Notes on Mathematical Models in Biology

J. Banasiak

University of KwaZulu-Natal, Durban

Contents

Chapter 1

Single species unstructured models

1.1 Models leading to single difference equations

1.1.1 Background

Many plants and animals breed only during a short, well-defined, breeding season.

(i) Monocarpic plants – flower once and the die. May be annual plants. Bamboos grow vegetatively for 20 years and then flower and die.

(ii) Semelparity.

Insects: die after lying eggs. Day-flies, cicads have 13 or 17 years cycles.

Fish: pacific salmon, European eel (lives 10-15 years in freshwater European lakes, migrates to Sargasso Sea, spawns and dies).

Birds: Greater Snow Geese, egg-lying between 8-20 June (peaks 12-17 June), practically all hatchings occur between 8-13 July.

Mammals: some marsupials ovulate once per year and produce a single litter. There occurs abrupt and total mortality of males after mating. Births are synchronised to within a day or two in population - related to the environment with predictable 'bloom' of insects in a given season of the year.

We have

- Cicads – non-overlapping generations – single difference equation

- Snow Geese - overlapping generations – age structured model to be discussed later.

1.1.2 Insects and other species

We focus on insect-type populations. Insects often have well-defined annual non-overlapping generations - adults lay eggs in spring/summer and then die. The eggs hatch into larvae which eat and grow and then overwinter in a pupal stage. The adults emerge from the pupae in spring. We take the census of adults in the breeding seasons. It is then natural to describe the population as the sequence of numbers

$$N_0, N_1, \ldots, N_k$$

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

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

$$N_{n+1} = f(N_n), \quad n = 0, 1, \ldots \tag{1.1.1}$$

Let us introduce the number R_0, which is the average number of eggs laid by an adult. R_0 is called the *basic reproductive ratio* or *intrinsic growth rate*. The simplest functional dependence in (1.1.1) is

$$N_{n+1} = R_0 N_n, \quad n = 0, 1, \ldots \tag{1.1.2}$$

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

Remark 1.1.1. The exponential (or Malthusian) equation (1.1.2) has a much larger range of applications. Even in the population theory, the generations can overlap. Looking at large populations in which individuals give birth to new offspring but also die after some time, we can treat population as a whole and assume that the population growth is governed by the average behaviour of its individual members. Thus, we make the following assumptions:

- Each member of the population produces in average the same number of offspring.
- Each member has an equal chance of dying (or surviving) before the next breeding season.
- The ratio of females to males remains the same in each breeding season

We also assume

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

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

(a) the number of individuals born in a particular breeding season is directly proportional to the population at the start of the breeding season, and

(b) the number of individuals who have died during the interval between the end of consecutive breeding seasons is directly proportional to the population at the start of the breeding season.

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

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

that is

$$N_{k+1} = (1 + \beta - \mu)N_k. \tag{1.1.3}$$

This equation reduces to (1.1.2) by putting $\mu = 1$ (so that the whole adult population dies) and $\beta = R_0$.

Equation (1.1.2) is easily solvable yielding

$$N_k = R_0^k N_0, \qquad k = 0, 1, 2 \ldots \tag{1.1.4}$$

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

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

$$N_{k+1} = R_0 S(N_k) N_k, \quad k = 0, 1, \ldots \tag{1.1.5}$$

which may be alternatively written as

$$N_{k+1} = F(N_k)N_k = f(N_k), \quad k = 0, 1, \ldots \tag{1.1.6}$$

where $F(N)$ is per capita production of a population of size N. Such models, with density dependent growth rate, lead to nonlinear equations. However, before introducing basic examples, we discuss typical types of behavior in such populations.

The survival rate S 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

- *No competition*: then $S(N) = 1$ for all N.
- *Contest competition*: here there is a finite number of units of resource. Each individual which obtains one of these units survives to breed, and produces R_0 offspring in the subsequent generations; all others die without producing offspring. Thus $S(N) = 1$ for $N \leq N_c$ and $S(N) = N_c/N$ for $N > N_c$ for some critical value N_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 survive and produce R_0 offspring in the next generation; if not, all die. Thus, $S(N) = 1$ for $N \leq N_c$ and $S(N) = 0$ if $N > N_c$ for a critical value N_c (different from the above).

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

1. Contest competition corresponds to *exact compensation*:

   $$\lim_{N\to\infty} f(N) = c \qquad (1.1.7)$$

   for some constant c (or $S(N) \sim cN^{-1}$ for large N). This describes the situation if the increased mortality compensates exactly any increase in numbers.

2. The other case is when

   $$S(N) \sim c/N^b, \quad N \to \infty. \qquad (1.1.8)$$

   Here we have

   Under-compensation if $0 < b < 1$ 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 contest, and scramble if b is large. Indeed, in the first case, $f(N)$ 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 $b > 1$ $f(N)$ tends to zero for large populations which indicates that the resources are over-utilized leading to eventual extinction.

We introduce most typical nonlinear models.

Beverton-Holt type models.
Let us look at the model (1.1.6)

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

where $F(N_k) = R_0 S(N_k)$. To exhibit compensatory behaviour, we should have $NS(N) \approx$ *const*. Also, for small N, $S(N)$ should be approximately 1 as we expect very small intra-species competition and thus the growth should be exponential with the growth rate R_0. A simple function of this form is

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

leading to

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

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

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

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

$$N_{k+1} = \frac{R_0 N_k}{1 + \frac{R_0 - 1}{K} N_k}. \tag{1.1.9}$$

As we said earlier, this model is compensatory.

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

$$N_{k+1} = \frac{R_0 N_k}{(1 + aN_k)^b}. \tag{1.1.10}$$

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 $x_k = aN_k$ reduces the number of parameters giving

$$x_{k+1} = \frac{R_0 x_k}{(1 + x_k)^b} \tag{1.1.11}$$

which will be analysed later.

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

$$N_{k+1} = N_k + R(N_k)N_k, \tag{1.1.12}$$

so that the increase in the population is given by $R(N) = R_0S(N)N$. Here we assume that no adults die (death can be incorporated by introducing factor $d < 1$ in front of the first N_k.

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

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

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

Constants R_0 and K are usually determined experimentally.

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

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

Substituting this formula into (1.1.12) yields the so-called discrete logistic equation

$$N_{k+1} = N_k + R_0N_k\left(1 - \frac{N_k}{K}\right), \tag{1.1.13}$$

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

While the above arguments may seem to be of *bunny-out-of-the-hat* type it could be justified by generalizing (1.1.3). Indeed, assume that the mortality β is not constant but equals

$$\beta = \mu_0 + \mu_1 N^\theta,$$

where μ_0 corresponds to death of natural caused and μ_1 could be attributed to cannibalism where one adult eats/kills on average μ_1 portion of the population. Then (1.1.3) can be written as

$$N_{k+1} = (1+\beta-\mu_0)N_k\left(1-\frac{N_k}{\frac{1+\beta-\mu_0}{\mu_1}}\right) \tag{1.1.14}$$

which is (1.1.13) with $R_0 = \beta-\mu_0$ and $K = 1+\beta-\mu_0/\mu_1$. A generalization of this equation, called the *Bernoulli equation* is

$$N_{k+1} = N_k + R_0N_k\left(1-\left(\frac{N_k}{K}\right)^{\theta}\right), \tag{1.1.15}$$

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

$$N_{k+1} = R_0N_k\left(1-\frac{N_k}{K}\right). \tag{1.1.16}$$

By substitution

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

we can reduce (1.1.13) to a simpler form

$$x_{n+1} = \mu x_n(1-x_n) \tag{1.1.17}$$

We observe that the logistic equation, especially with S given by (1.1.18) is an extreme example of the scramble competition.

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

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

$$S(N_k) = \left(1-\frac{N_k}{K}\right)_+. \tag{1.1.18}$$

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.

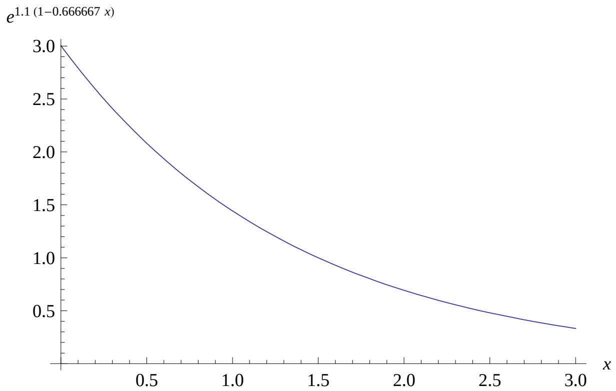

Figure 1.1: The function $f(x) = e^{r(1-x/K)}$

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

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

First we note that it is impossible for f to be less than -1 - this would mean that an individual could die more than once. We also need a decreasing f which is non-zero ($= R_0$) at 0. One such function can be recovered from the Beverton-Holt model, another simple choice is an exponential shifted down by 1:

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

which leads to

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

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

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

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

$$N_{k+1} = N_k e^{r(1-\frac{N_k}{K})}. \qquad (1.1.19)$$

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

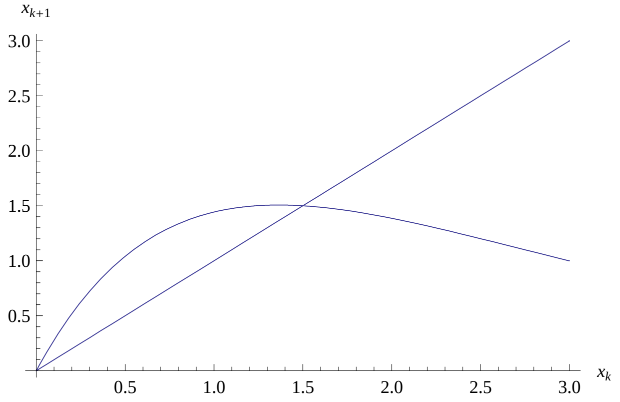

Figure 1.2: The relation $x_{n+1} = x_n e^{r(1-x_n/K)}$

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 $\Delta N/N < 0$ for $0 < N < L$ and $N > K$ and $\Delta N/N > 0$ for $L < N < K$. If

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

then the resulting difference equation is

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

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

$$N_{k+1} = N_k(1 + (L - N_k)(N_k - K). \qquad (1.1.20)$$

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

$$N_{k+1} = N_k \left(1 + \lambda \left(1 - \frac{N_k}{K} - \frac{A}{1 + BN_k}\right)\right) \qquad (1.1.21)$$

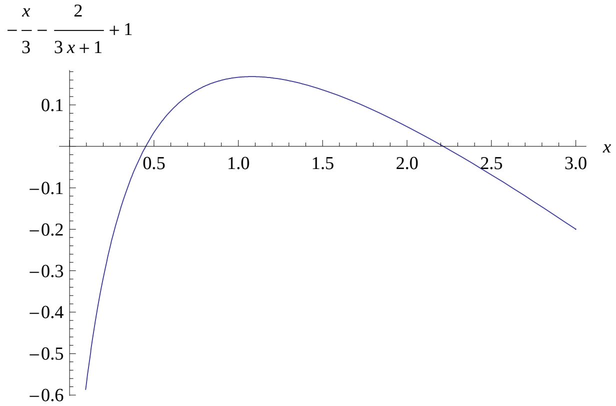

Figure 1.3: The function $1 - \frac{N_k}{K} - \frac{A}{1+BN_k}$

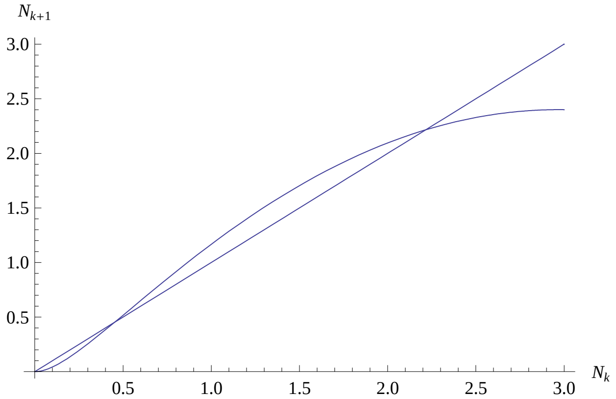

Figure 1.4: The relation $N_{k+1} = N_k + N_k f(N_k)$

where $\lambda > 0$ and

$$1 < A < \frac{(BK+1)^2}{4KB}, \qquad BK > 1. \tag{1.1.22}$$

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

1.2 Interlude: Continuous in time single species unstructured models I

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 $N(t)$ 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

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

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

$$N(t+\Delta t) - N(t) = \beta(t, N(t))N(t)\Delta t - \mu(t, N(t))N(t)\Delta t. \tag{1.2.1}$$

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

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

If we fix Δt and take it as a unit time interval and drop the dependence on t, then the above equation is exactly (1.1.6) with $F(N) = 1 + r(N)$. 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 (1.1.6):

$$\frac{dN}{dt} = r(t, N)N. \tag{1.2.2}$$

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

Exponential growth

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

$$\frac{dN}{dt} = rN. \tag{1.2.3}$$

which has a general solution given by

$$N(t) = N(t_0)e^{r(t-t_0)}, \tag{1.2.4}$$

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

To be able to give some numerical illustration to this equation we need the coefficient r and the population at some time t_0. We use the data of the U.S. Department of Commerce: it was estimated that the Earth population in 1965 was 3.34 billion and that the population was increasing at an average rate of 2% per year during the decade 1960-1970. Thus $N(t_0) = N(1965) = 3.34 \times 10^9$ with $r = 0.02$, and (1.2.4) takes the form

$$N(t) = 3.34 \times 10^9 e^{0.02(t-1965)}. \tag{1.2.5}$$

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

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

thus

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

and

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

This is an excellent agreement with the present observed value of the Earth population and also gives a good agreement with the observed data if we don't go too far into the past. On the other hand, if we try to extrapolate this model into a distant future, then we see that, say, in the year 2515, the population will reach 199980 $\approx$ 200000 billion. To realize what it means, let us recall that the Earth total surface area 167400 billion square meters,

Fig 1.1. Comparison of actual population figures (points) with those obtained from equation (1.2.5).

80% of which is covered by water, thus we have only 3380 billion square meters to our disposal and there will be only $0.16m^2$ ($40cm \times 40cm$) per person. Therefore we can only hope that this model is not valid for all times. 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.

Logistic equation

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

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K}\right), \tag{1.2.6}$$

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 (1.2.6) by taking in (1.2.1) constant birth rate β but introduce density dependent mortality rate

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

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

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

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

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

which is another form of (1.2.6).

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

$$\frac{dN}{dt} = (\beta - \mu_0)N - \mu_1 N^{\theta+1} \tag{1.2.7}$$

Let us focus on the logistic equation (1.2.6). 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 right-hand side vanishes for $N = 0$ and $N = K$ so that $N(t) = 0$ and $N(t) = K$ are equilibria. We shall focus on solutions with the initial condition $N(t_0) > 0$. Then, if $N(t_0) < K$, then $N(t)$ stays between 0 and K, exists for all $t > t_0$ and is an increasing function converging to K as $t \to \infty$. If $N(t_0) > K$, then the solution decreases, exists for all $t > 0$ and also tends to K as $t \to \infty$.

Hence, let us proceed with solving the related Cauchy problem

$$\begin{aligned} \frac{dN}{dt} &= rN\left(1 - \frac{N}{K}\right), \\ N(t_0) &= N_0 \end{aligned} \tag{1.2.8}$$

Separating variables and integrating we obtain

$$\frac{K}{r}\int\limits_{N_0}^{N} \frac{ds}{(K-s)s} = t - t_0.$$

To integrate the left-hand side we use partial fractions

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

which gives

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

From the considerations preceding (1.2.8), if $N_0 < K$, then $N(t) < K$ for any t, and if $N_0 > K$, then $N(t) > K$ for all $t > 0$. Therefore $(K-N_0)/(K-N(t))$ is always positive and

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

Exponentiating, we get

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

or

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

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

$$N(t)\left(N_0e^{r(t-t_0)}+K-N_0\right) = N_0Ke^{r(t-t_0)},$$

thus finally

$$N(t) = \frac{N_0K}{N_0+(K-N_0)e^{-r(t-t_0)}}. \tag{1.2.9}$$

Let us examine (1.2.9) 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

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

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

$$\frac{dN}{dt} = \frac{rN_0K(K-N_0)e^{-r(t-t_0)}}{(N_0+(K-N_0)e^{-r(t-t_0)})^2}$$

thus, if $N_0 < K$, 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

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

from which it follows that, if we start from $N_0 < K$, then the population curve is convex down for $N < K/2$ and convex up for $N > K/2$. Thus, as long as the population is small (less 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:

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

Fig 2.3 Logistic curves with $N_0 < K$ (dashed line) and $N_0 > K$ (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

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

We use this function with the value $N_0 = 3.34 \times 10^9$ at $t_0 = 1965$. The comparison is shown on Fig. 2.4.

From discrete to continuous models and back.

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

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

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.

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,

$$x((n+1)\Delta t, x_0) = x(\Delta t, x(n\Delta t, x_0)). \tag{1.2.10}$$

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

$$N' = rN, \qquad N(0) = N_0$$

having the solution

$$N(t) = N_0 e^{rt}.$$

The first discretization gives

$$N_{k+1} - N_k = rN_k$$

with the solution

$$N_k = (1+r)^k N_k$$

and this discretization gives perfect agreement with the discrete model (and $r = 1 + R_0$).

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

$$N_k = N(k) = e^{rk} N_0 = (e^r)^k N_0.$$

Comparing this equation with (1.1.4), we see that it corresponds to the discrete model with intrinsic growth rate

$$1 + R_0 = e^r.$$

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 $r = \ln(1 + R_0)$. However, the qualitative features are preserved.

Logistic growth

Consider now the logistic equation

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

The first type of discretization immediately produces the discrete logistic equation (1.1.13)

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

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

To use the time-one map discretization, we re-write (1.2.9) as

$$N(t) = \frac{N_0 e^{rt}}{1 + \frac{e^{rt}-1}{K}N_0}.$$

which, upon denoting $e^r = R_0$ gives the time-one map

$$N(1, N_0) = \frac{N_0 R_0}{1 + \frac{R_0-1}{K}N_0},$$

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

$$N_{k+1} = \frac{N_k R_0}{1 + \frac{R_0-1}{K}N_k},$$

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

1.2.1 Discrete models of seasonally changing population

So far we have considered models in which laws of nature are independent of time. In many real processes we have to take into account phenomena which depend on time such as seasons of the year. The starting point of modelling is as before the balance equation. If we denote by $B(t), D(t), E(t)$ and $I(t)$ rates of birth, death, emigration and immigration, so that e.g, the number of births in time interval $[t_1, t_2]$ equals $\int\limits_{t_1}^{t_2} B(s)ds$. Then, the change in the size of the population in this interval is

$$N(t_2) - N(t_1) = \int\limits_{t_1}^{t_2} (B(s) - D(s) + I(s) - E(s))ds,$$

or, in differential form

$$\frac{dN(t)}{dt} = B(t) - D(s) + I(t) - E(t).$$

Processes of birth, death and emigration are often proportional to the size of the population and thus it makes sense to introduce *per capita* coefficients so that $B(t) = b(t)N(t), D(t) = d(t)N(t), E(t) = e(t)N(t)$. Typically, it would be unreasonable to assume that immigration is proportional to the number of the target population (possibly rather to the inverse unless we consider processes like gold rush), so that we leave $I(t)$ unchanged and thus write the rate equation as

$$\frac{dN(t)}{dt} = (b(t) - d(s) + e(t))N(t) + I(t). \tag{1.2.11}$$

This equation provides good description of small populations in which birth and death coefficients are not influenced by the size of the population.

Our interest is in populations in which the coefficients change periodically e.g. with seasons of the year. We start with closed populations; that is we do not consider emigration and immigration. Then we define $\lambda(t) = b(t) - d(t)$ to be the net growth rate of the population and assume that it is a periodic function with period T. Under this assumption we introduce the average growth rate of the population by

$$\bar{\lambda} = \frac{1}{T}\int\limits_0^T \lambda(t)dt. \tag{1.2.12}$$

Thus, let us consider the initial value problem

$$\frac{dN(t)}{dt} = \lambda(t)N(t), \qquad N(t_0) = N_0, \tag{1.2.13}$$

where $\lambda(t)$ is a continuous periodic function with period T. Clearly, the solution is given by

$$N(t) = N_0 e^{\int\limits_{t_0}^{t} \lambda(s)ds}. \tag{1.2.14}$$

It would be tempting to believe that a population with periodically changing growth rate also changes in a periodic way. However, we have

$$r(t+T) := \int\limits_{t_0}^{t+T} \lambda(s)ds = \int\limits_{t_0}^{t} \lambda(s)ds + \int\limits_{t}^{t+T} \lambda(s)ds = r(t) + \int\limits_{0}^{T} \lambda(s)ds = r(t) + \bar{\lambda}T$$

so that

$$N(t+T) = N(t)e^{\bar{\lambda}T}$$

and we do not have periodicity in the solution. However, we may provide a better description of the evolution. Let us try to find what is 'missing' in the function r so that it is not periodic. Assume that $\tilde{r}(t) = r(t) + \phi(t)$, where ϕ is as yet an unspecified function, is periodic hence

$$\tilde{r}(t+T) = r(t+T)+\phi(t+T) = r(t)+\bar{\lambda}T+\phi(t+T) = \tilde{r}(t)+\bar{\lambda}T+\phi(t+T)-\phi(t)$$

thus

$$\phi(t+T) = \phi(t) - \bar{\lambda}T.$$

This shows that $\psi = \phi'$ is a periodic function. To reconstruct ϕ from its periodic derivative, first we assume that the average of ψ is zero. Then $F(t) = \int\limits_{t_0}^{t} \psi(s)ds$ is periodic. Indeed, $F(t+T) = \int\limits_{t_0}^{t+T} \psi(s)ds = F(t) + \int\limits_{t}^{t+T} \psi(s)ds = F(t) + \int\limits_{0}^{T} \psi(s)ds = F(t)$. Next, if the average of ψ is $\bar{\psi}$, then $\psi - \bar{\psi}$ has zero average. Indeed,

$$\int\limits_{t_0}^{t_0+T} (\psi(s) - \bar{\psi})ds = T\bar{\psi} - T\bar{\psi} = 0$$

Hence

$$\int\limits_{t_0}^{t} \psi(s)ds = g(t) + (t-t_0)\bar{\psi}$$

where $g(t)$ is a periodic function. Returning to function ϕ, we see that

$$\psi(t) = g(t) + c(t-t_0)$$

for some constant c and periodic function g. As we are interested in the simplest representation, we put $g(t) = 0$ and so $\psi(t)$ becomes a linear function and

$$-\bar{\lambda}T = \phi(t+T) - \phi(t) = c(t+T-t_0) - c(t-t_0)$$

and so $c = \bar{\lambda}$.

