1}(\mathcal{A}-\lambda_r)\mathbf{v}_r^j + \ldots\|\Big|
\end{aligned}$$

The coefficient $\|(\mathcal{A}-\lambda_r)^{m_j^r-1}\mathbf{v}_r^j\|$ is non-zero and the first term inside the absolute value bars converges to a finite limit $(1/(m_j^r-1)!)$ and the second to zero, hence $\|\mathcal{A}^n\mathbf{v}_r^j\|$ diverges to infinity. Thus, taking initial conditions of the form $\epsilon\mathbf{v}_r^l$ we see that we can take arbitrarily small initial condition, the resulting solution is unbounded and thus the zero solution is unstable.

Finally, if $|\lambda_1| > 1$, then argument as above gives instability of the zero solution. □

2.2 Stability by linearisation

Let us first note the following result.

Lemma 4.3. *If* $\mathbf{f}$ *has continuous partial derivatives of the first order in some neighbourhood of* $\mathbf{y}^0$*, then*

$$\mathbf{f}(\mathbf{x}+\mathbf{y}^0) = \mathbf{f}(\mathbf{y}^0) + \mathcal{A}\mathbf{x} + \mathbf{g}(\mathbf{x}) \tag{4.2.15}$$

where

$$\mathcal{A} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{y^0}) & \ldots & \frac{\partial f_1}{\partial x_n}(\mathbf{y^0}) \\ \vdots & & \vdots \\ \frac{\partial f_1}{\partial x_n}(\mathbf{y^0}) & \ldots & \frac{\partial f_n}{\partial x_n}(\mathbf{y^0}) \end{pmatrix},$$

and $\mathbf{g}(\mathbf{x})/\|\mathbf{x}\|$ *is continuous in some neighbourhood of* $\mathbf{y}^0$ *and vanishes at* $\mathbf{x} = \mathbf{y}^0$.

If $\mathcal{A}$ be the matrix defined above. We say that

$$\mathbf{y}_{t+1} = \mathcal{A}\mathbf{y}_t \tag{4.2.16}$$

is the linearization of (4.2.7) around an equilibrium $\mathbf{x}^*$.

We also note that to solve the nonhomogeneous system of equations

$$\mathbf{x}(n+1) = \mathcal{A}\mathbf{x}(n) + \mathbf{g}(n), \tag{4.2.17}$$

where $\mathbf{x} = (y_1, \ldots, y_k)$, $\mathbf{g} = (g_1, \ldots, g_k)$ and $\mathcal{A} = \{a_{ij}\}_{1\leq i,j\leq k}$. Exactly as in Subsection **??** we find that the solution to (4.2.17) satisfying the initial condition $\mathbf{x}(0) = \mathbf{x}^0$ is given by the formula

$$\mathbf{x}(n) = \mathcal{A}^n\mathbf{x}^0 + \sum_{r=0}^{n-1}\mathcal{A}^{n-r-1}\mathbf{g}(r). \tag{4.2.18}$$

We shall need a discrete version of Gronwall's lemma.

Lemma 4.4. *Let $z(n)$ and $h(n)$ be two sequences of real numbers, $n \geq n_0 > 0$ and $h(n) \geq 0$.. If*

$$z(n) \leq M\left(z(n_0) + \sum_{j=n_0}^{n-1} h(j)z(j)\right) \tag{4.2.19}$$

for some $M > 0$, then

$$z(n) \leq z(n_0)\prod_{j=n_0}^{n-1}(1 + Mh(j)) \tag{4.2.20}$$

$$z(n) \leq z(n_0)\exp\sum_{j=n_0}^{n-1} Mh(j) \tag{4.2.21}$$

Proof. Consider the equation

$$u(n) = M\left(u(n_0) + \sum_{j=n_0}^{n-1} h(j)u(j)\right), \quad u(n_0) = z(n_0).$$

From non-negativity, by induction we obtain $z(n) \leq u(n)$ for $n \geq n_0$. Hence

$$u(n+1) - u(n) = Mh(n)u(n)$$

or, equivalently,

$$u(n+1)) = (1 + Mh(n))u(n)$$

so

$$u(n) = u(n_0)\prod_{j=n_0}^{n-1}(1 + Mh(j))$$

which proves (4.2.20). The second follows from the formula $1 + Mh(j) \leq \exp(Mh(j))$. □

Theorem 4.5. *Assume that $\mathbf{f}$ is a C^1 function and $\mathbf{x}^*$ is an equilibrium point. If the zero solution of the linearised system (4.2.16) is asymptotically stable, then the equilibrium $\mathbf{x}^*$ is asymptotically stable.*

Proof. We have

$$\mathbf{x}(n+1) = \mathbf{f}(\mathbf{x}(n)) = \mathbf{f}(\mathbf{x}(n)) - \mathbf{f}(\mathbf{x}^*) + \mathbf{x}^*.$$

Denoting $\mathbf{y}(n) = \mathbf{x}(n) - \mathbf{x}^*$ and using Lemma 4.3 we obtain

$$\mathbf{y}(n+1) = \mathcal{A}\mathbf{y}(n) + \mathbf{g}(\mathbf{y}n),$$

so that, by (4.2.18),

$$\mathbf{y}(n) = \mathcal{A}^n\mathbf{y}(0) + \sum_{r=0}^{n-1}\mathcal{A}^{n-r-1}\mathbf{g}(\mathbf{y}(r)).$$

Since the condition (4.2.14) is equivalent to asymptotic stability of the linearized system, we get

$$\|\mathbf{y}(n)\| \leq K\eta^n\|\mathbf{y}(0)\| + K\eta^{-1}\sum_{r=0}^{n-1}\eta^{n-r}\|\mathbf{g}(\mathbf{y}(r))\|.$$

For a given $\epsilon > 0$, there is $\delta > 0$ such that $\|\mathbf{g}(\mathbf{y})\| < \epsilon\|\mathbf{y}\|$ whenever $\|\mathbf{y}\| < \delta$. So, as long as we can keep $\|\mathbf{y}(r)\| < \delta$ for $r \leq n-1$

$$\eta^{-n}\|\mathbf{y}(n)\| \leq K\|\mathbf{y}(0)\| + K\epsilon\sum_{r=0}^{n-1}\eta^{-r-1}\|\mathbf{y}(r)\|.$$

Applying the Gronwall inequality for $z(n) = \eta^{-n}\|\mathbf{y}(n)\|$ we obtain

$$\eta^{-n}\|\mathbf{y}(n)\| \leq \|\mathbf{y}(0)\| \prod_{j=0}^{n-1}(1+K\epsilon\eta^{-1}).$$

Thus

$$\|\mathbf{y}(n)\| \leq \|\mathbf{y}(0)\|(\eta+K\epsilon)^n.$$

Choose $\epsilon < (1-\eta)/K$ so that $\eta + K\eta < 1$. Thus, by induction, $\|\mathbf{y}(0)\| < \delta$, we have $\|\mathbf{y}(n)\| < \|\mathbf{y}(0)\| < \delta$ and the equilibrium is asymptotically stable.

□

It can be also proved that if $\rho(\mathcal{A}) > 1$, then the equilibrium is unstable but the proof is more involved.

3 Stability analysis of models

3.1 SIR model

Let us start with finding the equilibria of (4.1.6). These are solutions of

$$\begin{aligned} S &= F_1(S,I) = S - \frac{\alpha}{N}IS + \beta(N-S) \\ I &= F_21(S,I) = I(1-\gamma-\beta) + \frac{\alpha}{N}IS. \end{aligned} \tag{4.3.22}$$

$I = 0$ is a solution of this system with corresponding $S = N$ so this is a disease-free equilibrium. If $I \neq 0$, then dividing the second equation by I we find $S = N\delta/\alpha$ which yields $I = \beta N(\alpha-\delta)/\alpha\delta$ which is an endemic disease equilibrium. Thus

$$E_1^* = (N,0), \qquad E_2^* = \left(\frac{N\delta}{\alpha}, \frac{\beta N(\alpha-\delta)}{\alpha\delta}\right).$$

To find the Jacobian, we calculate

$$\begin{aligned} F_{1,S}(S,I) = 1 - \frac{\alpha}{N}I - \beta,\ F_{1,I}(S,I) = -\frac{\alpha}{N} \\ F_{2,S}(S,I) = \frac{\alpha}{N}I,\ F_{2,I}(S,I) = 1 - \delta + \frac{\alpha}{N}S, \end{aligned}$$

thus we have

$$J_{E_1^*} = \begin{pmatrix} 1-\beta & -\alpha \\ 0 & 1-\delta+\alpha \end{pmatrix}$$

and

$$J_{E_2^*} = \begin{pmatrix} 1-\frac{\alpha\beta}{\delta} & -\delta \\ \frac{\alpha\beta}{\delta}-\beta & 1 \end{pmatrix}.$$

To determine whether the magnitude of the eigenvalues is smaller or larger than 1 we could find the eigenvalues and directly compute their magnitude but this is in general time consuming and not always informative. There are other, easier methods.

Interlude: How to determine whether eigenvalues of a 2×2 matrix have magnitude less then 1 without solving the quadratic equation.

Consider the equation

$$\lambda^2 - B\lambda + A = 0$$

where B and A are real coefficients. The roots are given by

$$\lambda_{1,2} = \frac{B \pm \sqrt{B^2 - 4A}}{2}.$$

We consider two cases. First, let $B^2 - 4A > 0$ so that the roots are real. Then we must have

$$-2 - B < \sqrt{B^2 - 4A} < 2 - B$$

and

$$-2 - B < -\sqrt{B^2 - 4A} < 2 - B$$

Squaring the second inequality in the former expression, we obtain

$$1 - B + A > 0.$$

Similarly, squaring the first inequality in the second expression, we get

$$1 + B + A > 0.$$

Next, we get $2 - B > 0$ from the first and $-2 - B < 0$ from the second inequality, hence $|B| < 2$ and, since $B^2 - 4A \geq 0$, we have $A < 1$. Combining, we can write

$$|B| < 1 + A < 2 \tag{4.3.23}$$

Conversely, from (4.3.23) we get $-1 < B/2 < 1$ so that the midpoint between the roots is indeed inside $(-1, 1)$. Now, if $B > 0$, then we must only make sure that

$$\frac{B}{2} + \frac{\sqrt{B^2 - 4A}}{2} < 1.$$

This is equivalent to the following chain of inequalities (as $1 - B/2 > 0$)

$$\frac{\sqrt{B^2 - 4A}}{2} < 1 - \frac{B}{2} \iff \frac{B^2 - 4A}{4} < 1 - B + \frac{B^2}{4} \iff B < 1 + A$$

Similarly, if $B < 0$, then we must only make sure that

$$\frac{B}{2} - \frac{\sqrt{B^2 - 4A}}{2} > -1.$$

This is equivalent to the following chain of inequalities (as $1 + B/2 > 0$)

$$\frac{\sqrt{B^2 - 4A}}{2} < 1 + \frac{B}{2} \iff \frac{B^2 - 4A}{4} < 1 + B + \frac{B^2}{4} \iff -B < 1 + A.$$

Hence, (4.3.23) is sufficient.

Assume now that $4A - B^2 > 0$ so that the roots are complex conjugate. Since absolute values of complex conjugate numbers are equal, and A is the product of the roots, we must have $A < 1$ (in fact, $0 < A < 1$ from the condition on the discriminant). We must prove that

$$1 - |B| + A > 0. \tag{4.3.24}$$

But in this case, $|B| < 2\sqrt{A}$ so that if

$$1 - 2\sqrt{A} + A > 0$$

holds on $(0, 1)$, than (4.3.24) holds as well. But the former is nothing but $(1 - \sqrt{A})^2 > 0$ on this open interval. Hence, (4.3.24) is proved.

Conversely, assume (4.3.23) holds. Since in the first part we already proved that it yields the desired result if $4A - B^2 \leq 0$, we can assume that $4A - B^2 > 0$. This yields $A > 0$ and hence $0 < A < 1$ yields $\lambda\bar{\lambda} = |\lambda|^2 = A < 1$.

For a matrix

$$\mathcal{A} = \begin{pmatrix} a_{11} \; a_{12} \\ a_{21} \; a_{22} \end{pmatrix}$$

its eigenvalues are determined by solving

$$\begin{aligned} 0 &= det \begin{pmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{pmatrix} \\ &= \begin{pmatrix} a_{11} \; a_{12} \\ a_{21} \; a_{22} \end{pmatrix} - \begin{pmatrix} \lambda \; a_{12} \\ 0 \; a_{22} \end{pmatrix} - \begin{pmatrix} a_{11} \; 0 \\ a_{21} \; \lambda \end{pmatrix} \\ &= \lambda^2 - \lambda(a_{11} + a_{22}) + det\mathcal{A} = \lambda^2 - \lambda tr\, \mathcal{A} + det\mathcal{A} \end{aligned}$$

Hence, the condtion for stability can be expressed as

$$|tr\, \mathcal{A}| < 1 + det\, \mathcal{A} < 2 \tag{4.3.25}$$

Returning to our model, we find

$$|trJ_{E_1^*}| = |2 - \beta - \delta + \alpha| = 2 - \beta - \delta + \alpha$$

by assumptions on coefficients and

$$detJ_{E_1^*} = 1 - \delta + \alpha - \beta(1 - \delta + \alpha)$$

so that condition (4.3.25) can be written as

$$2 - \beta - \delta + \alpha < 2 - \delta + \alpha - \beta(1 - \delta + \alpha) < 2$$

Subtracting from both sides we obtain

$$0 < \beta(\delta - \alpha) < \delta - \alpha + \beta.$$

This gives $\delta - \alpha < 0$ while the second condition is automatically satisfied as $0 < \beta < 1$ and $(\delta - \alpha) > 0$ yields $\beta(\delta - \alpha) < \delta - \alpha < \beta + (\delta - \alpha)$. Hence, the equilibrium $(N, 0)$ is asymptotically stable if

$$\beta + \gamma > \alpha.$$

Consider the equilibrium at E_2^*. Here we have

$$|trJ_{E_2^*}| = \left|2 - \frac{\beta\alpha}{\delta}\right| = 2 - \frac{\beta\alpha}{\delta},$$

as $2(\gamma + \beta) - \beta\alpha = 2\gamma + \beta(2 - \alpha) > 0$, and

$$detJ_{E_2^*} = 1 - \frac{\beta\alpha}{\delta} + \alpha\beta - \delta\beta$$

so that condition (4.3.25) can be written as

$$2 - \frac{\beta\alpha}{\delta} < 2 - \frac{\beta\alpha}{\delta} + \alpha\beta - \delta\beta < 2$$

Subtracting from both sides, we get

$$0 < \beta(\alpha - \delta) < \frac{\beta\alpha}{\delta}$$

from where $\alpha - \delta > 0$. The second condition is equivalent to $(\alpha - \delta) < \alpha/\delta$; that is, $\delta(\alpha - \delta) < \alpha$ but this is always satisfied as $\delta < 1$. Hence, the endemic disease equilibrium

$$\left(\frac{N\delta}{\alpha}, \frac{\beta N(\alpha - \delta)}{\alpha\delta}\right)$$

is asymptotically stable provided

$$\alpha > \gamma + \delta.$$

We note that these conditions are consistent with the modelling process. The disease free equilibrium is stable if the infection rate is smaller than the rate of removal of infected individuals. On the other hand, in the opposite case we have an endemic disease.

3.2 Nicholson-Bailey model

Recall that the model is given by

$$\begin{aligned} N_{t+1} &= \lambda N_t e^{-aP_t}, \\ P_{t+1} &= cN_t(1-e^{-aP_t}). \end{aligned} \tag{4.3.26}$$

The equilibria are obtained by solving

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

This gives either trivial equilibrium $N = P = 0$ or

$$\lambda = e^{a\bar{P}};$$

that is,

$$\bar{P} = \frac{\ln \lambda}{a}, \tag{4.3.27}$$

and hence

$$\bar{N} = \frac{\lambda \ln \lambda}{(\lambda-1)ac}. \tag{4.3.28}$$

Clearly, $\lambda > 1$ for $\bar{N}$ to be positive. To analyse stability, we define

$$F(N,P) = Ne^{-aP}, \quad G(N,P) = cN(1-e^{-aP}).$$

Then, $F_N(N,P) = e^{-aP}, F_P(N,P) = -aNe^{-aP}$ and $G_N(N,P) = c(1-e^{-aP}), G_P(N,P) = cNe^{-aP}$ and we obtain the Jacobi matrix at $(0,0)$ as

$$\mathcal{A}|_{0,0} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

and

$$\mathcal{A}|_{\bar{N},\bar{P}} = \begin{pmatrix} 1 & -a\bar{N} \\ c\left(1-\frac{1}{\lambda}\right) & \frac{ca\bar{N}}{\lambda} \end{pmatrix}$$

In the subsequent considerations we use (4.3.25). At $(\bar{N},\bar{P})$. We obtain

$$\begin{aligned} tr\,\mathcal{A} &= 1 + \frac{\lambda}{\lambda-1}, \\ det\,\mathcal{A} &= \frac{ca\bar{N}}{\lambda} + ca\bar{N}\left(1-\frac{1}{\lambda}\right) = ca\bar{N} = \frac{\lambda\ln\lambda}{\lambda-1} \end{aligned}$$

We know that $\lambda > 1$. Consider the function

$$S(\lambda) = \lambda - 1 - \lambda \ln \lambda.$$

We have $S(1) = 0$, $S'(\lambda) = 1 - \ln\lambda - 1 = -\ln\lambda$ so that $S'(\lambda) < 0$ for $\lambda > 1$. Thus, $S(\lambda) < 0$ for $\lambda > 1$ and thus

$$\lambda \ln \lambda > \lambda - 1, \quad \lambda > 1.$$

Consequently,

$$det\,\mathcal{A} > 1$$

for any λ and the equilibrium is not stable.

Most natural parasitoid-host systems in nature are more stable than the Nicholson-Bailey seems to indicate and thus the model is not a satisfactory representation of real systems. We shall try to improve the system by modifying some parameters to see whether this could introduce stabilizing factors. We shall discuss the following modification:

In the absence of parasitoids, the host population grows to some limited density (determined by the carrying capacity K of the environment). Thus, the original system (4.3.26) would be amended as follows:

$$\begin{aligned} N_{t+1} &= \lambda(N_t)N_t e^{-aP_t}, \\ P_{t+1} &= cN_t(1-e^{-aP_t}), \end{aligned} \tag{4.3.29}$$

where for $\lambda(N_t)$ we might adopt

$$\lambda(N_t) = \exp r\left(1-\frac{N_t}{K}\right),$$

where $r > 0$. With this choice, we obtain a modified Nicholson-Bailey system

$$\begin{aligned} N_{t+1} &= N_t \exp\left(r\left(1-\frac{N_t}{K}\right) - aP_t\right), \\ P_{t+1} &= cN_t(1-\exp(-aP_t)), \end{aligned} \tag{4.3.30}$$

We simplify this system by introducing $n_t = N_t/K$ and $p_t = aP_t$. This converts (4.3.30) into

$$\begin{aligned} n_{t+1} &= n_t \exp\left(r(1-n_t) - p_t\right), \\ p_{t+1} &= Kcan_t(1-\exp(-p_t)), \end{aligned} \tag{4.3.31}$$

and, in what follows, we denote $Kca = C$.

The equilibria are obtained by solving solving

$$\begin{aligned} n &= n\exp\left(r\left(1-n\right)-p\right), \\ p &= Cn(1-\exp(-p)). \end{aligned}$$

We discard the trivial equilibrium $(0,0)$ so that we are left with

$$1 = \exp\left(r\left(1-n\right)-p\right), \tag{4.3.32}$$
$$p = Cn(1-\exp(-p)). \tag{4.3.33}$$

The equilibrium value $q = \bar{n} = \bar{N}/K$ is of interest in modelling as being the ratio of the steady-state host densities with and without parasitoid present. This gives

$$\bar{p} = r\left(1-\bar{n}\right) = r(1-q), \tag{4.3.34}$$
$$C\bar{n} = \frac{\bar{p}}{1-\exp(-\bar{p})}. \tag{4.3.35}$$

It is clear that one non-trivial equilibrium point is given by $\bar{n}_1 = 1$ $(\bar{N}_1 = K), \bar{P}_1 = 0$. Is there any other equilibrium point? To answer this question, we re-write (4.3.35) as

$$\bar{p} = C\bar{n}\left(1-\exp\left(-r\left(1-\bar{n}\right)\right)\right)$$

so that $\bar{n}$ satisfies

$$\frac{r\left(1-\bar{n}\right)}{C\bar{n}} = 1-\exp\left(r\left(1-\frac{\bar{N}}{K}\right)\right)$$

Define two functions

$$\begin{aligned} f_1(n) &= \frac{r\left(1-n\right)}{Cn} = \frac{r}{C}\left(\frac{1}{n}-1\right), \\ f_2(n) &= 1-\exp\left(-r\left(1-n\right)\right) \end{aligned}$$

First, we observe that, indeed, $f_1(1) = f_2(1) = 0$, which gives the equilibrium obtained above. Next,

$$f_1'(n) = -\frac{r}{Cn^2}, \qquad f_2'(n) = -r\exp\left(-r\left(1-n\right)\right)$$

hence both functions are decreasing for $n > 0$. Furthermore, $f_1'(1) = -\frac{r}{C}$ and $f_2'(K) = -r$. If we assume $C \geq 1$, then the graph of f_1 is below the graph of f_2 for n smaller than and close to $n = 1$. Furthermore, $f_2(0) = 1 - \exp(-r)$ and $f_1(N) \to +\infty$ as $n \to 0^+$. This implies the existence of at least one more equilibrium $(\bar{n}_2, \bar{p}_2)$. To show that there are no others, we find

$$f_1''(n) = \frac{2r}{Cn^3}, \qquad f_2''(n) = -r^2\exp\left(-r\left(1-n\right)\right)$$

so that f_1 is convex down and f_2 is convex down. In other words, $g(n) = f_1(n) - f_2(n)$ satisfies $g''(n) > 0$ which means that $g'(n)$ is strictly increasing and thus $g(n)$ can have at most two zeros. Thus, we found all possible equilibria of the system.

Let us focus on the last equilibrium describing coexistence of the parasitoid and the host. Let us consider stability of this equilibrium. The first step is to linearize the system around the equilibrium. To this end, we return to (4.3.30) and define

$$F(n,p) = n\exp\left(r\left(1-n\right)-p\right), G(n,p) = Cn(1-\exp(-p)),$$

and thus

$$\begin{aligned}
F_n(n,p) &= (1-rn)\exp\left(r\left(1-n\right)-p\right),\\
F_p(n,p) &= -n\exp\left(r\left(1-n\right)-p\right),\\
G_n(n,p) &= C(1-\exp(-p)),\\
G_p(n,p) &= Cn\exp(-p).
\end{aligned}$$

At $(\bar{n}_2, \bar{p}_2)$ we find, by (4.3.32),

$$\begin{aligned}
F_n(\bar{n}_2, \bar{p}_2) &= (1 - rn_2) = 1 - rq\\
F_p(\bar{n}_2, \bar{p}_2) &= -\bar{n}_2 = -q,
\end{aligned}$$

For the other two derivatives we find, by (4.3.33) and (4.3.32), that

$$1 - e^{-\bar{p}_2} = \frac{\bar{p}_2}{C\bar{n}_2} = \frac{r\left(1-\bar{n}_2\right)}{C\bar{n}_2} = \frac{r\left(1-q\right)}{Cq}$$

and thus

$$e^{-\bar{p}_2} = \frac{Cq - r(1-q)}{Cq}.$$

Hence

$$\begin{aligned}
G_n(\bar{n}_2, \bar{p}_2) &= \frac{r(1-q)}{q},\\
G_p(\bar{n}_2, \bar{p}_2) &= Cq - r(1-q),
\end{aligned}$$

Thus, the Jacobi matrix is given by

$$\begin{pmatrix} 1 - rq & -q \\ \frac{r(1-q)}{q} & Cq - r(1-q) \end{pmatrix}$$

The trace of the matrix is given by $1 - r + Cq$ and the determinant is

$$q(C(1-rq) + r^2(1-q)).$$

The condition for stability is

$$|1 - r + Cq| < q(C(1-rq) + r^2(1-q)) + 1 < 2.$$

By computer simulations it can be found that there is a range of parameters for which this equilibrium is stable.

5

McKendrick model

1 From discrete Leslie model to continuous McKendrick problem

In the classical Leslie model the census is taken in equal intervals taken, for convenience, to be also a unit of time. If the time between censuses and the length of each age class are instead taken to be $h > 0$ then, starting from some time t the Leslie model would take the form

$$\begin{pmatrix} x_0(t+h) \\ x_h(t+h) \\ x_{2h}(t+h) \\ \vdots \\ x_{(n-1)h}(t+h) \end{pmatrix} = \begin{pmatrix} f_0(h) & f_h(h) & \cdots & f_{(n-2)h}(h) & f_{(n-1)h}(h) \\ s_0(h) & 0 & \cdots & 0 & 0 \\ 0 & s_h(h) & \cdots & 0 & 0 \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & \cdots & s_{(n-2)h}(h) & 0 \end{pmatrix} \begin{pmatrix} x_0(t) \\ x_h(t) \\ x_{2h}(t) \\ \vdots \\ x_{(n-1)h}(t) \end{pmatrix}. \tag{5.1.1}$$

The maximal age of individuals $\omega = nh$ is thus divided into n age intervals $[0,h),[h,2h)\ldots[(n-1)h,nh)$ with the convention that if the age a of an individual is in $[kh,(k+1)h)$ is considered to be kh. In this definition, as in the discrete case, nobody actually lives till ω. Thus, $x_a(t)$ denotes the number of individuals of age a, $s_a = l_{a+h}/l_a$ is the probability of survival to the age of $a+h$ conditioned upon surviving up to age a with $l_0 = 1$ and $f_a = m_{a+h}s_h$ is the effective fecundity with m_{a+h} being the average fertility of females of age $a+h$. We note that $1 - s_a$ is the number of individuals who do not survive from a to $a+h$. We make the following assumptions and notation: for any $a \geq 0$

$$\lim_{h\to 0^+} s_a(h) = s_a(0) = 1, \tag{5.1.2}$$

$$\lim_{h\to 0^+} \frac{1-s_a(h)}{h} = \mu(a), \tag{5.1.3}$$

$$\lim_{h\to 0^+} \frac{f_a(h)}{h} = \beta(a). \tag{5.1.4}$$

To explain these notation, we note that probability of survival over a very short period of time should be close to 1, as in Eq. (5.1.2). Further, using subsection on the average life span, we note that if death rate μ is constant, then the probability of surviving over a short time interval h approximately is $s_a(h) = 1 - \mu h$ for any a and thus the limit in Eq. (5.1.3) can serve as a more general definition of the age dependant death rate. Similarly, if the average number of births per female over a unit time is a constant β then the number of births over h will be βh and the last equation gives the general definition of the age dependent birth rate which, moreover, is independent of the survival rate by Eq. (5.1.2).

Finally, we assume that there is a density function $n(a,t)$

$$x_a(t) = \int\limits_a^{a+h} n(\alpha,t)d\alpha. \tag{5.1.5}$$

We are going to derive a differential equation for n. Consider a fixed age $a = ih > 0$. From (5.1.1) we see that

$$x_{a+h}(t+h) = s_a(h)x_a(t), \qquad a = 0, h, \ldots, (n-2)h. \tag{5.1.6}$$

Using (5.1.5)

$$x_{a+h}(t+h) = \int\limits_{a+h}^{a+2h} n(\alpha, t+h)d\alpha = \int\limits_{a}^{a+h} n(\alpha+h, t+h)d\alpha,$$

thus (5.1.6) can be written as

$$\int\limits_{a}^{a+h} n(\alpha+h, t+h)d\alpha = s_a(h)\int\limits_{a}^{a+h} n(\alpha, t)d\alpha.$$

We re-write it as

$$\frac{1}{h^2}\left(\int\limits_{a}^{a+h} n(\alpha+h, t+h)d\alpha - \int\limits_{a}^{a+h} n(\alpha, t)d\alpha\right) = -\frac{1-s_a(h)}{h^2}\int\limits_{a}^{a+h} n(\alpha, t)d\alpha.$$

Assuming that the directional derivative

$$Dn(a,t) = \lim_{h\to 0^+}\frac{n(a+h, t+h) - n(a,t)}{h}$$

exists at each (a, t) and the limit is locally uniform in the sense that

$$n(a+h, t+h) = n(a,t) = Dn(a,t)h + R(a,t,h)$$

and each (a, t) has a neighbourhood such that

$$\frac{R(\alpha, \tau, h)}{h} \leq c(h)$$

for some $c(h)$ independent of (α, τ) in this neighbourhood and $c(h) \to 0$. Then we can write

$$\frac{1}{h^2}\left(\int\limits_{a}^{a+h} n(\alpha+h, t+h)d\alpha - \int\limits_{a}^{a+h} n(\alpha, t)d\alpha\right) = \frac{1}{h}\int\limits_{a}^{a+h} Dn(\alpha, t)d\alpha + \frac{1}{h}\int\limits_{a}^{a+h} R(\alpha, t, h)d\alpha.$$

If we assume that $Dn(\alpha, t)$ is continuous at $\alpha = a$, then passing to the limit we obtain from the fundamental theorem of calculus

$$\lim_{h\to 0^+}\frac{1}{h^2}\left(\int\limits_{a}^{a+h} n(\alpha+h, t+h)d\alpha - \int\limits_{a}^{a+h} n(\alpha, t)d\alpha\right) = Dn(a,t).$$

Similarly, since we assumed n to be continuous, using (5.1.3) we obtain

$$\lim_{h\to 0^+}\frac{1-s_a(h)}{h^2}\int\limits_{a}^{a+h} n(\alpha, t)d\alpha = \mu(a)n(a,t).$$

Combining, we obtain that at every point of continuity of n and Dn, n satisfies

$$Dn(a,t) = -\mu(a)n(a,t), \qquad a > 0, t > 0.$$

Assuming that the partial derivatives $\partial_t n, \partial_a n$ at (a, t) exist, we can further transform the last equation to

$$\partial_t n(a,t) + \partial_a n(a,t) = -\mu(a)n(a,t), \qquad a > 0, t > 0.$$

This is the most commonly used form of the equation for n though, as we shall see later, not the best for its analysis and, in fact, false in many cases as the differentiability assumptions are often not satisfied.

Now consider the class of neonates in the Leslie formulation:

$$x_0(t+h) = \sum_{j=0}^{n-1} f_{jh}(h)x_{jh}(t)$$

which can be rewritten as

$$\frac{1}{h}x_0(t+h) = \sum_{j=0}^{n-1} \frac{1}{h}f_{jh}(h)\frac{1}{h}x_{jh}(t)h.$$

Now, if n is continuous and f is differentiable at 0, then

$$\frac{1}{h}x_{jh}(t) = \frac{1}{h}\int\limits_{jh}^{(j+1)h} n(\alpha,t)d\alpha = n(jh+\theta_j h)$$

and

$$\frac{f_{jh}(h)}{h} = \beta(jh+\theta'_j h).$$

for some $0 < \theta_j, \theta'_j < 1$. Thus

$$n(\theta_j h, t) = \sum_{j=0}^{n-1} n(jh+\theta_j h)\beta(jh+\theta'_j h)h.$$

If we further assume that β is a continuous function, then the right hand side is the Riemann sum and we can pass to the limit as $h \to 0^+$ getting

$$n(0,t) = \int\limits_0^\omega \beta(\alpha)n(\alpha,t)d\alpha.$$

Thus, we arrived at the classical formulation of the McKendrick model

$$\partial_t n(a,t) + \partial_a n(a,t) = -\mu(a)n(a,t), \qquad a > 0, t > 0, \tag{5.1.7}$$

$$n(0,t) = \int\limits_0^\omega \beta(\alpha)n(\alpha,t)d\alpha, t > 0, \tag{5.1.8}$$

$$n(a,0) = n_0(a), \tag{5.1.9}$$

where the last equation provides the initial distribution of the population.

If $\omega < +\infty$ then we have to ensure that $n(a,t) = 0$ for $t \geq 0, a \geq \omega$, which can be done either by imposing an additional boundary condition on n or by introducing assumptions on the coefficients which ensure that no individual survives beyond ω. If $\omega = \infty$ then, instead of such an additional condition, we impose some requirements on the behaviour of the solution at ∞, e.g. that they are integrable over $[0,\infty)$.

2 Linear constant coefficient case

Before we embark on more advanced analysis of (5.1.7)–(5.1.9) let us get a taste of the structure of the problem by solving the simplest case with $\mu(a) = \mu$ and $\beta(a) = \beta$:

$$\partial_t n(a,t) + \partial_a n(a,t) = -\mu n(a,t). \tag{5.2.1}$$

coupled with the boundary condition

$$n(0,t) = \beta \int_0^\infty n(a,t)da,$$

and the initial condition

$$n(a,0) = n_0(a).$$

2.1 Interlude - solving first order partial differential equations with constant coefficients

Let us consider more general linear first order partial differential equation (PDE) of the form:

$$au_t + bu_x = 0, \quad t,x \in \mathbb{R} \tag{5.2.2}$$

where a and b are constants. This equation can be written as

$$D_{\boldsymbol{v}} u = 0, \tag{5.2.3}$$

where $\boldsymbol{v} = a\boldsymbol{j} + b\boldsymbol{i}$ ($\boldsymbol{j}$ and $\boldsymbol{i}$ are the unit vectors in, respectively, t and x directions), and $D_{\boldsymbol{v}} = \nabla u \cdot \boldsymbol{v}$ denotes the directional derivative in the direction of $\boldsymbol{v}$. This means that the solution u is a constant function along each line having direction $\boldsymbol{v}$, that is, along each line of equation $bt - ax = \xi$. Along each such a line the value of the parameter ξ remains constant. However, the solution can change from one line to another, therefore the solution is a function of ξ, that is the solution to Eq. (7.2.25) is given by

$$u(x,t) = f(bt - ax), \tag{5.2.4}$$

where f is an arbitrary differentiable function. Such lines are called the *characteristic lines* of the equation.

Example 5.1. To obtain a unique solution we must specify the initial value for u. Hence, let us consider the initial value problem for Eq. (7.2.25): find u satisfying both

$$\begin{aligned} au_t + bu_x &= 0 \quad x \in \mathbb{R}, t > 0, \\ u(x,0) &= g(x) \quad x \in \mathbb{R}, \end{aligned} \tag{5.2.5}$$

where g is an arbitrary given function. From Eq. (5.2.4) we find that

$$u(x,t) = g\left(-\frac{bt - ax}{a}\right). \tag{5.2.6}$$

We note that the initial shape propagates without any change along the characteristic lines, as seen below for the initial function $g = 1 - x^2$ for $|x| < 1$ and zero elsewhere. The speed $c = b/a$ is taken to be equal to 1.

Example 5.2. Let us consider a variation of this problem and try to solve the initial- boundary value problem

$$au_t + bu_x = 0 \quad x \in \mathbb{R}, t > 0,$$

$$u(x,0) = g(x) \quad x > 0, \tag{5.2.7}$$

$$u(0,t) = h(t) \quad t > 0, \tag{5.2.8}$$

for $a, b > 0$ From Example 5.1 we have the general solution of the equation in the form

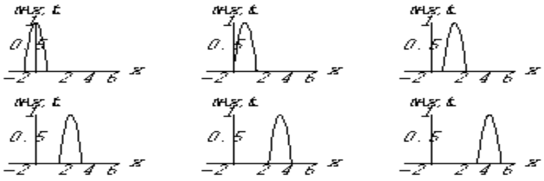

Fig. 4.1 The graph of the solution in Example 5.1

$$u(x,t) = f(bt - ax).$$

Putting $t = 0$ *we get* $f(-ax) = g(x)$ *for* $x > 0$*, hence* $f(x) = g(-x/a)$ *for* $x < 0$*. Next, for* $x = 0$ *we obtain* $f(bt) = h(t)$ *for* $t > 0$*, hence* $f(x) = h(x/b)$ *for* $x > 0$*. Combining these two equations we obtain*

$$u(x,t) = \begin{cases} g(-\frac{bt-ax}{a}) \text{ for } x > bt/a \\ h(\frac{bt-ax}{b}) \text{ for } x < bt/a \end{cases}$$

Now, let us consider what happens if $a = 1 > 0, b = -1 < 0$*. Then the initial condition defines* $f(x) = g(-x)$ *for* $x < 0$ *and the boundary condition gives* $f(x) = h(-x)$ *also for* $x < 0$*! Hence, we cannot specify both initial and boundary conditions in an arbitrary way as this could make the problem ill-posed.*

The physical explanation of this comes from the observation that since the characteristics are given by $\xi = x + t$*, the flow occurs in the negative direction and therefore the values at* $x = 0$ *for any* t *are uniquely determined by the initial condition. Therefore we see that to have a well-posed problem we must specify the boundary conditions at the point where the medium flows into the region.*

2.2 Solution of the McKendrick equation

First, let us simplify the equation (5.2.1) by introducing the integrating factor

$$\partial_t(e^{\mu a}n(a,t)) = -\partial_a(e^{\mu a}n(a,t))$$

and denote $u(a,t) = e^{\mu a}n(a,t)$. Then

$$u(0,t) = n(0,t) = \beta \int\limits_0^\infty e^{-\mu a}u(a,t)da$$

with $u(a,0) = e^{\mu a}n_0(a) =: u_0(a)$. Now, if we knew $\psi(t) = u(0,t)$, then

$$u(a,t) = \begin{cases} u_0(a-t), & t < a, \\ \psi(t-a), & a < t. \end{cases} \tag{5.2.9}$$

The boundary condition can be rewritten as

$$\psi(t) = \beta \int\limits_0^\infty e^{-\mu a} u(a,t)da = \beta \int\limits_0^t e^{-\mu a}\psi(t-a)da + \beta \int\limits_t^\infty e^{-\mu a} u_0(a-t)da$$

$$= \beta e^{-\mu t} \int\limits_0^t e^{\mu\sigma}\psi(\sigma)d\sigma + \beta e^{-\mu t} \int\limits_0^\infty e^{-\mu r} u_0(r)dr$$

which, upon denoting $\phi(t) = \psi(t)e^{\mu t}$ and using the original initial value, can be written as

$$\phi(t) = \beta \int\limits_0^t \phi(\sigma)d\sigma + \beta \int\limits_0^\infty n_0(r)dr. \tag{5.2.10}$$

Now, if we differentiate both sides, we get

$$\phi' = \beta\phi$$

which is just a first order linear equation. Letting $t = 0$ in (5.2.10) we obtain the initial value for ϕ: $\phi(0) = \beta \int\limits_0^\infty n_0(r)dr$. Then

$$\phi(t) = \beta e^{\beta t} \int\limits_0^\infty n_0(r)dr$$

and

$$\psi(t) = \beta e^{(\beta-\mu)t} \int\limits_0^\infty n_0(r)dr.$$

Then

$$n(a,t) = e^{-\mu a} u(a,t) = e^{-\mu t} \begin{cases} n_0(a-t), & t < a, \\ \beta e^{\beta(t-a)} \int\limits_0^\infty n_0(r)dr, & a < t. \end{cases}$$

Observe that

$$\lim_{a\to t^+} n(a,t) = e^{-\mu t} n_0(0)$$

and

$$\lim_{a\to t^-} n(a,t) = \beta e^{-\mu t} \int\limits_0^\infty n_0(r)$$

so that the solution is continuous, let alone differentiable only if the initial condition satisfies the following compatibility condition

$$n_0(0) = \beta \int\limits_0^\infty n_0(r)dr. \tag{5.2.11}$$

Thus, as we noted earlier, we must be very careful with using (5.1.7)-(5.1.9) in the differential form and interpreting the solution.

3 General linear McKendrick problem

The ideas used to solve the McKendrick case in this simple case also is used in more general situations but, unfortunately, the resulting integral equation (5.2.10) cannot be explicitly solved. Before, however, we discuss solvability of more general cases, let us introduce certain functions related to (5.1.7)-(5.1.9) which are relevant to the population dynamics.

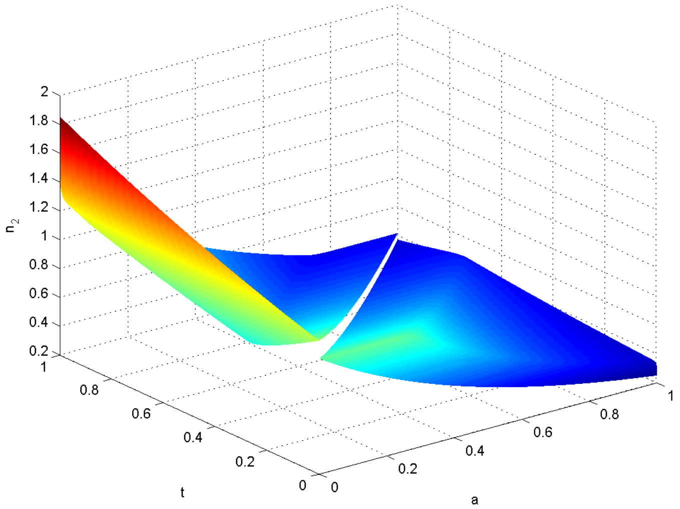

Fig. 5.1. Discontinuity of the population density $n(a,t)$.

3.1 Demographic parameters of the McKendrick problems

Consider again the general McKendrick problem

$$\begin{aligned}&\partial_t n(a,t) + \partial_a n(a,t) = -\mu(a)n(a,t)\\ &n(0,t) = \int\limits_0^\omega \beta(\alpha)n(\alpha,t)d\alpha,\\ &n(a,0) = n_0(a)\end{aligned}$$

We recall that $\beta(a)$ is the *age specific fertility* which can be defined as the number of newborns, in one time unit, coming from a single individual whose age is in the small time age interval $[a, a+da)$. So, the number of births coming from all individuals in the population aged between a_1 and a_2 in a one time unit is

$$\int\limits_{a_1}^{a_2} \beta(\alpha)n(\alpha,t)da$$

and we can define the *total birth rate* as

$$B(t) = \int\limits_0^\omega \beta(\alpha)n(\alpha,t)da$$

which gives the total number of newborns in a unit time (ω is the maximum age in the population).

Let us consider the death rate $\mu(a)$ which is average number of deaths per unit of population aged a. We can relate $\mu(a)$ to a number of vital characteristics of the population. Similarly to the discrete case, we introduce the *survival probability* $S(a)$ as the proportion of the initial population surviving to age a. We can relate μ and S by the following argument. Consider a population beginning with n_0 individuals of age 0. Then $n_0(a)S(a)(=n(a))$ is the average number of individuals that survived to age a. The decline in the population over a short age period $[a, a+da]$ is $n_0(a)S(a) - n_0(a)S(a+da)$. On the other hand, this decline can only be attributed to deaths: if the death rate is μ, then in this age interval we will have approximately $n_0(a)S(a)\mu(a)da$ deaths. Equating and passing to the limit as $da \to 0$ yields

$$\frac{dS}{da} = -S\mu$$

or

$$S(a) = S(0)e^{-\int\limits_0^a \mu(\sigma)d\sigma}$$

(compare with (2.2.6)). Since, however, the probability of surviving to age 0 is 1, we have

$$S(a) = e^{-\int\limits_0^a \mu(\sigma)d\sigma}. \tag{5.3.12}$$

We note that if no individuals can survive beyond ω, we must have $S(\omega) = 0$ or, equivalently,

$$\int\limits_0^\omega \mu(\sigma)d\sigma = \infty. \tag{5.3.13}$$

These considerations can be used, as before, to find average life span of individuals in the population. In fact, average life span is the mean value of the length of life in the population, which can be expressed as

$$L = \int\limits_0^\omega ap(a)da$$

where $p(a)$ is the probability (density) of an individual dying at age a. If we approximate the integral as the Riemann sum

$$L \approx \sum_i a_i p(a_i)\Delta a_i$$

where $p(a_i)$ is the probability that an individual survived till the age a_i and died at this age. Thus

$$p(a_i) = S(a_i)\mu(a_i).$$

We note that $S(a)\mu(a)$ is, indeed, a probability density. Thus

$$L = \int\limits_0^\omega a\mu(a)e^{-\int\limits_0^a \mu(s)ds}da = -\int\limits_0^\omega a\frac{d}{da}e^{-\int\limits_0^a \mu(s)ds}da = \int\limits_0^\omega S(a)da$$

where we used integration by parts and $S(\omega) = 0$.

Further, we introduce

$$K(a) = \beta(a)S(a) \tag{5.3.14}$$

which is called the *maternity function* and describes the rate of birth relative to the surviving fraction of the population and is the continuous equivalent to the coefficients $f_0, f_1, \ldots, f_{n-1}$. Further, we define

$$R = \int\limits_0^\omega \beta(a)S(a)da \tag{5.3.15}$$

and call it *net reproduction rate* of the population. It is the expected number of offspring produced by an individual during her reproductive life.

3.2 Solution of the linear McKendrick problem

One of the easiest way of analysing the general McKendrick model

$$\begin{aligned}
&\partial_t n(a,t) + \partial_a n(a,t) = -\mu(a)n(a,t)\\
&n(0,t) = \int\limits_0^\omega \beta(a)n(a,t)da,\\
&n(a,0) = n_0(a) \qquad (5.3.16)
\end{aligned}$$

is to reduce it to an integral equation in the same way as we proceeded in Section 2 though the technicalities are slightly more involved due to age dependence of the mortality and maternity functions. First, we simplify (5.3.16) by introducing the integrating factor

$$\partial_t \left(e^{\int\limits_0^a \mu(\sigma)d\sigma} n(a,t) \right) = -\partial_a \left(e^{\int\limits_0^a \mu(\sigma)d\sigma} n(a,t) \right) \qquad (5.3.17)$$

and denote $u(a,t) = e^{\int\limits_0^a \mu(\sigma)d\sigma} n(a,t)$. Then

$$u(0,t) = n(0,t) = \int\limits_0^\omega \beta(a)e^{-\int\limits_0^a \mu(\sigma)d\sigma} u(a,t)da = \int\limits_0^\omega K(a)u(a,t)da,$$

where we recognized that the kernel in the integral above is the maternity function introduced in (5.3.14). Further, $u(a,0) = e^{\int\limits_0^a \mu(s)ds} n_0(a) =: u_0(a)$. Also, the right hand side defines the total birth rate $B(t)$.

Now, if we knew $B(t) = u(0,t)$, then

$$u(a,t) = \begin{cases} u_0(a-t), & t < a, \\ B(t-a), & a < t. \end{cases} \qquad (5.3.18)$$

The boundary condition can be rewritten as

$$\begin{aligned}
B(t) &= \int\limits_0^\infty \beta(a)e^{-\int\limits_0^a \mu(\sigma)d\sigma} u(a,t)da\\
&= \int\limits_0^t \beta(a)e^{-\int\limits_0^a \mu(\sigma)d\sigma} B(t-a)da + \int\limits_t^\infty \beta(a)e^{-\int\limits_0^a \mu(\sigma)d\sigma} u_0(a-t)da\\
&= \int\limits_0^t K(t-a)B(a)da + \int\limits_0^\infty \beta(a+t)e^{-\int\limits_0^{a+t} \mu(\sigma)d\sigma} e^{\int\limits_0^a \mu(s)ds} n_0(a)da,
\end{aligned}$$

where to shorten notation we extended coefficients by zero beyond $a = \omega$. Summarizing, we arrived at the integral equation for the total birth rate

$$B(t) = \int\limits_0^t K(t-a)B(a)da + G(t) \qquad (5.3.19)$$

where

$$G(t) = \int\limits_0^\infty \beta(a+t)\frac{S(a+t)}{S(a)}n_0(a)da, \qquad (5.3.20)$$

is a known function. Explicitly, we have

$$B(t) = \int\limits_0^t K(t-a)B(a)da + \int\limits_0^{\omega-t} \beta(a+t)\frac{S(a+t)}{S(a)}n_0(a)da$$
$$= \int\limits_0^t K(t-a)B(a)da + \int\limits_t^{\omega} \beta(a)\frac{S(a)}{S(a-t)}n_0(a-t)da \tag{5.3.21}$$

for $0 \le t \le \omega$ and

$$B(t) = \int\limits_0^{\omega} K(t-a)B(a)da \tag{5.3.22}$$

for $t > \omega$. Unlike in Section 2, this equation cannot be solved explicitly and we have to use more abstract approach. For this we have to introduce a proper mathematical framework. As in the discrete case, the natural norm will be

$$\|n\|_1 = \int\limits_0^{\omega} |n(\alpha)|d\alpha$$

which in the current context, with $n \ge 0$ being the density of the population distribution with respect to age, is the total population. Thus, the state space is the space $X_0 = L_1([0,\omega))$ of Lebesgue integrable functions on $[0,\omega)$. Since we are dealing with functions of two variables, we often consider $(a,t) \to n(a,t)$ as a function $t \to u(t,\cdot)$, that is, for each t the value of this function is a function with a argument. For such functions, we consider the space $C([0,T], L_1([0,\omega]))$ of $L_1([0,\omega])$-valued continuous functions. For functions f bounded on $[0,\omega]$ we introduce $\|f\|_\infty = \sup_{0\le a\le\omega}|f(a)|$. We make the following assumptions.