Using this result we write

$$N(t) = N_0 e^{\int\limits_{t_0}^{t} \lambda(s)ds} = N_0 e^{\bar{\lambda}(t-t_0)}Q(t) \tag{1.2.15}$$

where

$$Q(t) = e^{\int\limits_{t_0}^{t} \lambda(s)ds - \bar{\lambda}(t-t_0)} \tag{1.2.16}$$

is a periodic function.

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

$$N(k) = N(t_0 + kT) = N_0 e^{\bar{\lambda}T} Q(t_0 + kT) = N_0 e^{\bar{\lambda}kT} Q(t_0) = N_0 [e^{\bar{\lambda}T}]^k,$$

which is the expected difference equation with growth rate given by $e^{\bar{\lambda}T}$.

Next let us consider an open population described by the equation

$$\frac{dN(t)}{dt} = \lambda(t)N(t) + c(t) \tag{1.2.17}$$

where $\lambda(t)$ and $c(t)$ are continuous and periodic functions with period T. The constant $\bar{\lambda}$ and the periodic function $Q(t)$ are defined by (1.2.12) and (1.2.16). As before our aim is find a periodic pattern (if such exists) in solutions to (1.2.17). From general theory of linear equations we find that

$$N(t) = e^{\int\limits_{t_0}^{t} \lambda(s)ds} N(t_0) + e^{\int\limits_{t_0}^{t} \lambda(s)ds} \int\limits_{t_0}^{t} e^{-\int\limits_{t_0}^{r} \lambda(s)ds} c(r)dr \tag{1.2.18}$$

is the general solution. For periodicity with period T it is necessary (but of course not sufficient) that $N(t_0) = N(t_0 + T)$. Let $\bar{N}$ be the solution satisfying this condition. Then

$$\bar{N}(t_0) = e^{\int\limits_{t_0}^{t_0+T} \lambda(s)ds} \bar{N}(t_0) + e^{\int\limits_{t_0}^{t_0+T} \lambda(s)ds} \int\limits_{t_0}^{t_0+T} e^{-\int\limits_{t_0}^{r} \lambda(s)ds} c(r)dr.$$

Assuming $\bar{\lambda} \neq 0$ where $\bar{\lambda}$ was defined in (1.2.12) and using the definition (1.2.16) we write the required initial condition as

$$\bar{N}(t_0) = \frac{e^{\bar{\lambda}T}}{1 - e^{\bar{\lambda}T}} \int\limits_{t_0}^{t_0+T} \frac{e^{-\bar{\lambda}(s-t_0)}c(r)}{Q(r)} dr.$$

Substituting this formula into (1.2.18) we get

$$\begin{aligned}
\bar{N}(t) &= \frac{e^{\bar{\lambda}T}e^{\int\limits_{t_0}^{t}\lambda(s)ds}}{1-e^{\bar{\lambda}T}}\int\limits_{t_0}^{t_0+T}\frac{e^{-\bar{\lambda}(r-t_0)}c(r)}{Q(r)}dr + e^{\int\limits_{t_0}^{t}\lambda(s)ds}\int\limits_{t_0}^{t}\frac{e^{-\bar{\lambda}(r-t_0)}c(r)}{Q(r)}dr\\
&= \frac{Q(t)}{1-e^{\bar{\lambda}T}}\left(e^{\bar{\lambda}(t-t_0+T)}\int\limits_{t_0}^{t_0+T}\frac{e^{-\bar{\lambda}(r-t_0)}c(r)}{Q(r)}dr + e^{\bar{\lambda}(t-t_0)}\int\limits_{t_0}^{t}\frac{e^{-\bar{\lambda}(r-t_0)}c(r)}{Q(r)}dr\right.\\
&\quad\left.-e^{\bar{\lambda}(t-t_0+T)}\int\limits_{t_0}^{t}\frac{e^{-\bar{\lambda}(r-t_0)}c(r)}{Q(r)}dr\right)\\
&= \frac{Q(t)}{1-e^{\bar{\lambda}T}}\left(\int\limits_{t}^{t_0+T}\frac{e^{\bar{\lambda}(t-r+T)}c(r)}{Q(r)}dr + \int\limits_{t_0}^{t}\frac{e^{\bar{\lambda}(t-r)}c(r)}{Q(r)}dr\right)\\
&= \frac{Q(t)}{1-e^{\bar{\lambda}T}}\left(\int\limits_{t-t_0}^{T}\frac{e^{\bar{\lambda}\sigma}c(t-\sigma)}{Q(t-\sigma)}d\sigma + \int\limits_{0}^{t-t_0}\frac{e^{\bar{\lambda}\sigma}c(t-\sigma)}{Q(t-\sigma)}d\sigma\right) = \frac{Q(t)}{1-e^{\bar{\lambda}T}}\int\limits_{0}^{T}\frac{e^{\bar{\lambda}\sigma}c(t-\sigma)}{Q(t-\sigma)}d\sigma,
\end{aligned}$$

where we used periodicity of c and Q. This property also shows that $\bar{N}$ is a periodic function. To provide a representation formula for arbitrary solution to (1.2.17) we recall that a general solution of a nonhomogeneous linear equation can be written as a sum of a particular solution to the inhomogeneous equation and the general solution of the homogeneous solution. Since $\bar{N}$ is a solution of the inhomogeneous equation and the solution of the homogeneous equation is given by (1.2.15), we can write

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

where the constant K satisfies $K = N(t_0) - \bar{N}(t_0)$. Hence, finally,

$$N(t) = (N(t_0) - \bar{N}(t_0))e^{\bar{\lambda}(t-t_0)}Q(t) + \bar{N}(t). \tag{1.2.19}$$

From this representation we easily find that if $\bar{\lambda} < 0$, then

$$\lim_{t\to\infty}(N(t) - \bar{N}(t)) = 0.$$

1.3 Methods of analysing single difference equations

The general form of a first order difference equation is

$$x(n+1) = f(n, x(n)), \tag{1.3.1}$$

where f is any function of two variables defined on $\mathbb{N}_0 \times \mathbb{R}$, where $\mathbb{N}_0 = \{0, 1, 2 \ldots\}$ is the set of natural numbers enlarged by 0. In theoretical considerations we write $x(n)$ instead of x_n - this will simplify the notation when dealing with systems of equations. In most cases we shall deal with autonomous equations

$$x(n+1) = f(x(n)), \tag{1.3.2}$$

1.3.1 Methods of solution

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,

$$x(n+1) = a(n)x(n) + g(n) \tag{1.3.3}$$

with the 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

$$x(n) = x(0)\prod_{k=0}^{n-1} a(k) + \sum_{k=0}^{n-1} g(k) \prod_{i=k+1}^{n-1} a(i) \tag{1.3.4}$$

where we adopted the convention that $\prod\limits_{n}^{n-1} = 1$. Similarly, to simplify notation, we agree to put $\sum\limits_{k=j+1}^{j} = 0$. To fully justify this formula, we shall use mathematical induction. Constructing (1.3.4) we have checked that the formula holds for a few initial values of the argument. Assume now that it is valid for n and consider

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

which proves that (1.3.4) is valid for all $n \in \mathbb{N}$.

Two special cases

There are two special cases of (1.3.3) that appear in many applications. In the first, the equation is given by

$$x(n) = ax(n) + g(n), \tag{1.3.5}$$

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 (1.3.4) takes the form

$$x(n) = a^n x(0) + \sum_{k=0}^{n-1} a^{n-k-1} g(k). \tag{1.3.6}$$

The second case is a simpler form of (1.3.5), given by

$$x(n) = ax(n) + g, \tag{1.3.7}$$

with g independent of n. In this case the sum in (1.3.6) can be evaluated in an explicit form giving

$$x(n) = \begin{cases} a^n x(0) + g\frac{a^n-1}{a-1} & \text{if} \quad a \neq 1, \\ x(0) + gn. \end{cases} \tag{1.3.8}$$

Example 1.3.1. It turns out that the Hassell equation with $b = 1$ can be solved explicitly. Let us recall this equation:

$$x(n+1) = \frac{R_0 x(n)}{1 + x(n)} \tag{1.3.9}$$

Writing

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

we see that the substitution $y(n) = 1/x(n)$ converts (1.3.9) to

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

Using (1.3.8) we find

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

if $R_0 \neq 1$ and

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

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

1.3.2 Equilibrium points

Most equations cannot be solved in closed form. One of the typical questions is the behaviour of the system after many iterations. The concept of equilibrium point and stability of it plays a central role in the study of dynamics of physical/biological systems.

Definition 1.3.2. A point x^* in the domain of f is said to be an equilibrium point of (1.3.2) if it is a fixed point of f; that is $f(x^*) = x^*$.

In other words, x^* is a constant solution of (1.3.2).

Graphically, an equilibrium point is the the x-coordinate of the point where the graph of f intersects the diagonal $y = x$. This is the basis of the cob-web method of finding equilibria and analyse their stability, which is described later.

In differential equations, an equilibrium cannot be reached in finite time. Difference equations do not share this property. This leads to the definition:

Definition 1.3.3. A point x in the domain of f is said to be an eventual equilibrium of (1.3.2) if there is an equilibrium point x^* of (1.3.2) and a positive integer r such that $x^* = f^r(x)$ and $f^{r-1}(x) \neq x^*$.

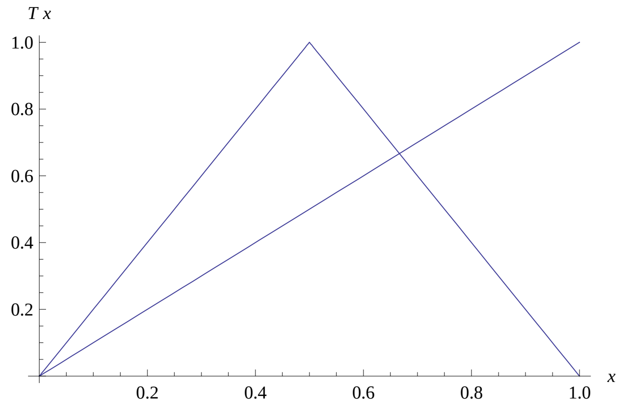

Figure 1.5: The tent map

Example 1.3.4. The Tent Map. Consider

$$x(n+1) = Tx(n)$$

where

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

There are two equilibrium points, 0 and 2/3. Looking for eventual equilibria is not as simple. Taking $x(0) = 1/8$, we find $x(1) = 1/4$, $x(2) = 1/2$, $x(3) = 1$ and $x(4) = 0$, and hence 1/8 (as well as $1/4, 1/2$ and 1) are eventual equilibria. It can be checked that all points of the form $x = n/2^k$, where $n, k \in \mathbb{N}$ satisfy $0 < n/2^k < 1$ are eventual equilibria.

Definition 1.3.5. (a) 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 $\epsilon >$ such that for any $\delta > 0$ there are x and n such that $|x - x^*| < \delta$ and $|f^n(x) - x^*| \geq \epsilon$.)

(b) The point x^* is called attracting if there is $\eta > 0$ such that

$$|x(0) - x^*| < \eta \text{ implies } \lim_{n\to\infty} x(n) = x^*.$$

If $\eta = \infty$, x^* is called a global attractor or globally attracting.

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

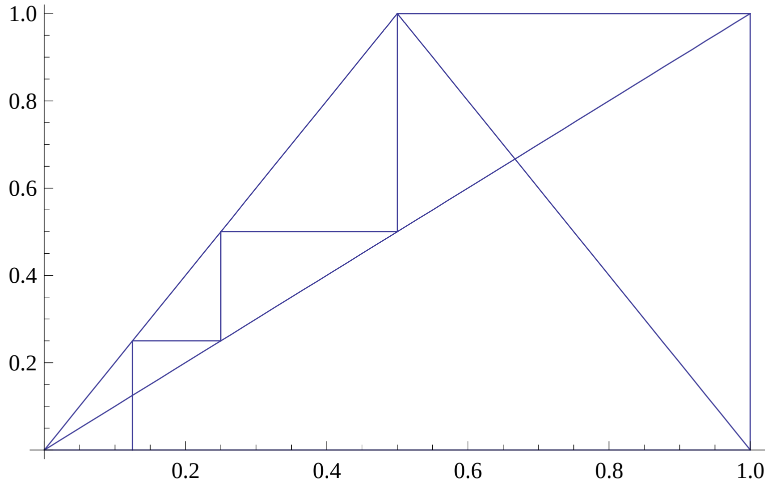

Figure 1.6: Eventual equilibrium $x = 1/8$ for the tent map.

Example 1.3.6. There are difference equations with attracting but not stable equilibria. Consider the equation

$$x(n+1) = G(x_n)$$

where

$$G(x) = \begin{cases} -2x & \text{for} \quad x < 1, \\ 0 & \text{for} \quad x \geq 1. \end{cases}$$

If $x_0 \geq 0$, $x(n) = 0$ for all n. If $x_0 < 1$, then $x(n) = (-2)^n x_0$ as long as $(-2)^n x_0 < 1$ and then $x(n) = 0$. Hence $x = 0$ is an attracting equilibrium point. However, taking $x_0 = 1/(-2)^k$, we see that x_0 can be arbitrarily close to 0 but $x(k) = 1$ (with $x(k+j) = 0$ for all $j \geq 1$).

It should be noted, however, that it can be proved that situation like that cannot happen for a continuous scalar G; for this we need at least a 2 dimensional case.

The Cobweb Diagrams

We start with an important graphical method for analysing the stability of equilibrium (and periodic) points of (1.3.2). 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 $x(0)$, we pinpoint the value $x(1)$ by drawing the vertical line through $x(0)$ so that it also intersects the graph of f at $(x(0), x(1))$. Next, draw a horizontal line from $(x(0), x(1))$ to meet the diagonal line $y = x$ at the point $(x(1), x(1))$. A vertical line drawn from the point $(x(1), x(1))$ will meet the graph of f at the point $(x(1), x(2))$. In this way we may find $x(n)$. *Analytic*

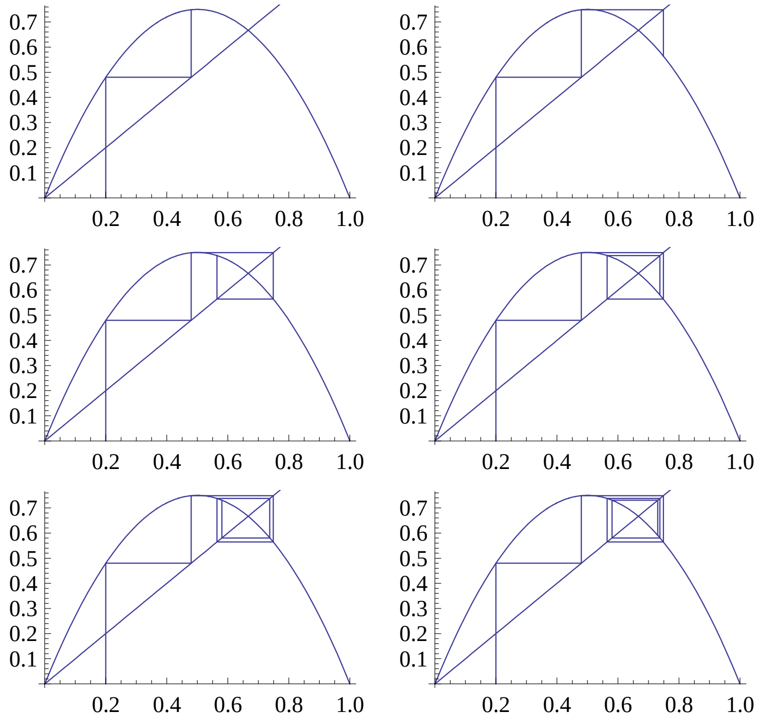

Figure 1.7: Cobweb diagram of a logistic difference equation

criterion for stability

Theorem 1.3.7. *Let x^* be an equilibrium point of the difference equation*

$$x(n+1) = f(x(n)) \tag{1.3.10}$$

where f is continuously differentiable at x^*. Then:

(i) If $|f'(x^*)| < 1$, then x^* is asymptotically stable;

(ii) If $|f'(x^*)| > 1$, then x^* is unstable.

Proof. Suppose $|f'(x^*)| < M < 1$. Then $|f'(x)| \leq M < 1$ over some interval $J = (x^* - \gamma, x^* + \gamma)$ by the property of local preservation of sign for continuous function. Now, we have

$$|x(1) - x^*| = |f(x(0)) - f(x^*)|.$$

By the Mean Value Theorem, there is $\xi \in [x(0), x^*]$ such that

$$|f(x(0)) - f(x^*)| = |f'(\xi)||x(0) - x^*|.$$

Hence

$$|f(x(0)) - f(x^*)| \leq M|x(0) - x^*|,$$

and therefore

$$|x(1) - x^*| \leq M|x(0) - x^*|.$$

Since $M < 1$, the inequality above shows that $x(1)$ is closer to x^* than $x(0)$ and consequently $x(1) \in J$. By induction,

$$|x(n) - x^*| \leq M^n|x(0) - x^*|.$$

For given ϵ, define $\delta = \epsilon/2M$. Then $|x(n) - x^*| < \epsilon$ for $n > 0$ provided $|x(0) - x^*| < \delta$ (since $M < 1$). Furthermore $x(n) \to x^*$ 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 $J = (x^* - \epsilon, x^* + \epsilon)$ on which $|f'(x)| \geq M > 1$. Take arbitrary $\delta > 0$ smaller than ϵ and x satisfying $|x - x^*| < \delta$. Using again the Mean Value Theorem

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

for some ξ between x^* and x so that

$$|f(x) - x^*| \geq M|x - x^*|.$$

If $f(x)$ is outside J, then we are done. If not, we can repeat the argument getting $|f^2(x) - x^*| \geq M^2|x - x^*|$, that is, $f^2(x)$ which is further away from x^* than $f(x)$. If it is in J we can continue the procedure till $|f^n(x) - x^*| \geq M^n|x - x^*| > \epsilon$ for some n. □

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

What happens if the equilibrium point is non-hyperbolic. We start with the case $f'(x^*) = 1$.

Theorem 1.3.8. *Let x^* be an isolated equilibrium with $f'(x^*) = 1$. Then*

(i) If $f''(x^*) \neq 0$, then x^* is unstable.

(ii) If $f''(x^*) = 0$ and $f'''(x^*) > 0$, then x^* is unstable.

(iii) If $f''(x^*) = 0$ and $f'''(x^*) < 0$, then x^* is asymptotically stable.

Proof. (i) If $f''(x^*) \neq 0$, then there is an interval $J = (x^* - \eta, x^* + \eta)$ such that either $f''(x) > 0$ or $f''(x) < 0$ on J. Thus over J the function f is concave up (in the first case) or concave down (in the second). Writing

$$f(x) - f(x^*) = (x - x^*) + \frac{1}{2}f''(\xi)(x - x^*)^2$$

for some ξ between $x \in J$ and x^* and noting that $f(x^*) = x^*$, we get

$$f(x) - x = \frac{1}{2}f''(\xi)(x - x^*)^2$$

which shows that $f(x)$ is above $y = x$ in the first case and below in the second for all $x \in J$. Let us concentrate on $f'(x^*) > 0$ and take arbitrary $J \ni x(0) > x^*$. Then $f'(x) > 1$ for x in a one-sided neighbourhood $(x^*, x^* + \eta)$. We can the following modification of the argument from the proof of Theorem 1.3.7: Putting $x(1) = f(x(0))$ we find that either $x_1 \notin J$, in which case the proof ends, or $J' \ni x(1) > x(0)$ (as otherwise there would be a point $y \in (x(0), x(1)) \subset J$ with $f'(y) = 1$. Considering now $J' = (x(0), x^* + \eta)$, we see that $f'(x) \geq M$ for all $x \in J'$ and some $M > 1$. Hence either $x(2) = f(x(1)) > x^* + \eta$, in which case the proof is over, or $|x(2) - x(1)| = |f(x(1)) - f(x(0))| \geq M|x(1) - x(0)|$. By induction, if $x(n) \in J'$, then either $x(n + 1) = f(x(n)) > x^* + \eta$ or $|x(n + 1) - x(1)| = |f(x(n)) - f(x(0))| \geq M^{n-1}|x(1) - x(0)|$ and for some n the iterate $x(n) > x^* + \eta$, thus ending the proof of instability of x^*.

To prove (ii) and (iii), we have as above

$$f(x) - x = \frac{1}{6}f'''(\xi)(x - x^*)^3$$

for some ξ between x and x^*. Considering first (ii), we find $J = (x^* - \eta, x^* + \eta)$ over which $f'''(x) > 0$ and thus $f(x) > x$ for $x^* < x < x^* + \eta$ and $f(x) < x$ for $x^* - \eta < x < x^*$. Let as fix $x(0) \in (x^*, x^* + \eta)$. Using $f''(x^*) = 0$, we have

$$f''(x) = \int\limits_{x^*}^{x} f'''(s)ds > 0$$

for any $x \in (x^*, x^* + \eta)$ and is strictly increasing, therefore is bounded away from 0 on $[x_0, x^* + \eta)$. Hence $f'(x) \geq M > 1$ on $[x(0), x^* + \eta)$ and we can use the first part of the proof.

If (iii) is satisfied, we find find $J = (x^* - \eta, x^* + \eta)$ over which $f'''(x) < 0$ and thus $f(x) < x$ for $x^* < x < x^* + \eta$ and $f(x) > x$ for $x^* - \eta < x < x^*$. hence, in particular, there is no other equilibrium in J. Since we can assume that $f' > 0$ over J, we also have $f(x) > x* = f(x^*)$ for $x^* < x < x^* + \eta$ and $f(x) < x* = f(x^*)$ on $x^* - \eta < x < x^*$. Thus, the iterations produce a sequence $x^* < \ldots f(x(n)) < f(x(n-1)) < \ldots < f(x(0)) < x(0)$ in the first case and $x(0) < f(x(0)) < \ldots < f(x(n-1)) < f(x(n)) < \ldots < x^*$. Hence, both sequences converge to, say, l_1 and l_2. Using now continuity of f we find

$$l_1 = \lim_{n\to\infty} x(n+1) = \lim_{n\to\infty} f(x(n)) = f(\lim_{n\to\infty} x(n)) = f(l_1).$$

Thus, l_1 is a fixed point in J and, since x^* is the only fixed point in J, $l_1 = x^*$. The same argument applies for x^*. □

The case of $f'(x^*) = -1$ is dealt with in the following theorem.

Theorem 1.3.9. *Suppose that for an equilibrium point x^* we have $f'(x^*) = -1$ and*

$$S(x^*) = -f'''(x^*) - \frac{3}{2}(f''(x^*))^2. \tag{1.3.11}$$

Then x^* is asymptotically stable if $S(x^*) < 0$ and unstable if $S(x^*) > 0$.

Proof. Consider the equation

$$x(n+1) = f(f(x(n)) = g(x(n)) \tag{1.3.12}$$

We observe first that if x^* is an equilibrium point of (1.3.2), that it is also an equilibrium point of (1.3.12). Second, if such an x^* is asymptotically stable (unstable) for (1.3.12), then it is also asymptotically stable (unstable) for (1.3.2). Indeed, consider stability. For any ϵ we find $\delta < \epsilon$ such that from $|x - x^*| < \delta$ it follows that $|f(x) - x^*| < \epsilon$. Next, from stability for f^2 it follows that for this δ, we can find $\delta_1 < \delta$ such that $|f^{2n}(x) - x^*| < \delta < \epsilon$. Thus $|f^{2n+1}(x) - x^*| = |f(f^{2n}(x)) - x^*| < \epsilon$ whenever $|x - x^*| < \delta_1$. In the same way we prove asymptotic stability. The instability statement is obvious. Now

$$g'(x) = f'(f(x))f'(x)$$

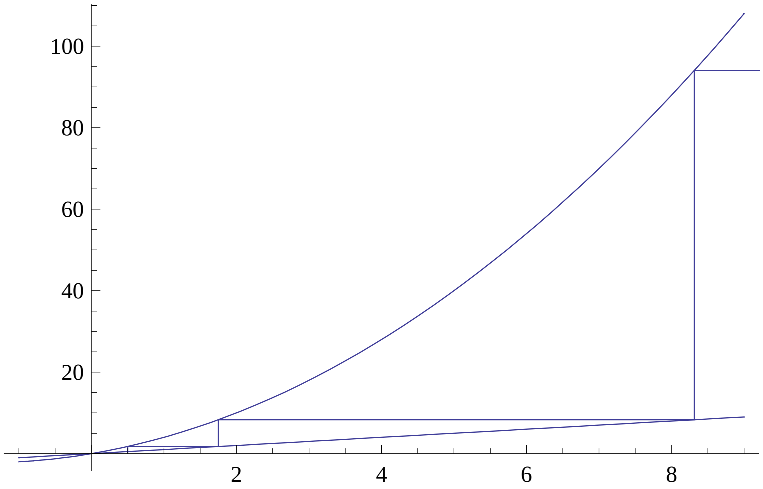

Figure 1.8: Unstable character of the equilibrium $x = 0$. Initial point $x_0 = 0.5$

so $g'(x^*) = 1$ and we are in the situation of the previous theorem. Further,

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

and, since $f(x^*) = x^*$ and $f'(x^*) = -1$,

$$g''(x^*) = 0.$$

Using the chain rule once again, we find

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

□

Example 1.3.10. 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.

Remark 1.3.11. Analysing cob-web diagrams (or otherwise) we observe that we can provide a further fine-tuning of the stability. Clearly, if $f'(x^*) < 0$, then the solution behaves in an oscillatory way around x^* and if $f'(x^*) > 0$, it is monotonic. Indeed, consider (in a neighbourhood of x^* where $f'(x) < 0$)

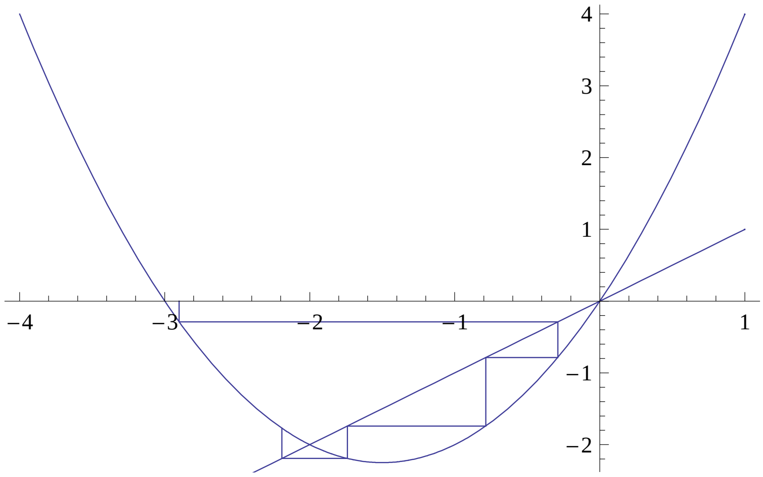

Figure 1.9: Stable character of the equilibrium $x = -2$. Initial point $x_0 = -2.9$

$f(x) - f(x^*) = f(x) - x^* = f'(\xi)(x - x^*)$, where ξ is between x^* and x. Since $f' < 0$, $f(x) > x^*$ if $x < x^*$ and $f(x) < x^*$ if $x > x^*$, which means that each iteration move the point to the other side of x^*. If $|f'| < 1$ over this interval, then $f^n(x)$ converge to x^* in an oscillatory way, while if $|f'| > 1$, the iterations will move away from the interval, also in an oscillatory way.

Based on on this observation, we may say that the equilibrium is oscillatory unstable or stable if $f'(x^*) < -1$ or $-1 < f'(x^*) < 0$, respectively, and monotonically stable or unstable depending on whether $0 < f'(x^*) < 1$ or $f'(x^*) > 1$, respectively.

Periodic points and cycles

Definition 1.3.12. Let b be in the domain of f. Then:

(i) b is called a periodic point of f if $f^k(b) = b$ for some $k \in \mathbb{N}$. The periodic orbit of b, $O(b) = \{b, f(b), f^2(b), \ldots, f^{k-1}(b)\}$ 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 1.3.13. The Tent Map revisited. Consider

$$x(n+1) = T^2 x(n)$$

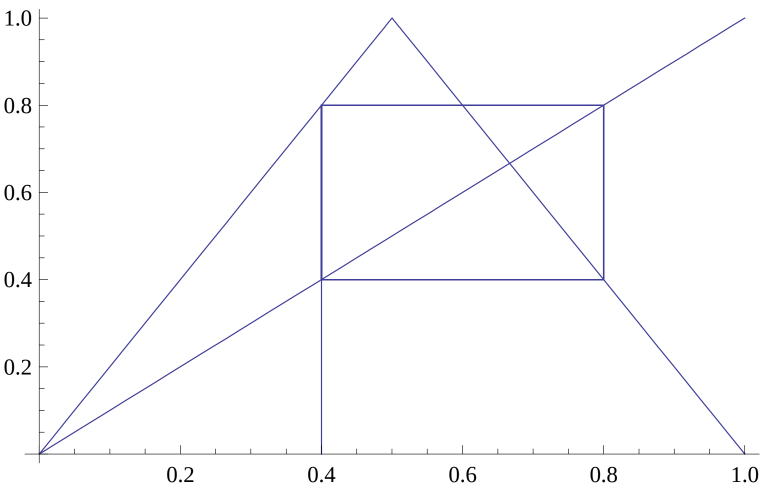

Figure 1.10: 2-cycle for the tent map

where we have

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

There are four equilibrium points, $0, 0.4, 2/3$ and 0.8, two of which are equilibria of T. Hence $\{0,4,0.8\}$ is the only 2-cycle of T. $x^* = 0.8$ is not stable. Calculation for T^3 shows that $\{2/7, 4/7, 6/7\}$ is a 3-cycle. There is a famous theorem by Šarkowski (rediscovered by Li and Yorke) that if a map has a 3-cycle, then it has k-cycles for arbitrary k. This is one of symptoms of chaotic behaviour.

Definition 1.3.14. Let b be a k-periodic point of f. Then b is said to be:

(i) stable if it is a stable fixed point of f^k;

(ii) asymptotically stable if it is an asymptotically stable fixed point of f^k;

(iii) unstable if it is an unstable fixed point of f^k.

It follows that if b is k-periodic, then every point of its k-cycle $\{x(0) = b, x(1) = f(b), \ldots, x(k-1) = f^{k-1}(b)\}$ is also k-periodic. This follows from $f^k(f^r(b)) = f^r(f^k(b)) = f^r(b)$, $r = 0, 1, \ldots, k-1$. Moreover, each such point possesses the same stability property as b. Here, the stability of b means that $|f^{nk}(x) - b| < \epsilon$ for all n, provided x is close enough to b. To prove the statement, we have to show that for any ϵ there is δ such that

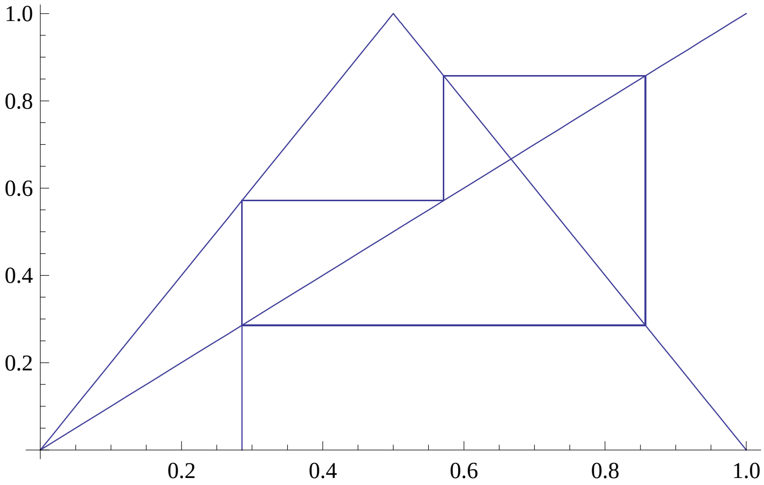

Figure 1.11: 3-cycle for the tent map

$|f^{nk}(x)-f^r(b)|<\epsilon$ for any fixed $r=0,1,\ldots,k-1$ and $n\in\mathbb{N}$, if $|x-f^r(b)|<\delta$. Let us take arbitrary $\epsilon>0$. From continuity of f (at thus of f^k), there is δ_1 such that $|x-f^r(b)|<\delta_1$ implies, by $f^{k+r}(b)=f^r(f^k(b))=f^r(b)$, that

$$|f^k(x)-f^r(b)|=|f^k(x)-f^{k+r}(b)|<\epsilon. \tag{1.3.13}$$

With the same ϵ, using continuity of f^r we find δ_2 such that $|f^r(z)-f^r(b)|<\epsilon$, provided $|z-b|<\delta_2$. For this δ_2, we find δ_3 such that if $|y-b|<\delta_3$, then $|f^{nk}(y)-b|<\delta_2$ for any n. Hence, for $|y-b|<\delta_3$, taking $z=f^{nk}(y)$, we obtain

$$|f^{r+nk}(y)-f^r(b)|<\epsilon \tag{1.3.14}$$

for any n. On the other hand, for this δ_3 we find δ_4 such that if $|x-f^r(b)|<\delta_4$, then $|f^{k-r}(x)-f^k(b)|=|f^{k-r}(x)-b|<\delta_3$ and, using $y=f^{k-r}(x)$ in (1.3.14), we obtain

$$|f^{(n+1)k}(x)-f^r(b)|<\epsilon \tag{1.3.15}$$

for any $n\geq 1$. Taking $|x-f^r(b)|<\delta_5=\min\{\delta_4,\delta_1\}$ and combining (1.3.13) with (1.3.15), we get

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

for any $n\geq 1$.

The definition together with Theorem 1.3.7 yield the following classification of stability of k-cycles.

Theorem 1.3.15. *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*

$$|f'(x(0))f'(x(1))\dots f'(x(k-1))| < 1.$$

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

$$|f'(x(0))f'(x(1))\dots f'(x(k-1))| > 1.$$

Proof. Follow from Theorem 1.3.7 by the Chain Rule applied to f^k. □

The Logistic Equation and Bifurcations Consider the logistic equation

$$x(n+1) = \mu x(n)(1-x(n)), \quad x \in [0,1], \mu > 0 \tag{1.3.16}$$

which arises from iterating $F_\mu(x) = \mu x(1-x)$. To find equilibrium point, we solve

$$F_\mu(x^*) = x^*$$

which gives $x^* = 0, (\mu-1)/\mu$.

We investigate stability of each point separately.

(a) For $x^* = 0$, we have $F'_\mu(0) = \mu$ and thus $x^* = 0$ is asymptotically stable for $0 < \mu < 1$ and unstable for $\mu > 1$. To investigate the stability for $\mu = 1$, we find $F''_\mu(0) = -2 \neq 0$ and thus $x^* = 0$ is unstable in this case. However, instability comes from negative values of x which we discarded from the domain. If we restrict our attention to the domain $[0,1]$, then $x^* = 0$ is stable. Such points are called *semi-stable*.

(b) The equilibrium point $x^* = (\mu-1)/\mu$ belongs to the domain $[0,1]$ only if $\mu > 1$. Here, $F'((\mu-1)/\mu) = 2-\mu$ and $F''((\mu-1)/\mu) = -2\mu$. Thus, using Theorems 1.3.7 and 1.3.8 we obtain:

  (i) x^* is asymptotically stable if $1 < \mu \leq 3$,

  (ii) x^* is unstable if $3 < \mu$.

We observe further that for $1 < \mu < 2$ the population approaches the carrying capacity monotonically from below. However, for $2 < \mu \leq 3$ the population can go over the carrying capacity but eventually stabilizes around it.

What happens for $\mu = 3$? Consider 2-cycles. We have $F^2_\mu(x) = \mu^2 x(1-x)(1-\mu x(1-x))$ so that we are looking for solutions to

$$\mu^2 x(1-x)(1-\mu x(1-x)) = x$$

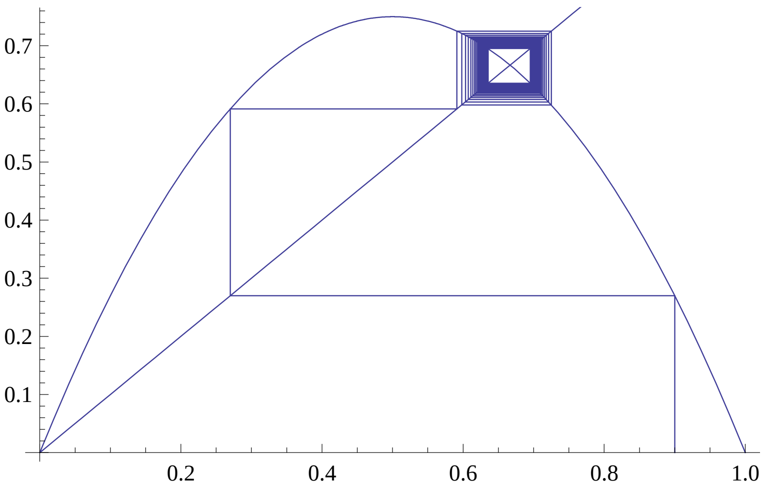

Figure 1.12: Asymptotically stable equilibrium $x = 2/3$ for $\mu = 3$.

We can re-write this equation as

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

To simplify the considerations, we observe that any equilibrium is also a 2-cycle (and any k-cycle for that matter). Thus, we can divide this equation by x and $x - (\mu-1)/\mu$, getting

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

Solving this quadratic equation, we obtain 2-cycle

$$\begin{aligned} x(0) &= \frac{(1+\mu) - \sqrt{(\mu-3)(\mu+1)}}{2\mu} \\ x(1) &= \frac{(1+\mu) + \sqrt{(\mu-3)(\mu+1)}}{2\mu}. \end{aligned} \tag{1.3.17}$$

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

The stability of 2-cycles can be determined by Theorem 1.3.15. We have $F'(x) = \mu(1-2x)$ so the 2-cycle is stable provided

$$-1 < \mu^2(1-2x(0))(1-2x(1)) < 1.$$

Using Viete's formulae we find that the above yields

$$-1 < \mu^2 + 2\mu + 4 < 1$$

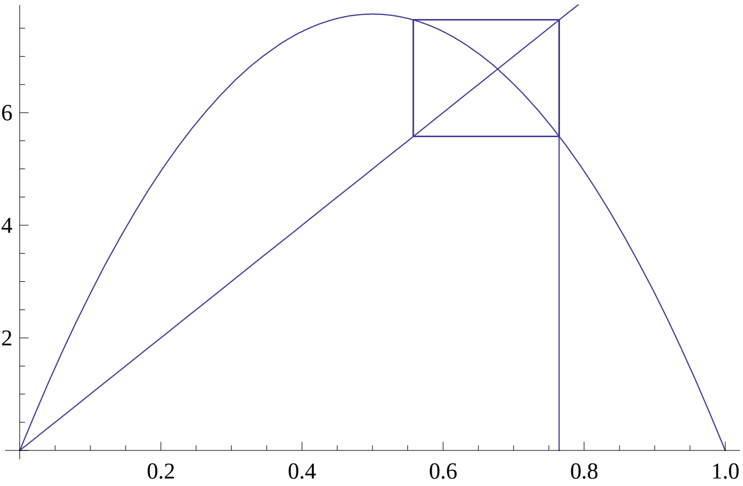

Figure 1.13: 2-cycle for $x \approx 0.765$ and $\mu = 3.1$.

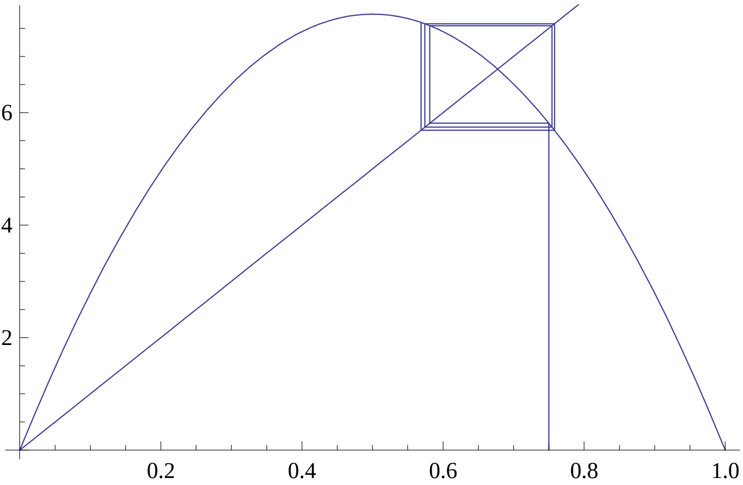

Figure 1.14: Asymptotic stability of the 2-cycle for $x \approx 0.765$ and $\mu = 3.1$.

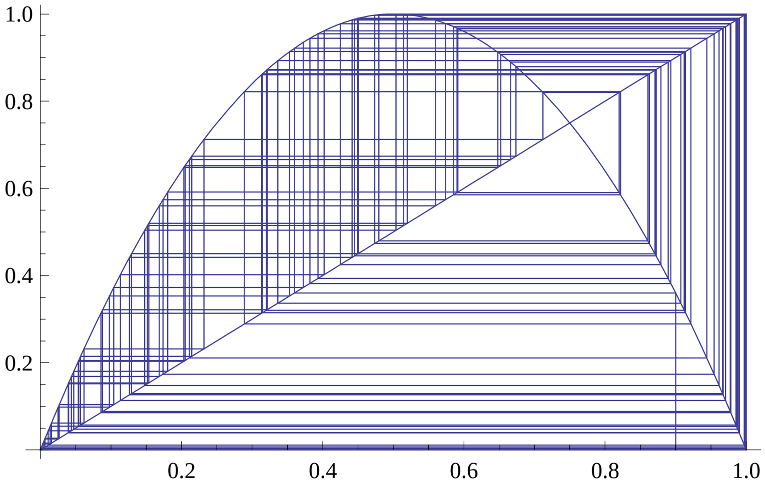

Figure 1.15: Chaotic orbit for $x = 0.9$ and $\mu = 4$.

and solving this we see that this is satisfied if $\mu < -1$ or $\mu > 3$ and $1-\sqrt{6} < \mu < 1+\sqrt{6}$ which yields $3 < \mu < 1+\sqrt{6}$.