(i)
$$\beta \ge 0 \text{ is bounded on } [0,\omega], \tag{5.3.23}$$

(ii)
$$0 \le \mu \in L_1([0,\omega']) \text{ for any } \omega' < \omega \tag{5.3.24}$$

with
$$\int\limits_0^{\omega} \mu(\alpha)d\alpha = \infty, \tag{5.3.25}$$

(iii)
$$0 \le n_0 \in L_1([0,\omega]). \tag{5.3.26}$$

Now, if (5.3.23)-(5.3.26) are satisfied, then we can show that K is a non-negative bounded function which is zero for $t \ge \omega$ and G is a continuous function which also is zero for $t \ge \omega$. If, additionally

$$n_0 \in W^{1,1}([0,\omega]) \quad \text{and} \quad \mu n_0 \in L_1([0,\omega]), \tag{5.3.27}$$

(here by W^1_1 we denote the Sobolev space of functions from L_1 with generalized derivatives in L_1) then G is differentiable with bounded derivative. Indeed, let us look at G for $t < \omega$

$$G(t) = \int\limits_t^{\omega} \beta(a)\frac{S(a)}{S(a-t)}n_0(a-t)da = \int\limits_t^{\omega} \beta(a)e^{-\int\limits_{a-t}^{a}\mu(s)ds}n_0(a-t)da$$

If we formally differentiate using the Leibnitz rule, we get

$$G'(t) = -\beta(t)S(t)n_0(0) + \int\limits_t^{\omega} \beta(a)e^{-\int\limits_{a-t}^{a}\mu(s)ds}\mu(a-t)n_0(a-t)da$$
$$+ \int\limits_t^{\omega} \beta(a)e^{-\int\limits_{a-t}^{a}\mu(s)ds}\mu(a-t)n_0'(a-t)da$$

so we see that for existence of the integrals we need integrability of μn_0 and differentiability of n_0. Then we can prove the main result

Theorem 5.3. *If (5.3.23)-(5.3.26) are satisfied, then (5.3.19) has a unique continuous and non-negative solution. If, additionally, (5.3.27) is satisfied, then B is differentiable with B' bounded on bounded intervals.*

Proof. We define iterates

$$\begin{aligned} B_0(t) &= G(t), \\ B^{k+1}(t) &= G(t) + \int\limits_0^t K(t-s)B^k(s)ds. \end{aligned} \tag{5.3.28}$$

Take $T > 0$. Then, for any $t \in [0, T]$ we have

$$|B^1(t) - B^0(t)| = \int\limits_0^t |K(t-s)F(s)|ds \leq tK_mF_m$$

where $K_m = \sup_{0\leq t\leq T} |K(s)|$ and $L_m = \sup_{0\leq t\leq T} |F(s)|$. Then

$$|B^2(t) - B^1(t)| \leq K_m \int\limits_0^t |B^1(s) - B^0(s)|ds \leq \frac{K_m^2 F_m}{2} t^2$$

and, by induction,

$$|B^{k+1}(t) - B^k(t)| \leq K_m \int\limits_0^t |B^k(s) - B^{k-1}(s)|ds \leq \frac{K_m^{k+1} F_m}{(k+1)!} t^{k+1}. \tag{5.3.29}$$

Further

$$\lim_{k\to\infty} B^{k+1}(t) = G(t) + \lim_{k\to\infty} \sum_{i=0}^k (B^{i+1}(t) - B^i(t))$$

with

$$\sup_{0\leq t\leq T} \left| \sum_{i=0}^k (B^{i+1}(t) - B^i(t)) \right| \leq \sum_{i=0}^k \sup_{0\leq t\leq T} \left| B^{i+1}(t) - B^i(t) \right| \leq F_m \sum_{i=0}^k \frac{(TK_m)^{k+1}}{(k+1)!}.$$

The series on the right hand side converges to $F_m e^{TK_m}$ and thus $(B^k(t))_{k\geq 0}$ converges uniformly to a continuous solution B of (5.3.19). Uniqueness follows by the Gronwall inequality.

If, in addition, (5.3.27) is satisfied, then B^k can be differentiated with respect to t and

$$V^k := \frac{d}{dt} B^k$$

satisfy the recurrence

$$V^{k+1}(t) = F'(t) + K(t)F(0) + \int\limits_0^t K(t-s)V^k(s)ds$$

which converges uniformly to some continuous function V which, by the theorem of uniform convergence of derivatives, must be the derivative of B. □

Once we have B, we can recover n by (5.4.46) and back substitution

$$n(a,t) = e^{-\int\limits_0^a \mu(\sigma)d\sigma} u(a,t) = \begin{cases} \frac{S(a)}{S(a-t)} n_0(a-t), & t < a, \\ S(a)B(t-a), & a < t. \end{cases} \tag{5.3.30}$$

Thus, if (5.3.27) is satisfied in addition to (5.3.23)-(5.3.26), then it is easy to see that n defined above satisfies the equation (5.1.7) everywhere except the line $a = t$. Along this line we have, as before

$$\lim_{a\to t^+} n(a,t) = S(0)n_0(0) = n_0(0)$$

and

$$\lim_{a\to t^-} n(a,t) = S(0)B(0) = \int\limits_0^\omega \beta(a)n_0(a)da$$

and, to ensure at least continuity of the solution we need to assume the compatibility condition

$$n_0(0) = \int\limits_0^\omega \beta(a)n_0(a)da. \tag{5.3.31}$$

We note that if a function is continuous at a point and differentiable in both one sided neighbourhoods, then it is a Lipschitz function and it is in fact differentiable almost everywhere (in the sense that the function can be recovered from its derivative). On the other hand, if a function has a jump at a point, then its derivative at this point is of a Dirac delta type. Thus, we can state that if (5.3.31) is satisfied, then the solution is continuous and satisfies (5.1.7) almost everywhere. If we do not assume (5.3.31) then we can still claim that the solution satisfies

$$Dn(a,t) = \lim_{h\to 0^+} \frac{n(a+h,t+h) - n(a,t)}{h} = -\mu(a)n(a,t), \qquad a > 0, t > 0.$$

Furthermore, both the birth rate B and the solution n itself grow at most at an exponential rate. Consider again (5.3.19)

$$B(t) = \int\limits_0^t K(t-a)B(a)da + G(t).$$

with G given by (5.3.20).

$$S(a) = e^{-\int\limits_0^a \mu(\sigma)d\sigma}.$$

and $K(a) = \beta(a)S(a)$ we see that $K(t) \leq \|\beta\|_\infty$ and $G(t) \leq \|\beta\|_\infty \|n_0\|_1$ so that

$$B(t) \leq \max_{0\leq a\leq\omega} \beta(a) \int\limits_0^t B(s)ds + \max_{0\leq a\leq\omega} \beta(a) \int\limits_0^\omega n_0(s)ds =: \|\beta\|_\infty \int\limits_0^t B(s)ds + \|\beta\|_\infty \|n_0\|_1,$$

which, by Gronwall's inequality, yields

$$B(t) \leq \|\beta\|_\infty \|n_0\|_1 e^{t\|\beta\|_\infty}. \tag{5.3.32}$$

This gives the estimate for n:

$$\begin{aligned} \|n(\cdot,t)\|_1 &\leq \int\limits_0^t B(t-s)S(s)ds + \int\limits_t^\infty \frac{S(s)}{S(s-t)} n_0(s-t)ds \\ &\leq \|\beta\|_\infty \|n_0\|_1 \left(\int\limits_0^t e^{(t-s)\|\beta\|_\infty} ds + 1 \right), \end{aligned}$$

where we used $S(s)/S(s-t) \leq 1$. Then, by integration

$$\|n(\cdot,t)\|_1 \leq \|n_0\|_1 + \|n_0\|_1 e^{t\|\beta\|_\infty}(1 - e^{-t\|\beta\|_\infty}) = \|n_0\|_1 e^{t\|\beta\|_\infty}. \tag{5.3.33}$$

4 Long time behaviour of the solution

4.1 Warm up - the constant coefficients case

Consider again the constant coefficient problem

$$\begin{aligned}
&\partial_t n(a,t) + \partial_a n(a,t) = -\mu n(a,t)\\
&n(0,t) = \beta \int\limits_0^\infty n(a,t)da,\\
&n(a,0) = n_0(a)
\end{aligned}$$

with the solution

$$n(a,t) = e^{-\mu t}\begin{cases} n_0(a-t), & t<a,\\ \beta e^{\beta(t-a)} \int\limits_0^\infty n_0(r)dr, & a<t\end{cases}$$

and ask what happens with the population as $t \to \infty$. Clearly, for large t we can consider only the second part of the solution

$$n(a,t) = \beta N_0 e^{-\beta a} e^{t(\beta-\mu)}, \quad a<t,$$

where $N_0 = \int\limits_0^\infty n_0(r)dr$. Denote by $r = \beta - \mu$ the net growth rate. We see that if $r = 0$, we have

$$n(a,t) = \beta N_0 e^{-\mu a}, \quad a<t,$$

and one can surmise that

$$n(a,t) \approx \beta N_0 e^{-\beta a},$$

for large t and all $a > 0$. If we assume that n_0 is bounded, this can be easily checked. Indeed

$$n(a,t) = \begin{cases} e^{-\mu t} n_0(a-t), & t<a,\\ \beta e^{-\mu a}N_0, & a<t\end{cases} = \beta e^{-\mu a} N_0 + \begin{cases} -\beta e^{-\mu a}N_0 + e^{-\mu t}n_0(a-t), & t<a,\\ 0, & a<t\end{cases}$$

and, since for $t<a$, $e^{-\mu a} < e^{-\mu t}$, we have

$$|-\beta e^{-\mu a}N_0 + e^{-\mu t}n_0(a-t)| \leq Ce^{-\mu t}$$

where $C = \max\{\beta N_0, \sup|n_0|\}$. In other words

$$n(a,t) = \beta e^{-\mu a}N_0 + O(e^{-\mu t}). \tag{5.4.34}$$

So we see that for large t the solution has the shape of $\beta e^{-\mu a}$, independent of the initial data, multiplied by the scalar N_0. Thus, the shape of the solution is practically not affected by the initial age distribution. In other words, the age distribution of the population after long time is the same independently of the initial age distribution.

Even if $r \neq 0$, we can write

$$e^{-rt}n(a,t) = \beta N_0 e^{-\beta a}, \quad a<t,$$

and, as before, Indeed

$$e^{-rt}n(a,t) = \begin{cases} e^{-\beta t} n_0(a-t), & t<a,\\ \beta e^{-\beta a}N_0, & a<t\end{cases} = \beta e^{-\mu a} N_0 + \begin{cases} -\beta e^{-\beta a}N_0 + e^{-\beta t}n_0(a-t), & t<a,\\ 0, & a<t\end{cases}$$

and, since for $t<a$, $e^{-\beta a} < e^{-\beta t}$, we have

$$|-\beta e^{-\beta a}N_0 + e^{-\beta t}n_0(a-t)| \leq Ce^{-\beta t}$$

where $C = \max\{\beta N_0, \|n_0\|_\infty\}$. In other words

$$n(a,t) = N_0 e^{rt} \beta e^{-\beta a} + O(e^{-\mu t}). \tag{5.4.35}$$

where we used $e^{rt}e^{-\beta t} = e^{(\beta-\mu)t}e^{-mt} = e^{-\mu t}$. Hence, the population is described by the Malthusian part $N_0 e^{rt}$, which is independent of the age profile of the population, multiplied by the age profile $\beta e^{-\beta a}$. The profile is called the *stable age distribution* and the property described above is called *asynchronous exponential growth property.* In what follows we shall prove that this property holds for general McKendrick model. However, before we move to more general models, we provide another way of deriving the stable age distribution. Let us consider the eigenvalue problem for (5.2.1)

$$\begin{aligned} \lambda n(a) + n_a(a) &= -\mu n(a) \\ n(0) &= \beta \int\limits_0^\infty n(a) da. \end{aligned} \tag{5.4.36}$$

The first equation is simply a linear equation with the general solution

$$n(a) = Ce^{-(\mu+\lambda)a}$$

while the nonlocal initial condition yields

$$1 = \beta \int\limits_0^\infty e^{-(\mu+\lambda)a} da$$

where we cancelled the constant C. This is an example of the Lotka renewal equation. In our case, we solve it explicitly. Integration gives

$$1 = \frac{\beta}{\mu+\lambda} \tag{5.4.37}$$

or

$$\lambda = \beta - \mu = r$$

and

$$n(a) = Ce^{-\beta a}.$$

So, the unique eigenvalue of (5.4.41) is (in this case) precisely the net growth rate. This eigenvalue is simple and the corresponding eigenvector is the stable age distribution. As we shall see, this is not a coincidence.

4.2 Long time behaviour–general case

By (5.3.32), we can apply the Laplace transform to analyse (5.3.19). The Laplace transform of an exponentially bounded integrable function f is defined by

$$\hat{f}(\lambda) = (\mathcal{L}f)(\lambda) = \int\limits_0^\infty e^{-\lambda t} f(t) dt,$$

and $\hat{f}$ is defined and analytic in a right half-plane (determined by the rate of growth of f) of the complex plane $\mathbb{C}$. In the case of B, (5.3.32) shows that $\hat{B}(\lambda)$ is analytic in $\Re\lambda > \|m\|_\infty$. For our applications it is also important to note that if the f is only non-zero over a finite interval $[a,b]$, then its Laplace transform is defined and analytic everywhere in $\mathbb{C}$. Such functions are called entire. Moreover, we also use $\hat{f}(\lambda) \to 0$ as $|\lambda| \to \infty$ in any strip contained in the domain of analyticity of $\hat{f}$ (the proof of this fact is nontrivial).

We use the property of the Laplace transform that the convolution is transformed into the algebraic product of transforms; that is, for the convolution

$$(f*g)(t) = \int\limits_0^t f(t-s)g(s)ds = \int\limits_0^t f(s)g(t-s)ds,$$

using the definition of the Laplace transform and changing the order of integration, we obtain

$$[\mathcal{L}(f*g)](\lambda) = (\mathcal{L}f)(\lambda)\cdot(\mathcal{L}g)(\lambda). \tag{5.4.38}$$

With this result, (5.3.19) yields

$$\hat{B}(\lambda) = \hat{B}(\lambda)\hat{K}(\lambda) + \hat{G}(\lambda). \tag{5.4.39}$$

Hence,

$$\hat{B}(\lambda) = \frac{\hat{G}(\lambda)}{1-\hat{K}(\lambda)} = \hat{G}(\lambda) + \frac{\hat{G}(\lambda)\hat{K}(\lambda)}{1-\hat{K}(\lambda)} \tag{5.4.40}$$

As we noted above, $\hat{G}$ is an entire function so the only singularities of $\hat{B}$ are due to zeroes of $1-\hat{K}$. Since $\hat{K}$ is an entire function, these zeroes are isolated of finite order (thus giving rise to poles of $\hat{B}$ and with no finite accumulation point. However, there may be infinitely many of them and this requires some care with handling the inverse. It is known that if $\hat{f}$ is the Laplace transform of a continuous function f, then

$$f(t) = \frac{1}{2\pi i}\int\limits_{c-i\infty}^{c+i\infty} e^{\lambda t}\hat{f}(\lambda)d\lambda$$

where integration is carried along any vertical line in the domain of analyticity of $\hat{f}$.

Let us look closer at the equation

$$\hat{K}(\lambda) = 1, \tag{5.4.41}$$

or, explicitly,

$$\int\limits_0^\infty \beta(a)e^{-\lambda a - \int\limits_0^a \mu(\sigma)d\sigma}da = 1, \tag{5.4.42}$$

where $\lambda \in \mathbb{C}$.

Remark 5.4. We observe that (5.4.42) is a continuous copy of the discrete renewal equation (3.4.105) if one replaces λ of the latter by e^{λ}. However, as we shall see below, continuous case does not admit any cyclic behaviour.

Theorem 5.5. *Equation (5.4.41) has exactly one real root, $\lambda = \lambda_0$, of algebraic multiplicity 1. All other roots λ_j of (5.4.41) occur as complex conjugates (real root is its own conjugate). Moreover, $\Re\lambda_j < \lambda_0$ for any j, there could be only denumerable number of them and, in each strip $a < \Re\lambda < b$, there is at most a finite number of them.*

Proof. We introduce the real function

$$\psi(\lambda) = \int\limits_0^\infty e^{-\lambda a}K(a)da$$

for $\lambda \in \mathbb{R}$. We note that this function is well defined on $\mathbb{R}$ since K is non zero only on a finite interval. Also, because of this, it is continuous and differentiable, see Remark 5.6 below. Then

$$\lim_{\lambda\to-\infty}\psi(\lambda) = \infty,$$
$$\lim_{\lambda\to\infty}\psi(\lambda) = 0.$$

Moreover,

$$\psi'(\lambda) = -\int\limits_{\alpha}^{\beta} ae^{-\lambda a}K(a)da < 0,$$

$$\psi''(\lambda) = \int\limits_{\alpha}^{\beta} a^2e^{-\lambda a}K(a)da > 0,$$

so that ψ is strictly decreasing and concave up function. Since it is continuous, it takes on every positive value exactly once. Thus, in particular, there is exactly one real value λ_* satisfying (5.4.41).

Suppose $\lambda = u + iv$ is a root of (5.4.41). Then

$$1 = \int\limits_{0}^{\infty} e^{-va}(\cos(-ua) + i\sin(-ua))K(a)da$$

and, taking the real and imaginary part,

$$\int\limits_{0}^{\infty} e^{-va}K(a)\cos ua\, da = 1,$$

$$\int\limits_{0}^{\infty} e^{-va}K(a)\sin va\, da = 0.$$

We observe that these two equations are invariant under the change $v \to -v$ so that $\bar{\lambda} = u - iv$ also satisfies (5.4.41).

To prove the second part, we note that, since the variable a is continuous, there must be a range of a, say, $[\alpha, \beta]$ over which $\cos ua < 1$. Thus,

$$\int\limits_{0}^{\infty} e^{-va}K(a)da > \int\limits_{0}^{\infty} e^{-va}K(a)\cos ua\, da = 1.$$

However

$$\int\limits_{\alpha}^{\beta} e^{-\lambda_* a}K(a)da = 1,$$

and direct comparison of these two integrals yields $\lambda_* > v = \Re\lambda$.

The last part follows from the fact that since $\hat{K} - 1$ is an entire function, in each bounded set there can be only finitely many zeros of it, by the principle of isolated zeros. Thus, there could be no more than denumerable amount of them in $\mathbb{C}$. Finally, since $\hat{K} \to 0$ as $|\lambda| \to \infty$ in any strip, we also see that there can be only finitely many of them in any vertical strip. □

Remark 5.6. In the proof above, the continuity of ψ is a consequence of the boundeness of the support of definition of K. In general, if we allow K to be nonzero on $[0, \infty)$, then the above statement is not true. Consider $K(a) = c(1 + a^2)^{-1}$ with $c < 2/\pi$. Then

$$\psi(\lambda) = c\int\limits_{0}^{\infty} \frac{e^{-\lambda a}}{1 + a^2}dt$$

then $\psi(\lambda) < 1$ for $\lambda \geq 0$ but $\psi(\lambda) = \infty$ for $\lambda < 0$ and $\psi(\lambda) < 1$ for all $\lambda \geq 0$ and Eq. (5.4.41) has no real solution.

In general, if $\omega = \infty$, one has to prove that the range of ψ contains 1. For instantce, in the constant coefficient case ψ is given by (5.4.37)

$$\psi(\lambda) = \frac{m}{\lambda + \mu}$$

and though it is discontinuous at $\lambda = -\mu$, its range for $\lambda \in (-\mu, \infty)$ is $\mathbb{R}$ and the argument holds.

Observe that the function ψ crosses the a axis at

$$R := \psi(0) = \int\limits_0^\infty K(a)da \tag{5.4.43}$$

which is precisely the net reproductive rate. R must exceed 1 for λ_* to be positive, $R = 1$ if and only if $\lambda_* = 0$ and, finally, $R < 1$ if and only if $\lambda_* < 0$.

Next we shall show that the sign of λ_* indeed determines the long time behaviour of the population.

Let us consider the second term in the last formula of (5.4.40)

$$\hat{H}(\lambda) = \frac{\hat{G}(\lambda)\hat{K}(\lambda)}{1 - \hat{K}(\lambda)}.$$

We noted that $\hat{G}(\lambda)$ and $\hat{K}(\lambda)$ tend to zero as $|\lambda| \to \infty$ in any half plane $\Re\lambda > \delta$, $\delta \in \mathbb{R}$. Furthermore, on any line $\{\sigma + iy;\ y \in \mathbb{R}\}$ which does not meet any root of (5.4.41), we have $\inf_{y\in\mathbb{R}} |1 - \hat{K}(\sigma+iy)| > 0$ and

$$\int\limits_{-\infty}^{\infty} \left| \frac{\hat{G}(\sigma+iy)\hat{K}(\sigma+iy)}{1 - \hat{K}(\sigma+iy)} \right| dy < \infty. \tag{5.4.44}$$

This follows from the fact that any finitely supported function, multiplied by $e^{-\sigma t}$ is an L_2 function and thus its Laplace transform, treated as the Fourier transform, is in L_2 with respect to y. Then the result follows from the Plancherel theorem.

Inverting $\hat{H}(\lambda)$ we have

$$H(t) = \frac{1}{2\pi i} \int\limits_{\sigma - i\infty}^{\sigma + i\infty} \frac{\hat{G}(\sigma+iy)\hat{K}(\sigma+iy)}{1 - \hat{K}(\sigma+iy)} e^{(\sigma+iy)t} dy$$

for any $\sigma > \lambda_*$. Hence

$$B(t) = G(t) + H(t).$$

To estimate $H(t)$ we note that, by properties of $\hat{H}$, we can shift the line of integration to $\{\sigma_1 + iy;\ y \in \mathbb{R}\}$ where $\Re\lambda_1 < \sigma_1 < \lambda_*$ and λ_1 is the eigenvalue with the largest real part less than λ_*. Then the Cauchy theorem gives

$$H(t) = H_1(t) + H_2(t)$$

where

$$\begin{aligned} H_1(t) &= res_{\lambda=\lambda_*} \frac{e^{\lambda t}\hat{G}(\lambda)\hat{K}(\lambda)}{1 - \hat{K}(\lambda)} = \lim_{\lambda\to\lambda_*} (\lambda - \lambda_*) \frac{e^{\lambda t}\hat{G}(\lambda)\hat{K}(\lambda)}{1 - \hat{K}(\lambda)} \\ &= e^{\lambda_* t} \frac{\hat{G}(\lambda_*)K(\lambda_*)}{-\hat{K}'(\lambda_*)} = B_0 e^{\lambda_* t} \end{aligned}$$

with

$$B_0 = \frac{\int_0^\infty e^{-\lambda_* a} G(a)da}{\int_0^\infty a e^{-\lambda_* a} K(a)da}$$

and

$$H_2(t) = \frac{1}{2\pi i} \int\limits_{\sigma_1 - i\infty}^{\sigma_1 + i\infty} \frac{\hat{G}(\sigma_1 + iy)\hat{K}(\sigma_1 + iy)}{1 - \hat{K}(\sigma_1 + iy)} e^{(\sigma_1 + iy)t} dy.$$

The function H_2 satisfies the estimate

$$|H_2(t)| \leq \frac{e^{\sigma_1 t}}{2\pi} \int\limits_{\sigma_1 - i\infty}^{\sigma_1 + i\infty} \left| \frac{\hat{G}(\sigma_1 + iy)\hat{K}(\sigma_1 + iy)}{1 - \hat{K}(\sigma_1 + iy)} \right| dy = B_1 e^{\sigma_1 t}.$$

Here B_1 is a constant. Thus, we arrived at the representation

$$B(t) = e^{\lambda_* t} B_0 + G(t) + e^{\sigma_1 t} B_1.$$

However, remembering that $G(t) = 0$ for $t \geq 0$, we can write

$$B(t) = B_0 e^{\lambda_* t} \left(1 + \frac{e^{-\lambda_*} G(t)}{B_0} + e^{(\sigma_1 - \lambda_*)t} \frac{B_1}{B_0}\right) = B_0 e^{\lambda_* t} \left(1 + \Omega(t)\right) \tag{5.4.45}$$

where $\Omega(t) \to 0$ as $t \to \infty$, provided $B_0 \neq 0$.