In similar fashion we can determine that for $\mu_1 = 1+\sqrt{6}$ the 2-cycle is still attracting but becomes unstable for $\mu > \mu_1$.

Remark 1.3.16. To find 4-cycles, we solve $F_\mu^4(x)$. However, in this case algebra becomes unbearable and one should resort to a computer. It turns out that there is 4-cycle when $\mu > 1+\sqrt{6}$ which is attracting for $1+\sqrt{6} < \mu < 3.544090\ldots =: \mu_2$. When $\mu = \mu_2$, then 2^2-cycle bifurcates into a 2^3-cycle, which is stable for $\mu_2 \leq \mu \leq \mu_3 := 3.564407..$ Continuing, we obtain a sequence of numbers $(\mu_n)_{n\in\mathbb{N}}$ such that the 2^n-cycle bifurcates into 2^{n+1}-cycle passing through μ_n. In this particular case, $\lim_{n\to\infty}\mu_n = \mu_\infty = 3.57....$ A remarkable observation is

Theorem 1.3.17. *(***Feigenbaum, 1978***) For sufficiently smooth families F_μ of mapping of an interval into itself, the number*

$$\delta = \lim_{n\to\infty}\frac{\mu_n - \mu_{n-1}}{\mu_{n+1}-\mu_n} = 4.6692016...$$

in general does not depend on the family of maps, provided they have single maximum.

This theorem expresses the fact that the bifurcation diagrams for such maps are equivalent to the bifurcation diagram of a unique mapping for which it is exactly self-similar.

What happens for μ_∞? Here we find a densely interwoven region with both periodic and chaotic orbits. In particular, a 3-cycle appears and, as we

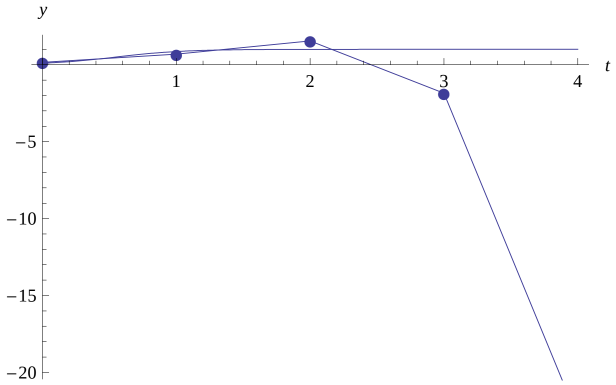

Figure 1.16: Comparison of solutions to (1.3.18) with $a = 4$ and (1.3.19).

mentioned earlier, period 3 implies existence of orbits of any period. We can easily prove that 3-cycles appear if $\mu = 4$. Consider first $F_4(x)$. We have $F_4(0) = F_4(1) = 0$ and $F_4(0.5) = 1$. This shows that $F_4^2(0.5) = F_4(1) = 0$. From the Darboux property, there are $a_1 \in (0, 0.5)$ and $a_2 \in (0.5, 1)$ such that $F_4(a_i) = 0.5$ and $F_4^2(a_i) = 1$. Thus we have graph with two peaks at 1 and attaining zero in between. This shows that $F_4^2(x) = x$ has four solutions, two of which are (unstable) equilibria and two are (unstable) 2-cycles. Repeating the argument there is $b_1 \in (0, a_1)$ such that $F_4(b_1) = a_1$ (since the graph is steeper than that of $y = x$) and thus $F_4^3(b_1) = F_4^2(a_1) = 1$. Similarly, we get 3 other points in which $F_4^3 = 1$ and clearly $F_4^3(a_i) = F_4^3(0.5) = 0$. This means that $y = x$ meets $F_4^3(x)$ at 8 points, two of which are equilibria (2-cycles are not 3-cycles). So, we obtain two 3-cycles.

Euler scheme for the logistic equation Consider the logistic differential equation

$$y' = ay(1-y), \qquad y(0) = y_0. \tag{1.3.18}$$

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

$$y(n+1) = y(n) + 4y(n)(1-y(n)) \tag{1.3.19}$$

is chaotic and thus the latter cannot be used for numerical calculations of (1.3.18) as the solutions to (1.3.18) are monotonic. This is shown in Fig. 1.16. Let us, however, write down the complete Euler scheme:

$$y(n+1) = y(n) + a\Delta t y(n)(1-y(n)), \tag{1.3.20}$$

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

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

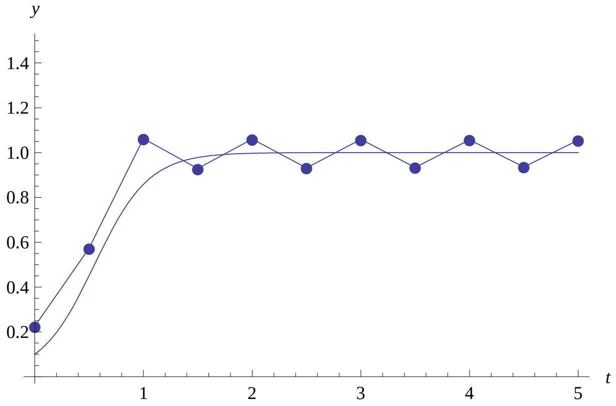

Figure 1.17: Comparison of solutions to (1.3.18) with $a = 4$ and (1.3.22) with $\mu = 3$ ($\Delta t = 0.5$).

Substitution

$$x(n) = \frac{a\Delta t}{1 + a\Delta t} y(n) \tag{1.3.21}$$

reduces (1.3.20) to

$$x(n+1) = \mu x(n)(1 - x(n)). \tag{1.3.22}$$

Thus, the parameter μ which controls the long time behaviour of solutions to the discrete equation (1.3.22) depends on Δt and, by choosing a suitably small Δt we can get solutions of (1.3.22) to mimic the behaviour of solutions to (1.3.18). Indeed, by taking $1+a\Delta t < 3$ we obtain convergence of solutions $x(n)$ to the equilibrium

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

which, reverting (1.3.21), gives the discrete approximation $y(n)$ which converges to 1, as the solution to (1.3.18). However, as seen on Fig 1.17, this convergence is not monotonic which shows that the approximation is rather poor. This can be remedied by taking $1 + a\Delta t < 2$ in which case the qualitative features of $y(t)$ and $y(n)$ are the same, see Fig. 1.18). We note that above problems can be also solved by introducing the so-called non-standard difference schemes which consists in replacing the derivatives and/or nonlinear terms by more sophisticated expressions which, though equivalent when the time step goes to 0 produce, nevertheless, qualitatively different discrete picture. In the case of the logistic equation such a non-standard scheme can be constructed replacing y^2 not by $y^2(n)$ but by $y(n)y(n+1)$.

$$y(n+1) = y(n) = a\Delta t(y(n) - y(n)y(n+1)).$$

In general, such a substitution yields an implicit scheme but in our case the

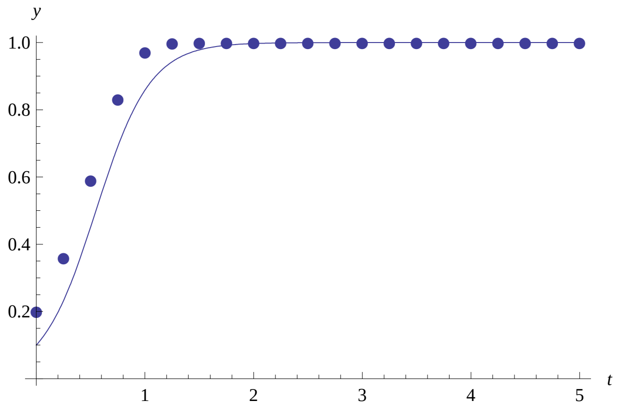

Figure 1.18: Comparison of solutions to (1.3.18) with $a = 4$ and (1.3.22) with $\mu = 2$ ($\Delta t = 0.25$).

resulting recurrence can be solved for $y(n+1)$ producing

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

and we recognize the Beverton-Holt-Hassel equation with $R_0 = 1 + a\Delta t$ (and $K = 1$).

The Beverton-Holt-Hassell equation We conclude with a brief description of stability of equilibrium points for the Hassell equation.

Let us recall the equation

$$x(n+1) = f(x_n, R_0, b) = \frac{R_0 x_n}{(1 + x_n)^b}.$$

Writing

$$x^*(1 + x^*)^b = R_0 x^*$$

we find steady state $x^* = 0$ and we observe that if $R_0 \leq 1$, then this is the only steady state (at least for positive values of x). If $R_0 > 1$, the there is another steady state given by

$$x^* = R_0^{1/b} - 1.$$

Evaluating the derivative, we have

$$f'(x^*, R_0, b) = \frac{R_0}{(1 + x^*)^b} - \frac{R_0 b x^*}{(1 + x^*)^{b+1}} = 1 - b + \frac{b}{R_0^{1/b}}$$

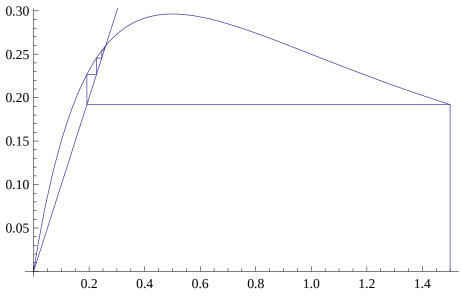

Figure 1.19: Monotonic stability of the equilibrium for the Beverton-Holt model with $b = 3$ and $R_0 = 2$; see Eqn (1.3.23).

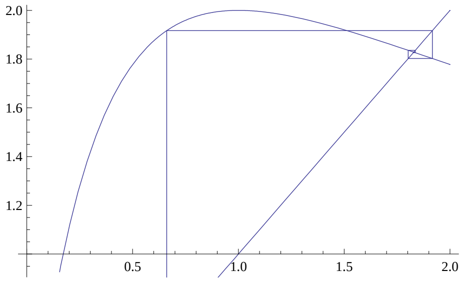

Figure 1.20: Oscillatory stability of the equilibrium for the Beverton-Holt model with $b = 2$ and $R_0 = 8$; see Eqn (1.3.24).

Clearly, with $R_0 > 1$, we always have $f' < 1$, so for the monotone stability we must have

$$1 - b + \frac{b}{R_0^{1/b}} > 0$$

and for oscillatory stability

$$-1 < 1 - b + \frac{b}{R_0^{1/b}} < 0.$$

Solving this inequalities, we obtain that the borderlines between different behaviours are given by

$$R_0 = \left(\frac{b}{b-1}\right)^b \tag{1.3.23}$$

and

$$R_0 = \left(\frac{b}{b-2}\right)^b. \tag{1.3.24}$$

Let us consider existence of 2-cycles. The second iteration of the map

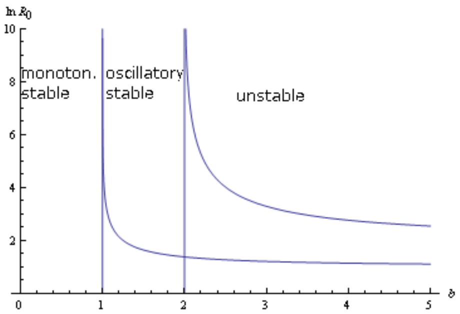

Figure 1.21: Regions of stability of the Beverton-Holt model described by (1.3.23) and (1.3.24)

$$Hx = \frac{R_0 x}{(1+x)^b}$$

is given by

$$H(H(x)) = \frac{R_0^2 x (1+x)^{b^2-b}}{((1+x)^b + R_0 x)^b}$$

so that 2-cycles can be obtained from $H(H(x)) = x$ which can be rewritten as

$$xR_0^2(1+x)^{b^2-b} = x((1+x)^b + R_0x)^b,$$

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

$$(1+x)R_0^{\frac{2}{b}} = (1+x)^b + R_0 x.$$

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

$$f(z) = z^b - z^{b-1}R_0^{\frac{2}{b}} + R_0 z - R_0.$$

Clearly we have $f(R_0^{\frac{1}{b}}) = 0$ as any equilibrium of H is also an equilibrium of H^2. First let us consider $1 < b < 2$ (the case $b = 1$ yields explicit solution (see Example 1.3.1) whereas the case $b = 2$ can be investigated directly and is referred to the tutorial problems).

We have

$$f'(z) = bz^{b-1} - (b-1)z^{b-2}R_0^{\frac{2}{b}} + R_0$$

and

$$f''(z) = (b-1)z^{b-3}(bz + (2-b)R_0^{\frac{2}{b}})$$

and we see that $f'' > 0$ for all $z > 0$. Furthermore, $f(0) = -R_0 < 0$. Hence, the region Ω bounded from the left by the axis $z = 0$ and lying above the

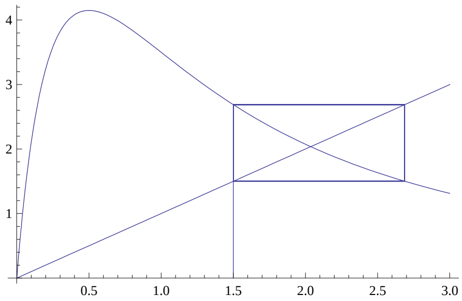

Figure 1.22: 2-cycles for the Beverton-Holt model with $b = 3$ and $R_0 = 28$; see Eqn (1.3.24).

graph of f for $z > 0$ is convex. Thus, the z axis, being transversal to the axis $z = 0$ cuts the boundary of Ω in exactly two points, one being $(0, 0)$ and the other $(R_0^{\frac{1}{b}}, 0)$. Hence, there are no additional equilibria of H^2 and therefore H does not have 2-cycles for $b \leq 2$.

Let us consider $b > 3$ (the case $b = 3$ is again referred to tutorials). In this case f has exactly one inflection point

$$z_i = \frac{b-2}{b} R_0^{\frac{2}{b}}$$

The fact that the equilibrium $x* = R_0^{\frac{1}{b}} - 1$ loses stability at $R_0 = (b/b - 2)^b$ suggests that a 2-cycle can appear when R_0 increases passing through this point. Let us first discuss the stable region $R_0 \leq (b/b - 2)^b$. Then

$$z_i \leq \frac{b}{b-2} < 1,$$

that is, the inflection point occurs in the nonphysical region $x = z - 1 < 0$. For $z = 1$ we have $f(1) = 1 - R_0^{\frac{2}{b}} < 0$ and we can argue as above, using the line $z = 1$ instead of the axis $z = 0$. Thus, when the equilibrium $x* = R_0^{\frac{1}{b}} - 1$ is stable, there are no 2-cycles. Let us consider the case with $R_0 > (b/b - 2)^b$. At the equilibrium we find

$$\begin{aligned} f'(R_0^{\frac{1}{b}}) &= bR_0^{\frac{b-1}{b}} - (b-1)R_0^{\frac{b-2}{b}} R_0^{\frac{2}{b}} + R_0 \\ &= bR_0^{\frac{b-1}{b}} - (b-2)R_0 = R_0(bR_0^{-\frac{1}{b}} - (b-2)) \end{aligned}$$

and $f'(R_0^{\frac{1}{b}}) > 0$ provided $R_0 > (b/b - 2)^b$. So, f takes negative values for $z > R_0^{\frac{1}{b}}$ but, on the other hand, $f(z)$ tends to $+\infty$ for $z \to \infty$ and therefore there must be $z^* > R_0^{\frac{1}{b}}$ for which $f(z^*)$. Since $R_0^{\frac{1}{b}} - 1$ and 0 were the only equilibria of H, z^* must give a 2-cycle.

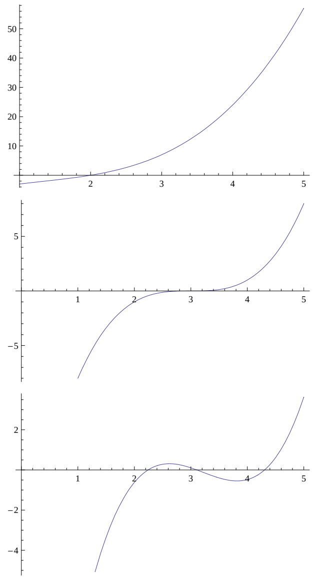

Figure 1.23: Function f for $b = 3$ and, from top to bottom, $R_0 = 8, 27, 30$ Notice the emergence of 2-cycles represented here by new zeros of f besides $z = \sqrt[3]{R_0}$.

With much more, mainly computer aided, work we can establish that, as with the logistic equation, we obtain period doubling and transition to chaos.

Experimental results are in quite good agreement with the model. Most models fell into the stable region. It is interesting to note that laboratory populations are usually less stable then the field ones. This is because scramble for resources is confined and more homogeneous and low density-independent mortality (high R_0). Also, it is obvious that high reproductive ratio R_0 and highly over-compensating density dependence (large b) are capable of provoking periodic or chaotic fluctuations in population density. This can be demonstrated mathematically (before the advent of mathematical theory of chaos it was assumed that these irregularities are of stochastic nature) and is observed in the fluctuations of the population of the Colorado beetle.

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

1.4 A comparison of stability results for differential and difference equations

Let us consider a phenomenon is a static environment which can be described in both continuous and discrete time. In the first case we have an (autonomous) differential equation

$$y' = f(y), \qquad y(0) = y_0, \tag{1.4.1}$$

and in the second case a difference equation

$$y(n+1) = g(y(n)), \qquad y(0) = y_0. \tag{1.4.2}$$

In all considerations of this section we assume that both f and g are sufficiently regular functions so as not to have any problems with existence, uniqueness etc.

First we note that while in both cases y is the number of individuals in the population, the equations (1.4.1) and (1.4.2) refer to two different aspects of the process. In fact, while (1.4.1) describes the (instantaneous) rate of change of the population's size, (1.4.2) give the size of the population after each cycle. To be more easily comparable, (1.4.2) should be written as

$$y(n+1) - y(n) = -y(n) + g(y(n)) =: \bar{f}(y(n)), \qquad y(0) = y_0, \tag{1.4.3}$$

which would describe the rate of change of the population size per unit cycle. However, difference equations typically are written and analysed in the form (1.4.2).

Let us recall the general result describing dynamics of (1.4.1). As mentioned above, we assume that f is at least a Lipschitz continuous function on $\mathbb{R}$ and the solutions exist for all t. An equilibrium solution is any solution $y(t) \equiv y$ satisfying $f(y) = 0$.

Theorem 1.4.1. *(i) If y_0 is not an equilibrium point, then $y(t)$ never equals an equilibrium point.*
(ii) All non-stationary solutions are either strictly decreasing or strictly increasing functions of t.
(iii) For any $y_0 \in \mathbb{R}$, the solution $y(t)$ either diverges to $+\infty$ or $-\infty$, or converges to an equilibrium point, as $t \to \infty$.

From this theorem it follows that if f has several equilibrium points, then the stationary solutions corresponding to these points divide the (t, y) plane into strips such that any solution remains always confined to one of them. If we look at this from the point of phase space and orbits, first we note that

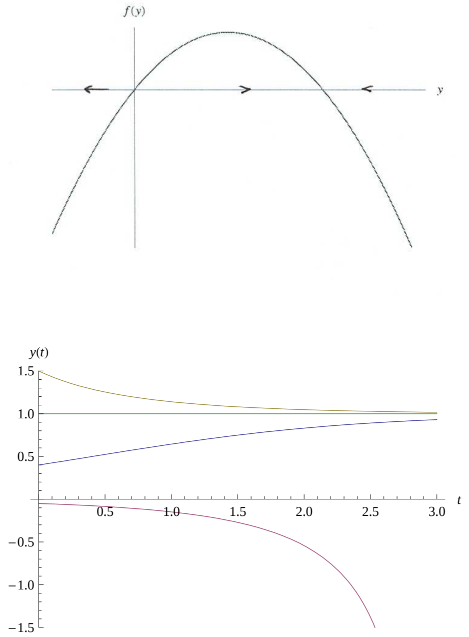

Figure 1.24: Monotonic behaviour of solutions to (1.4.1) depends on the right hand side f of the equation.

the phase space in the 1 dimensional case is the real line $\mathbb{R}$, divided by equilibrium points and thus and orbits are open segments (possibly stretching to infinity) between equilibrium points.

Furthermore, we observe that if $f(y) > 0$, then the solution $y(t)$ is increasing at any point t when $y(t) = y$; conversely, $f(y) < 0$ implies that the solution $y(t)$ is decreasing when $y(t) = y$. This also implies that any equilibrium point y^* with $f'(y^*) < 0$ is asymptotically stable and with $f'(y^*) > 0$ is unstable; there are no stable, but not asymptotically stable, equilibria.

If we look now at the difference equation (1.4.2), then at first we note some similarities. Equilibria are defined as

$$g(y) = y,$$

(or $\bar{f}(y) = 0$) and, as in the continuous case we compared f with zero, in the discrete case we compare $g(x)$ with x: $g(y) > y$ means that $y(n+1) = g(y(n)) > y(n)$ so that the iterates are increasing while if $g(x) < x$, then they are decreasing. Also, stability of equilibria is characterized in a similar way: if $|g'(y^*)| < 1$, then y^* asymptotically stable and if $|g'(y^*)| > 1$, then y^* unstable. In fact, if $g'(y^*) > 0$, then we have exact equivalence: y^* is stable provided $\bar{f}'(y^*) < 0$ and unstable if $\bar{f}'(y^*) > 0$. Indeed, in such a case, if we

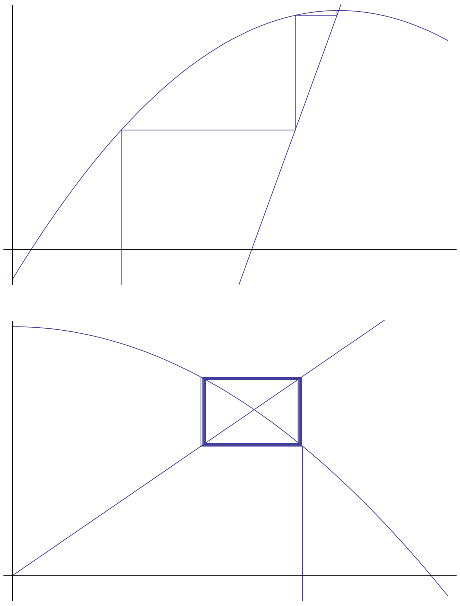

Figure 1.25: Change of the type of convergence to the equilibrium from monotonic if $0 < g'(y^*) < 1$ to oscillatory for $-1 < g'(y^*) < 0$.

start on a one side of an equilibrium y^*, then no iteration can overshot this equilibrium as for, say $y < y^*$ we have $f(y) < f(y^*) = y^*$. Thus, as in the continuous case, the solutions are confined to intervals between successive equilibria.

However, similarities end here as the dynamics of difference equation is much richer that that of the corresponding differential equation as the behaviour of the solution near an equilibrium is also governed the sign of g itself.

First, contrary to Theorem 1.4.1 (i), solutions can reach an equilibrium in a finite time, as demonstrated in Example 1.3.4.

Further, recalling Remark 1.3.11, we see that if $-1 < g'(y^*) < 0$, then the solution can overshoot the equilibrium creating damped oscillations towards equilibrium, whereas any reversal of the direction of motion is impossible in autonomous scalar differential equations. Also, as we have seen, difference equations may have periodic solutions which are precluded from occurring in the continuous case. Finally, no chaotic behaviour can occur in scalar differential equations (partly because they do not admit periodic solutions abundance of which is a signature of chaos). In fact, in can be proved that chaos in differential equations may occur only if the dimension of the state space exceeds 3.

Chapter 2

Structured populations leading to systems of linear difference equations

So far we have neglected any population structure by implicitly assuming that any member of the population is equally likely to die or give birth. While for some populations, like insects, this leads to good results, for most it is certainly not true: very young and very old individuals typically are more likely to die and less likely to give birth, etc. We start with revisiting the classical Fibonacci's problem of rabbits.

2.1 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 formulate a mathematical model, we assume that each pair consists of one male and one female. We also assume that the monthly census of the population is taken just before births for this month take place. Denoting by $y(k,n)$ the number of k-months old pairs of rabbits at time n and by $y(n)$ the total number of rabbits at time n, we find $y(n) = \sum_{k=1}^{\infty} y(k,n)$. It should be noted that the series above is always finite but depends on the initial condition; that is, how old were the pairs in the initial population.

Since no rabbits ever die, all those who were k months old in month n, are $k+1$ months old in the month $n+1$; that is,

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

for $n,k \geq 0$. There are as many one month old pairs at month $n+1$ as there were pairs of at least two months old pairs at time n:

$$y(1,n+1) = y(2,n) + y(3,n) + \ldots.$$

Further, we assume that the first pair is juvenile, one month old. The initial condition are thus the initial conditions are $y(1,0) = 1, y(2,0) = 0$ and $y(k,0) = 0$ for $k > 2$. Thus, for $n \geq 0$, we use the above formulae we may write

$$\begin{aligned} y(n+2) &= y(1,n+2) + y(2,n+2) + \ldots, \\ &= (y(2,n+1) + y(3,n+1) + \ldots) + (y(1,n+1) + y(2,n+1) + \ldots, \\ &= (y(1,n) + y(2,n) + \ldots) + y(n+1) \\ &= y(n) + y(n+1). \qquad (2.1.1) \end{aligned}$$

Our single initial condition is not sufficient for solving this equation. However, it is easy to see that we have $y(1,1) = 0, y(2,1) = 1$, $y(3,1) = 0$ and $y(k,1) = 0$ for $k > 3$ or, in other words, $y(1) = 1$. The resulting initial value problem

$$y(n+2) = y(n+1) + y(n), \quad y(0) = 1, y(0) = 1 \qquad (2.1.2)$$

can be easily solved using general theory of linear difference equations. The characteristic equation $r^2 - r - 1 = 0$ has the roots given by

$$r_{\pm} = \frac{1 \pm \sqrt{5}}{2}$$

and the general solution is

$$y(n) = C_+ r_+^n + C_- r_-^n.$$

Using the initial conditions we find

$$y(n) = \left(\frac{\sqrt{5}+1}{2\sqrt{5}}\right) r_+^n + \left(\frac{\sqrt{5}-1}{2\sqrt{5}}\right) r_-^n.$$

We further note that since $r_+ > r_-$, we have

$$\lim_{n\to\infty} r_+^{-n} y_n = \left(\frac{\sqrt{5}+1}{2\sqrt{5}}\right).$$

So,

$$y(n) \approx \left(\frac{\sqrt{5}+1}{2\sqrt{5}}\right) r_+^n \tag{2.1.3}$$

for large n.

Our aim here is not to reheat the classical Fibonacci equation but rather use it to explain a more general structure which will be discussed below. To make the present considerations useful for it, we will present an alternative approach to the Fibonacci equation.

2.2 Leslie matrices

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

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

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

$$\begin{aligned} v_1(n+1) &= v_1(n) + v_0(n) \\ v_0(n+1) &= v_1(n) \end{aligned} \tag{2.2.1}$$

or, in a more compact form

$$\mathbf{v}(n+1) = \mathcal{L}\mathbf{v}(n) := \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \mathbf{v}(n). \tag{2.2.2}$$

The solution can be found by iterations

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

How do we generalize this? Assume that we are tracking only females and not pairs and that census is taken immediately before the reproductive period. Further, assume that there is an oldest age class n and if no individual can stay in an age class for more than one time period (which is **not** the case for Fibonacci rabbits). We introduce the survival rate s_i and the age dependent maternity function m_i; that is, s_i is probability of survival from age $i-1$ to age i (or conditional probability of survival of an individual to age i provided it survived till $i-1$–this point of view will be explored later), and each individual of age i produces m_i offspring in average. Hence, s_1m_i is the average number of offspring produced by each individual of age i which survived to the census. 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

$$\mathcal{L} := \begin{pmatrix} s_1m_1 & s_1m_2 & \cdots & s_1m_{n-1} & s_1m_n \\ s_2 & 0 & \cdots & 0 & 0 \\ 0 & s_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & s_n & 0 \end{pmatrix}, \tag{2.2.3}$$

The matrix of the form (2.2.3) is referred to as a *Leslie matrix*.

Remark 2.2.1. If the census is taken immediately after the reproduction then the structure of the Leslie matrix is the same but the interpretation of coefficients is slightly different. If we fix our attention on class i, in the previous case we had v_i individuals immediately before the reproduction period, these v_i individuals produced m_iv_i class 1 individuals, $s_1m_iv_i$ of survived to the next census. In this case, we have v_i individuals immediately after the reproduction period, s_iv_i of them survives as class i individuals till the next reproduction moment producing $m_is_iv_i$ individuals of class 1. Thus, the Leslie matrix will take the form

$$\mathcal{L} := \begin{pmatrix} s_1m_1 & s_2m_2 & \cdots & s_{n-1}m_{n-1} & s_nm_n \\ s_2 & 0 & \cdots & 0 & 0 \\ 0 & s_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & s_n & 0 \end{pmatrix}, \tag{2.2.4}$$

Mathematically, both approaches are equivalent and, to avoid notational complications, Leslie matrices often are written as

$$\mathcal{L} := \begin{pmatrix} f_1 & f_2 & \cdots & f_{n-1} & f_n \\ s_2 & 0 & \cdots & 0 & 0 \\ 0 & s_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & s_n & 0 \end{pmatrix}, \tag{2.2.5}$$

where f_i are referred to as the age specific *fertility*.

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

$$\mathcal{L} := \begin{pmatrix} f_1+\tau_1 & f_2 & \cdots & f_{n-1} & f_n \\ s_2 & \tau_2 & \cdots & 0 & 0 \\ 0 & s_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & \cdots & s_n & \tau_n \end{pmatrix}, \tag{2.2.6}$$

Such matrices are called *Usher matrices.*

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

$$\mathcal{L} := \begin{pmatrix} \tau_1 & f_2 & 0 \\ s_2 & \tau_2 & 0 \\ 0 & s_3 & \tau_3 \end{pmatrix}, \tag{2.2.7}$$

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 $i = n$.

Leslie matrices fit into a more general mathematical structure describing evolution of populations divided in states, or subpopulations, not necessarily related to age. For example, we can consider clusters of cells divided into classes with respect to their size, cancer cells divided into classes on the basis of the number of copies of a particular gene responsible for its drug resistance, or a population divided into subpopulations depending on the geographical patch they occupy in a particular moment of time. Let us suppose we have n states. Each individual in a given state j contributes on average 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 $d_j \Delta t$,

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

$$\mathbf{v}_{k+1} = \mathcal{A}\mathbf{v}_k, \tag{2.2.8}$$

where

$$\mathcal{A} := \left(\begin{array}{ccccc} a_{11} & a_{12} & \cdots & a_{1\,n-1} & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2\,n-1} & a_{2n} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{n\,n-1} & a_{nn} \end{array}\right). \tag{2.2.9}$$

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 2.2.2. 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 offspring with telomer of length i by a parent of the same class is

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

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

$$\mathcal{A} := \left(\begin{array}{ccccc} 0 & d_1 & 0 & \cdots & 0 \\ 0 & r_1 & d_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & r_n \end{array}\right). \tag{2.2.10}$$

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.

A particular version of (2.2.9) 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

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

or

$$p_{ii} = \sum_{\substack{j=1\\ j\neq i}}^{n} p_{ij}, \quad i = 1, \ldots, n, \tag{2.2.11}$$

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

$$\begin{aligned} N(k+1) &= \sum_{1\leq i\leq n} v_i(k+1) = \sum_{1\leq i\leq n} \left(\sum_{1\leq j\leq n} p_{ij} v_j(k) \right) \\ &= \sum_{1\leq j\leq n} v_j(k) \left(\sum_{1\leq i\leq n} p_{ij} \right) = \sum_{1\leq j\leq n} v_j(k) = N(k). \end{aligned} \tag{2.2.12}$$

Example 2.2.3. Suppose a forest is composed of two species of trees, with A_k and B_k denoting the number of each species in the forest in year k. When a tree dies, a new one grows in its place but may be of a different species. Suppose that A are long living, with only 1% dying in average per year; on the other hand, an average of 5% of B trees dies each year. However, it is more likely that a free spot will be taken over by a B tree, say, 75% of free spots goes to the species B and only 25% to the species A. This can be expressed as

$$\begin{aligned} A_{n+1} &= 0.99A_n + 0.25 \times 0.01A_n + 0.25 \times 0.05B_n \\ B_{n+1} &= 0.75 \times 0.01A_n + 0.95B_n + 0.75 \times 0.05B_n \end{aligned}$$

or

$$\begin{aligned} A_{n+1} &= 0.9925A_n + 0.125B_n \\ B_{n+1} &= 0.0075A_n + 0.9875B_n, \end{aligned}$$

that is

$$\left(\begin{array}{c} A_{n+1} \\ B_{n+1} \end{array}\right) = \left(\begin{array}{cc} 0.9925 & 0.125 \\ 0.0075 & 0.9875 \end{array}\right)\left(\begin{array}{c} A_n \\ B_n \end{array}\right), \qquad (2.2.13)$$

We see that each column sums to 1 expressing the fact that that the spot belonging to each species must belong to some other species in the next round.

Markov processes will be discussed later. Here, after necessary preliminaries concerning general spectral properties of matrices, we shall focus on implications of the fact that all matrices discussed above have non-negative entries.

2.2.1 Interlude - 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 do to interactions with individuals in all other classes + the number of individuals who migrated to class i from all other classes - the number of individuals who migrated from class i to all other classes,

or, mathematically,

$$\begin{aligned} v_i(t+\Delta t) &= v_i(t) - d_i \Delta t v_i(t) + \sum_{j=1}^{n} b_{ij} \Delta t v_j(t) \\ &= \sum_{\substack{j=1\\ j\neq i}}^{n} \left(p_{ij}\Delta t v_j(t) - p_{ji}\Delta t v_i(t)\right), \quad i=1,\ldots,n. \end{aligned} \tag{2.2.14}$$

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

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

Using this notation in (2.2.14), dividing by Δt and passing to the limit with $\Delta t \to 0$ we obtain

$$v_i'(t) = \sum_{j=1}^{n} q_{ij} v_j(t), \quad ,i = 1,\ldots,n, \tag{2.2.15}$$

or

$$\mathbf{v}' = \mathcal{Q}\mathbf{v}, \tag{2.2.16}$$

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

$$\begin{aligned}
\frac{dN}{dt} &= \sum_{1\leq i\leq n}\frac{dv_i(t)}{dt} = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} q_{ij}v_j(t)\right)\\
&= \sum_{i=1}^{n} q_{ii}v_i(t) + \sum_{i=1}^{n}\left(\sum_{\substack{j=1\\ j\neq 1}}^{n} q_{ij}v_j(t)\right)\\
&= -\sum_{i=1}^{n} v_i(t)\left(\sum_{\substack{j=1\\ j\neq i}}^{n} p_{ji}\right) + \sum_{i=1}^{n}\left(\sum_{\substack{j=1\\ j\neq i}}^{n} p_{ij}v_j(t)\right)\\
&= -\sum_{i=1}^{n} v_i(t)\left(\sum_{\substack{j=1\\ j\neq i}}^{n} p_{ji}\right) + \sum_{j=1}^{n} v_j(t)\left(\sum_{\substack{i=1\\ i\neq j}}^{n} p_{ij}\right)\\
&= -\sum_{i=1}^{n} v_i(t)\left(\sum_{\substack{j=1\\ j\neq i}}^{n} p_{ji}\right) + \sum_{i=1}^{n} v_i(t)\left(\sum_{\substack{j=1\\ j\neq i}}^{n} p_{ji}\right) = 0, \qquad (2.2.17)
\end{aligned}$$

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

Remark 2.2.4. 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

$$\begin{aligned}
\mathbf{v}(k+1) - \mathbf{v}(k) &= (-I + \mathcal{P})\mathbf{v}(k) \qquad (2.2.18)\\
&= \begin{pmatrix} -1+p_{11} & p_{12} & \cdots & p_{1n}\\ p_{21} & -1+p_{22} & \cdots & p_{2n}\\ \vdots & \vdots & \cdots & \vdots\\ p_{n1} & p_{n2} & \cdots & -1+p_{nn} \end{pmatrix}\mathbf{v}(k),
\end{aligned}$$

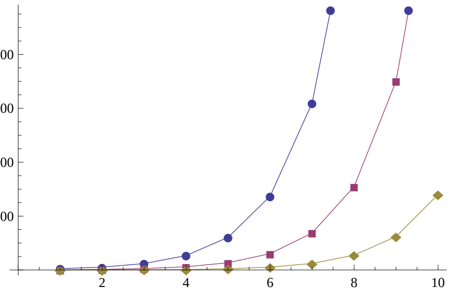

Figure 2.1: Evolution of $v_1(k)$ (circles), $v_2(k)$ (squares) and $v_3(k)$ (rhombuses) for the initial distribution $\overset{\circ}{\mathbf{v}}= (1,0,3)$ and $k = 1,\ldots,20$.

so that the 'increment' matrix has the property that each row adds up to zero due to (2.2.11). 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 zero. In fact, they can be arbitrary numbers and represent probability rates with $p_{ij}\Delta t$ being approximate interstate transition probabilities.

2.3 Long time behaviour of structured population models

As usual, we are interested in the long time behaviour of solutions. Before we embark on mathematical analysis, let us consider two numerical examples which indicate what we should expect from the models.

Example 2.3.1. Population in discrete time
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. 2.1 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. 2.2) we see that the

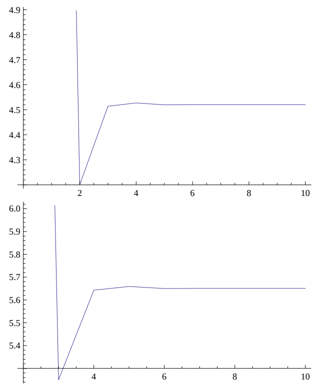

Figure 2.2: Evolution of $v_1(k)/v_2(k)$ (top) and $v_2(k)/v_3(k)$ (bottom) for the initial distribution $\overset{\circ}{\mathbf{v}}= (1, 0, 3)$ and $k = 1, \ldots, 20$.

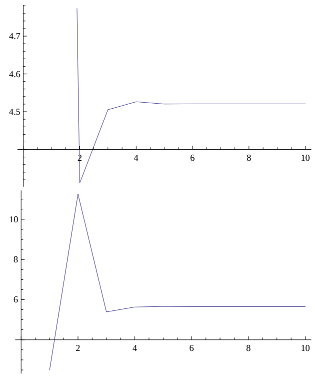

Figure 2.3: Evolution of $v_1(k)/v_2(k)$ (top) and $v_2(k)/v_3(k)$ (bottom) for the initial distribution $\overset{\circ}{\mathbf{v}}= (2, 1, 4)$ and $k = 1, \ldots, 20$.

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

$$\mathbf{v}(k) \approx f(k)\mathbf{e}. \tag{2.3.1}$$

Let us consider another initial condition, say, $\overset{\circ}{v}= (2, 1, 4)$ and do the same comparison. It turns out that the ratios stabilize at the same level which further suggest that $\mathbf{e}$ does not depend on the initial condition so that (2.3.1) can be refined to

$$\mathbf{v}(k) \approx f_1(k)g(\overset{\circ}{\mathbf{v}}))\mathbf{e}, \quad k \to \infty \tag{2.3.2}$$

where g is a linear function. Anticipation 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

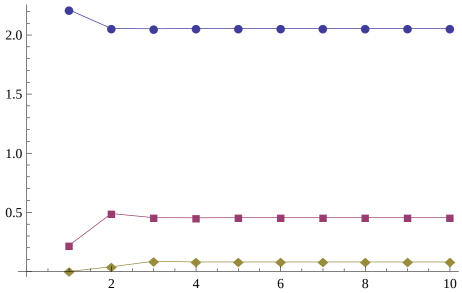

Figure 2.4: Evolution of $v_1(k)/\lambda^k)$ (circles), $v_2(k)/\lambda^k$ (squares)and $v_3(k)/\lambda^k$ (rhombuses) for the initial distribution $\overset{\circ}{\mathbf{v}}= (1, 0, 3)$ and $k = 1, \dots, 20$.

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. 2.4.

The next example shows that structured population models in continuous time have the same property.

Example 2.3.2. Population in continuous time.
Consider the following problem

$$\frac{d\mathbf{v}}{dt} = \mathcal{A}\mathbf{v}, \tag{2.3.3}$$

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 $\overset{\circ}{\mathbf{v}}= (2, 1, 4)$.

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. 2.6 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. 2.7, $f(t) = e^{0.288153t}$ and the number 0.288153 is the largest eigenvalue of $\mathcal{A}$.

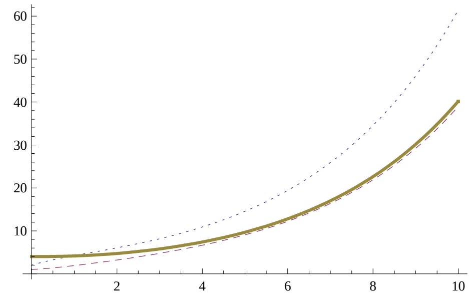

Figure 2.5: Solutions $v_1(t)$ (dotted), $v_2(t)$ (dashed) and $v_3(t)$ (continuous) for the initial condition $\overset{\circ}{\mathbf{v}}= (2, 1, 4)$

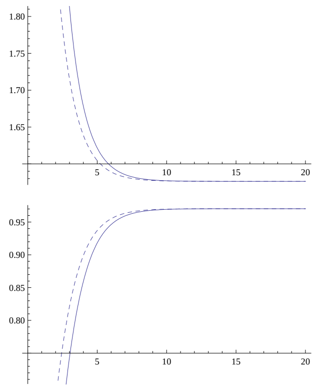

Figure 2.6: Evolution of $v_1(t)/v_2(t)$ (top) and $v_2(t)/v_3(t)$ (bottom) for the initial distributions $\overset{\circ}{\mathbf{v}}= (1, 0, 3)$ (continuous line) and $\overset{\circ}{\mathbf{v}}= (2, 1, 4)$ (dashed line).

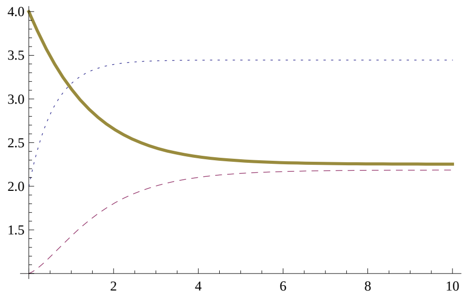

Figure 2.7: Evolution $v_1(t)/e^{0.288153t}$ (dotted), $v_2(t)/e^{0.288153t}$ (dashed) and $v_3(t)/e^{0.288153t}$ (continuous) for the initial condition $\overset{\circ}{\mathbf{v}}= (2, 1, 4)$.