Now, $B_0 = 0$ if and only if $G(t) = 0$ for all $t \geq 0$ but then, from uniqueness, $B(t) = 0$ for all t. Let us interpret this condition. We have

$$0 = G(t) = \int\limits_0^\infty \beta(a + t) \frac{S(a + t)}{S(a)} n_0(a) da$$

which, by positivity of n_0, is possible only if

$$\beta(a + t) n_0(a) = 0$$

for $a \in [0, \omega]$ and $t \geq 0$. This occurs only if the support of β is to the left of the support of n_0 (as the support of $\beta(\cdot + t)$ moves to the left as t increases). In other words, this case occurs only if the original population is too old to become fertile. In this case

$$n(a, t) = \begin{cases} n_0(a - t)\frac{S(a)}{S(a-t)}, & t < a, \\ 0, & a < t. \end{cases} \tag{5.4.46}$$

Otherwise, we can write

$$n(a, t) = \begin{cases} n_0(a - t)\frac{S(a)}{S(a-t)}, & t < a, \\ B_0 e^{\lambda_*(t-a)} \left(1 + \Omega(t - a)\right) S(a), & a < t. \end{cases} \tag{5.4.47}$$

Now, in the case $\omega < +\infty$ we see that for $t \geq \omega$ we have

$$n(a, t) = B_0 e^{\lambda_*(t-a)} \left(1 + \Omega(t - a)\right) S(a)$$

and we identify the stable age distribution

$$n_\infty(a) = e^{-\lambda_* a - \int\limits_0^a \mu(s) ds}.$$

so that

$$\lim_{t \to \infty} e^{-\lambda_* t} n(a, t) = e^{-\lambda_* a - \int\limits_0^a \mu(s) ds}$$

on $[0, \omega]$ (provided the supports of n_0 and β meet).

Finally, we noted in (5.4.43) that $\lambda_* > 0$, $\lambda_* = 0$ and $\lambda_* < 0$ if and only if, respectively, $R > 1, R = 1$ and $R < 1$. Thus, the population is growing if $R > 1$, it is stable if $R = 1$ and it decays if $R < 1$ (again if supports of n_0 and β meet), in accordance with the interpretation of the parameter R.

6

Basic nonlinear models

1 Typical nonlinear extensions

1.1 Density dependent vital demographic coefficients

In Subsection 2 we introduced the age structured McKendrick population model

$$\begin{aligned}
&\partial_t n(a,t) + \partial_a n(a,t) = -\mu(a,t)n(a,t), \qquad t, a > 0\\
&n(0,t) = \int\limits_0^{\omega} n(a,t)\beta(a,t)da, \quad t > 0,\\
&n(a,0) = n_0(a), \quad a > 0,
\end{aligned}$$

where μ is the death rate, β is the maternity rate and $\omega \leq \infty$ is the maximum age of individuals in the population.

In many cases the assumption that μ and β do not depend on the population is a serious oversimplification. A more realistic is the system

$$\begin{aligned}
&\partial_t n(a,t)\partial t + \partial_a n(a,t)\partial a = -\mu(a,t,N_{\gamma_1})n(a,t), \qquad t, a > 0\\
&n(0,t) = \int\limits_0^{\omega} n(a,t)\beta(a,t,N_{\gamma_2})da, \quad t > 0,\\
&n(a,0) = n_0(a), \quad a > 0
\end{aligned} \tag{6.1.1}$$

where

$$N_{\gamma_i}(t) = \int\limits_0^{\infty} \gamma_i(a)n(a,t)da, \qquad i = 1, 2,$$

is the a weighted total population at time t. The weights γ_i account for the fact that the death and maternity rates may be more sensitive on the density of population at particular ages.

This makes (6.1.1) a (badly) nonlinear equation which only can be solved explicitly only in very special cases by using a technique often referred to as *linear chain trick*. Later we shall describe some cases yielding to more straightforward analysis but first we discuss another nonlinear extension.

1.2 An epidemiological system with age structure

Standard epidemiological models treat the population as homogeneous apart from the differences due to the disease. Then, for the description of the epidemics the population is divided into three main classes: susceptibles (individuals who are not sick and can be infected), infectives (individuals who have the disease and can infect others) and removed (individuals who were infective but recovered and are now immune, dead or isolated). Depending on the disease, other classes can be introduces to cater e.g., for the latent period of the disease. We denote by $S(t), I(t), R(t)$ the number of individuals in the classes above. By

$$S(t) + I(t) + R(t) = N(t)$$

we denote the total population size. In many models it is assumed that the population size is constant disregarding thus vital dynamics such as births and deaths. Thus, the total population is a conserved quantity and the relevant conservation law can be written as

$$\begin{aligned} S' &= -\lambda S + \delta I, \\ I' &= \lambda S - (\gamma + \delta) I, \\ R' &= \gamma I \end{aligned} \tag{6.1.2}$$

with $S(0) = S_0, I(0) = I_0, R(0) = R_0$ and $S_0 + I_0 + R_0 = N$. The parameter λ is the force of infection, δ is the recovery rate and γ is the recovery/removal rate. While δ and γ are usually taken to be constant, the force of infection requires a constitutive law. The simplest is the law of mass action

$$\lambda = c\phi \frac{I}{N}, \tag{6.1.3}$$

where c is the contact rate (the number of contacts that a single individual has with other individuals in the population per unit time, ϕ is the infectiveness; that is, the probability that a contact with an infective will result in infection and I/N is the probability that the contacted individual is infective. In what follows we shall denote $k = c\phi/N$.

There are many other assumptions underlying this model: that the population is homogenous, that no multiple infections are possible, that an infected individual immediately become infective, etc.

Concerning the nature of the disease the basic distinction is between those which are not lethal and do not impart immunity (influenza, common cold) and those which could be caught only once (leading to death or immunity) such as measles or AIDS. In the first case, $\gamma = 0$ and the model is referred to as an SIS model and in the second $\delta = 0$ and the model is called an SIR model.

In many cases the rate of infection significantly varies with age and thus it is important to consider the age structure of the population. Thus we expect the interaction of the vital dynamics and the infection mechanism to produce a nontrivial behaviour.

To introduce the model we note again that, in absence of the disease, the age-dependent density of the population $n(a,t)$ would be the solution of the linear model introduced in (5.1.7)–(5.1.9). However, because of the epidemics, the population is partitioned into the three classes: susceptibles, infectives and removed, represented by their respective age densities $s(a,t), i(a,t)$ and $r(a,t)$. Now, if we look at the population of susceptibles, than we see that it is losing individuals at the rate $\lambda(a,t)s(a,t)$ and gaining at the rate $\delta(a)i(a,t)$, where we have taken into account that the infection force and the cure rate are age dependent. Similarly, the source terms for the other two classes are given by the (age dependent) terms of the (6.1.2) model. This leads to the system

$$\begin{aligned} \partial_t s(a,t) + \partial_a s(a,t) + \mu(a) s(a,t) &= -\lambda(a,t) s(a,t) + \delta(a) i(a,t), \\ \partial_t i(a,t) + \partial_a i(a,t) + \mu(a) i(a,t) &= \lambda(a,t) s(a,t) - (\delta(a) + \gamma(a)) i(a,t), \\ \partial_t r(a,t) + \partial_a r(a,t) + \mu(a) r(a,t) &= \gamma(a) i(a,t). \end{aligned} \tag{6.1.4}$$

$$\begin{aligned} s(0,t) &= \int_0^\omega \beta(a)(s(a,t)+(1-q)i(a,t)+(1-w)r(a,t))da, \\ i(0,t) &= q\int_0^\omega \beta(a)i(a,t)da, \\ r(0,t) &= w\int_0^\omega \beta(a)r(a,t)da, \end{aligned} \tag{6.1.5}$$

where $q \in [0,1]$ and $w \in [0,1]$ are the vertical transmission coefficients of infectiveness and immunity, respectively. The system is complemented by initial conditions $s(a,0) = s_0(a), i(a,0) = i_0(a)$ and $r(a,0) = r_0(a)$. We remark that here we assumed that the death and birth coefficients are not significantly affected by the disease. In particular, if we assume that the solution of (6.2.52) exists in such a way that all terms are separately well defined, then adding the equations together we obtain that the total population density $n(a,t) = s(a,t) + i(a,t) + r(a,t)$ satisfies

$$\begin{aligned} \partial_t n(a,t) + \partial_a n(a,t) + \mu(a)n(a,t) = 0,& \\ n(0,t) &= \int_0^\omega \beta(a)n(a,t)da, \\ n(a,0) &= n_0(a) = s_0(a) + i_0(a) + r_0(a), \end{aligned}$$

that is, the disease does not change the global picture of the evolution of the population, as expected from the model.

Finally, we have to specify a constitutive relation for the force of infection λ. This usually is given by the equation

$$\lambda(a,t) = K_0(a)i(a,t) + \int_0^\omega K(a,s)i(s,t)ds, \tag{6.1.6}$$

where the two terms on the right hand side are called the intracohort and intercohort terms, respectively. The intracohort term describes the situation in which individuals only can be infected by those of their own age while the intercohort term describes the case in which they can be infected by individuals of any age.

Analysis of this problem requires abstract theory which will be developed later.

1.3 An exactly solvable nonlinear model - a linear chain trick

Consider

$$\begin{aligned} &\partial_t n(a,t) + \partial_a n(a,t) = -\mu(N(t))n(a,t), \qquad t > 0, 0 < a < \omega 0 \\ &n(0,t) = \int_0^\omega n(a,t)\beta(N(t))da, \quad t > 0, \\ &n(a,0) = n_0(a), \quad 0 < a < \omega. \end{aligned} \tag{6.1.7}$$

We assume that there is a nonnegative classical solution n to this problem defined for $t \in [0, t_{max})$ such that $\int_0^\omega \partial_t n(a,t)da = \partial_t \int_0^\omega n(a,t)da$ and $n(\omega,t) = 0$ for all $t \in [0, t_{max})$. Then, by these assumptions,

$$\int\limits_0^\omega \partial_a n(a,t)da = n(\omega,t) - n(0,t) = 0 - \beta(N(t)) \int\limits_0^\omega n(a,t)da = -\beta(N(t))N(t)$$

and we obtain an ordinary differential equation for the total population

$$\begin{aligned} \frac{dN}{dt} &= N(\beta(N) - \mu(N)), \\ N(0) &= N_0 = \int\limits_0^\omega n_0(a)da. \end{aligned} \tag{6.1.8}$$

This is a simple separable equation which can be solved explicitly or in quadratures, depending on the form of μ and β. An example allowing for explicit solution will be discussed later. Here we shall address two more theoretical issues.

First, let us ask ourselves whether this procedure produces the solution of the original problem. To wit, suppose N is a the solution to (6.1.8) defined on an interval $[0, t_{max})$. Substituting this (known) N into (6.1.7) we obtain a linear boundary value problem which can be explicitly solved, yielding a solution, say, u. Is this a solution of the original nonlinear problem (6.1.7)? We observe that u solves

$$\begin{aligned} &\partial_t u(a,t) + \partial_a u(a,t) = -\mu(N(t))n(a,t), \qquad t > 0, 0 < a < \omega \\ &u(0,t) = \int\limits_0^\omega u(a,t)\beta(N(t))da, \quad t > 0, \\ &u(a,0) = n_0(a), \quad 0 < a < \omega. \end{aligned} \tag{6.1.9}$$

where N is a given function, whereas n is a solution to

$$\begin{aligned} &\partial_t n(a,t) + \partial_a n(a,t) = -\mu\left(\int\limits_0^\omega n(a,t)da\right) n(a,t), \qquad t > 0, 0 < a < \omega \\ &n(0,t) = \int\limits_0^\omega n(a,t)\beta\left(\int\limits_0^\omega n(a,t)da\right)da, \quad t > 0, \\ &n(a,0) = n_0(a), \quad 0 < a < \omega. \end{aligned} \tag{6.1.10}$$

Integrating (6.1.9) we find that $U(t) = \int\limits_0^\omega u(a,t)da$ satisfies

$$\begin{aligned} \frac{dU}{dt} &= U(\beta(N) - \mu(N)), \\ U(0) &= N_0. \end{aligned} \tag{6.1.11}$$

Using the fact that N is a given solution to (6.1.8), the factor $(\beta(N) - \mu(N))$ in the latter and in (6.1.11) is the same, $U - N$ satisfies

$$\begin{aligned} \frac{d(U-N)}{dt} &= (U-N)(\beta(N) - \mu(N)), \\ U(0) - N(0) &= 0, \end{aligned}$$

which is a linear equation and thus uniquely solvable. Hence, $U(t) = N(t)$ on $[0, t_{max})$. Thus u satisfies

$$\begin{aligned}
&\partial_t u(a,t) + \partial_a u(a,t) = -\mu\left(\int\limits_0^\omega u(a,t)da\right) n(a,t), \qquad t > 0, 0 < a < \omega\\
&u(0,t) = \int\limits_0^\omega u(a,t)\beta\left(\int\limits_0^\omega u(a,t)da\right) da, \quad t > 0,\\
&u(a,0) = n_0(a), \quad 0 < a < \omega.
\end{aligned}$$

and hence $u(a,t) = n(a,t)$ provided (6.1.7) is uniquely solvable (which is a separate issue.)

As the second issue we consider the long time behaviour. Assume that there is a unique nontrivial equilibrium for (6.1.8) denoted by N^*. This can be ensured if e.g. $\beta(N)$ is a decreasing and $\mu(N)$ an increasing function with $\mu(0) < \beta(0)$ or if $\mu(0) = \beta(0)$ and $\beta'' - \mu'' < 0$. Under this assumptions N^* is the global attractor, that is $N(t) \to N^*$ as $t \to \infty$. Is there a corresponding stationary density $n^*(a)$ such that

$$N^* = \int\limits_0^\infty n^*(a)da?$$

A stationary density is a solution to

$$\begin{aligned}
&\partial_a n(a) = -\mu\left(\int\limits_0^\omega n(a)da\right) n(a), \qquad 0 < a < \omega 0\\
&n(0) = \int\limits_0^\omega n(a)\beta\left(\int\limits_0^\omega n(a)da\right). \qquad (6.1.12)
\end{aligned}$$

We can follow the argument for the time dependent solution. Assuming $u(\omega) = 0$ and integrating the first equation over $[0,\omega)$, we find

$$-N\beta\left(N\right) = -\mu\left(N\right)N$$

so that the total population at the stationary state is the equilibrium population. Further, arguing as in the time dependent case, we see that the nonlinear, nonlocal equation (6.1.12) for the stationary state n^* can be replaced by the linear one

$$\begin{aligned}
&\partial_a n(a) = -\mu\left(N^*\right) n(a), \qquad 0 < a < \omega\\
&n(0) = N^*\beta(N^*), \qquad (6.1.13)
\end{aligned}$$

which can be solved explicitly giving

$$n^*(a) = N^*\beta(N^*)e^{-\mu(N^*)}. \qquad (6.1.14)$$

This can be considered as the stable age profile of the population. However, at this moment we do not know whether $n(a,t) \to n^*(a)$ as $t \to \infty$.

Example 6.1. Consider

$$\begin{aligned}
&\partial_t n(a,t) + \partial_a n(a,t) = -\mu_0 N(t) n(a,t), \qquad t > 0, 0 < a < \omega\\
&n(0,t) = \beta_0 \int\limits_0^\omega n(a,t)N(t)da, \quad t > 0,\\
&n(a,0) = n_0(a), \quad 0 < a < \omega, \qquad (6.1.15)
\end{aligned}$$

where μ_0, β_0 are positive constants.

Let us integrate the first and the second equation with respect to a*. Using the considerations from the begining of the subsection, we obtain*

$$\begin{aligned}\frac{dN}{dt} &= N^2(\beta_0 - \mu_0),\\ N(0) = N_0 &= \int\limits_0^{\omega} n_0(a)da.\end{aligned} \tag{6.1.16}$$

There are two cases to consider.

a) $\mu_0 = \beta_0$*.*
In this case, $N(t) = N_0$ *for all* $0 \le t \le t_{max}$ *and (6.1.15) turns into*

$$\begin{aligned}&\partial_t n(a,t) + \partial_a n(a,t) = -\mu_0 N_0 n(a,t), \qquad t > 0, 0 < a < \omega\\ &n(0,t) = \beta_0 N_0^2, \quad t > 0,\\ &n(a,0) = n_0(a), \quad a > 0\end{aligned} \tag{6.1.17}$$

which can be transformed to the form (5.4.46) with the boundary condition given explicitly. We could use the formula which we derived there but here we can use the fact that

$$n(0,t) = \int\limits_0^{\infty} n(a,t)\beta_0 N_0 da = \beta_0 N_0 \int\limits_0^{\infty} n(a,t)da = \beta_0 N_0 N(t) = \beta_0 N_0^2$$

and we deal with a straightforward initial boundary value problem

$$\begin{aligned}&\frac{\partial n(a,t)}{\partial t} + \frac{\partial n(a,t)}{\partial a} = -\mu_0 N_0 n(a,t), \qquad t, a > 0\\ &n(0,t) = \beta_0 N_0^2, \quad t > 0,\\ &n(a,0) = n_0(a), \quad a > 0\end{aligned} \tag{6.1.18}$$

the solution of which is given by (5.4.46)

$$\begin{aligned}n(a,t) = e^{-\mu_0 N_0 a} u(a,t) &= e^{-\mu_0 N_0 a}\begin{cases} e^{\mu_0 N_0 (a-t)} n_0(a-t), & t < a,\\ \beta_0 N_0^2, & a < t\end{cases}\\ &= \begin{cases} e^{-\mu_0 N_0 t} n_0(a-t), & t < a,\\ e^{-\mu_0 N_0 a}\beta_0 N_0^2, & a < t.\end{cases}\end{aligned} \tag{6.1.19}$$

b) $\mu_0 \ne \beta_0$*.*
In this case the solution of (6.1.16) is given by

$$\frac{1}{N_0} - \frac{1}{N(t)} = (\beta_0 - \mu_0)t$$

or

$$N(t) = \frac{N_0}{1 - (\beta_0 - \mu_0)N_0 t}. \tag{6.1.20}$$

At this moment we note that the evolution of the population is determined by the sign of $r_0 = \beta_0 - \mu_0$ *which plays the role of the net growth rate: if* $r_0 > 0$ *(that is, if the maternity coefficient exceeds the mortality rate), the population will explode at* $t = 1/r_0 N_0$*. On the other hand, if* $r_0 < 0$ *(that is, if the maternity coefficient is lower than the mortality rate), the population will exists for all time gradually decaying to 0. However, in both cases we can give explicit formulae for* $n(a,t)$*. Indeed, similarly to a) we arrive at the straightforward initial boundary value problem*

$$\begin{aligned}&\partial_t n(a,t)+\partial_a n(a,t) = -\mu_0 N(t)n(a,t), \qquad t,a>0\\ &n(0,t)=\beta_0 N(t)^2, \quad t>0,\\ &n(a,0)=n_0(a), \quad a>0 \end{aligned} \tag{6.1.21}$$

where $N(t)$ is given (6.1.20). However, contrary to (6.1.18) (and (5.2.1)), the coefficient μ is dependent on time and thus the reduction must be made with more care. We begin by noting that

$$\frac{d}{dt}e^{\mu_0\int\limits_0^t N(s)ds} = \mu_0 N(t)e^{\mu_0\int\limits_0^t N(s)ds}$$

and thus the first equation of (6.1.21) can be written as

$$\frac{\partial}{\partial t}\left(n(a,t)e^{\mu_0\int\limits_0^t N(s)ds}\right)+\frac{\partial}{\partial a}\left(n(a,t)e^{\mu_0\int\limits_0^t N(s)ds}\right)=0.$$

If we denote

$$u(a,t)=n(a,t)e^{\mu_0\int\limits_0^t N(s)ds}$$

then

$$u(a,0)=n(a,0)e^{\mu_0\int\limits_0^0 N(s)ds}=n_0(a)$$

and

$$u(0,t)=n(0,t)e^{\mu_0\int\limits_0^t N(s)ds}=\beta_0 N(t)^2 e^{\mu_0\int\limits_0^t N(s)ds}$$

so that, using again (5.4.46), we obtain

$$\begin{aligned} n(a,t) = e^{-\mu_0\int\limits_0^t N(s)ds}u(a,t) &= e^{-\mu_0\int\limits_0^t N(s)ds}\begin{cases} n_0(a-t), & t<a,\\ \beta_0 N(t-a)^2 e^{\mu_0\int\limits_0^{t-a} N(s)ds}, & a<t\end{cases}\\ &=\begin{cases} e^{-\mu_0\int\limits_0^t N(s)ds} n_0(a-t), & t<a,\\ \beta_0 N(t-a)^2 e^{-\mu_0\int\limits_{t-a}^t N(s)ds}, & a<t\end{cases}\end{aligned} \tag{6.1.22}$$

as long as N is defined. For our particular case we have

$$\int N(s)ds = N_0\int\frac{ds}{1-r_0N_0t}=\ln\left(\frac{1}{1-N_0r_0t}\right)^{\frac{1}{r_0}}+C$$

for $0\le t<1/r_0N_0$ if $r_0>0$ and for $0\le t<\infty$ otherwise. Hence, for such t

$$e^{-\mu_0\int\limits_0^t N(s)ds}=(1-r_0N_0t)^{\frac{\mu_0}{r_0}}$$

and

$$e^{-\mu_0\int\limits_{t-a}^t N(s)ds}=\left(\frac{1-r_0N_0t}{1-r_0N_0(t-a)}\right)^{\frac{\mu_0}{r_0}}.$$

Thus

$$n(a,t)=\begin{cases}(1-r_0N_0t)^{\frac{\mu_0}{r_0}}n_0(a-t), & t<a,\\ \frac{\beta_0N_0^2}{(1-r_0N_0(t-a))^2}\left(\frac{1-r_0N_0t}{1-r_0N_0(t-a)}\right)^{\frac{\mu_0}{r_0}}, & a<t.\end{cases}$$

We note that in this example we cannot talk about a stable age profile, as either there is no equilibrium total population or every number is an equilibrium (if $\mu_0\neq\beta_0$).

1.4 Another example of linear chain trickery

Consider a class of organisms of potentially infinite longevity which can reproduce immediately after birth and whose age-specific birth rate decreases exponentially with age. This results in the system

$$\begin{aligned} &\partial_t n(a,t) + \partial_a n(a,t) = -\mu(N(t))n(a,t), \qquad t>0, 0<a<\infty \\ &n(0,t) = \beta_0 \int\limits_0^\infty e^{-\gamma a} n(a,t)da, \quad t>0, \\ &n(a,0) = n_0(a), \quad 0<a<\infty, \end{aligned} \tag{6.1.23}$$

where $\gamma, \beta > 0$ are constants and μ is a strictly monotonic function. Assuming, as before, that there is a nonnegative classical solution n to this problem defined for $t \in [0,\infty)$ such that $\int_0^\infty \partial_t n(a,t)da = \partial_t \int_0^\infty n(a,t)da$ and $\lim_{a\to\infty} n(a,t) = 0$ for all $t \in [0,\infty)$, we integrate the equation with respect to a over $[0,\infty)$

$$\int\limits_0^\infty \partial_t n(a,t) + \lim_{a\to\infty} n(a,t) - n(0,t) = -\mu(N(t)) \int\limits_0^\infty n(a,t)da$$

to obtain

$$N' = B - \mu(N)N, \tag{6.1.24}$$

where

$$B(t) = n(0,t) = \beta_0 \int_0^\infty e^{-\gamma a} n(a,t)da.$$

To close the system, we multiply the equation by the maternity function $\beta_0 e^{-\gamma a}$

$$\partial_t \left(\beta_0 e^{-\gamma a} n(a,t)\right) + \beta_0 e^{-\gamma a} \partial_a n(a,t) = -\mu(N(t))\beta_0 e^{-\gamma a} n(a,t)$$

and integrate over $[0,\infty)$ so that

$$B'(t) + \int\limits_0^\infty \beta_0 e^{-\gamma a} \partial_a n(a,t)da = -\mu(N(t))B(t). \tag{6.1.25}$$

The integral can be transformed by integration by parts

$$\int\limits_0^\infty e^{-\gamma a} \partial_a n(a,t)da = -n(0,t) + \gamma \int\limits_0^\infty e^{-\gamma a} n(a,t)da = -\beta_0 B(t) + \gamma B(t)$$

so that (6.1.26) can be written as

$$B'(t) - \beta_0 B(t) + \gamma B(t) = -\mu(N(t))B(t). \tag{6.1.26}$$

Altogether, this time we have obtained a system of differential equations

$$\begin{aligned} N' &= B - \mu(N)N, \\ B' &= (\beta_0 - \gamma - \mu(N))B \end{aligned} \tag{6.1.27}$$

This system can be solved in some special cases; otherwise we have to resort to a phase-plane analysis.