2.3.1 Spectral properties of a matrix.

To explain and be able to predict similar behaviour in population models, first we discuss basic facts concerning eigenvalues and eigenvectors of a matrix. Let us start with discrete systems.

Systems of difference equations I. We are interested in solving

$$\mathbf{y}(k+1) = \mathcal{A}\mathbf{y}(k), \tag{2.3.4}$$

where $\mathcal{A}$ is an $n \times n$ matrix $\mathcal{A} = \{a_{ij}\}_{1\leq i,j\leq n}$; that is

$$\mathcal{A} = \left(\begin{array}{ccc} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{array}\right),$$

and $\mathbf{y}(k) = (y_1(k), \dots, y_n(k))$.

Eq. (2.3.4) is usually supplemented by the initial condition $\mathbf{y}(0) = \mathbf{y^0}$. It is obvious, by induction, to see that the solution to (2.3.4) is given by

$$\mathbf{y}(k) = \mathcal{A}^k\mathbf{y^0}, k = 1, 2, \dots. \tag{2.3.5}$$

The problem with (2.3.5) 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 $\mathbf{v^1}, \dots, \mathbf{v^n}$ are linearly independent vectors, then also $\mathcal{A}\mathbf{v^1}, \dots, \mathcal{A}\mathbf{v^n}$ are linearly independent. Since $\mathbb{R}^n$ is n-dimensional, it is enough to find n linearly independent vectors $\mathbf{v^i}$, $i = 1, \dots, n$ 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 $c_1, \dots, c_n$ such that

$$\mathbf{x^0} = c_1\mathbf{v^1} + \dots + c_n\mathbf{v^n}.$$

Precisely, let $\mathcal{V}$ be the matrix having vectors $\mathbf{v^i}$ as its columns

$$\mathcal{V} = \left(\begin{array}{ccc} | & \dots & | \\ \mathbf{v^1} & \dots & \mathbf{v^n} \\ | & \dots & | \end{array}\right). \tag{2.3.6}$$

Note, that $\mathcal{V}$ is invertible as the vectors $\mathbf{v^i}$ are linearly independent. Denoting $\mathbf{c} = (c_1, \dots, c_n)$, we obtain

$$\mathbf{c} = \mathcal{V}^{-1}\mathbf{x^0}. \tag{2.3.7}$$

Thus, for an arbitrary $\mathbf{x^0}$ we have

$$\mathcal{A}^n\mathbf{x^0} = \mathcal{A}^n(c_1\mathbf{v^1} + \ldots + c_2\mathbf{v^n}) = c_1\mathcal{A}^n\mathbf{v^1} + \ldots + c_k\mathcal{A}^n\mathbf{v^n}. \qquad (2.3.8)$$

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

$$\mathcal{A}^k\mathbf{x^0} = \mathcal{A}_n\mathbf{c} = \mathcal{A}_k\mathcal{V}^{-1}\mathbf{x^0}. \qquad (2.3.9)$$

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 (2.3.8) 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 (2.3.7), or

$$\mathcal{A}^k\mathbf{x^0} = \begin{pmatrix} | & \ldots & | \\ \lambda_1^k\mathbf{v^1} & \ldots & \lambda_n^k\mathbf{v^n} \\ | & \ldots & | \end{pmatrix}\mathcal{V}^{-1}\mathbf{x_0} \qquad (2.3.10)$$

Systems of differential equations I. Considerations of the previous paragraph to some extent can be repeated for systems of differential equations

$$\mathbf{y}' = \mathcal{A}\mathbf{y}, \qquad (2.3.11)$$

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 (2.3.11) is considered together with the following initial conditions

$$\mathbf{y}(t_0) = \overset{\circ}{\mathbf{y}}. \qquad (2.3.12)$$

The question of solvability and uniqueness of (2.3.11), (2.3.12) 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 2.3.3.

i. There exists one and only one solution of the initial value problem (2.3.11), (2.3.12), which is defined for all $t \in \mathbb{R}$.

ii. The set $\mathbf{X}$ of all solutions to (2.3.11) is a linear space of dimension n.

iii. If $\mathbf{y_1}(t),\ldots,\mathbf{y_k}(t)$ are linearly independent solutions of (2.3.11) and let $t_0 \in \mathbb{R}$ be an arbitrary number. Then, $\{\mathbf{y_1}(t),\ldots,\mathbf{y_k}(t)\}$ form a linearly independent set of functions if and only if $\{\mathbf{y_1}(t_0),\ldots,\mathbf{y_k}(t_0)\}$ 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 2.3.3 implies that there is matrix $\mathcal{E}(t)$ such that the solution $\mathbf{y}(t)$ can be represented as

$$\mathbf{y}(t) = \mathcal{E}(t)\overset{\circ}{\mathbf{y}} \tag{2.3.13}$$

which satisfies $\mathcal{E}(0) = \mathcal{I}$ (the identity matrix). Then we follow as in the discrete case 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 $\overset{\circ}{\mathbf{x}}\in \mathbb{R}^n$ 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 $\mathbf{c} = (c_1,\ldots,c_n)$,

$$\mathbf{c} = \mathcal{V}^{-1}\mathbf{x^0}, \tag{2.3.14}$$

where $\mathcal{V}$ was defined in (2.3.6). Thus, for an arbitrary $\overset{\circ}{\mathbf{y}}$ we have

$$\mathcal{E}(t)\overset{\circ}{\mathbf{y}}= \mathcal{E}(t)(c_1\mathbf{v^1} + \ldots + c_2\mathbf{v^n}) = c_1\mathcal{E}(t)\mathbf{v^1} + \ldots + c_k\mathcal{E}(t)\mathbf{v^n}. \tag{2.3.15}$$

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

$$\mathcal{E}(t)\overset{\circ}{\mathbf{y}}= \mathcal{E}_\mathbf{v}(t)\mathbf{c} = \mathcal{E}_\mathbf{v}(t)\mathcal{V}^{-1}\overset{\circ}{\mathbf{y}}\ . \tag{2.3.16}$$

Hence, again, the problem lies in finding linearly independent vectors $\mathbf{v^i}$, $i = 1,\ldots,k$, on which powers of $\mathcal{E}$ 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 (2.3.11) if and only if

$$\mathcal{A}\mathbf{v} = \lambda\mathbf{v}, \tag{2.3.17}$$

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 2.3.3, these solutions are linearly independent if and only if the eigenvectors $\mathbf{v^j}$ are linearly independent in $\mathbb{R}^n$. Thus, if we can find n linearly independent eigenvectors of $\mathcal{A}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ (not necessarily distinct), then the general solution of (2.3.24) is of the form

$$\mathbf{y}(t) = c_1 e^{\lambda_1 t}\mathbf{v^1} + \ldots + c_n e^{\lambda_n t}\mathbf{v^n}. \tag{2.3.18}$$

with $c_1, \ldots, c_n$ given by (2.3.7), or

$$\mathcal{E}(t)\overset{\circ}{\mathbf{y}}= \begin{pmatrix} | & \ldots & | \\ e^{\lambda_1 t}\mathbf{v^1} & \ldots & e^{\lambda_n t}\mathbf{v^n} \\ | & \ldots & | \end{pmatrix} \mathcal{V}^{-1}\overset{\circ}{\mathbf{y}}. \tag{2.3.19}$$

Unfortunately, in many cases there is 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

$$\mathcal{A}\mathbf{v} = \lambda\mathbf{v}. \tag{2.3.20}$$

Such a solution is called an *eigenvector* of $\mathcal{A}$. The set of eigenvectors corresponding to a given eigenvalue is a vector subspace. Eq. (2.3.20) 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

$$det(\mathcal{A} - \lambda\mathcal{I}) = \begin{vmatrix} a_{11} - \lambda & \ldots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \ldots & a_{nn} - \lambda \end{vmatrix} = 0. \tag{2.3.21}$$

Evaluating the determinant we obtain a polynomial in λ of degree n. This polynomial is also called the *characteristic polynomial* of the system (2.3.11). We shall denote this polynomial by $p(\lambda)$. From algebra we know that there are exactly n, possibly complex, roots of $p(\lambda)$. Some of them may be multiple, so that in general $p(\lambda)$ factorizes into

$$p(\lambda) = (\lambda_1 - \lambda)^{n_1} \cdot \ldots \cdot (\lambda_k - \lambda)^{n_k}, \tag{2.3.22}$$

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 (2.3.22) 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 differential equations is to find solutions to equations of the form $(\mathcal{A} - \lambda_i\mathcal{I})^k\mathbf{v} = 0$ for $1 < k \leq n_i$, where n_i is the algebraic multiplicity of λ_i. Precisely speaking, if λ_i has algebraic multiplicity n_i and if

$$(\mathcal{A} - \lambda_i\mathcal{I})\mathbf{v} = 0$$

has only $\nu_i < n_i$ linearly independent solutions, then we consider the equation

$$(\mathcal{A} - \lambda_i\mathcal{I})^2\mathbf{v} = 0.$$

Clearly all solutions of the preceding equation (eigenvectors) solve this equation but there is at least one more independent solution so that we have at least $\nu_i + 1$ independent vectors (note that these new vectors are no longer eigenvectors). If the number of independent solutions is still less than n_i, then we consider

$$(\mathcal{A} - \lambda_i\mathcal{I})^3\mathbf{v} = 0,$$

and so on, till we get a sufficient number of them. Note, that to make sure that in the step j we select solutions that are independent of the solutions obtained in step $j-1$ it is enough to find solutions to $(\mathcal{A} - \lambda_i\mathcal{I})^j\mathbf{v} = 0$ that satisfy $(\mathcal{A} - \lambda_i\mathcal{I})^{j-1}\mathbf{v} \neq 0$.

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

$$\begin{aligned}\mathcal{A}^k\mathbf{v^j} &= (\lambda_i\mathcal{I} + \mathcal{A} - \lambda_i\mathcal{I})^k\mathbf{v^j} = \sum_{r=0}^{k}\lambda_i^{k-r}\begin{pmatrix} k \\ r\end{pmatrix}(\mathcal{A} - \lambda_i\mathcal{I})^r\mathbf{v^j} \\ &= \Big(\lambda_i^k\mathcal{I} + k\lambda_i^{k-1}(\mathcal{A} - \lambda_i\mathcal{I}) + \ldots \\ &\quad + \frac{k!}{(j-1)!(k-j+1)!}\lambda_i^{k-j+1}(\mathcal{A} - \lambda_i\mathcal{I})^{j-1}\Big)\mathbf{v^j}, \qquad (2.3.23)\end{aligned}$$

where

$$\begin{pmatrix} k \\ r\end{pmatrix} = \frac{k!}{r!(k-r)!}$$

is the Newton symbol. It is important to note that (2.3.23) 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 2.3.4. 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

$$\begin{array}{rcl} (\mathcal{A} - 4\mathcal{I})^2 \mathbf{v} & = & \begin{pmatrix} 0 & 1 & 2 \\ 0 & -2 & -4 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} 0 & 1 & 2 \\ 0 & -2 & -4 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \\ & = & \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \end{array}$$

Any vector solves this equation so that we have to take a vector that is not an eigenvalue. Possibly the simplest choice is

$$\mathbf{v^3} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Thus, by (2.3.23)

$$\begin{array}{rcl} \mathcal{A}^k \mathbf{v^3} & = & \left(4^k \mathcal{I} + k4^{k-1}(\mathcal{A} - 4\mathcal{I})\right) \mathbf{v^3} \\ & = & \left(\begin{pmatrix} 4^k & 0 & 0 \\ 0 & 4^k & 0 \\ 0 & 0 & 4^k \end{pmatrix} + k4^{k-1} \begin{pmatrix} 0 & 1 & 2 \\ 0 & -2 & -4 \\ 0 & 1 & 2 \end{pmatrix} \right) \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \\ & = & \begin{pmatrix} 2k4^{k-1} \\ -k4^k \\ 4^k + 2k4^{-1} \end{pmatrix}. \end{array}$$

To find explicit expression for $\mathcal{A}^k$ we use (2.3.9). 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

$$\begin{aligned} \Re(\lambda^n\mathbf{v}) &= r^n(\cos n\phi\Re\mathbf{v} - \sin n\phi\Im\mathbf{v}), \\ \Im(\lambda^n\mathbf{v}) &= r^n(\sin n\phi\Re\mathbf{v} + \cos n\phi\Im\mathbf{v}). \end{aligned}$$

Example 2.3.5. 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 (2.3.9). 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

$$\mathbf{y}' = \mathcal{A}\mathbf{y}. \tag{2.3.24}$$

As before, our goal is to find n linearly independent solutions of (2.3.24). 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 (2.3.19). 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 (2.3.24) 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

$$e^{\mathcal{A}} = \mathcal{I} + \mathcal{A} + \frac{1}{2}\mathcal{A}^2 + \frac{1}{3!}\mathcal{A}^3 + \ldots + \frac{1}{k!}\mathcal{A}^k + \ldots. \tag{2.3.25}$$

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 (2.3.25) 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

$$\begin{aligned} e^{\lambda \mathcal{I}} &= \mathcal{I} + \lambda \mathcal{I} + \frac{\lambda^2}{2}\mathcal{I} + \frac{\lambda^3}{3!}\mathcal{I} + \ldots + \frac{\lambda^k}{k!} + \ldots \\ &= \left(1 + \lambda + \frac{\lambda^2}{2} + \frac{\lambda^3}{3!} + \ldots + \frac{\lambda^k}{k!} + \ldots\right)\mathcal{I} \\ &= e^{\lambda}\mathcal{I}. \end{aligned} \tag{2.3.26}$$

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

$$e^{\mathcal{A}+\mathcal{B}} = e^{\mathcal{A}}e^{\mathcal{B}} \tag{2.3.27}$$

provided the matrices $\mathcal{A}$ and $\mathcal{B}$ commute: $\mathcal{A}\mathcal{B} = \mathcal{B}\mathcal{A}$. Furthermore, defining a function of t by

$$e^{t\mathcal{A}} = \mathcal{I} + t\mathcal{A} + \frac{t^2}{2}\mathcal{A}^2 + \frac{t^3}{3!}\mathcal{A}^3 + \ldots + \frac{t^k}{k!}\mathcal{A}^k + \ldots\ , \tag{2.3.28}$$

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 (2.3.26) and (2.3.27)

$$e^{t\mathcal{A}}\mathbf{v} = e^{t(\mathcal{A}-\lambda\mathcal{I})}e^{t\lambda\mathcal{I}}\mathbf{v} = e^{\lambda t}e^{t(\mathcal{A}-\lambda\mathcal{I})}\mathbf{v}.$$

Secondly, if $(\mathcal{A}-\lambda\mathcal{I})^m\mathbf{v} = \mathbf{0}$ for some m, then

$$(\mathcal{A}-\lambda\mathcal{I})^r\mathbf{v} = \mathbf{0}, \tag{2.3.29}$$

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

$$e^{t\mathcal{A}}\mathbf{v} = e^{\lambda t}e^{t(\mathcal{A}-\lambda\mathcal{I})}\mathbf{v} = e^{\lambda t}\left(\mathbf{v} + t(\mathcal{A}-\lambda\mathcal{I})\mathbf{v} + \ldots + \frac{t^{m-1}}{(m-1)!}(\mathcal{A}-\lambda\mathcal{I})^{m-1}\mathbf{v}\right). \tag{2.3.30}$$

Thus, to find all solutions to $\mathbf{y}' = \mathcal{A}\mathbf{y}$ it is sufficient to find n independent vectors $\mathbf{v}$ satisfying (2.3.29) 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 (2.3.29) is equal to 1. Consequently, the sum in (2.3.30) terminates after the first term and we obtain

$$\mathbf{y_1}(t) = e^{\lambda_1 t}\mathbf{v^1}$$

in accordance with (2.3.18). 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 2.3.6. Let us mention here that the exponential function $e^{t\mathcal{A}}$ has been introduced just as a guideline, to explain how the formula (2.3.30) was arrived at. Once we have this formula, we can directly check that it gives a solution to (2.3.24). 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)\\
&\quad=\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)\\
&\qquad+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)\\
&\quad=\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)\\
&\qquad+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)\\
&\quad=\lambda e^{\lambda t}\frac{t^{m-1}}{(m-1)!}(\mathcal{A}-\lambda\mathcal{I})^{m-1}\mathbf{v}\\
&\qquad+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=\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}\\
&\qquad+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)\\
&\qquad-e^{\lambda t}\frac{t^{m-1}}{(m-1)!}(\mathcal{A}-\lambda\mathcal{I})^{m}\mathbf{v}\\
&\qquad+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)\\
&\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)
\end{aligned}$$

where we used $(A-\lambda\mathcal{I})^m\mathbf{v}=0$.

We illustrate the theory on a few examples.

Example 2.3.7. 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 2.3.5, $\bar{\lambda}$ 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

$$\begin{aligned}
\mathbf{y^1}(t) &= e^{\xi t}(\cos\omega t\,\Re\mathbf{v} - \sin\omega t\,\Im\mathbf{v})\\
\mathbf{y^2}(t) &= e^{\xi t}(\cos\omega t\,\Im\mathbf{v} + \sin\omega t\,\Re\mathbf{v}) \qquad (2.3.31)
\end{aligned}$$

Example 2.3.8. 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:

$$\begin{aligned} e^{(1+i)t} \begin{pmatrix} 0 \\ i \\ 1 \end{pmatrix} &= e^t(\cos t + i \sin t) \left(\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + i \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right) \\ = \quad e^t & \left(\cos t \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} - \sin t \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + i \sin t \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + i \cos t \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right) \\ = \quad e^t & \begin{pmatrix} 0 \\ -\sin t \\ \cos t \end{pmatrix} + ie^t \begin{pmatrix} 0 \\ \cos t \\ \sin t \end{pmatrix}. \end{aligned}$$

Thus, we obtain two real solutions

$$\begin{aligned} \mathbf{y^1}(t) &= e^t \begin{pmatrix} 0 \\ -\sin t \\ \cos t \end{pmatrix} \\ \mathbf{y^2}(t) &= e^t \begin{pmatrix} 0 \\ \cos t \\ \sin t \end{pmatrix} \end{aligned}$$

and the general solution to our original system is given by

$$\mathbf{y}(t) = C_1 e^t \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + C_2 e^t \begin{pmatrix} 0 \\ -\sin t \\ \cos t \end{pmatrix} + C_3 e^t \begin{pmatrix} 0 \\ \cos t \\ \sin t \end{pmatrix}.$$

We can check that all these solutions are independent as their initial values

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix},$$

are independent. To find the solution to our initial value problem we set $t = 0$ and we have to solve for C_1, C_2 and C_3 the system

$$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = C_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + C_3 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix}.$$

Thus $C_1 = C_2 = C_3 = 1$ and finally

$$\mathbf{y}(t) = e^t \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + e^t \begin{pmatrix} 0 \\ -\sin t \\ \cos t \end{pmatrix} + e^t \begin{pmatrix} 0 \\ \cos t \\ \sin t \end{pmatrix} = e^t \begin{pmatrix} 1 \\ \cos t - \sin t \\ \cos t + \sin t \end{pmatrix}.$$

The last example deals with multiple eigenvalues.

Example 2.3.9. Find three linearly independent solutions of the differential equation

$$\mathbf{y}' = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} \mathbf{y}.$$

To obtain the eigenvalues we calculate the characteristic polynomial

$$\begin{aligned} p(\lambda) &= det(\mathcal{A} - \lambda\mathcal{I}) = \begin{vmatrix} 1-\lambda & 1 & 0 \\ 0 & 1-\lambda & 0 \\ 0 & 0 & 2-\lambda \end{vmatrix} \\ &= (1-\lambda)^2(2-\lambda) \end{aligned}$$

so that $\lambda_1 = 1$ is eigenvalue of multiplicity 2 and $\lambda_2 = 2$ is an eigenvalue of multiplicity 1.

(i) $\lambda = 1$: We seek all non-zero vectors such that

$$(\mathcal{A} - 1\mathcal{I})\mathbf{v} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

This implies that $v_2 = v_3 = 0$ and v_1 is arbitrary so that we obtain the corresponding solutions

$$\mathbf{y^1}(t) = C_1 e^t \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

However, this is only one solution and $\lambda_1 = 1$ has algebraic multiplicity 2, so we have to look for one 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

$$\begin{aligned}\mathbf{y^2}(t) &= e^t\left(\mathcal{I}+t(\mathcal{A}-\mathcal{I})\right)\begin{pmatrix}0\\1\\0\end{pmatrix} = e^t\left(\begin{pmatrix}0\\1\\0\end{pmatrix} + t\begin{pmatrix}0&1&0\\0&0&0\\0&0&1\end{pmatrix}\begin{pmatrix}0\\1\\0\end{pmatrix}\right)\\ &= e^t\begin{pmatrix}0\\1\\0\end{pmatrix} + te^t\begin{pmatrix}1\\0\\0\end{pmatrix} = e^t\begin{pmatrix}t\\1\\0\end{pmatrix}\end{aligned}$$

(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

$$\mathbf{y^3}(t) = C_3e^{2t}\begin{pmatrix}0\\0\\1\end{pmatrix}.$$

Thus we have found three linearly independent solutions.