Constant death rate

We begin with a simple case when $\mu(N) = \mu$ is a constant and thus we have to take $\omega = \infty$. Then we obtain the decoupled system

$$\begin{aligned} N' &= B - \mu N, \\ B' &= (\beta_0 - \gamma - \mu)B. \end{aligned}$$

This gives

$$B(t) = B_0 e^{(\beta_0-\gamma-\mu)t},$$

where

$$B_0 = B(0) = \beta_0 \int\limits_0^\infty e^{-\gamma a} n_0(a)da.$$

Then the first equation becomes a linear nonhomogeneous equation

$$N' = -\mu N + B_0 e^{(\beta_0-\gamma-\mu)t}$$

Using the integrating factor, we have

$$(Ne^{\mu t})' = B_0 e^{(\beta_0-\gamma)t};$$

that is

$$N(t) = N_0 e^{-\mu t} + \frac{B_0}{\beta_0-\gamma}\left(e^{(\beta_0-\gamma-\mu)t} - e^{-\mu t}\right),$$

provided $\beta_0 \neq \gamma$. Note, that finding N is not necessary for finding n. Indeed, in our case (6.1.23) takes the form

$$\begin{aligned} &\partial_t n(a,t) + \partial_a n(a,t) = -\mu n(a,t), \qquad t > 0, 0 < a < \infty \\ &n(0,t) = B(t) = B_0 e^{(\beta_0-\gamma-\mu)t}, \quad t > 0, \\ &n(a,0) = n_0(a), \quad 0 < a < \infty. \end{aligned} \tag{6.1.28}$$

Then, as in (5.4.46), we get

$$n(a,t) = e^{-\mu t}\begin{cases} n_0(a-t), & t < a, \\ \beta_0 e^{(\beta_0-\gamma)(t-a)} \int\limits_0^\infty e^{-\gamma r} n_0(r)dr, & a < t. \end{cases}$$

General death rate

In general system (6.1.23) cannot be explicitly solved and the best thing we can hope for is to determine asymptotic behaviour of the solution. We shall use phase plane analysis to to this. The isoclines are given by

$$\begin{aligned} B &= \mu(N)N, \\ 0 &= (\beta_0 - \gamma - \mu(N))B \end{aligned} \tag{6.1.29}$$

and the stationary points are at the intersection of these isoclines. We immediately see that we have trivial equilibrium $(0,0)$ and a nontrivial one are given by

$$\mu(N) = \beta_0 - \gamma, \qquad B = \mu(N)N.$$

Then, clearly a nontrivial equilibrium exists if $\beta_0 - \gamma$ is in the range of μ; then it is unique by monotonicity of μ. Assume this is the case and denote the equilibrium by (N^*, B^*). As usual, we begin by looking at stability of the equilibrium. The Jacobian is given by

$$\mathcal{J} = \begin{pmatrix} -\mu'(N)N - \mu(N), & 1 \\ -\mu'(N)B, & (\beta_0 - \gamma - \mu(N)) \end{pmatrix}$$

and

$$\mathcal{J}(N^*, B^*) = \begin{pmatrix} -\mu'(N^*)N^* - \mu(N^*), & 1 \\ -\mu'(N^*)B^*, & 0 \end{pmatrix}$$

The characteristic equation is

$$\lambda^2 + (\mu'(N^*)N^* + \mu(N^*))\lambda + \mu'(N^*)B^* = 0.$$

Under natural biological assumption $\mu' > 0$, using Vieta's formulae, we see that if there are two real roots, then there are of the same sign with their sum negative which yields that they are both negative. If they are complex, then they are complex conjugate and thus the real part is negative. Hence the equilibrium is asymptotically stable.

We can get a better understanding by considering the line

$$B = (\beta_0 - \gamma)N.$$

It is an invariant manifold. Indeed, consider $x(t) = B(t) - (\beta_0 - \gamma)N(t)$. Then

$$\begin{aligned} x' &= B'(t) - (\beta_0 - \gamma)N' \\ &= (\beta_0 - \gamma - \mu(N))B - (\beta_0 - \gamma)(B - \mu(N)N) = -\mu(N)(B - (\beta_0 - \gamma)N) = -\mu(N)x. \end{aligned}$$

Treating N as a known function, we see that x is a solution of a linear equation. Thus, if (B, N) is on the manifold, it will stay there. Moreover, if $x \neq 0$, then it will decrease to 0 monotonically (even exponentially provided $\mu(N)$ is separated from zero). On the manifold the equation reduces to

$$\frac{dN}{dt} = ((\beta_0 - \gamma) - \mu(N))N$$

which produces logistic like behaviour. It is also clear that there cannot be closed orbits or damped oscillations around the nontrivial equilibrium as such orbits would have to cut the invariant manifold which is impossible. We can get a global phase portrait if we note that since $\mu' > 0$, for the existence of the stationary state N^* we must have $\mu(0) < \beta_0 - \gamma$. This shows that

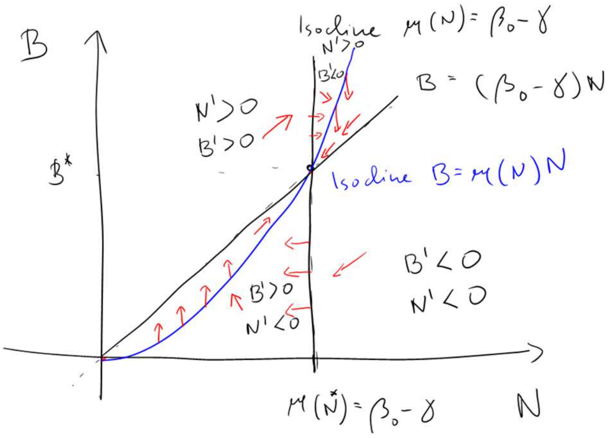

Fig. 6.1. Phase plane (N, B)

$$\frac{d}{dN}(N\mu(N))|_{N=0} = \mu(0) < \beta_0 - \gamma$$

and thus the isocline $B = \mu(N)N$ enters the first quadrant below the invariant manifold $B = (\beta_0 - \gamma)N$. Furthermore, since along this isocline the vector field is of the form $(0, (\beta_0 - \gamma - \mu(N))B)$ and along the invariant manifold is $(\beta_0 - \gamma - \mu(N))N(1, (\beta_0 - \gamma))$ we see that the only intersection of the isocline and the invariant manifold can occur at the equilibrium (N^*, B^*). Hence, the isocline and the invariant manifold must be situated with respect to each other as on Fig. 1.4. From the picture we see that the sectors between the isocline and the invariant manifold are trapping regions and any trajectory starting outside them must eventually get there by monotonicity, the fact that $B = 0$ is also an invariant manifold and the fact that any finite limit point of a trajectory must be an equilibrium

Finally, using Eq. (6.1.14), we find that the stable age profile in this case will be

$$n^*(a) = (\beta_0 - \gamma)N^* e^{-(\beta_0-\gamma)a}.$$

As explained earlier, the total population $N(t)$ converges to N^* which is the total population of the stationary density n^* but proving that the densities converge requires a separate theory.

2 Solvability of the SIR/SIS system with age structure

Let us recall the system

$$\begin{aligned}\partial_t s(a,t) + \partial_a s(a,t) + \mu(a)s(a,t) &= -\lambda(a,t)s(a,t) + \delta(a)i(a,t),\\ \partial_t i(a,t) + \partial_a i(a,t) + \mu(a)i(a,t) &= \lambda(a,t)s(a,t) - (\delta(a) + \gamma(a))i(a,t),\\ \partial_t r(a,t) + \partial_a r(a,t) + \mu(a)r(a,t) &= \gamma(a)i(a,t).\end{aligned} \tag{6.2.30}$$

with the boundary conditions

$$\begin{aligned}s(0,t) &= \int\limits_0^\omega \beta(a)(s(a,t) + (1-q)i(a,t) + (1-w)r(a,t))da,\\ i(0,t) &= q\int\limits_0^\omega \beta(a)i(a,t)da,\\ r(0,t) &= w\int\limits_0^\omega \beta(a)r(a,t)da,\end{aligned} \tag{6.2.31}$$

where $q \in [0,1]$ and $w \in [0,1]$ are the vertical transmission coefficients of infectiveness and immunity, respectively, and the initial conditions

$$s(a,0) = s_0(a), \quad i(a,0) = i_0(a), \quad r(a,0) = r_0(a). \tag{6.2.32}$$

To deal with solvability of this system we have to provide a more general framework for the original McKendrick problem.

2.1 Solution of the McKendrick as a semidynamical system

We considered the problem

$$\begin{aligned}&\partial_t n(a,t) + \partial_a n(a,t) = -\mu(a)n(a,t)\\ &n(0,t) = \int\limits_0^\omega \beta(a)n(a,t)da,\\ &n(a,0) = n_0(a)\end{aligned} \tag{6.2.33}$$

and proved that, under assumptions (5.3.23)–(5.3.26), it has a solution

$$n(a,t) = \begin{cases} \frac{S(a)}{S(a-t)} n_0(a-t), & t < a, \\ S(a)B(t-a), & a < t, \end{cases} \tag{6.2.34}$$

where B is the solution of the Volterra equation

$$\begin{aligned} B(t) &= \int\limits_0^t K(t-a)B(a)da + \int\limits_0^{\omega-t} \beta(a+t)\frac{S(a+t)}{S(a)} n_0(a)da \\ &= \int\limits_0^t K(t-a)B(a)da + \int\limits_t^{\omega} \beta(a)\frac{S(a)}{S(a-t)} n_0(a-t)da \end{aligned} \tag{6.2.35}$$

for $0 \leq t \leq \omega$ and

$$B(t) = \int\limits_0^{\omega} K(t-a)B(a)da \tag{6.2.36}$$

for $t > \omega$. The solution satisfies the estimate

$$\|n(\cdot,t)\|_{L_1([0,\omega])} \leq \|n_0\|_{L_1([0,\omega])} e^{t\|\beta\|_\infty}. \tag{6.2.37}$$

Unique solvability of (6.2.33) in the sense of formula (6.2.51) allows for the dynamical systems approach to the problem. This amounts to looking at the solution not through individual trajectories but more globally, as a family of mappings of the state space into itself, parametrized by time. Let us recall that our state space is $X = L_1([0,\omega])$ of all population densities with finite total population. Then, for any $n_0 \in X$ we define

$$[T(t)n_0](a) = n(a,t) \tag{6.2.38}$$

where n is the solution defined by (6.2.51). From (6.2.37) we see that $\{T(t)\}_{t\geq 0}$ is a family of linear bounded operators on X with at most exponential growth in time. If what follows we investigate the properties of $\{T(t)\}_{t\geq 0}$.

Earlier we observed that, in general, the solution (6.2.51) is discontinuous along the characteristic $a = t$. Here we prove that the condition (5.3.31) is indeed essential for regularity of the solution.

Theorem 6.2. *Let us assume that (5.3.23)–(5.3.26) are satisfied. Further let*

$$n_0 \in W_1^1([0,\omega]), \quad \mu n_0 \in L_1([0,\omega]) \tag{6.2.39}$$

and

$$n_0(0) = \int\limits_0^{\omega} \beta(a)n_0(a)da. \tag{6.2.40}$$

Then $n \in C([0,\omega]\times\mathbb{R}_+)$, $\mu(\cdot)n(\cdot,t) \in L_1([0,\omega])$ for any $t > 0$, $n \in W_1^1([0,\omega]\times\mathbb{R}_+)$, that is, $\partial_t n, \partial_a n$ exist almost everywhere on $[0,\omega]\times\mathbb{R}_+$ and (6.2.33) is satisfied almost everywhere.

Proof. First we note that functions from $W_1^1([0,\omega])$ are (absolutely) continuous and thus, by Theorem 5.3 and formula (6.2.51), n is continuous outside the characteristic $t = a$. Using argument as in (5.3.31) and the relevant assumption we find that n is continuous on $[0,\omega]\times\mathbb{R}_+$. Similarly, if (6.2.39) is satisfied then, again by Theorem 5.3, B is differentiable almost everywhere and thus n is differentiable almost everywhere away from the characteristic $t = a$. Then the continuity along $t = a$ shows that $n \in W_1^1([0,\omega]\times\mathbb{R}_+)$. To simplify the argument, let as consider $\omega = \infty$, $\Delta_1 = \{(a,t);\ t > a\}$ and $\Delta_2 = \{(a,t);\ t < a\}$ and functions $f_i \in W_1^1(\Delta_i)$, $i = 1,2$ satisfying $f_1(t,t) = f_2(t,t)$. Define F by $F|_{\Delta_1} = f_1$ and $F|_{\Delta_2} = f_2$. Let $\phi \in C_0^\infty(\mathbb{R}_+^2)$. Then

$$\begin{aligned}\int\limits_{\mathbb{R}_+^2} F\phi_a dadt &= \int\limits_{\Delta_1} f_1\phi_a dadt + \int\limits_{\Delta_2} f_2\phi_a dadt\\ &= \int\limits_0^\infty\left(\int\limits_0^t f_1\phi_a da\right)dt + \int\limits_0^\infty\left(\int\limits_t^\infty f_2\phi_a da\right)dt\\ &= -\int\limits_0^\infty\left(\int\limits_0^t f_{1,a}\phi da\right)dt + \int\limits_0^\infty\left(\int\limits_t^\infty f_{2,a}\phi da\right)dt + \int\limits_0^\infty (f_1(t,t)-f_2(t,t))\phi(t,t)dt\\ &= -\int\limits_{\mathbb{R}_+^2}\tilde{F}\phi dadt\end{aligned}$$

where $\tilde{F}$ is given by $\tilde{F}|_{\Delta_1} = f_{1,a}$ and $\tilde{F}|_{\Delta_2} = f_{2,a}$. Thus, $\tilde{F} = F_a$ is the generalized derivative of F and is integrable since the its components are. Analogous argument holds for the derivative with respect to t.

Finally, we have

$$\begin{aligned}\int\limits_0^\omega \mu(a)n(a,t)da &= \int\limits_0^{\min\{t,\omega\}} \mu(a)B(t-a)S(a)da + \int\limits_{\min\{t,\omega\}}^\omega \mu(a)\frac{S(a)}{S(a-t)}n_0(t-a)da\\ &\leq \max_{s\in[0,\omega]}|B(s)| \int\limits_0^{\min\{t,\omega\}} \mu(a)S(a)da + e^{\int\limits_0^\omega \mu(s)ds}\max_{s\in[0,\omega]}|n_0(s)| \int\limits_{\min\{t,\omega\}}^\omega \mu(a)S(a)da\\ &\leq \max_{s\in[0,\omega]}|B(s)| + e^{\int\limits_0^\omega \mu(s)ds}\max_{s\in[0,\omega]}|n_0(s)|\end{aligned}$$

where we used the fact that μS is a probability density. □

Lemma 6.3. *The set*

$$\mathcal{D} = \{\phi \in W_1^1([0,\omega]);\ \mu\phi \in L_1([0,\omega]), \phi(0) = \int\limits_0^\omega \beta(a)\phi(a)da\} \qquad (6.2.41)$$

is dense in $L_1([0,\omega])$.

Proof. Let $f \in L_1([0,\omega])$. Fix $\epsilon > 0$ and take $\phi \in C_0^\infty((0,\omega))$ such that

$$\|f-\phi\|_{L_1([0,\omega])} \leq \epsilon. \qquad (6.2.42)$$

Since ϕ is of compact support, $\mu\phi \in L_1([0,\omega])$ is automatically satisfied. We have to construct ϕ_ϵ satisfying (6.2.40) and close to ϕ in L_1. For a given $\delta > 0$

$$\phi_\delta(a) = \phi(a) + \alpha_\delta e^{-a/\delta},$$

where α_δ is an undetermined constant. We require

$$\phi_\delta(0) = \phi(0) + \alpha_\delta = \int\limits_0^\omega \beta(a)\phi(a)da + \alpha_\delta\int\limits_0^\omega e^{-a/\delta}\beta(a)da. \qquad (6.2.43)$$

Now,

$$|\int\limits_0^\omega e^{-a/\delta}\beta(a)da| \leq \delta \sup_{a\in[0,\omega]} |\beta(a)| \int\limits_0^{\omega/\delta} e^{-s}ds =: b\delta$$

and therefore, for sufficiently small δ, (6.2.43) can be solved for α_δ with

$$\alpha_\delta = \frac{\int\limits_0^\omega \beta(a)\phi(a)da - \phi(0)}{1 - \int\limits_0^\omega e^{-a/\delta}\beta(a)da},$$

and

$$|\alpha_\delta| \leq \frac{\sup\limits_{a\in[0,\omega]} |\beta(a)| \|\phi\|_{L_1([0,\omega])} + \max\limits_{a\in[0,\omega]} |\phi|}{1-\delta b} \leq C\|\phi\|_{W_1^1([0,\omega])}.$$

where C is a independent of δ for sufficiently small δ. Hence

$$\|\phi - \phi_\delta\|_{L_1([0,\omega])} \leq |\alpha_\epsilon| \int\limits_0^\omega e^{-a/\delta}da \leq |\alpha_\delta| \int\limits_0^\omega e^{-a/\delta}da \leq \delta C\|\phi\|_{W_1^1([0,\omega])}.$$

Thus, for a given $\phi \in C_0^\infty((0,\omega))$ satisfying (6.2.42) we can choose $\delta = \epsilon/C\|\phi\|_{W_1^1([0,\omega])}$ so that

$$\|f - \phi_\delta\|_{L_1([0,\omega])} \leq 2\epsilon$$

with $\phi_\delta \in \mathcal{D}$. □

Proposition 6.4. *If (5.3.23)–(5.3.26) are satisfied, then the solution n defined by (6.2.51) satisfies*

$$n \in C([0,T], L_1([0,\omega])). \tag{6.2.44}$$

In other words, the family $\{T(t)\}_{t\geq 0}$ is strongly continuous.

Proof. Clearly, if $\phi \in C([0,\omega] \times [0,T])$, then $\phi \in C([0,T], L_1([0,\omega])$ by uniform continuity and bounded domain of integration. Let n be the solution with initial condition $n_0 \in L_1([0,\omega])$ and consider a sequence $\phi_k \in \mathcal{D}$ such that $\phi_k \to n_0$ in L_1. Let further n_k be the solution with initial condition ϕ_k. Then, by (6.2.37),

$$\sup_{t\in[0,T]} \|n(\cdot,t) - n_k(\cdot,t)\|_{L_1([0,\omega])} \leq= e^{T\|\beta\|_\infty} \|n_0 - \phi_k\|_{L_1([0,\omega])}.$$

which shows that $n \in C([0,T], L_1([0,\omega])$. □

Proposition 6.5. *The family $\{T(t)\}_{t\geq 0}$ has the semigroup property, that is, for any $n_0 \in X$ and $t_1, t_2 \geq 0$ we have*

$$T(t+\tau)n_0 = T(\tau)(T(t)n_0). \tag{6.2.45}$$

Proof. Let $\{T(t)\}_{t\geq 0}$ be associated with (6.2.33) and $\{U(t)\}_{t\geq 0}$ be associated with the McKendrick problem with $\mu = 0$ and $\beta(a)$ replaced by $K(a) = \beta(a)S(a)$. By (5.3.17) we have $[T(t)n_0](a) = [S(a)U(t)S^{-1}(a)n_0]$ and thus it is enough to prove (6.2.45) for $\{U(t)\}_{t\geq 0}$, that is, for the solution

$$u(a,t) = [U(t)u_0] = \begin{cases} u_0(a-t), & t < a, \\ B(t-a), & a < t. \end{cases} \tag{6.2.46}$$

where

$$B(t) = \int\limits_0^t K(t-a)B(a)da + \int\limits_t^\infty K(a)u_0(a-t)da$$

and, again, we consider $\omega = \infty$ (by extending the coefficients by 0 beyond ω. Further, denote by B_f the unique solution to the above equation with u_0 replaced by f. Then we have

$$\begin{aligned}[U(\tau)(U(t)u_0)](a) &= \begin{cases} [U(t)u_0](a-\tau) & \text{for } a > \tau, \\ B_{U(t)u_0}(\tau - a) & \text{for } a < \tau \end{cases} \\ &= \begin{cases} u_0(a-t-\tau) & \text{for } a-\tau > t, \\ B_{u_0}(t+\tau-a)) & \text{for } \tau < a < t+\tau \\ B_{U(t)u_0}(\tau-a) & \text{for } a < \tau, \end{cases}\end{aligned}$$

where we used

$$[U(t)u_0)](a-\tau) = \begin{cases} u_0(a-(t+\tau)) & \text{for } a-\tau > t, \\ B_{u_0}(t+\tau-a)) & \text{for } \tau < a < t+\tau. \end{cases}$$

Now, for $r > 0$,

$$B_{U(t)u_0}(r) = \int_0^r K(r-s)B_{U(t)u_0}(s)ds + \int_r^\infty K(\alpha)[U(t)u_0](\alpha - r)d\alpha.$$

We transform the right hand side as follows

$$\begin{aligned}&\int_0^r K(r-s)B_{U(t)u_0}(s)ds + \int_r^\infty K(\alpha)[U(t)u_0](\alpha-r)d\alpha \qquad (6.2.47)\\ &= \int_0^r K(t+r-(t+s))B_{U(t)u_0}(s)ds + \int_r^{t+r} K(\alpha)[U(t)u_0](\alpha-r)d\alpha + \int_{t+r}^\infty K(\alpha)[U(t)u_0](\alpha-r)d\alpha \\ &= \int_t^{t+r} K(t+r-\sigma)B_{U(t)u_0}(\sigma-t)d\sigma + \int_r^{t+r} K(\alpha)[U(t)u_0](\alpha-r)d\alpha + \int_{t+r}^\infty K(\alpha)[U(t)u_0](\alpha-r)d\alpha \\ &= \int_t^{t+r} K(t+r-\sigma)B_{U(t)u_0}(\sigma-t)d\sigma + \int_r^{t+r} K(\alpha)B_{u_0}(t-(\alpha-r))d\alpha + \int_{t+r}^\infty K(\alpha)u_0(\alpha-(r+t))d\alpha \\ &= \int_t^{t+r} K(t+r-\sigma)B_{U(t)u_0}(\sigma-t)d\sigma + \int_0^t K(t+r-v)B_{u_0}(v)dv + \int_{t+r}^\infty K(\alpha)u_0(\alpha-(r+t))d\alpha \\ &= \int_0^{t+r} K(t+r-\sigma)F(\sigma)d\sigma + \int_{t+r}^\infty K(\alpha)u_0(\alpha-(r+t))d\alpha, \qquad (6.2.48)\end{aligned}$$

where $F(\sigma) = B_{U(t)u_0}(\sigma - t)$ for $t \leq \sigma \leq t+r$ and $F(\sigma) = B_{u_0}(s)$ for $0 \leq \sigma < t$. Next, we apply the following argument. Consider the Volterra equation

$$f(t) = \int_{t_0}^t \Phi(t-s)f(s)ds + g(t)$$

for $t \geq t_0 \geq 0$ and g is a known function. Let this equation be uniquely solvable for arbitrary t_0 and arbitrary g and denote $f(t, t_0, g)$ the solution. Then

$$f(t,0,g) = f\left(t, t_0, g + \int_0^{t_0} \Phi(t-s)f(s,0,g)ds\right).$$

Indeed,

$$f(t) \equiv \int_0^t \Phi(t-s)f(s)ds + g(t) \equiv \int_{t_0}^t \Phi(t-s)f(s)ds + \int_0^{t_0} \Phi(t-s)f(s)ds + g(t).$$

On the other hand, there is a solution $\tilde{f}$ to

$$\tilde{f}(t) = \int_{t_0}^t \Phi(t-s)\tilde{f}(s)ds + \int_0^{t_0} \Phi(t-s)f(s)ds + g(t),$$

but we know that f already satisfies this equation. By uniqueness, f and $\tilde{f}$ must coincide.

Going back to our problem, there is a unique solution to

$$F(t+r) = \int_0^{t+r} K(t+r-\sigma)F(\sigma)d\sigma + \int_{t+r}^{\infty} K(\alpha)u_0(\alpha-(r+t))d\alpha$$

and, using the penultimate line of (6.2.48), we see that $B_{u_0}(t+r) = B_{U(t)u_0}(r)$ or, returning to the original variables, $B_{u_0}(t+\tau-a) = B_{U(t)u_0}(\tau-a)$. Hence

$$[U(\tau)(U(t)u_0)](a) = \begin{cases} u_0(a-t-\tau) & \text{for } a > t+\tau, \\ B_{u_0}(t+\tau-a)) & \text{for } a < t+\tau \end{cases} = [U(\tau+t)u_0)](a)$$

□

The previous results show that $\{T(t)\}_{t\geq 0}$ is a strongly continuous dynamical systems, or a strongly continuous semigroup, that is, it is a family of bounded linear operators satisfying, for any $x \in X$,

1. $T(t+\tau)x = T(t)T(\tau)x, \quad t,\tau \geq 0$;
2. $T(0)x = x$;
3. $\lim\limits_{t\to 0^+} T(t)x = x$.