2.3.2 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:

$$y(k+n)+a_1y(k+n-1)+\ldots+a_ny(k)=0, \qquad n\geq 0 \tag{2.3.32}$$

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 (2.3.32) can be written as a system of first order equations of dimension n. We let

$$\begin{aligned}
z_1(k) &= y(k),\\
z_2(k) &= y(k+1)=z_1(k+1),\\
z_3(k) &= y(k+2)=z_2(k+1),\\
\vdots &\ \vdots\ \vdots,\\
z_n(k) &= y(k+n-1)=z_{n-1}(k-1),
\end{aligned} \tag{2.3.33}$$

hence we obtain the system

$$\begin{aligned}
z_1(k+1) &= z_2(k),\\
z_2(k+1) &= z_3(k),\\
\vdots &\ \vdots\ \vdots,\\
z_{n-1}(k+1) &= z_n(k),\\
z_n(k+1) &= -a_nz_1(k)-a_2z_2(k)\ldots-a_1z_n(k),
\end{aligned}$$

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 $\mathcal{A}$ is often called the companion matrix of the equation (2.3.32). 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

$$\begin{vmatrix} -\lambda & 1 & 0 & \dots & 0 \\ 0 & -\lambda & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -a_n & -a_{n-1} & -a_{n-2} & \dots & -a_1-\lambda \end{vmatrix}$$
$$= (-1)^n(\lambda^n + a_1\lambda^{n-1} + \ldots + a_n) = 0.$$

We note that the characteristic polynomial of the companion matrix can be obtained by just replacing $y(k+n-i)$ in (2.3.32) 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 (2.3.33), 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 2.3.10. Consider the Fibonacci equation (2.1.2):

$$y(k+2) = y(k+1) + y(k) \tag{2.3.34}$$

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

$$\begin{array}{rcl} z_1(k+1) &=& z_2(k), \\ z_2(k+1) &=& z_1(k) + z_2(k); \end{array}$$

note that it is not the same form as (2.2.2). The eigenvalues of the matrix

$$\mathcal{A} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$$

are obtained by solving the equation

$$\begin{vmatrix} -\lambda & 1 \\ 1 & 1-\lambda \end{vmatrix} = \lambda^2 - \lambda - 1 = 0;$$

they are $\lambda_{1,2} = \frac{1\pm\sqrt{5}}{2}$. Since the eigenvalues are distinct, we immediately obtain that the general solution of (2.3.34) is given by

$$y(n) = c_1\left(\frac{1+\sqrt{5}}{2}\right)^n + c_2\left(\frac{1-\sqrt{5}}{2}\right)^n.$$

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 $c_2 = -3\sqrt{5}/5$.

Example 2.3.11. Gambler's ruin A gambler plays a sequence of games against an adversary. The probability that the gambler wins R 1 in any given game is q and the probability of him losing R 1 is $1 - q$. He quits the game if he either wins a prescribed amount of N rands, or loses all his money; in the latter case we say that he has been ruined. Let $p(n)$ denotes the probability that the gambler will be ruined if he starts gambling with n rands. We build the difference equation satisfied by $p(n)$ using the following argument. Firstly, note that we can start observation at any moment, that is, the probability of him being ruined with n rands at the start is the same as the probability of him being ruined if he acquires n rands at any moment during the game. If at some moment during the game he has n rands, he can be ruined in two ways: by winning the next game and ruined with $n+1$ rand, or by losing and then being ruined with $n - 1$ rands. Thus

$$p(n) = qp(n+1) + (1-q)p(n-1). \tag{2.3.35}$$

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

$$p(n+2) - \frac{1}{q}p(n+1) + \frac{1-q}{q}p(n) = 0, \tag{2.3.36}$$

with $n = 0, 1 \ldots, N$. We supplement (2.3.36) with the (slightly untypical) side (boundary) conditions $p(0) = 1$ and $p(N) = 0$.

The characteristic equation is given by

$$\lambda^2 - \frac{1}{q}\lambda + \frac{1-q}{q} = 0$$

and the eigenvalues are $\lambda_1 = \frac{1-q}{q}$ and $\lambda_2 = 1$. Hence, if $q \neq 1/2$, then the general solution can be written as

$$p(n) = c_1 + c_2\left(\frac{1-q}{q}\right)^n$$

and if $q = 1/2$, then $\lambda_1 = \lambda_2 = 1$ and

$$p(n) = c_1 + c_2 n.$$

To find the solution for the given boundary conditions, we denote $Q = (1-q)/q$ so that for $q \neq 1/2$

$$\begin{array}{rcl} 1 & = & c_1 + c_2, \\ 0 & = & c_1 + Q^N c_2, \end{array}$$

from where

$$c_2 = \frac{1}{1-Q^N}, \qquad c_1 = -\frac{Q^N}{1-Q^N}$$

and

$$p(n) = \frac{Q^n - Q^N}{1-Q^N}.$$

Analogous considerations for $q = 1/2$ yield

$$p(n) = 1 - \frac{n}{N}.$$

For example, if $q = 1/2$ and the gambler starts with $n = 20$ rands with the target $N = 1000$, then

$$p(20) = 1 - \frac{20}{1000} = 0,98,$$

that is, his ruin is almost certain.

In general, if the gambler plays a long series of games, which can be modelled here as taking $N \to \infty$, then he will be ruined almost certainly even if the game is fair ($q = \frac{1}{2}$).

Higher order differential equations can be dealt with in the same manner. Indeed, any nth order linear equation

$$y^{(n)} + a_{n-1}y^{(n-1)} + \ldots + a_1 y' + a_0 y = 0 \tag{2.3.37}$$

can be written as a linear system of n first order equations by introducing new variables $z_1 = y$, $z_2 = y' = z_1'$, $z_3 = y'' = z_2', \ldots z_n = y^{(n-1)} = z_{n-1}'$ so that $z_n' = y^{(n)}$ and (2.3.37) turns into

$$\begin{array}{rcl} z_1' & = & z_2, \\ z_2' & = & z_3, \\ \vdots & & \vdots \\ z_n' & = & -a_{n-1}z_n - a_{n-2}z_{n-1} - \ldots - a_0 z_1. \end{array}$$

Note that if (2.3.37) was supplemented with the initial conditions $y(t_0) = y_0, y'(t_0) = y_1, \ldots y^{(n-1)} = y_{n-1}$, then these conditions will become natural initial conditions for the system as $z_1(t_0) = y_0, z_2(t_0) = y_1, \ldots z_n(t_0) = y_{n-1}$. All comments made for higher order difference equations are thus valid one must remember, however, to change λ_i^k for $e^{\lambda_i t}$ in the set of fundamental solutions and use $\{e^{\lambda_i t}, te^{\lambda_i t}, \ldots, t^{n_i-1}e^{\lambda_i t}\}$.

2.3.3 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 2.3.12. *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

$$(\mathcal{A} - \lambda_i\mathcal{I})\mathbf{v}^r = \mathbf{v}^{r'}, \tag{2.3.38}$$

where $(\mathcal{A} - \lambda_i\mathcal{I})^{k-1}\mathbf{v}^{r'} = 0$ and

$$(\mathcal{A} - \lambda_i\mathcal{I})^{k-1}\mathbf{v}^r = \mathbf{v}^{r'}, \tag{2.3.39}$$

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

$$0 = (\mathcal{A} - \lambda\mathcal{I})^3\mathbf{v}^\rho = (\mathcal{A} - \lambda\mathcal{I})(\mathcal{A} - \lambda\mathcal{I})^2\mathbf{v}^\rho \tag{2.3.40}$$

and since $\mathbf{v}^\rho$ in this range does not satisfy $(\mathcal{A} - \lambda\mathcal{I})^2\mathbf{v}^\rho = 0$, we may put

$$(\mathcal{A} - \lambda\mathcal{I})^2\mathbf{v}^\rho = \mathbf{v}^j \tag{2.3.41}$$

for some $1 \leq j \leq \nu_i$; that is, for some eigenvector $\mathbf{v}^j$. Alternatively, we can write (2.3.40) 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,

$$(\mathcal{A} - \lambda\mathcal{I})\mathbf{v}^\rho = v^{\rho'} \tag{2.3.42}$$

for some ρ' between $\nu_i + 1$ and r'. By induction, we obtain a basis of E_λ consisting of vectors satisfying (2.3.41) 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 2.3.13. *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

$$\begin{aligned}\mathcal{A}^k\mathbf{x} &= (\mathcal{A}-\lambda_i\mathcal{I}+\lambda_i\mathcal{I})^k\mathbf{x} = \sum_{j=1}^{n_i} a_j(\mathcal{A}-\lambda_i\mathcal{I}+\lambda_i\mathcal{I})^k\mathbf{v}^j \\ &= \sum_{j=1}^{n_i} a_j \sum_{r=0}^{k} \lambda_i^{k-r}\left(\begin{array}{c} k \\ r \end{array}\right)(\mathcal{A}-\lambda_i\mathcal{I})^r\mathbf{v^j}\end{aligned}$$

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 (2.3.30) we obtain

$$e^{t\mathcal{A}}\mathbf{x} = e^{\lambda_i t}\sum_{j=1}^{n_i} a_j e^{t(\mathcal{A}-\lambda\mathcal{I})}\mathbf{v^j} = e^{\lambda_i t}\sum_{j=1}^{n_i} a_j \sum_{r=0}^{r_j}\frac{t^{r-1}}{(r-1)!}(\mathcal{A}-\lambda\mathcal{I})^{r-1}\mathbf{v^j}. \tag{2.3.43}$$

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

$$\mathcal{A}^k \overset{\circ}{\mathbf{x}} = \left(\begin{array}{ccc} | & \dots & | \\ \mathcal{A}^k\mathbf{v^1} & \dots & \mathcal{A}^k\mathbf{v^n} \\ | & \dots & | \end{array}\right)\mathcal{V}^{-1}\overset{\circ}{\mathbf{x}} \tag{2.3.44}$$

and

$$e^{t\mathcal{A}} \overset{\circ}{\mathbf{x}} = \left(\begin{array}{ccc} | & \dots & | \\ e^{t\mathcal{A}}\mathbf{v^1} & \dots & e^{t\mathcal{A}}\mathbf{v^n} \\ | & \dots & | \end{array}\right)\mathcal{V}^{-1}\overset{\circ}{\mathbf{x}}, \tag{2.3.45}$$

where

$$\mathcal{V} = \left(\begin{array}{ccc} | & \dots & | \\ \mathbf{v^1} & \dots & \mathbf{v^n} \\ | & \dots & | \end{array}\right). \tag{2.3.46}$$

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

$$\mathcal{P}_i = \begin{pmatrix} 0 & \ldots & | & \ldots & 0 \\ 0 & \ldots & \mathbf{v^i} & \ldots & 0 \\ 0 & \ldots & | & \ldots & 0 \end{pmatrix} \begin{pmatrix} | & \ldots & | \\ \mathbf{v^1} & \ldots & \mathbf{v^n} \\ | & \ldots & | \end{pmatrix}^{-1}. \tag{2.3.47}$$

and note that, for $\mathbf{x} = c_1\mathbf{v^1} + \ldots + c_n\mathbf{v^n}$, $\mathcal{P}_i\mathbf{x} = c_i\mathbf{v^i}$; that is, $\mathcal{P}_i$ selects the part of $\mathbf{x}$ along $\mathbf{v^i}$. It is easy to see, that

$$\mathcal{P}_i^2 = \mathcal{P}_i, \qquad \mathcal{P}_i\mathcal{P}_j = 0, \tag{2.3.48}$$

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

$$\mathcal{P}_{\lambda_i} = \sum_{j;\ \mathbf{v^j}\in E_{\lambda_i}} \mathcal{P}_j, \tag{2.3.49}$$

where $\mathcal{P}_i$, then such operators again will be projections. This follows from (2.3.48) 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

$$\mathcal{I} = \sum_{\lambda\in\sigma(\mathcal{A})} \mathcal{P}_\lambda, \tag{2.3.50}$$

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 $e^{t\mathcal{A}}$,

$$\mathcal{A}^k = \sum_{\lambda\in\sigma(\mathcal{A})} \lambda^k\mathcal{P}_\lambda, \tag{2.3.51}$$

and

$$e^{t\mathcal{A}} = \sum_{\lambda\in\sigma(\mathcal{A})} e^{\lambda t}\mathcal{P}_\lambda, \tag{2.3.52}$$

which is another way of writing (2.3.10) and (2.3.19), respectively.

In general case, we use (2.3.50) to write

$$\mathcal{A}\mathbf{x} = \sum_{\lambda\in\sigma(A)} \mathcal{A}\mathcal{P}_\lambda\mathbf{x}, \qquad (2.3.53)$$

where, by Corollary 2.3.13, we have $\mathcal{A}\mathcal{P}_\lambda\mathbf{x} \in E_\lambda$. Thus, using (2.3.48), we get $\mathcal{P}_{\lambda_i}\mathcal{A}\mathcal{P}_{\lambda_j} = 0$ for $i \neq j$. Using (2.3.49) 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, (2.3.53) 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 (2.3.51) and (2.3.52), we observe that, by Corollary 2.3.13, 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

$$\mathcal{A}^k\mathbf{x} = \sum_{\lambda\in\sigma(\mathcal{A})} \mathcal{A}^k\mathcal{P}_\lambda\mathbf{x} = \sum_{\lambda\in\sigma(\mathcal{A})} \lambda^k\mathbf{p}_\lambda(k)\mathbf{x}, \qquad (2.3.54)$$

and

$$e^{t\mathcal{A}}\mathbf{x} = \sum_{\lambda\in\sigma(\mathcal{A})} e^{\lambda t}\mathcal{P}_\lambda\mathbf{x} = \sum_{\lambda\in\sigma(\mathcal{A})} e^{\lambda t}\mathbf{q}_\lambda(t)\mathbf{x}, \qquad (2.3.55)$$

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 (2.3.54) and (2.3.55), 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 2.3.1 and 2.3.2 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

$$\mathcal{A}^k \overset{\circ}{\mathbf{x}}\approx c_1\lambda_1^k\mathbf{v^1}$$

for large k in discrete time or, in continuous time,

$$e^{t\mathcal{A}} \overset{\circ}{\mathbf{x}}\approx c_1e^{\lambda_1 t}\mathbf{v^1},$$

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 2.3.14. *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*

$$< \mathbf{v}^*, \mathbf{v} >= 0 \tag{2.3.56}$$

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 2.3.12, $(\mathcal{A} - \lambda\mathcal{I})\mathbf{v} = \mathbf{v}'$, 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
$$\mathcal{A}^k\mathbf{x} \approx \lambda_1^k < \mathbf{v}^*, \mathbf{x} > \mathbf{v} \tag{2.3.57}$$
for large k in discrete time or, in continuous time, by
$$e^{t\mathcal{A}}\mathbf{x} \approx e^{\lambda_1 t} < \mathbf{v}^*, \mathbf{x} > \mathbf{v} \tag{2.3.58}$$
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 2.3.7.

Example 2.3.15. 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

$$\mathbf{y}(t) = e^{3t} \left(C_2 \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} + C_1 e^{-2t} \begin{pmatrix} -1 \\ 4 \\ 1 \end{pmatrix} + C_3 e^{-5t} \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} \right). \tag{2.3.59}$$

we see that the dominant eigenvalue is $\lambda_2 = 3$ and for large time

$$\mathbf{y}(t) \approx e^{3t} C_2 \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \tag{2.3.60}$$

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 (2.3.59) 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 (2.3.60) will not be realized if $\overset{\circ}{\mathbf{x}}$ is orthogonal to $\mathbf{v_2}^*$.

Example 2.3.16. 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 2.1: $v_1(0) = 0, v_0(0) = 1$, 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 (2.1.3).

The question whether any matrix with nonnegative entries in discrete time and with positive off-diagonal entries gives rise to such behaviour and, if not, what models lead to AEG, is much more delicate and requires invoking the Frobenius-Perron theorem which will be discussed in the next section. In the meantime we consider an example which shows that some Leslie matrices exhibit different behaviour.

Example 2.3.17. Consider a Leslie matrix given by

$$\mathcal{L} = \begin{pmatrix} 0 & 0 & 3 \\ 0.5 & 0 & 0 \\ 0 & 0.4 & 0 \end{pmatrix}$$

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. 2.8. The picture is completely different from that obtained in Example 2.3.1. We observe some pattern but the ratios do not tend to a fixed limit but oscillate, as shown in Fig. 2.9. 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.

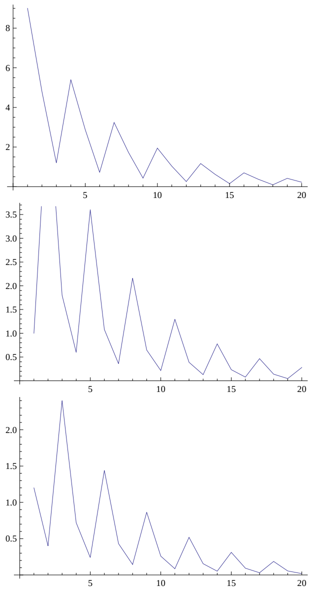

Figure 2.8: Evolution of $y_1(k)$ (top) and $y_2(k)$ (middle) and $y_3(k)$ (bottom) for the initial distribution $\overset{\circ}{\mathbf{v}}= (2,3,4)$ and $k = 1, \ldots, 10$.

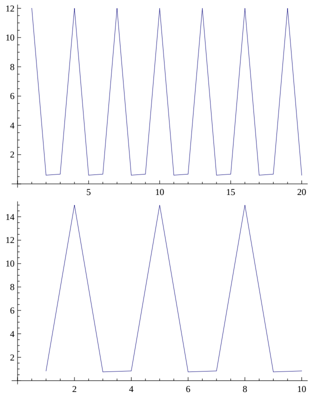

Figure 2.9: Evolution of $y_1(k)/y_2(k)$ (top) and $y_2(k)/y_3(k)$ (bottom) for the initial distribution $\overset{\circ}{\mathbf{v}}= (2,3,4)$ and $k = 1, \ldots, 20$.

2.4 Frobenius-Perron theorem

2.4.1 The issue of positivity

We have to extend the notion of positivity to vectors and functions. We say that a vector $\mathbf{x} = (x_1, \ldots, x_n)$ is non-negative, resp. positive, resp. strictly positive, if for all $i = 1, \ldots, n$, $x_i \geq 0$, resp. $x_i \geq 0$ with $\mathbf{x} \neq 0$, resp. $x_i > 0$. We denote these as $\mathbf{x} \geq 0, \mathbf{x} > 0$ and $\mathbf{x} >> 0$, respectively.

Analogous definitions can be given for positivity of scalar and vector valued functions and sequences; that is, e.g. a function $[a, b] \ni t \to \mathbf{f}(t) \in \mathbb{R}^n$ is called nonnegative if $f_i(t) \geq 0$ for any $t \in [a, b]$ and $i = 1, \ldots, n$.

Similarly, we say that a matrix $\mathcal{A} = \{a_{ij}\}_{1\leq i,j\leq n}$ is non-negative and write $\mathcal{A} \geq 0$ if $a_{ij} \geq 0$ for all $i, j = 1, \ldots, n$.

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 (1.1.18), 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 2.4.1. *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 2.4.2. *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 (2.3.28)

$$e^{t\mathcal{A}} = \mathcal{I} + t\mathcal{A} + \frac{t^2}{2}\mathcal{A}^2 + \frac{t^3}{3!}\mathcal{A}^3 + \ldots + \frac{t^k}{k!}\mathcal{A}^k + \ldots,$$

and the results of the previous proposition we see that $e^{t\mathcal{A}} \geq 0$. 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

$$e^{t\mathcal{A}} = e^{-rt}e^{t\tilde{\mathcal{A}}} \geq 0. \tag{2.4.1}$$

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 $\mathcal{E}(t)\mathbf{e}_i = (\epsilon_{1i}(t), \ldots, \epsilon_{ii}(t), \ldots, \epsilon_{in}(t))$. 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 $j \neq i$. □

We identify the $n \times n$ matrix $\mathcal{A}$ with the linear operator on $X = \mathbb{R}^n$ associated with it.

2.4.2 Norm of an operator

We define a norm of an operator $\mathcal{A}$ by

$$\|\mathcal{A}\| = \sup_{\|\mathbf{x}\|\neq 0} \frac{\|\mathcal{A}\mathbf{x}\|}{\|\mathbf{x}\|} = \sup_{\|\mathbf{x}=1\|\neq 0} \|\mathcal{A}\mathbf{x}\|. \tag{2.4.2}$$

Further, the spectral radius of $\mathcal{A}$ is defined as

$$r(A) = \sup_{\lambda \in \sigma(\mathcal{A})} |\lambda|. \tag{2.4.3}$$

From the definition, $\|\mathcal{A}\| \geq r(A)$.

Consider solving the equation

$$\mathbf{x} - \mathcal{A}\mathbf{x} = \mathbf{y}. \tag{2.4.4}$$

If we try iterations, like in the proof of Picard's theorem, we obtain the scheme

$$\mathbf{x}_{k+1} = \mathbf{y} + \mathcal{A}\mathbf{x}_k;$$

that is,

$$\mathbf{x}_{k+1} = \mathbf{y} + \mathcal{A}\mathbf{y} + \ldots + \mathcal{A}^{k+1}\mathbf{y}$$

which suggest that the solution can be obtained by

$$\mathbf{x} = \sum_{k=0}^{\infty} \mathcal{A}^k \mathbf{y}. \tag{2.4.5}$$

If the series converges then, by direct calculation using linearity and continuity of $\mathcal{A}$, it is a unique solution to (2.4.4). The series converges, in particular, if $\|\mathcal{A}\| < 1$. We provide a better characterization of the criteria of convergence of this series below. The series is called the Neumann series.

2.4.3 The resolvent

We identify the $n \times n$ matrix $\mathcal{A}$ with the linear operator on $X = \mathbb{R}^n$ associated with it.