An operator A is called the generator of the semigroup $\{T(t)\}_{t\geq 0}$ if

$$Ax = \lim_{h\to 0^+} \frac{T(h)x - x}{h} \tag{6.2.49}$$

whenever the limit exists in X. The set of such $x \in X$ is called the domain of A and denoted $D(A)$. Typically A is unbounded and $D(A) \neq X$. It follows that if $x \in D(A)$, then $t \to u(t,x) = T(t)x$ is differentiable, $T(t)x \in D(A)$ for all $t \geq 0$ and

$$\frac{d}{dt}u(t,x) = Au(t,x), \qquad u(t,x) = x \tag{6.2.50}$$

that is $\{T(t)\}_{t\geq 0}$ gives solutions to the Cauchy problem (6.2.50) and $u(t,x)$ is the semiflow associated with this equation due to the semigroup property

$$u(t+\tau,x) = T(t+\tau)x = T(t)T(\tau)x = u(t,u(\tau,x)).$$

In the case of the McKendrick problem, it can be proved that $A = -\partial_a - \mu$ with $D(A) = \mathcal{D}$ but the proof is outside the scope of these notes.

Finally we note the following fact which will be of importance in the analysis of nonlinear problems.

Proposition 6.6. *Under assumptions (5.3.23)–(5.3.26), if n_0 is bounded, then $n(t,a)$ is bounded on $[0,\omega] \times [0,T]$ for any $0 \leq T < \infty$.*

Proof. From the formula

$$n(a,t) = \begin{cases} \frac{S(a)}{S(a-t)} n_0(a-t), & t < a, \\ S(a)B(t-a), & a < t, \end{cases} \tag{6.2.51}$$

we see that since $S(a)/S(a-t) \le 1$, $S(a) \le 1$, the part in $t < a$ is bounded by the boundedness of n_0. Also, we have proved that the iterates (5.3.28) defining B converge uniformly on each bounded time interval and thus $B(t-a)$ is bounded on $[0,\omega] \times [0,T]$ (if $\omega = \infty$, then we observe that in any case, $a < t \le T$). □

2.2 Linear systems

To address the solvability of the McKendrick epidemiological system we first look at its linear part

$$\begin{aligned} \partial_t s(a,t) + \partial_a s(a,t) + \mu(a)s(a,t) - \delta(a)i(a,t) &= 0, \\ \partial_t i(a,t) + \partial_a i(a,t) + \mu(a)i(a,t) + (\delta(a)+\gamma(a))i(a,t) &= 0, \\ \partial_t r(a,t) + \partial_a r(a,t) + \mu(a)r(a,t) - \gamma(a)i(a,t) &= 0. \end{aligned} \tag{6.2.52}$$

with the boundary conditions

$$\begin{aligned} s(0,t) &= \int_0^\omega \beta(a)(s(a,t) + (1-q)i(a,t) + (1-w)r(a,t))da, \\ i(0,t) &= q\int_0^\omega \beta(a)i(a,t)da, \\ r(0,t) &= w\int_0^\omega \beta(a)r(a,t)da, \end{aligned} \tag{6.2.53}$$

and the initial conditions

$$s(a,0) = s_0(a), \quad i(a,0) = i_0(a), \quad r(a,0) = r_0(a). \tag{6.2.54}$$

The problem is an example of a more general vector McKendrick system

$$\partial_t \mathbf{n} = \mathcal{S}\mathbf{n} + \mathcal{M}\mathbf{n}, \tag{6.2.55}$$

$$\mathbf{n}(t,a) = (n_1(a,t), \dots, n_N(a,t))$$

and $n_i(a,t)$ is the population density at time t of individuals in patch i and being of age a (here $\mathbf{n} = (s,i,r)$). Further,

$$\mathcal{S}\mathbf{n} = -\partial_a \mathbf{n} = (-\partial_a n_1, \dots, -\partial_a n_N) \tag{6.2.56}$$

describes aging, $\mathcal{M}(a) = \{\mu_{ij}(a)\}_{1\le i,j\le N}$ is the mortality/projection matrix. In our case

$$\mathcal{M} = \begin{pmatrix} -\mu & \delta & 0 \\ 0 & -(\mu+\delta+\gamma) & 0 \\ 0 & \gamma & -\mu \end{pmatrix} = \begin{pmatrix} -\mu & 0 & 0 \\ 0 & -\mu & 0 \\ 0 & 0 & -\mu \end{pmatrix} + \begin{pmatrix} 0 & \delta & 0 \\ 0 & -(\delta+\gamma) & 0 \\ 0 & \gamma & 0 \end{pmatrix} \tag{6.2.57}$$

where the first matrix describes death which is an intrapatch phenomenon and the second refers to migrations between patches and is thus a Kolmogorov matrix. We require that our general $\mathcal{M}$ has the same structure, that is, it can be written as

$$\mathcal{M} = \operatorname{diag}\{-\mu_1, \dots, -\mu_N\} + \mathcal{Q}$$

where $\mathcal{Q}$ is a Kolmogorov matrix, that is, it is positive diagonal and the sum of entries in each column is 0. This structure has important consequences as far as the asymptotic properties are concerned.

This system is supplemented by the McKendrick boundary condition

$$[\gamma \mathbf{n}](t) = \mathbf{n}(t,0) = [\mathcal{B}\mathbf{n}](t) = \int\limits_0^\infty B(a)\mathbf{n}(t,a)da, \tag{6.2.58}$$

where γ denotes the operator of taking the trace at $a = 0$ and $B(a) = \{\beta_{ij}(a)\}_{1\leq i,j\leq N}$ is the fertility matrix. We note that births may be interpatch phenomenon like in our case

$$B = \begin{pmatrix} \beta & \beta(1-q) & \beta(1-w) \\ 0 & \beta q & 0 \\ 0 & 0 & \beta w \end{pmatrix} \tag{6.2.59}$$

The initial condition is given by

$$\mathbf{n}|_{t=0} = \mathbf{n}(0,a) = \overset{\circ}{\mathbf{n}}(a). \tag{6.2.60}$$

The natural phase space for the problem is $X = L_1([0,\omega],\mathbb{R}^N)$. We denote by X_+ the subset of X consisting of vectors $\ltimes$ which are coordinate-wise nonnegative almost everywhere. Further, $X_\infty = L_\infty([0,\omega],\mathbb{R}^N)$ By adapting our earlier consideration concerning the scalar case, or by a more functional analytic approach, we can prove the solvability result as follows. Let us denote by $\mathcal{V}_\mathcal{M}(a,b)$ the fundamental solution matrix of the equation $\mathbf{z}'_a(a) = \mathcal{M}(a)\mathbf{z}(a)$; that is, $\mathbf{z}(a) = \mathcal{V}_\mathcal{M}(a)\mathbf{z}_0$ satisfies the above equation with $\mathbf{z}(b) = \mathbf{z}_0$ ($\mathcal{V}_\mathcal{M}$ plays the role of the integrating factor in the scalar case). Since the columns of $\mathcal{V}_\mathcal{M}(a,0)$ are linearly independent for any a, the inverse $\mathcal{V}_\mathcal{M}^{-1}(a,0)$ always exists and thus $\mathcal{V}_\mathcal{M}(a,b) = \mathcal{V}_\mathcal{M}(a,0)\mathcal{V}_\mathcal{M}^{-1}(b,0)$. With this, we can write the solution to (6.2.55)-(6.2.58) as

$$\mathbf{n}(a,t) = \begin{cases} \mathcal{V}_\mathcal{M}(a,a-t)\,\overset{\circ}{\mathbf{n}}, & a > t, \\ (\mathcal{V}_\mathcal{M}(a,0)\boldsymbol{\psi})(t-a) & a < t, \end{cases} \tag{6.2.61}$$

where $\boldsymbol{\psi}$ satisfies the Volterra equation

$$\boldsymbol{\psi}(t) = \int\limits_0^t (B(a)\mathcal{V}_\mathcal{M}(a,0)\boldsymbol{\psi})(t-a)da + \int\limits_t^\infty B(a)\mathcal{V}_\mathcal{M}(a,a-t)\,\overset{\circ}{\mathbf{n}}\,(a-t)da$$

Let us define by $\mathcal{A}$ the realization of $-diag\{\partial_a\} - \mathcal{M}$ on the domain

$$D_A = \{\mathbf{n} \in (W^{1,1}(\mathbb{R}_+))^N, \mathbf{n}(0) = \mathcal{B}\mathbf{n}\}.$$

Then

Theorem 6.7. *$\mathcal{A}$ generates a strongly continuous semigroup $\{\mathcal{T}(t)\}_{t\geq 0}$ such that*

$$||\mathcal{T}(t)|| \leq e^{(\,\bar{b}-\underline{m}\,)t},$$

where $\bar{b} := \sup\limits_{a\in\mathbb{R}_+} ||B(a)||$ and $\underline{m} := \inf\limits_{j,a} \mu_j(a)$. Furthermore, if $\overset{\circ}{\mathbf{n}}\in X_+$, then $\mathbf{n}(t,\cdot) \in X_+$ and if $\overset{\circ}{\mathbf{n}}\in X_\infty \cap X_1$, then $\mathbf{n}(t,\cdot) \in X_\infty \cap X_1$.

2.3 The nonlinear system

With the notation of the previous section, the problem (6.2.30)–(6.2.32) can be written in compact form

$$\begin{aligned}\partial_t \mathbf{n} &= \mathcal{A}\mathbf{n} + \mathfrak{F}(\mathbf{n}), \quad t > 0,\\ \mathbf{n}|_{a=0} &= \mathcal{B}\mathbf{n},\\ \mathbf{n}|_{t=0} &= \overset{\circ}{\mathbf{n}},\end{aligned} \tag{6.2.62}$$

where $\mathbf{n} = (s, i, r)$, $\mathcal{A} = \mathcal{S} + \mathcal{M}$ with $\mathcal{S}$ and $\mathcal{M}$ defined by (6.2.56) and (6.2.57). Further, $\mathcal{B}$ is defined by (6.2.59) and $\mathfrak{F}$ is a nonlinear perturbation

$$\mathfrak{F}((s, i, r)) = \begin{pmatrix} -\lambda \; 0 \; 0 \\ \lambda \;\; 0 \; 0 \\ 0 \;\; 0 \; 0 \end{pmatrix} \begin{pmatrix} s \\ i \\ r \end{pmatrix} \tag{6.2.63}$$

where the force of infection depends on the solution through the formula

$$\lambda(a, t) = K_0(a)i(a, t) + \int\limits_0^\omega K(a, s)i(s, t)ds, \tag{6.2.64}$$

where $K_0(a)$ and $K(a, s)$ are known functions. How to deal with such problems? First we note that the boundary condition is really a part of the definition of the domain of $\mathcal{A}$ and thus, if we find a solution to (6.2.62), then it must satisfy $u \in D_A$ and automatically satisfies the boundary condition. Thus, for a time being we shall ignore it.

The main idea is to use the fact that we can solve the linear version of (6.2.62)

$$\begin{aligned}\partial_t \mathbf{u} &= \mathcal{A}\mathbf{u} + \mathbf{f}(t), \quad t > 0,\\ \mathbf{u}|_{t=0} &= \overset{\circ}{\mathbf{u}},\end{aligned} \tag{6.2.65}$$

where f is a given function, and the solution is given by the Duhamel formula

$$\mathbf{u}(t) = \mathcal{T}(t)\,\overset{\circ}{\mathbf{u}} + \int\limits_0^t \mathcal{T}(t - s)\mathbf{f}(s)ds.$$

So, if we knew the solution $\mathbf{n}$ to (6.2.62), then it would be given by

$$\mathbf{n}(t) = \mathcal{T}(t)\,\overset{\circ}{\mathbf{n}} + \int\limits_0^t \mathcal{T}(t - s)\mathfrak{F}(\mathbf{n}(s))ds. \tag{6.2.66}$$

Even if we do not know the solution, then (6.2.66) offers a simplification of (6.2.62) by not involving unbounded operator $\mathcal{A}$. Of course, a solution to (6.2.66) is not necessarily a solution to (6.2.62) but at least it seems to be step in right direction in the sense that any continuous solution to (6.2.62) must be a solution to (6.2.66).

The problem (6.2.66) can be solved by Picard's iterations, similar to (5.3.28)

$$\begin{aligned}\mathbf{n}_0(t) &= \overset{\circ}{\mathbf{n}},\\ \mathbf{n}_{k+1}(t) &= \mathcal{T}(t)\,\overset{\circ}{\mathbf{n}} + \int\limits_0^t \mathcal{T}(t - s)\mathfrak{F}(\mathbf{n}_k(s))ds,\end{aligned} \tag{6.2.67}$$

however, handling the nonlinearity $\mathfrak{F}$ requires more care.

Let as recall that by $Y = C([t_0, T], X)$ we denoted the space of continuous functions $[t_0, T] \ni t \to \mathbf{u}(t) \in X$, where $-\infty < t_0 < T < \infty$. We define the norm in Y by

$$\|\mathbf{u}(\cdot)\|_Y = \sup_{t \in [t_0, T]} \|\mathbf{u}(t)\|_X.$$

Let us assume that $\mathfrak{F}$ satisfies the global Lipschitz condition on X, that is, there is $L > 0$ such that for any $\mathbf{u}, \mathbf{v} \in X$

$$\|\mathfrak{F}(\mathbf{u}) - \mathfrak{F}(\mathbf{v})\|_X \leq L\|\mathbf{u} - \mathbf{v}\|. \tag{6.2.68}$$

Then, returning to (6.2.67) we obtain, for any $t \in [0, T]$,

$$\begin{aligned} \|\mathbf{n}_{k+1}(t) - \mathbf{n}_k(t)\|_X &\leq \int\limits_0^t \|\mathcal{T}(t-s)(\mathfrak{F}(\mathbf{n}_k(s)) - \mathfrak{F}(\mathbf{n}_{k-1}(s)))\|_X ds \\ &\leq Me^{\omega T} Lt\|\mathbf{n}_k - \mathbf{n}_{k-1}\|_Y \end{aligned} \tag{6.2.69}$$

which, by induction as in (5.3.29), yields

$$\sup_{t\in[0,T]} \|\mathbf{n}_{k+1}(t) - \mathbf{n}_k(t)\|_X \leq \frac{(Me^{\omega T}L)^k}{k!}\|\mathbf{n}_1 - \overset{\circ}{\mathbf{n}}\|_Y$$

and, as before, this shows that $(\mathbf{n}_k)_{k\in\mathbb{N}}$ converges to a continuous solution to (6.2.66) defined on the whole interval $[0, T]$ for any $T < \infty$. Such solutions are called global.

If we consider another solution $\mathbf{v}$ to (6.2.66) with the initial condition $\overset{\circ}{\mathbf{v}}$, then

$$\begin{aligned} \|\mathbf{n}(t) - \mathbf{v}(t)\|_X &\leq Me^{\omega T}\| \overset{\circ}{\mathbf{n}} - \overset{\circ}{\mathbf{v}} \|_X + \int\limits_0^t \|\mathcal{T}(t-s)(\mathfrak{F}(\mathbf{n}(s)) - \mathfrak{F}(\mathbf{v}(s)))\|_X ds \\ &\leq Me^{\omega T}\| \overset{\circ}{\mathbf{n}} - \overset{\circ}{\mathbf{v}} \|_X + Me^{\omega T}L\int\limits_0^t \|\mathbf{n}(s) - \mathbf{v}(s)\|_X ds \end{aligned}$$

and Gronwall's inequality gives

$$\|\mathbf{n}(t) - \mathbf{v}(t)\|_X \leq Me^{\omega T}e^{MLTe^{\omega T}}\| \overset{\circ}{\mathbf{n}} - \overset{\circ}{\mathbf{v}} \|_X, \quad 0 \leq t \leq T,$$

so that we obtain that the solution is (Lipschitz) continuous with respect to the initial data and is unique (by putting $\overset{\circ}{\mathbf{n}}=\overset{\circ}{\mathbf{v}}$).

However, it is easy to see that even for a simple nonlinearities such as $\mathfrak{F}(u) = u^2$ we have

$$|\mathfrak{F}(u) - \mathfrak{F}(v)| = |(u + v)||u - v|$$

and thus $\mathfrak{F}$ is Lipschitz continuous as long as we restrict u and v to satisfy $|u|, |v| \leq K$ for some constant K. Thus, a quadratic nonlinearity is Lipschitz continuous,but not globally, as the Lipschitz constant depends on bounds for u and v. Functions like this are called locally Lipschitz. Precisely, $\mathfrak{F}$ is said to satisfy a local Lipschitz condition on X if for any $c > 0$ there is $L = L_c$ such that

$$\|\mathfrak{F}(\mathbf{u}) - \mathfrak{F}(\mathbf{v})\|_X \leq L_c\|\mathbf{u} - \mathbf{v}\|_X \tag{6.2.70}$$

whenever $\|u\|_X, \|v\|_X \leq c$.

In such a case we cannot use directly (6.2.69) as the constant L changes with the iterates and can grow to infinity. We can, however, use this argument if we make sure that all the iterates will stay in a fixed bounded set of X. We can prove the following result.

Theorem 6.8. *Let $\mathfrak{F} : X \to X$ be a locally Lipschitz function. If $\mathcal{A}$ is the generator of a semigroup $\{\mathcal{T}(t)\}_{t\geq 0}$, then for every $\overset{\circ}{\mathbf{n}}\in X$ and every $t_0 \in \mathbb{R}$ there is $t_{max} > t_0$ such that the Cauchy problem*

$$\begin{aligned} \partial_t \mathbf{n} &= \mathcal{A}\mathbf{n} + \mathfrak{F}(\mathbf{n}), \quad t > 0, \\ \mathbf{n}|_{t=t_0} &= \overset{\circ}{\mathbf{n}}, \end{aligned} \tag{6.2.71}$$

has a unique mild solution $\mathbf{n}$ *on* $[t_0, t_{max})$. *Moreover, if* $t_{max} < \infty$, *then*

$$\lim_{t \to t_{max}} \|\mathbf{n}(t)\|_X = \infty.$$

Proof. The proof consists in finding $\delta > 0$ and a closed subset of $C([t_0, t_0 + \delta], X)$ such that the iterates (6.2.67) stay in this set or, in other words, that the mapping

$$[\mathbf{F}(\mathbf{n})](t) = \mathcal{T}(t - t_0) \overset{\circ}{\mathbf{n}} + \int\limits_{t_0}^{t} \mathcal{T}(t - s)\mathfrak{F}(\mathbf{n}(s))ds,$$

is a self map of this set. Hence, let us take some $0 < \delta < 1$ and denote $M_0 = \sup_{0 \le t \le t_0 + 1} \|\mathcal{T}(t)\|$, $K_0 > M_0\|\mathbf{n}_0\|_0 + 1$. Then, for $t_0 \le t \le t_0 + \delta$ and $\mathbf{n} \in B(0, K_0) \subset C([t_0, t + \delta], X)$

$$\begin{aligned}\|[\mathbf{F}(\mathbf{n})](t)\|_X &\le M_0\|\mathbf{n}_0\|_X + M_0 \int\limits_{t_0}^{t} \|\mathfrak{F}(\mathbf{n}(s)) - \mathfrak{F}(0)\|_X ds + M_0(t - t_0)\|\mathfrak{F}(0)\|_X \\ &\le M_0\|\mathbf{n}_0\|_X + \delta M_0 L_{K_0} K_0 + M_0\delta\|\mathfrak{F}(0)\|_X \\ &\le M_0(\|\mathbf{n}_0\|_X + \delta(L_{K_0}K_0 + \|\mathfrak{F}(0)\|_X)).\end{aligned}$$

Hence we see that if take the length of the time interval

$$\delta = \min\left\{1, \frac{1}{M_0(L_{K_0}K_0 + \|\mathfrak{F}(0)\|_X)}\right\}, \tag{6.2.72}$$

then

$$\|[\mathbf{F}(\mathbf{n})](t)\|_X \le K_0$$

and the $k + 1$ iterate is in $B(0, K_0)$ provided the kth one is. Then on $B(0, K_0)$ the function $\mathfrak{F}$ is globally Lipschitz and the iterates converge to the mild solution of (6.2.71).

From what we just proved, it follows that if $\mathbf{n}$ is a mild solution on a closed interval $[t_0, t_0 + \tau]$, then it can be extended onto $[t_0, t_0 + \tau + \delta]$ with $\delta > 0$ defined by (6.2.72) with t_0 replaced by $t_0 + \tau$. Of course, δ can decrease with every step. Let $[t_0, t_{max})$ be the maximal interval to which the solution can be extended by such a procedure. If $t_{max} < \infty$, then $\|\mathbf{n}(t)\|_X \to \infty$ as $t \to t_{max}$. Indeed, otherwise we would have a sequence $t_n \to t_{max}$ such that $\|\mathbf{n}(t_n)\|_X \le C$ for some constant C. However, at each t_n we can extend the solution to the interval $[t_n, t_n + \delta]$ with δ independent of t_n (δ depends only on C and the linear semigroup). For t_n sufficiently close to t_{max} we would then have $t_n + \delta > t_{max}$ which contradicts the definition of t_{max}. □

The above theorem does not address the question whether our mild solution is the solution to (6.2.71), that is, whether it can be differentiated and whether it belongs to D_A and thus satisfies the boundary conditions. Full discussion of regularity is beyond the scope of these notes. We only note that for $\mathbf{n}$ to be a classical solution to (6.2.71) it suffices that $\overset{\circ}{\mathbf{n}} \in D_A$ and $\mathbf{n} \to \mathfrak{F}(\mathbf{n})$ be continuously differentiable.

Let us return to the epidemiological problem (6.2.62) with $\mathfrak{F}$ given by (6.2.63)–(6.2.64). Our state space is $X = L_1([0, \infty))^3 = L_1([0, \infty), \mathbb{R}^3)$ and

$$\|\mathbf{n}\|_X = \|(s, i, r)\|_X = \|s\|_{L_1([0,\infty))} + \|i\|_{L_1([0,\infty))} + \|r\|_{L_1([0,\infty))}.$$

To simplify discussion, we only consider the intercohort infection and disregard r. Then

$$\mathfrak{F}(\mathbf{n}_1) - \mathfrak{F}(\mathbf{n}_2) =$$

$$\begin{pmatrix} -\int\limits_0^\infty Ki_1 da \ 0 \\ \int\limits_0^\infty Ki_1 da \ \ 0 \end{pmatrix} \begin{pmatrix} s_1 \\ i_1 \end{pmatrix} - \begin{pmatrix} -\int\limits_0^\infty Ki_2 da \ 0 \\ \int\limits_0^\infty Ki_2 da \ \ 0 \end{pmatrix} \begin{pmatrix} s_2 \\ i_2 \end{pmatrix}$$

$$\begin{pmatrix} -s_1 \int\limits_0^\infty Ki_1 da + s_2 \int\limits_0^\infty Ki_2 da \\ s_1 \int\limits_0^\infty Ki_1 da - s_2 \int\limits_0^\infty Ki_2 da \end{pmatrix}$$

It is now easy to see that $\mathfrak{F}$ is locally Lipschitz continuous as

$$\begin{aligned} \|\mathfrak{F}(\mathbf{n}_1) - \mathfrak{F}(\mathbf{n}_2)\|_X &\leq 2 \int\limits_0^\infty \left| s_1(a) \int\limits_0^\infty K(\alpha) i_1(\alpha) d\alpha - s_2(a) \int\limits_0^\infty K(\alpha) i_2(\alpha) d\alpha \right| da \\ &\leq 2 \left(\int\limits_0^\infty |s_1(a) - s_2(a)| \int\limits_0^\infty K(\alpha)|i_1(\alpha)| d\alpha da + \int\limits_0^\infty |s_2(a)| \int\limits_0^\infty K(\alpha)|i_1(\alpha) - i_2(\alpha)| d\alpha da \right) \\ &\leq C\|(s_1 - s_2, i_1 - i_2)\|_X, \end{aligned}$$

where

$$C = 2 \sup_{a \in \mathbb{R}_+} K(a) \max\{\|i_1\|_{L_1([0,\infty))}, \|s_2\|_{L_1([0,\infty))}\}.$$

Hence, the problem has a unique solution defined at least on some interval $[0, \delta]$.

We note that the case with intracohort infection the situation is slightly different as the product is of two integrable functions not necessarily is integrable. However, from Proposition 6.6, we know that if the initial conditions $\overset{\circ}{s}, \overset{\circ}{i} \in X_{1,\infty} := L_1([0,\infty)) \cap L_\infty([0,\infty))$, then the solution also is in this space and so is the product is.

$$\begin{aligned} \|\mathfrak{F}(\mathbf{n}_1) - \mathfrak{F}(\mathbf{n}_2)\|_{X_{1,\infty}} &\leq 2 \int\limits_0^\infty |K(a)(i_1(a)s_1(a) - s_2(a)i_2(a))| \, da \\ &\leq 2 \sup_{a \in \mathbb{R}_+} K(a) \int\limits_0^\infty (|s_1(a) - s_2(a)||i_1(a)| da + |s_2(a)||i_1(a) - i_2(a)|) \, da \\ &\leq C\|(s_1 - s_2, i_1 - i_2)\|_{X_{1,\infty}}, \end{aligned}$$

where

$$C = 2 \sup_{a \in \mathbb{R}_+} K(a) \max\{\|i_1\|_{L_\infty([0,\infty))}, \|s_2\|_{L_\infty([0,\infty))}\},$$

and, as above, there exists a mild solution to (6.2.62) on some interval $[0, \delta]$.

Can this solution be extended to $[0, \infty)$? We observe that if $i, s \geq 0$ then in, say, intercohort case,

$$\|\mathbf{n}(t)\|_X = \int\limits_0^\infty (|i(a,t)| + |s(a,t)|) da = \int\limits_0^\infty (i(a,t) + s(a,t)) da = \int\limits_0^\infty n(a,t) da$$

where n is the solution of the equation obtained by adding together (6.2.52)–(6.2.53). Since we know that n exists for all t, $\|\mathbf{n}(t)\|_X$ would be bounded for any finite t and thus would be extendable to $[0, \infty)$. If we look at the iterates (6.2.67), we see that, since $\overset{\circ}{\mathbf{n}} \geq 0$ and the linear semigroup $\{\mathcal{T}(t)\}_{t \geq 0}$ preserves positivity, the iterates, and thus the solution, will be positive if $\mathfrak{F}(\mathbf{u}) \geq 0$ for $\mathbf{u} \geq 0$. However, clearly $\mathfrak{F}$ is not positive.

To solve this problem, we observe that the iterations in Theorem 6.8 are performed on a fixed ball in X (or $X_{1,\infty}$). This means that in the iterations we can always assume that the argument $\mathbf{n}$ of $\mathfrak{F}$ satisfies $\|\mathbf{n}\| \leq C$ with respective norm, for some constant C.

Then we observe that the problem (6.2.71) is equivalent to

$$\partial_t \mathbf{n} = (\mathcal{A}\mathbf{n} - \omega \mathbf{n}) + (\omega \mathbf{n} + \mathfrak{F}(\mathbf{n})) = \mathcal{A}_\omega \mathbf{n} + \mathfrak{F}_\omega(\mathbf{n})$$

for any $\omega \in \mathbb{R}$. It is easy to see that the semigroup generated by $\mathcal{A}_\omega$ is $\{\mathcal{T}_\omega(t)\}_{t\geq 0} = \{e^{-\omega t}\mathcal{T}(t)\}_{t\geq 0}$. The semigroup $\{T_\omega(t)\}_{t\geq 0}$ also preserves positivity. Therefore $\mathbf{n}$ is the mild solution to

$$\mathbf{n}(t) = \mathcal{T}_\omega(t)\, \overset{\circ}{\mathbf{n}} + \int_0^t \mathcal{T}_\omega(t-s)\mathfrak{F}_\lambda(\mathbf{n}(s))\, ds, \quad 0 \leq t < \delta. \tag{6.2.73}$$

In our case we have

$$\mathfrak{F}_\omega(\mathbf{n}) = \begin{pmatrix} -\int\limits_0^\infty K(a)i(a)da \; 0 \\ \int\limits_0^\infty K(a)i(a)da \;\; 0 \end{pmatrix} \begin{pmatrix} s \\ i \end{pmatrix} + \omega \begin{pmatrix} s \\ i \end{pmatrix}$$

$$\begin{pmatrix} -s(a)\int\limits_0^\infty K(a)i(a)da + \omega s(a) \\ s(a)\int\limits_0^\infty K(a)i(a)da + \omega i(a) \end{pmatrix}$$

and we see that if we take $\omega > C \sup_{a\in\mathbb{R}_+} K(a)$, then $\mathfrak{F}_\omega(\mathbf{n}) \geq 0$ for any $\mathbf{n} \geq 0$ satisfying $\|\mathbf{n}\| \leq C$. Thus all iterates are nonnegative and thus the solution is nonnegative. By the earlier argument, we have global solvability of the age structured epidemiological problem.

7

Appendices

1 Appendix A: Solvability of differential equations and Picard iterates

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

$$y' = f(t, y), \tag{7.1.1}$$

where f is a given function of two variables.

Several comments are in place here. Firstly, even though in such a simplified form, equation (7.1.1) in general has no closed form solution, that is, it is impossible to write the solution in the form

$$y(t) = combination\ of\ elementary\ functions\ like \sin t, \cos t, \ln t, polynomials...$$

Example 7.1. A trivial example is the equation

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

We know 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.

If a solution to a given equation can be written in terms of integrals of elementary functions (as above), then we say that the equation is *solvable in quadratures.* 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. However, there are many instances when we cannot solve an equation even in quadratures. How do we know then that the equation has a solution? The answer is that if the right hand side of the equation, that is the function f, is continuous, then there is at least one solution to (7.1.1). This result is called the Peano Theorem and involves some more advanced calculus. Thus, we can safely talk about solutions to ODEs of the form (7.1.1) even without knowing their explicit formulae.

Another important problem is related to the uniqueness of solutions, that is, whether there is only one solution to a given ODE. A quick reflection shows that clearly not: for the simplest equation

$$y' = 0,$$

the solutions are

$$y(t) = C,$$

where C is an arbitrary constant; thus there are infinitely many solutions. The uniqueness question, however, hasn't been properly posed. In fact, what we are looking for is usually a solution passing through a specified point.

Example 7.2. Assume that a point is moving along the horizontal line with speed given by $v(t) = t$. Find the position of the point at $t = 5$. To solve this problem let us recall that $v(t) = \frac{ds}{dt}$ where s is the distance travelled. Thus the problem results in an equation of the type discussed above:

$$v(t) = \frac{ds}{dt} = t$$

and

$$s(t) = 0.5t^2 + C$$

where C is an arbitrary constant. Hence $s(5) = 12.5 + C$ and there is no proper answer. In this physical setting the original question is clearly wrongly posed. What we need to give the proper answer is the information about the position of the point at some other time t, say, $t = 1$. If we know that (with respect to a fixed origin) $s(1) = 2$, then also $s(1) = 0.5 + C$ and $C = 1.5$. Therefore $s(5) = 12.5 + 1.5 = 14$.

From this example (and from physical or other considerations) it follows that if we are interested in getting a unique answer, we not only need the equation (which reflects usually some natural law) but also the state of the system (that is, the value of the solution) at some specified point. Thus, the complete *Cauchy* or *initial value* problem would be to solve

$$\begin{aligned} y' &= f(t, y), \quad \text{for all} \quad t \in [t_1, t_2] \\ y(t_0) &= y_0, \quad \text{for some } t_0 \in [t_1, t_2]. \end{aligned} \tag{7.1.2}$$

Once again we emphasize that to solve (7.1.2) is to find a continuously differentiable function $y(t)$ such that

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

Example 7.3. Check that the function $y(t) = \sin t$ is the 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.

Returning to our uniqueness question we ask whether the problem (7.1.2) has always a unique solution. The answer is negative.

Example 7.4. The Cauchy problem

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

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

However, there is a large class of functions f for which (7.1.2) has exactly one solution. Before we formulate the main result of this section, we must introduce necessary definitions. We say that a function g defined on an interval $[a, b]$ is Lipschitz continuous if there is a constant L such that

$$|g(x_1) - g(x_2)| \leq L|x_1 - x_2|, \tag{7.1.3}$$

for any $x_1, x_2 \in [a, b]$. In particular, if g is continuously differentiable on a closed rectangle $[a, b]$, the derivative g' is bounded there by, say, a constant M and then, by the Mean Value Theorem, we have for $x_1, x_2 \in [a, b]$

$$|g(x_1) - g(x_2)| = |g'(\theta)(x_1 - x_2)| \leq M|x_1 - x_2|, \tag{7.1.4}$$

where θ is a point between x_1 and x_2, so that continuously differentiable functions are Lipschitz continuous. However, e.g., $g(x) = |x|$ is Lipschitz continuous but not continuously differentiable.

Theorem 7.5. *(Picard's theorem) Let f be continuous in the rectangle $R : |t - t_0| \leq a, |y - y_0| \leq b$ for some $a, b > 0$ and satisfy the Lipschitz condition with respect to y uniformly in t: for any $(t, y_1), (t, y_2) \in R$*

$$|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2|. \tag{7.1.5}$$

Denote

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

and define $\alpha = \min\{a, b/M\}$. Then the initial value problem (7.1.2) has exactly one solution at least on the interval $t_0 - \alpha \leq t \leq t_0 + \alpha$.

Remark 7.6. All continuous functions $f(t, y)$ having continuous partial derivative $\frac{\partial f}{\partial y}$ in some neighbourhood of (t_0, y_0) give rise to (7.1.2) with exactly one solution (at least close to t_0). In fact, if f is continuously differentiable with respect to y on some closed rectangle R so that the derivative is bounded there, say, by a constant L:

$$\left|\frac{\partial f}{\partial y}(t, y)\right| \leq L$$

for all $(t, y) \in R$ which, by (7.1.4), gives Lipschitz continuity w.r.t y

We split the proof of this result into several steps. First we shall prove a general result known as Gronwall's lemma.

Lemma 7.7. *If $f(t)$, $g(t)$ are continuous and nonnegative for $t \in [t_0, t_0 + \alpha]$, $\alpha > 0$, and $c > 0$, then*

$$f(t) \leq c + \int_{t_0}^{t} f(s)g(s)ds \tag{7.1.6}$$

on $[t_0, t_0 + \alpha]$ implies

$$f(t) \leq c \exp\left(\int_{t_0}^{t} g(s)ds\right) \tag{7.1.7}$$

for all $[t_0, t_0 + \alpha]$.

If f satisfies (7.1.6) with $c = 0$, then $f(t) = 0$ on $[t_0, t_0 + \alpha]$.

Proof. Define $F(t) = c + \int_{t_0}^{t} f(s)g(s)ds$ on $[t_0, t_0 + \alpha]$. Then $F(t) \geq f(t)$ and $F(t) > 0$ on this interval. Differentiating, we get $F'(t) = f(t)g(t)$ and therefore

$$\frac{F'(t)}{F(t)} = \frac{f(t)g(t)}{F(t)} \leq \frac{f(t)g(t)}{f(t)} = g(t).$$

However, the left-hand side is equal to $\frac{d}{dt} \ln F(t)$ so that

$$\frac{d}{dt}\ln F(t) \le g(t),$$

or, integrating

$$\ln F(t) - \ln F(t_0) \le \int\limits_{t_0}^{t} g(s)ds.$$

Since $F(t_0) = c$, we obtain

$$F(t) \le c\exp\left(\int\limits_{t_0}^{t} g(s)ds\right)$$

but since, as we observed, $f(t) \le F(t)$, we have

$$f(t) \le c\exp\left(\int\limits_{t_0}^{t} g(s)ds\right),$$

which proves the first part. If $c = 0$, then we cannot use the above argument directly, as it would involve taking logarithm of zero. However, if

$$f(t) \le \int\limits_{t_0}^{t} f(s)g(s)ds$$

then

$$f(t) \le c + \int\limits_{t_0}^{t} f(s)g(s)ds$$

for any $c > 0$ and so

$$0 \le f(t) \le c\exp\left(\int\limits_{t_0}^{t} g(s)ds\right)$$

for any $c > 0$ but this yields $f(t) = 0$ for all t in $[t_0, t_0 + \alpha]$. ■

Gronwall's inequality can be used to show that, under the assumptions of Picard's theorem, there can be at most one solution to the Cauchy problem (7.1.2). Let $y_1(t)$ and $y_2(t)$ be two solutions of the Cauchy problem (7.1.2) on R with the same initial condition y_0, that is $y_1'(t) \equiv f(t, y_1(t))$, $y_1(t_0) = y_0$ and $y_2'(t) \equiv f(t, y_2(t))$, $y_2(t_0) = y_0$. Then $y_1(t_0) - y_2(t_0) = 0$ and

$$y_1'(t) - y_2'(t) = f(t, y_1(t)) - f_2(t, y_2(t)).$$

Integrating and using the condition at t_0 we see that

$$y_1(t) - y_2(t) = \int\limits_{0}^{t} (f(t, y_1(s)) - f_2(t, y_2(s)))ds.$$

Using next (7.1.5) we have

$$\begin{aligned}|y_1(t) - y_2(t)| &= \left|\int\limits_{t_0}^{t} (f(t, y_1(s)) - f_2(t, y_2(s)))ds\right| \le \int\limits_{t_0}^{t} |f(t, y_1(s)) - f_2(t, y_2(s))|ds \\ &\le L\int\limits_{t_0}^{t} |y_1(s)) - y_2(s)|ds, \qquad (7.1.8)\end{aligned}$$

thus we can use the second part of Gronwall's lemma to claim that $|y_1(t)-y_2(t)| = 0$ or $y_1(t) = y_2(t)$ for all t satisfying $|t - t_0| < a$.

The proof of the existence is much more complicated. Firstly, we convert the Cauchy problem (7.1.2) to an integral equation by integrating both sides of the equation in (7.1.2) and using the initial condition to get

$$y(t) = y_0 + \int\limits_{t_0}^{t} f(s, y(s))ds. \tag{7.1.9}$$

If $y(t)$ is a differentiable solution to (7.1.2), then of course (7.1.9) is satisfied. On the other hand, if $y(t)$ is a continuous solution to (7.1.9), then it is also differentiable, and we see that by differentiating (7.1.9) we obtain the solution of (7.1.2). Thus, we shall concentrate on finding continuous solutions to (7.1.9). The approach is to define the so-called Picard's iterates by

$$\begin{aligned} y_0(t) &= y_0, \\ y_n(t) &= \int\limits_{t_0}^{t} f(s, y_{n-1}(s))ds, \end{aligned} \tag{7.1.10}$$

and proving that they converge to the solution.

As a first step, we shall show that the iterates remain in the rectangle R. Namely, if M, a, b, α are defined as in the formulation of Picard's theorem and y_n is defined as in (7.1.10), then for any n

$$|y_n(t) - y_0| \le M|t - t_0| \tag{7.1.11}$$

for $|t - t_0| \le \alpha$. Note that (7.1.11) means that y_n is sandwiched between lines $y_0 \pm M(t - t_0)$ and the wedge created by these lines is always inside the rectangle R if $|t - t_0| < \alpha$.

To prove (7.1.11), we note that for $n = 0$ the estimate (7.1.11) is obvious, so to proceed with induction, we shall assume that it is valid for some $n > 0$ and, taking $n + 1$, we have

$$|y_{n+1}(t) - y_0| = \left|\int\limits_{t_0}^{t} f(s, y_n(s))ds\right| \le \int\limits_{t_0}^{t} |f(s, y_n(s))|ds.$$

However, by the remark above and the induction assumption, $y_n(s)$ is in R as long as $|t - t_0| \le \alpha$ and thus $|f(s, y_n(s))| \le M$. Thus easily

$$|y_{n+1}(t) - y_0| \le M|t - t_0|.$$

In the next step we shall show that the sequence of Picard's iterates converges. To make things simpler, we shall convert the sequence into a series the convergence of which is easier to establish. To this end we write

$$y_n(t) = y_0 + (y_1(t) - y_0) + (y_2(t) - y_1(t)) + \ldots + (y_n(t) - y_{n-1}(t)) \tag{7.1.12}$$

and try to show that

$$\sum_{n=0}^{\infty} |y_{n+1}(t) - y_n(t)| < +\infty$$

for any $|t - t_0| \le \alpha$, which would give the convergence of the series.

We use induction again. Assume that $t > t_0$, the analysis for $t < t_0$ being analogous. Firstly, proceeding as in (7.1.8), we observe that for $n > 1$

$$|y_n(t) - y_{n-1}(t)| \le \int\limits_{t_0}^{t} |f(s, y_{n-1}(s)) - f(s, y_{n-2}(s))|ds \le L\int\limits_{t_0}^{t} |y_{n-1}(s) - y_{n-2}(s)|ds. \tag{7.1.13}$$

Now, for $n = 1$ we obtain

$$|y_1(t) - y_0| \leq M(t - t_0)$$

and for $n = 2$

$$|y_2(t) - y_1(t)| \leq \int\limits_{t_0}^{t} |f(s, y_1(s)) - f(s, y_0)|ds \leq L \int\limits_{t_0}^{t} |y_1(s) - y_0|ds$$
$$\leq LM \int\limits_{t_0}^{t} (s - s_0)ds = \frac{ML}{2}(t - t_0)^2.$$

This justifies the induction assumption

$$|y_n(t) - y_{n-1}(t)| \leq \frac{ML^{n-1}}{n!}(t - t_0)^n$$

and by (7.1.13)

$$|y_{n+1}(t) - y_n(t)| \leq \int\limits_{t_0}^{t} |f(s, y_n(s)) - f(s, y_{n-1}(s))|ds \leq L \int\limits_{t_0}^{t} |y_n(s) - y_{n-1}(s)|ds$$
$$\leq \frac{ML^n}{n!} \int\limits_{t_0}^{t} (s - t_0)^n = \frac{ML^n}{(n+1)!}(t - t_0)^{n+1}.$$

Now, because $|t - t_0| < \alpha$, we see that

$$|y_n(t) - y_{n-1}(t)| \leq \frac{ML^{n-1}}{n!}\alpha^n$$

so that

$$\sum_{n=0}^{\infty} |y_{n+1}(t) - y_n(t)| \leq \sum_{n=1}^{\infty} \frac{ML^{n-1}}{n!}\alpha^n = \frac{M}{L}(e^{\alpha L} - 1)$$

which is finite. Thus, the sequence $y_n(t)$ converges for any t satisfying $|t - t_0| \leq \alpha$. Let us denote the limit by $y(t)$. By (7.1.12) we obtain

$$y(t) = y_0 + (y_1(t) - y_0) + (y_2(t) - y_1(t)) + \ldots + (y_n(t) - y_{n-1}(t)) + \ldots = y_0 + \sum_{n=0}^{\infty} (y_{n+1}(t) - y_n(t)) \quad (7.1.14)$$

and so

$$|y(t) - y_n(t)| = \left| \sum_{j=n}^{\infty} (y_{j+1}(t) - y_j(t)) \right| \leq M \sum_{j=n}^{\infty} L^j \frac{(t - t_0)^{j+1}}{(j+1)!}$$
$$\leq \frac{M}{L} \sum_{j=n}^{\infty} \frac{L^{j+1}\alpha^{j+1}}{(j+1)!} \quad (7.1.15)$$

where the tail of the series is convergent to zero. It is clear that the left hand side does not depend on t as long as $|t - t_0| \leq \alpha$. This fact can be used to show the continuity of the limit function $y(t)$. Firstly, we observe that, by induction, $y_n(t)$ is continuous if $y_{n-1}(t)$ is. In fact, we have

$$|y_n(t_1) - y_n(t_2)| \leq \left| \int\limits_{t_0}^{t_1} f(s, y_n(s))ds - \int\limits_{t_0}^{t_2} f(s, y_n(s))ds \right| \leq \left| \int\limits_{t_1}^{t_2} f(s, y_n(s))ds \right|$$
$$\leq M|t_2 - t_1|.$$

Next, let t_1 and t_2 be arbitrary numbers satisfying $|t_i - t_0| \leq \alpha$ for $i = 1, 2$. By (7.1.15) we can find n so large that

$$\frac{M}{L}\sum_{j=n}^{\infty}\frac{L^{j+1}\alpha^{j+1}}{(j+1)!} < \epsilon/3$$

so that

$$|y(t_i) - y_n(t_i)| < \epsilon/3, \qquad i = 1, 2.$$

Since $y_n(t)$ is continuous, we can find $\delta > 0$ such that if $|t_1 - t_2| < \delta$, then

$$|y_n(t_1) - y_n(t_2)| < \epsilon/3.$$

Combining, we see that whenever $|t_1 - t_2| < \delta$ and $|t_i - t_0| < \alpha$ for $i = 1, 2$, we have

$$|y(t_1) - y(t_2)| \leq |y(t_1) - y_n(t_1)| + |y_n(t_1) - y_n(t_2)| + |y_n(t_2) - y(t_2)| \leq \epsilon,$$

so that $y(t)$ is continuous.

The last step is to prove that the obtained function is indeed the solution of the Cauchy problem in the integral form (7.1.9)

$$y(t) = y_0 + \int_{t_0}^{t} f(s, y(s))ds.$$

Since by construction

$$y_{n+1}(t) = y_0 + \int_{t_0}^{t} f(s, y_n(s))ds$$

we obtain

$$y(t) = \lim_{n\to\infty} y_{n+1}(t) = y_0 + \lim_{n\to\infty}\int_{t_0}^{t} f(s, y_n(s))ds$$

so that we have to prove that

$$\lim_{n\to\infty}\int_{t_0}^{t} f(s, y_n(s))ds = \int_{t_0}^{t} f(s, y(s))ds. \tag{7.1.16}$$

Firstly, note that the right-hand side is well-defined as y is a continuous function, f is continuous so that the composition $f(s, y(s))$ is continuous and the integral is well-defined. Thus, we can write, by (7.1.15),

$$\begin{aligned}\left|\int_{t_0}^{t} f(s, y_n(s))ds - \int_{t_0}^{t} f(s, y(s))ds\right| &\leq \int_{t_0}^{t} |f(s, y_n(s)) - f(s, y(s))|ds \leq L\int_{t_0}^{t} |y_n(s) - y(s)|ds\\ &\leq L\frac{M}{L}\sum_{j=n}^{\infty}\frac{L^{j+1}\alpha^{j+1}}{(j+1)!}\int_{t_0}^{t} ds \leq M\alpha\sum_{j=n}^{\infty}\frac{L^{j+1}\alpha^{j+1}}{(j+1)!}.\end{aligned}$$

As before, the sum above approaches zero as $n \to \infty$ and therefore (7.1.16), and the whole theorem, is proved. ■

We illustrate the use of this theorem on several examples.

Example 7.8. We have seen in Example 7.4 that there are two solutions to the problem

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

In this case $f(t,y) = \sqrt{y}$ *and* $f_y = 1/2\sqrt{y}$*; obviously* f_y *is not continuous in any rectangle* $|t| \leq a$, $|y| \leq b$ *and we may expect troubles.*

Another example of nonuniqueness is offered by

$$\begin{aligned} y' &= (\sin 2t)y^{1/3}, \quad t \geq 0 \\ y(0) &= 0, \end{aligned} \tag{7.1.17}$$

Direct substitution shows that we have 3 different solutions to this problem: $y_1 \equiv 0, y_2 = \sqrt{8/27}\sin^3 t$ *and* $y_3 = -\sqrt{8/27}\sin^3 t$.

Example 7.9. Show that the solution $y(t)$ *of the initial value problem*

$$\begin{aligned} y' &= t^2 + e^{-y^2}, \\ y(0) &= 0, \end{aligned}$$

exists for $0 \leq t \leq 0.5$*, and in this interval,* $|y(t)| \leq 1$.

Let R *be the rectangle* $0 \leq t \leq 0.5, |y| \leq 1$*. The function* $f(t,y) = t^2 + e^{-y^2}$ *is continuous and has continuous derivative* f_y*. We find*

$$M = \max_{(t,y)\in R} |f(t,y)| \leq (1/2)^2 + e^0 = 5/4,$$

thus the solution exists and is unique for

$$0 \leq t \leq \min\{1/2, 5/4\} = 1/2,$$

and of course in this interval $|y(t)| \leq 1$.

Example 7.10. 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* $f_y(t,y) = 2y$ *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

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

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 sometimes does not give the best possible answer - that is why it is sometimes called "the local existence theorem".

Example 7.11. Suppose that $|f(t,y)| \leq K$ *in the whole plane* $\mathbb{R}^2$*. Show that the solution of the initial value problem*

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

where t_0 *and* y_0 *are arbitrary, exists for all* $t \in \mathbb{R}$.

Let R be the rectangle $|t-t_0| \leq a$, $|y-y_0| \leq b$ for some a, b. The quantity M is given by

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

and the quantity

$$|t-t_0| \leq \alpha = \min\{a, \frac{b}{K}\},$$

can be made as large as we wish by choosing a and b sufficiently large. Thus the solution exists for all t.

To be able to extend the class of functions f for which the solution is defined on the whole real line we must introduce the concept of the continuation of the solution.

Remark 7.12. 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 (7.1.2) 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 $[t_0-\alpha, t_0+\alpha+\alpha']$.

Example 7.13. Using such a patchwork technique, global existence can be proved for a larger class of right-hand sides in (7.1.2). Assume that the function f is globally Lipschitz that is

$$|f(t,y_1) - f(t,y_2)| \leq L|y_1 - y_2|$$

for all $t, y_1, y_2 \in \mathbb{R}$, where the constant L is independent of t and y_1, y_2. Let $y(t)$ be the solution passing through (t_0, y_0). It is defined on the interval $|t-t_0| \leq \alpha$ with $\alpha = \min\{a, b/M\}$. Here, a and b can be arbitrarily large as f is defined on the whole $\mathbb{R}^2$. Let as fix $a = 1/(L+1)$. Lipschitz continuity yields

$$|f(t,y) - f(t,y_0)| \leq L|y-y_0| \leq Lb$$

for $(t,y) \in R$. Thus,

$$|f(t,y)| \leq Lb + |f(t,y_0)| \leq Lb + \max_{t_0-1/(L+1)\leq t\leq t_0+1/(L+1)} |f(t,y_0)| = Lb + m(t_0,y_0)$$

so that

$$M \leq Lb + m(t_0,y_0)$$

and

$$\frac{b}{M} \geq \frac{b}{Lb + m(t_0,y_0)} = \frac{1}{L + m(t_0,y_0)/b}.$$

For any fixed t_0, y_0 we can select b large enough so that $m(t_0,y_0)/b \leq 1$ and thus, for such a b

$$\alpha = \frac{1}{L+1}.$$

The solution therefore is defined at least on the interval $[t_0-\alpha, t_0+\alpha]$, where $\alpha = \frac{1}{L+1}$, and the length of the interval of existence is **independent of** t_0 **and** y_0*. Next we shall use the method of small steps. We take $t_{1,0}$, $t_{1,0} = t_0 + 0.9\alpha$ with corresponding $y_{1,0} = y(t_{1,0})$ as a new Cauchy data. By the above there is a solution of this Cauchy problem that is defined on $[t_{1,0}-\alpha, t_{1,0}+\alpha] = [t_0 - 0.1\alpha, t_0 + 0.9\alpha + \alpha]$ and by uniqueness the two solutions coincide on $[t_0 - 0.1\alpha, t_0 + \alpha]$ and therefore by gluing them together we obtain a the solution of the original Cauchy problem defined on $[t_0 - \alpha, t_0 + 0.9\alpha + \alpha]$. Continuing this way we eventually cover the whole real line with solutions as each time we make a step of constant length.*

Note that the crucial rôle here was played by the fact that the numerator and denominator of the fraction b/M grew at the same rate. The procedure described above would be impossible if $f(y)$ grew faster than linearly as $y \to \infty$, like in Example 7.10. There, $b/M = b/(1+b^2)$ and if we enlarge b, then the possible time step will become smaller and there is a possibility that the time steps will sum up to a finite time determining the maximal interval of existence of the solution as in Example 7.10.

Picard's theorem ensures existence for only a bounded interval $|t - t_0| \leq \alpha$, where in general α depends on the initial condition (through the rectangle R). In most applications it is important to determine whether the solution exists for all times, as discussed in the previous two examples. To be able to discuss this question we shall introduce the *maximal interval of existence* of a solution of the differential equation.

Next we present a powerful result allowing to assess whether a local solution to (7.1.2) can be extended to a global one; that is, defined for all t (and also providing an alternative proof of the result in the example above). First we have to introduce new terminology. We say that $[t_0, t_0+\alpha^*)$ is the maximal interval of existence for a solution $y(t)$ to (7.1.2) if there is no solution $y_1(t)$ on a longer time interval $[t_0, t_0 + \alpha^+)$ where $\alpha^+ > \alpha^*$ satisfying $y(t) = y_1(t)$ for $t \in [t_0, t_0+\alpha^*)$. In other words, we cannot extend $y(t)$ beyond $t_0 + \alpha^*$ so that it remains a solution of (7.1.2).

Theorem 7.14. *Assume that f in (7.1.2) satisfies the assumptions of Picard's theorem on $\mathbb{R}^2$. The solution $y(t)$ of (7.1.2) has a finite maximal interval of existence $[t_0, t_0+\alpha^*)$ if and only if*

$$\lim_{t \to t_0+\alpha^*} |y(t)| = \infty. \tag{7.1.18}$$

Proof. Clearly, if (7.1.18) is satisfied, then $y(t)$ cannot be extended beyond $t_0 + \alpha^*$. On the other hand, assume that (7.1.18) does not hold. Let us reflect what it means. The meaning of (7.1.18) is that for any K there is t_K such that for any $t_0 + \alpha^* > t \geq t_K$ we have $|y(t)| \geq K$. Thus, by saying that (7.1.18) does not hold, we mean that there is K such that for any $t < t_0 + \alpha^*$ there is $t < t' < t_0 + \alpha^*$ with $|y(t')| < K$. In particular, there is a sequence $(t_n)_{n\in\mathbb{N}}$ such that $t_n \to t_0 + \alpha^*$ we have $|y_n| := |y(t_n)| < K$. Consider Cauchy problems

$$y' = f(t,y), \quad y(t_n) = y_n. \tag{7.1.19}$$

Since $|y_n|$ are bounded by K and f satisfies the conditions of the Picard theorem on $\mathbb{R}^2$, we can consider the above problem in rectangles $R_n = \{(t,y);\ |t-t_n| < a, |y - y_n| < b\}$ for some fixed a, b. Moreover, all R_ns are contained in the rectangle $R = \{(t,y);\ t_0 - a < t < t_0 + \alpha^*, -K - b < y < K + b\}$ and the solutions of the problems (7.1.19) are defined on intervals $(t_n - \alpha, t_n + \alpha)$ where $\alpha = \min\{a, b/M\}$ and M can be taken as $\max_{(t,y)\in R} |f(t,y)|$ and is independent of n. If α^* was finite, then we could find t_n with $t_0 + \alpha^* - t_n < \alpha$ so that the solution could be continued beyond $t_0 + \alpha^*$ contradicting the assumption that $[t_0, t_0 + \alpha^*)$ is the maximal interval of existence. □

Example 7.15. This result allows to give another proof of the fact that solutions of (7.1.2) with globally Lipschitz right-hand side are defined on the whole line. In fact, using Gronwall's lemma, we obtain

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

If $y(t)$ is not defined for all t, then by the previous remark, $|y(t)|$ becomes unbounded as $t \to t_{max}$ for some t_{max}. Denoting

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

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

$$|y(t)| \le c + L\int\limits_{t_0}^{t} |y(s)|ds$$

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

$$|y(t)| \le c \exp Lt \le c \exp Lt_{max}$$

contradicting thus the definition of t_{max}.

2 Appendix B: Cayley-Hamilton theorem and dimension of generalized eigenspaces

Let us recall that the characteristic polynomial of a square $d \times d$ matrix $\mathcal{A}$ is defined as

$$p_{\mathcal{A}}(\lambda) = det(\mathcal{A} - \lambda\mathcal{I}) = \begin{vmatrix} a_{11}-\lambda \dots & a_{1d} \\ \vdots & \vdots \\ a_{1d} & \dots a_{nn}-\lambda \end{vmatrix}. \tag{7.2.20}$$

Evaluating the determinant we obtain that

$$p_{\mathcal{A}}(\lambda) = p_0 + p_1\lambda + \ldots + (-1)^d\lambda^d$$

is a polynomial in λ of degree d. Then

Theorem 7.16. *(Cayley-Hamilton). We have*

$$p_{\mathcal{A}}(\mathcal{A}) = p_0\mathcal{I} + p_1\mathcal{A} + \ldots + (-1)^d\mathcal{A}^d = \mathbf{0}. \tag{7.2.21}$$

Proof. Let $\lambda \notin \sigma(\mathcal{A})$, that is, λ is not an eigenvalue of $\mathcal{A}$. Then there exists an inverse $(\lambda\mathcal{I} - \mathcal{A})^{-1}$ which can be expressed as

$$(\lambda\mathcal{I} - \mathcal{A})^{-1} = \frac{\mathcal{C}(\lambda)}{\det(\lambda\mathcal{I} - \mathcal{A})}$$

where $C(\lambda)$ is the matrix composed of the minors of $(\lambda\mathcal{I}-\mathcal{A})$, that is, determinants of the matrices obtained by deleting the i-th row and the j-th column of $\lambda\mathcal{I}-\mathcal{A}$, multiplied by $(-1)^{i+j}$. Thus, the entries of $\mathcal{C}(\lambda)$ are polynomials of degree $d-1$ and therefore

$$\mathcal{C}(\lambda)=\lambda^{d-1}\mathcal{C}_{d-1}+\ldots+\mathcal{C}_0,$$

where $\mathcal{C}_i$, $i=1,\ldots,d-1$, are $d\times d$ matrices. In other words, we can write

$$(\lambda\mathcal{I}-\mathcal{A})(\lambda^{d-1}\mathcal{C}_{d-1}+\ldots+\mathcal{C}_0)=p_0\mathcal{I}+p_1\lambda\mathcal{I}+\ldots+(-1)^d\lambda^d\mathcal{I}. \tag{7.2.22}$$

Formally, if we replace λ by $\mathcal{A}$ on both sides of (7.2.22), we will get the statement of the theorem. However, it is not immediately clear that we can extend the equality from $\mathbb{C}$ to the space of matrices so we argue as follows. Since we have polynomials on both sides of (7.2.22), the (matrix) coefficients of like powers of λ on both sides must be equal. Thus we obtain

$$\begin{aligned}-\mathcal{A}\mathcal{C}_0&=p_0\mathcal{I},\\ \mathcal{C}_0-\mathcal{A}\mathcal{C}_1&=p_1\mathcal{I},\\ \vdots\ \vdots\ &\ddots,\\ \mathcal{C}_{d-1}&=(-1)^d\mathcal{I}.\end{aligned}$$

Now, if we multiply the first row of the above table by $\mathcal{A}^0$, the second by $\mathcal{A}^1$, and so on, with the last multiplied by $\mathcal{A}^d$, and add them together, we will get

$$\begin{aligned}\mathbf{0}&=-\mathcal{A}\mathcal{C}_0+\mathcal{A}\mathcal{C}_0-\mathcal{A}^2\mathcal{C}_1+\ldots+\mathcal{A}^d\mathcal{C}_{d-1}=(\mathcal{A}-\mathcal{A})(\mathcal{A}^{d-1}\mathcal{C}_{d-1}+\ldots+\mathcal{C}_0)\\ &=p_0\mathcal{I}+p_1\mathcal{A}+\ldots+(-1)^d\mathcal{A}^d,\end{aligned}$$

which ends the proof. □

The main aim of the second part of this appendix is to prove the following result.

Theorem 7.17. *Let X be a complex, d-dimensional vector space and let $A : X\to X$ be a linear operator. Then X is the direct sum of the generalized eigenspaces of A and the dimension of each generalized eigenspace is equal to the algebraic multiplicity of the corresponding eigenvalue.*

Proof. Consider an operator $T : X\to X$ and, for $j\in\mathbb{N}_0=\mathbb{N}\cup\{0\}$, define the subspaces

$$\begin{aligned}&N_j(T)=Ker\,T^j=\{\mathbf{x}\in X;\ T^j\mathbf{x}=0\},\quad N=\bigcup_{j\in\mathbb{N}_0}N_j(T),\\ &M_j(T)=Im\,T^j=\{\mathbf{y}\in X;\ T^j\mathbf{x}=\mathbf{y}\text{ for some }\mathbf{x}\in X\},\quad M=\bigcap_{j\in\mathbb{N}_0}M_j(T).\end{aligned}$$

It is easy to see that both families of sets are nested:

$$\begin{gathered}\mathbf{0}=N_0(T)\subseteq N_1\subseteq\ldots\subseteq N_j\subseteq\ldots\subseteq N;\\ X=M_0(T)\supseteq M_1\supseteq\ldots\supseteq M_j\supseteq\ldots\supseteq M.\end{gathered}$$

We observe that whenever $N_j\neq N_{j+1}$, then there is an element $\mathbf{z}\in N_{j+1}$ which is linearly independent of N_j. Indeed, otherwise we would have

$$\mathbf{z}=\sum_k\alpha_k\mathbf{x}_k,\quad \alpha_k\in\mathbb{C},\mathbf{x}_k\in N_j$$

and, by linearity of T, and thus T^j,

$$T^j\mathbf{z}=\sum_k\alpha_kT^j\mathbf{x}_k=\mathbf{0},$$

so $\mathbf{z} \in N_j(T)$. Similarly, if $M_j \neq M_{j+1}$, then there is an element $\mathbf{y} \in M_j$ which is linearly independent of M_{j+1}. Otherwise we would have

$$\mathbf{y} = \sum_k \beta_k \mathbf{y}_k, \quad \beta_k \in \mathbb{C}, \mathbf{y}_k \in M_{j+1}$$

and, by linearity of T, and thus T^j, for some $\mathbf{z}_k \in X$,

$$\mathbf{y} = \sum_k \beta_k T^{j+1}\mathbf{z}_k = T^{j+1}\left(\sum_k \beta_k \mathbf{z}_k\right),$$

so $\mathbf{y} \in M_{j+1}(T)$. This shows that there are r, m such that

$$\begin{aligned} N_j(T) &= N_r(T), \quad j \geq r, \\ M_j(T) &= N_m(T), \quad j \geq m. \end{aligned}$$

This follows since X is finite dimensional and, as we proved above, if we had infinitely many different N_js or M_js, then we would have infinitely many linearly independent elements in X. Consequently,

$$N(T) = N_r(T), \qquad M(T) = M_m(T)$$

We observe that both $N(T)$ and $M(T)$ are invariant under T. Indeed, from the definition, it follows that if $T^j\mathbf{x} = \mathbf{0}$ for $j > r$, then there exists $i \leq r$ such that $T^i\mathbf{x} = \mathbf{0}$. Thus, if $\mathbf{x} \in N(T)$, then $T^j\mathbf{x} = \mathbf{0}$ for some $0 \leq j \leq r$. If $j = 0$, then $\mathbf{x} = \mathbf{0}$ and $\mathbf{0} = T\mathbf{x} \in N(T)$. If $j \geq 1$, then we can write $\mathbf{0} = T^j\mathbf{x} = T^{j-1}T\mathbf{x}$ and thus $T\mathbf{x} \in N(T)$. Similarly, $if \mathbf{y} \in M(T)$, then $\mathbf{y} = T^j\mathbf{x}_j$, $x_j \in X$, $j = 0, \ldots, m$ and $T\mathbf{y} = T^{j+1}\mathbf{x}_j$. If $j = 0, \ldots, m-1$, and $T^{j+1}\mathbf{x}_j \in M(T)$ and $T^{m+1}\mathbf{x}_m \in M(T)$ as $M_{m+1}(T) = M_m(T)$. Furthermore, $T|_{N(T)}$ is a nilpotent operator, that is, $T^r\mathbf{x} = 0$ for any $\mathbf{x} \in N(T)$. Then we have

$$X = N(T) \oplus M(T).$$

Indeed, since $TM(T) = M(T)$, T is surjective onto $M(T)$. It is a standard result in the theory of systems of linear equations that if a $k \times k$ equation always has a solution, then this solution is unique. This can be ascertained by considering a matrix representation of $T|_{M(T)}$. Then the fact that the system always has a solution means that the columns of the matrix span the whole space (and the solution is formed of the coefficient of the expansion of the right hand side in this basis). But sice these are k vectors in a k-dimensional space, the columns must form a basis, so that the solution is unique. Thus $T|_{T(M)}$ is invertible. Then, since $T^rM(T) = M(T)$, $T^r|_{M(T)}$ is also invertible and thus $T^r\mathbf{x} \neq \mathbf{0}$ for any $M(T) \ni \mathbf{x} \neq \mathbf{0}$. Thus $M(T) \cap N(T) = \{\mathbf{0}\}$. On the other hand, for any $\mathbf{v} \in X$, we have $T^m\mathbf{v} =: \mathbf{y} \in M(T)$. Since $T^m|_{M(T)}$ is invertible, we have $T^m\mathbf{v} = T^m\mathbf{z}$ for some $\mathbf{z} \in M(T)$. Writing $\mathbf{v} = \mathbf{z} + (\mathbf{v} - \mathbf{z})$ gives the desired decomposition as $T^m(\mathbf{v} - \mathbf{z}) = \mathbf{0}$ yields $\mathbf{v} - \mathbf{z} \in N(T)$.

To proceed, let $\lambda_1, \ldots, \lambda_q$ be the distinct eigenvalues of T. We define

$$\begin{aligned} N_k &:= N(\lambda_k I - T) = \bigcup_{i \geq 0} Ker(\lambda_k I - T)^j = \bigcup_{i=0}^{r_k} Ker(\lambda_k I - T)^j, \\ M_k &:= M(\lambda_k I - T) = \bigcap_{j \geq 0} Im(\lambda_k I - T)^j = \bigcap_{j=0}^{m_k} Im(\lambda_k I - T)^j, \end{aligned}$$

for some r_k, m_k, $k = 1, \ldots, q$. As before, these space are invariant under T. Indeed, since T commutes with $\lambda_k I - T$, if $(\lambda_k I - T)^j\mathbf{x} = 0$ for some j, then $(\lambda_k I - T)^j T\mathbf{x} = T(\lambda_k I - T)^j\mathbf{x} = \mathbf{0}$. Similarly, if $\mathbf{y} \in M_k$, then $\mathbf{y} = (\lambda_k I - T)^j\mathbf{x}_j$ for any j and thus $T\mathbf{y} = T(\lambda_k I - T)^j\mathbf{x}_j = (\lambda_k I - T)^j T\mathbf{x}_j$.

We want to prove

$$X = N_1 \oplus \ldots \oplus N_q. \tag{7.2.23}$$

From the previous part,

$$X = N_1 \oplus M_1.$$

The next part of the proof is inductive, with induction on the dimension d of the space X. The dimension $d = 1$ is trivial and $d = 1$ is also obvious with $M_1 = \{\mathbf{0}\}$. Let $d > 1$ and assume that the result is valid for any space with smaller dimension. In particular, the decomposition holds for M_1 and $T|_{M_1}$. Therefore it is enough to show that

$$N((\lambda_k I - T)|_{M_1}) = N(\lambda_k I - T) = N_k \tag{7.2.24}$$

for $k > 1$.

We begin with showing that

$$N_k \cap Ker(\lambda_k I - T) = \{\mathbf{0}\}. \tag{7.2.25}$$

Indeed, let $\mathbf{x} \neq \mathbf{0}$ satisfies $(\lambda_1 I - T)\mathbf{x} = 0$. Then

$$(\lambda_k I - T)\mathbf{x} = (\lambda_k - \lambda_1)\mathbf{x}$$

and

$$(\lambda_k I - T)^j \mathbf{x} = (\lambda_k - \lambda_1)^j \mathbf{x} \neq \mathbf{0}$$

for any j and hence $\mathbf{x} \notin N_k$. As under T, N_k is invariant under $\lambda_1 I - T$, that is

$$(\lambda_1 - T)N_k \subset N_k.$$

Since by (7.2.25), $\lambda_1 I - T$ is one-to-one on N_k, we must have

$$(\lambda_1 - T)N_k = N_k$$

for any k. This shows that $(\lambda_1 - T)^j N_k = N_k$ for any $j \geq 0$ so that $N_k \subset M_1$ for $k > 1$. This shows that $\lambda_2, \ldots, \lambda_q$ are eigenvalues of $T|_{M_2}$ and, since eigenvalues of the latter must be eigenvalues of T, the spectrum of $T|_{M_2}$ consists of $\lambda_2, \ldots, \lambda_q$. Since it is clear that $N_k((\lambda_k I - T)|_{M_1}) \subset N_k$, we obtain (7.2.24) and (7.2.23) is proved.

To complete the proof of the theorem, we have to prove that $N_k = Ker(\lambda_k I - T)^{n_k} = E_{\lambda_k}$ where n_k is the algebraic multiplicity of λ_k and E_{λ_k} is the generalized (associated) eigenspace corresponding to λ_k. First we observe that due to (7.2.23) the characteristic polynomial $p_T(\lambda)$ factorizes as

$$p_T(\lambda) = p_{T|_{N_1}}(\lambda) \cdot \ldots \cdot p_{T|_{N_q}}(\lambda).$$

This is due to the fact that by change of coordinates T can be transformed to a block diagonal form where each block at the diagonal corresponds to one of the operators $T|_{N_k}$ and the determinant does not change under similarity transform. On the other hand, λ_k is the only eigenvalue of $T|_{N_k}$ so $p_{T|_{N_k}}(\lambda) = (\lambda_k - \lambda)^{\dim N_k}$. Hence, $\dim N_k = n_k$. Since clearly $E_{\lambda_k} \subseteq N_k$, we need to prove the inverse inclusion. Let $N_k = Ker(\lambda_k I - T)^{r_k}$, then there is $\mathbf{x} \in N_k$ such that $(\lambda_k I - T)^{r_k}\mathbf{x} = \mathbf{0}$ and $(\lambda_k I - T)^{r_k - 1}\mathbf{x} \neq 0$. However, by the Cayley-Hamilton theorem, $(\lambda_k I - T)^{n_k}\mathbf{x} = \mathbf{0}$ for all $\mathbf{x} \in N_k$. Thus $r_k - 1 < n_k$ or $r_k \leq n_k$. Thus $N_k \subseteq E_{\lambda_k}$. □

3 Appendix C: The Perron-Frobenius theorem

References

1. J. Banasiak and M. Lachowicz, *Methods of small parameter in mathematical biology and other applied sciences*, in preparation.
2. J. Banasiak, *Mathematical Modelling in One Dimension. An Introduction via Difference and Differential Equations*, Cambridge University Press, Cambridge, 2013.
3. Å. Brännström, D.J.T Sumpter, The role of competition and clustering in population dynamics, *Proc. R. Soc. B*, **272**, (2005), 2065-2072.
4. F. Brauer and C. Castillo-Chávez, *Mathematical Models in Population Biology and Epidemiology*, Texts in Appl. Math. 40, Springer, New York, 2001.
5. M. Braun, *Differential Equations with Applications*, 3 wyd., Springer, New York, 1983.
6. N. F. Britton, *Essential Mathematical Biology*, Springer, London, 2003.
7. R. Courant, F. John, *Introduction to Calculus and Analysis I*, Springer, Berlin, 1999.
8. R. S. Cantrell and C. Cosner, *Spatial Ecology via Reaction-Diffusion Equations*, Wiley, Chichester, 2003.
9. J. M. Cushing, *An Introduction to Structured Population Dynamics*, Conference Series in Applied Mathematics 71, SIAM, Philadelphia, 1998.
10. J. Cronin, *Ordinary Differential Equations*, 3rd ed., Chapman & Hall/CRC, Boca Raton, 2008
11. S. Elaydi, *Introduction to Difference Equations*, 3 wyd., Springer, New York, 2003.
12. F. R. Gantmacher, *Applications of the theory of matrices*, Interscience, New York, 1959.
13. P. Glendinning, *Stability, instability, and chaos: an introduction to the theory of nonlinear differential equations*, Cambridge University Press, Cambridge, 1994.
14. M. Hirsch and S. Smale, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, San Diego, 1974.
15. M. Iannelli, *Mathematical Theory of Age-Structured Population Dynamics,* C.N.R. Giardini, Pisa, 1994.
16. M. Kot, *Elements of Mathematical Ecology*, Cambridge University Press, Cambridge, 2003.
17. K. Lorenz, *The Natchez of Southwest Mississippi.* In: B. G. McEwan (Ed.) Indians of the Greater Southeast: Historical Archaeology and Ethnohistory, 142–177, University Press of Florida, Gainesville, 2000.
18. D. G. Luenberger, *Introduction to Dynamic Systems. Theory, Models, and Applications,* Wiley, New York, 1979.
19. C. D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, 2000.
20. T. Royama, *Analytical population dynamics*, Chapman& Hall, London, 1992.

21. J. L. Schiff, *The Laplace Transformation: Theory and Applications*, Springer, New York, 1999.

22. S. H. Strogatz, *Nonlinear Dynamics and Chaos*, Addison-Wesley, Reading, 1994.

23. J.R. Swanton, Indian Tribes of the Lower Mississippi Valley and Adjacent Coast of the Gulf of Mexico, *Bureau of American Ethnology Bulletin 43. Smithsonian Institution, Washington, D.C.*, 1911.

24. H.R. Thieme, *Mathematics in Population Biology*, Princeton University Press, Princeton and Oxford, 2003.

Index