The spectrum $\sigma(\mathcal{A})$ is the set of $\lambda \in \mathbb{C}$, for which $\lambda I - \mathcal{A}$ is not invertible. The spectrum of $\mathcal{A}$ consists of finitely many (at most n) points. The (open) complement of $\sigma(\mathcal{A})$ is called the resolvent set of $\mathcal{A}$. In other words, the resolvent set $\rho(\mathcal{A})$ is the set for which there exists the inverse matrix

$$\mathcal{R}(\lambda) = (\lambda I - \mathcal{A})^{-1}.$$

We observe that the entries of $\mathcal{R}(\lambda)$ are given by fractions

$$(\mathcal{R}(\lambda))_{ij} = M_{ij}(\lambda)/det(\lambda I - \mathcal{A})$$

where $M_{ij}(\lambda)$ are corresponding minors of $(\lambda I - \mathcal{A})$. Thus, $\mathcal{R}(\lambda)$ is a rational function of λ with poles exactly at the points of $\sigma(\mathcal{A})$. Thus, $R(\lambda)$ is bounded if and only if $\lambda \in \rho(\mathcal{A})$.

The resolvent satisfies the *resolvent identity*

$$\mathcal{R}(\lambda) - \mathcal{R}(\mu) = (\mu - \lambda)\mathcal{R}(\lambda)\mathcal{R}(\mu) \tag{2.4.6}$$

whenever $\lambda, \mu \in \rho(\mathcal{A})$.

The resolvent also can be characterized by the Neumann series. In fact, using (2.4.5) we can write

$$(\lambda I - \mathcal{A})^{-1} = \lambda^{-1}\left(I - \lambda^{-1}\mathcal{A}\right)^{-1} = \lambda^{-1}\sum_{k=0}^{\infty}\mathcal{A}^n\lambda^{-n}.$$

We recognize the series as the Laurent series of the function $\lambda \to (\lambda I - \mathcal{A})^{-1}$ at ∞. Using the Cauchy-Hadamard criterion, the series converges if $\lambda > \overline{\lim_{n\to\infty}}\|\mathcal{A}^n\|^{1/n}$ and diverges inside this circle. It can be proved that the upper limit in this expression can be replaced by the normal limit. On the other hand, the Cauchy formula for the coefficients of the Laurent expansion shows that the radius of convergence is determined by the first singularity of the function staring from the centre of expansion. Hence, in this case the radius of convergence coincides with the spectral radius; that is

$$r(\mathcal{A}) = \lim_{n\to\infty}\|\mathcal{A}^n\|^{1/n}.$$

2.4.4 Positive vectors and matrices

We say that a vector $\mathbf{x}$ (resp. a matrix $\mathcal{A}$) are non-negative, if $x_i \geq 0$ for $1 \leq i \leq n$ (resp. $a_{ij} \geq 0$ for $1 \leq i, j \leq n$). Similarly, $\mathbf{x} > 0$ is all entries are strictly positive and the same convention applies to matrices. $\mathbf{x} \geq \mathbf{y}$ (resp. $\mathbf{x} > \mathbf{y}$) if $x_i \geq y_i$ (resp. $x_i > y_i$) with the analogous notation for matrices. The absolute value is defined as $|\mathbf{x}| = (|x_i|, \ldots, |x_n|\}$, etc. Clearly

$$|\mathcal{A}\mathbf{x}| \leq |\mathcal{A}||\mathbf{x}|.$$

It follows that $\|\mathbf{x}\| = \||\mathbf{x}|\|$ and consequently

$$\|\mathcal{A}\| = \||\mathcal{A}|\|.$$

For nonnegative matrices we obtain

$$|\mathcal{R}(\lambda)| = \left|\lambda^{-1}\sum_{k=0}^{\infty}\mathcal{A}^n\lambda^{-n}\right| \leq |\lambda|^{-1}\sum_{k=0}^{\infty}\mathcal{A}^n|\lambda|^{-n} = \mathcal{R}(|\lambda|). \tag{2.4.7}$$

Theorem 2.4.3. *The spectral radius of a positive matrix $\mathcal{A}$ is an eigenvalue of $\mathcal{A}$.*

Proof. We know that there is an eigenvalue λ_0 with $|\lambda_0| = r(\mathcal{A})$. Consider the real sequence $\lambda_n = r(\mathcal{A}) + 1/n$ which converges to $r(\mathcal{A})$. Clearly, $\lambda_n \in \rho(\mathcal{A})$ thus $\mathcal{R}(\lambda_n)$ are well defined. If we consider the sequence $(\mu_n)_{n\in\mathbb{N}}$ defined as $\mu_n = \lambda_n\lambda_0/|\lambda_0|$. Then $\rho(\mathcal{A}) \ni \mu_n \to \lambda_0$ with $|\mu_n| = \lambda_n$. Using (2.4.7) we get

$$\mathcal{R}(\lambda_n) = \mathcal{R}(|\mu_n|) \geq |\mathcal{R}(\mu_n)|.$$

If $r(A)$ belonged to the resolvent set of $\mathcal{A}$, we would have $\mathcal{R}(\lambda_n)$ converging to a finite limit. However, $|\mathcal{R}(\mu_n)|$ is unbounded as λ_0 is a pole. This contradiction proves the theorem. □

This result can be significantly strengthen but for this we need to better understand the structure of the resolvent. Since $r := r(\mathcal{A})$ is an eigenvalue, we can expand $\mathcal{R}(\lambda)$ into the Laurent series around $\lambda = r$:

$$\mathcal{R}(\lambda) = \sum_{k=-h}^{\infty} A_k(\lambda - r)^k. \tag{2.4.8}$$

Since $\mathcal{R}(\lambda)$ is a rational function, $\lambda = r$ is a pole and thus the principal part of the Laurent series terminates at finite place, denoted here by $-h$. The coefficient $A_h \neq 0$ and is given by

$$A_{-h} = \lim_{\lambda\to r^+} (\lambda - r)^h\mathcal{R}(\lambda). \tag{2.4.9}$$

Lemma 2.4.4. *The coefficients A_k satisfy the relation*

$$A_kA_m = \begin{cases} -A_{k+m+1} & \text{for} \quad k, m \geq 0, \\ 0 & \text{for} \quad k < 0, m \geq 0, \\ A_{k+m+1} & \text{for} \quad k, m < 0. \end{cases} \tag{2.4.10}$$

Proof. We may assume $r = 0$. We have the Cauchy formula

$$A_k = \frac{1}{2\pi i}\int\limits_C \frac{\mathcal{R}(\lambda)}{\lambda^{k+1}}d\lambda$$

where C is a circle enclosing 0, but not other eigenvalues, traversed counterclockwise.

Using two such circles C_1 and C_2 with radii $r_1 < r_2$, we obtain, using the resolvent identity,

$$\begin{aligned}
A_k A_m &= \frac{-1}{2^2\pi^2} \int\limits_{C_1} \int\limits_{C_2} \frac{\mathcal{R}(\lambda)\mathcal{R}(\mu)}{\lambda^{k+1}\mu^{m+1}} d\lambda d\mu \\
&= \frac{1}{2^2\pi^2} \int\limits_{C_1} \int\limits_{C_2} \frac{\mathcal{R}(\lambda)-\mathcal{R}(\mu)}{\lambda^{k+1}\mu^{m+1}(\lambda-\mu)} d\lambda d\mu \\
&= \frac{1}{2\pi i} \int\limits_{C_1} \frac{\mathcal{R}(\lambda)}{\lambda^{k+1}} \left(\frac{1}{2\pi i} \int\limits_{C_2} \frac{d\mu}{\mu^{m+1}(\mu-\lambda)} \right) d\lambda \\
&\quad + \frac{1}{2\pi i} \int\limits_{C_2} \frac{\mathcal{R}(\mu)}{\mu^{m+1}} \left(\frac{1}{2\pi i} \int\limits_{C_1} \frac{d\lambda}{\lambda^{k+1}(\mu-\lambda)} \right) d\mu
\end{aligned}$$

However, we have general formulae for integration counterclockwise along the circle $|z| = r$

$$\frac{1}{2\pi i} \int\limits_{C} \frac{dz}{z^j(z-w)} dz = \begin{cases} 0 & \text{if} \quad j \leq 0, |w| > r, \\ -\frac{1}{w^j} & \text{if} \quad j > 0, |w| > r, \\ \frac{1}{w^j} & \text{if} \quad j \leq 0, |w| < r, \\ 0 & \text{if} \quad j > 0, |w| < r. \end{cases}$$

and so the lemma follows by inspection of the formula for $A_k A_m$. E.g., if $k < 0$ and $m \geq 0$, then both final integrals are 0. If both $k, m < 0$, then

$$\frac{1}{2\pi i} \int\limits_{C_2} \frac{d\mu}{\mu^{m+1}(\mu-\lambda)} = \frac{1}{\lambda^{m+1}}$$

while the inner integral in the second term is 0, giving

$$A_k A_m = \frac{1}{2\pi i} \int\limits_{C_1} \frac{\mathcal{R}(\lambda)}{\lambda^{k+m+1}} d\lambda.$$

□

With this auxiliary result we can prove the following improvement of Theorem 2.4.3

Proposition 2.4.5. *The eigenspace $E_{r(\mathcal{A})}$ contains non-negative vectors.*

Proof. We use (2.4.8) composed with A_{-h}

$$\mathcal{R}(\lambda)A_{-h} = \sum_{k=-h}^{\infty} (\lambda - r)^k A_k A_{-h}.$$

Now, if $k \geq 0$, then $A_k A_{-h} = 0$ by the lemma. Otherwise, since $-l-h+1 \leq -h$, we obtain $A_{-l}A_{-h} = A_{-l-h+1} = 0$ unless $-l = 1$ as A_{-h} is the lowest non-zero term of the Laurent expansion. Thus

$$\mathcal{R}(\lambda)A_{-h} = (\lambda - r)^{-1}A_{-1}A_{-h} = (\lambda - r)^{-1}A_{-h}$$

or equivalently

$$\mathcal{A}A_{-h} = rA_{-h}$$

Since A_{-h} is non-negative and non-zero, each non-zero column of A_{-h} is a non-negative eigenvector of $\mathcal{A}$. □

These two results haven't answered the basic question: is there a dominant and uniquely determined pattern in long time evolution of a system governed by a non-negative matrix. In fact, there can be multiple eigenvectors corresponding to the eigenvalue $\lambda = r(\mathcal{A})$ as well as there is a possibility of complex eigenvalues on the circle $|\lambda| = r(\mathcal{A})$ which would give rise to oscillatory behaviour.

To make this step we have to demand more from the matrix $\mathcal{A}$. Let us at first assume that $\mathcal{A}$ is positive; that is, all entries are greater than 0. Under this assumption we have

Theorem 2.4.6. *Assume that $\mathcal{A} > 0$. Then $r(\mathcal{A}) > 0$ is a simple eigenvalue; that is, the eigenspace $E_{r(\mathcal{A})}$ is one-dimensional and is spanned by a positive eigenvector. Furthermore, $r(\mathcal{A})$ is dominant; that is, all other eigenvalues are smaller in absolute value.*

Proof. Assume $\mathcal{A} > 0$ and let $\mathbf{v}$ be a non-negative eigenvector belonging to the spectral radius (necessarily non-zero). Then $r(\mathcal{A})\mathbf{v} = \mathcal{A}\mathbf{v} > 0$, hence $r(\mathcal{A}) > 0$ and $\mathbf{v} > 0$. Thus, every non-zero (and, clearly, non-negative) column of A_{-h} is a strictly positive eigenvector belonging to $r(\mathcal{A})$. On the other hand, $A_{-h}A_{-h} = A_{-2h+1} = 0$ if $h > 1$ which is impossible as A_{-h} has strictly positive columns. Thus $h = 1$ - the pole at $\lambda = r(\mathcal{A})$ has order 1.

Furthermore, the non-zero elements of the range of A_{-1} (spanned by non-zero columns of A_{-1}) are eigenvectors of $\mathcal{A}$ belonging to $\lambda = r(\mathcal{A})$. Conversely, let $\mathbf{v}$ be an eigevector of $\mathcal{A}$ associated with $r(\mathcal{A})$; that is, $\mathcal{A}\mathbf{v} = r(\mathcal{A})\mathbf{v}$. Hence, for λ close to $r(\mathcal{A})$ we obtain

$$\mathcal{R}(\lambda)\mathbf{v} = (\lambda - r(\mathcal{A}))^{-1}\mathbf{v}$$

and, by the uniqueness of the Laurent expansion,

$$A_{-1}\mathbf{v} = r(\mathcal{A})\mathbf{v},$$

hence the eigenspace of $\mathcal{A}$ associated with $\lambda = r(\mathcal{A})$ coincides with the range of A_{-1} and its dimension is 1. In fact, if $\mathbf{v}^1$ and $\mathbf{v}^2$ be two non-zero (hence necessarily positive) columns which are not scalar multiples of each other, then there would be first t for which $\mathbf{v}^1 - t\mathbf{v}^2$ would have one entry equal to zero an other non-negative. However, such a combination is a non-negative eigenvector and, by the preceding argument, must be strictly positive. So, the eigenspace is one-dimensional.

This argument does not rule out $r(\mathbf{A})$ having associated eigenvectors. For this, there would be $\mathbf{v} = (r(\mathcal{A})I - \mathcal{A})\mathbf{x}$ belonging to $r(\mathcal{A})$. However, then

$$\mathbf{v} = A_{-1}\mathbf{v} = A_{-1}(r(\mathcal{A})I - \mathcal{A})\mathbf{x} = (r(\mathcal{A})I - \mathcal{A})A_{-1}\mathbf{x} = 0$$

as the range of A_{-1} consists of eigenvectors. However, $\mathbf{v} \neq 0$ which shows that $r(\mathcal{A})$ is a simple eigenvalue.

Finally, consider $\mathcal{A} - \epsilon I$ for ϵ small enough so that $\mathcal{A} - \epsilon I > 0$. Clearly, the largest positive eigenvalue is $r(\mathcal{A}) - \epsilon$ which, by Theorem 2.4.3, is its spectral radius. Hence, all other eigenvalues of $\mathcal{A} - \epsilon I$ are within the (possibly closed) circle with radius $r(\mathcal{A}) - \epsilon$. Translating this circle back, we obtain that all eigenvalues of $\mathcal{A}$ are within the circle with centre at ϵ and radius $r(\mathcal{A}) - \epsilon$ which is tangent to the circle $|\lambda| = r(\mathcal{A})$ only at $r(\mathcal{A})$. □

We note that if λ is an eigenvalue of $\mathcal{A}$, then λ^k is an eigenvalue of $\mathcal{B} = \mathcal{A}^k$ with the same eigenvector. This follows from

$$(\lambda^k I - \mathcal{A})\mathbf{x} = \left(\sum_{i=0}^{k-1} \lambda^i \mathcal{A}^{k-i}\right)(\lambda I - \mathcal{A})\mathbf{x}. \qquad (2.4.11)$$

Conversely, to any $\mu \in \sigma(\mathcal{B})$ there corresponds $\lambda \in \sigma(\mathcal{A})$ such that $\lambda^k = \mu$. In fact, denote by μ_j, $j = 0, 1, \ldots, k-1$ the k-th roots of μ so that

$$\mathcal{A}^k - \mu I = \prod_{j=0}^{k-1} (\mathcal{A} - \mu_j I).$$

Let $\mathbf{v}_0$ an eigenvector of $\mathcal{A}$ belonging to μ and consider $\mathbf{v}_1 = (\mathcal{A} - \mu_0 I)\mathbf{v}_0$. If $\mathbf{v}_1 = 0$, then we are done, if $\mathbf{v}_1 \neq 0$, then we consider $\mathbf{v}_2 = (\mathbf{A} - \mu_1 I)\mathbf{v}_1$. Since after k steps we obtain zero, there must be k_0 such that $\mathbf{v}_{k_0} \neq 0$ but $(\mathcal{A} - \mu_{k_0} I)\mathbf{v}_{k_0} = 0$ so that μ_{k_0} is an eigenvalue of $\mathcal{A}$.

Theorem 2.4.7. *The statement of Theorem 2.4.6 remains valid if $\mathcal{A}$ in non-negative with $\mathcal{A}^k > 0$ for some k.*

Proof. Denote $\mathcal{B} = \mathcal{A}^k > 0$. Clearly $r(\mathcal{B}) = r(\mathcal{A})^k$. Then $\mathcal{B}$ has a simple positive eigenvalue equal to its spectral radius with associated positive eigenvector with all other ($n-1$, properly counted) eigenvalues having strictly smaller moduli. By the comment above, $r(\mathcal{A})$ must be also a strictly dominant eigenvalue. Indeed, if λ is another eigenvalue with $|\lambda| = r(\mathcal{A})$, then

$|\lambda^k| = r(\mathcal{B})$ and there are two possibilities. First, that $\lambda^k \neq r(\mathcal{B})$ but this would contradict $r(\mathbf{B})$ being dominant. If $\lambda^k = r(\mathbf{B})$ then eigenvectors corresponding to $r(\mathcal{A})$ and λ would be eigenvectors of $\mathcal{B}$ by (2.4.11) and, since eigenvectors corresponding to different eigenvalues are linearly independent, we would have the eigenspace corresponding to $r(\mathcal{B})$ at least two-dimensional.

Moreover, if the eigenspace corresponding to $r(\mathcal{A})$ was multi-dimensional, then also the eigenspace of $\mathcal{B}$ corresponding to $\lambda = r(\mathcal{B})$ would be multi-dimensional. Finally, if there was a vector $\mathbf{x}$ such that

$$\mathbf{v} = (r(\mathcal{A})I - \mathcal{A})\mathbf{x}$$

was an eigenvector, then, using the fact that no other k-th root of $r(\mathcal{B})$ is an eigenvalue of $\mathcal{A}$, we obtain that $(\lambda^k - \mathcal{A}^k)\mathbf{x} = \mathbf{v}' \neq 0$. On the other hand, by commutativity, $(\lambda^k - \mathcal{A}^k)^2\mathbf{x} = 0$, which contradicts simplicity of λ^k. □

Remark 2.4.8. Frobenius-Perron Theorem often is formulated in terms of irreducibility and primitivity of matrices: these concepts are more 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. A *loop* is a path from i to itself. 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. 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 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.

An irreducible matrix is called *primitive* if the greatest common divisor divisor of all loops is 1; otherwise it is called *imprimitive*. In population dynamics, a matrix is imprimitive if the population has a single reproductive stage. E.g., the matrix

$$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

describing a semelparous population is imprimitive.

For irreducible matrices the first part of Theorem 2.4.6 is valid; that is, $r(\mathcal{A})$ is a simple positive eigenvalue with strictly positive eigenvector. To see this, we note that the eigenvectors are non-negative. Take such an eigenvector $0 \neq \mathbf{v} \geq 0$ and consider

$$\mathcal{A}\mathbf{v} = r(\mathcal{A})\mathbf{v}.$$

Let $v_i \neq 0$. If $r(\mathcal{A}) = 0$, then, since all entries are non-negative, $\mathcal{A}\mathbf{v} = 0$ yields $a_{ji} = 0$ for any j and thus there can be no path leading from state i to any other state. Similarly, if $\mathbf{v}$ is non-negative but not strictly positive, then $v_i > 0$ and $v_j = 0$ for some i, j. Suppose there is a path from i to j, then, for some m, $(A^m)_{ji} > 0$ and thus $v_j > 0$ which is a contradiction. Thus, $\mathbf{v} > 0$ and all eigenvectors in this case are positive hence the proof goes as in the first part of Theorem 2.4.6. However, one cannot rule out eigenvalues on the spectral circle.

Primitive matrices can have zero entries. E.g., the matrix

$$\begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

is primitive.

It can be proved that an irreducible matrix $\mathcal{A}$ is primitive if and only if $\mathcal{A}^k > 0$ for some k.

2.4.5 Examples

Let us consider the Leslie matrix

$$\mathcal{L} := \begin{pmatrix} f_1 & f_2 & \cdots & f_{n-1} & f_n \\ s_2 & 0 & \cdots & 0 & 0 \\ 0 & s_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & s_n & 0 \end{pmatrix}, \tag{2.4.12}$$

and find under what conditions the population exhibits asynchronous exponential growth.

Let us first assume that $f_j > 0$ for $j = 1, \ldots, n$, that is, any age group is capable of reproduction. Let us consider arbitrary initial state j. Then there is an arc between j and 1 ($a_{1j} = f_j > 0$) and then from state 1 one can reach any state i in exactly $i - 1$ steps ($s_2 s_3 \cdot \ldots \cdot s_i$). Thus, there is a path joining j and i of length i which still depends on the target state. However, there is an arc from 1 to itself, so we can wait at 1 for any number of steps. In particular we can wait for $n - i$ steps so that j can be connected with i is n steps. In other words

$$s_i s_{i-1} \cdot \cdots \cdot s_3 s_2 f_1 \cdot \ldots \cdot f_1 f_j > 0$$

where f_1 occurs $n-i$ times. Hence $\mathcal{L}^n > 0$.

This result assumes too much - typically young individuals cannot reproduce. If fact, it can be strengthened - it suffices to assume that $f_{n-1} > 0$ and $f_n > 0$. First we observe that $[\mathcal{L}^{n-1}]_{11} >$ and $[\mathcal{L}^n]_{11} > 0$ as starting from 1 we return to 1 either by following $1 \to 2 \to \ldots n-1 \to 1$ (which gives loop of length $n-1$) $1 \to 2 \to \ldots n-1 \to n \to 1$ (which gives loop of length n). Next, to pass from state j to state i with $j \geq i$ we take $n-j$ steps to reach n, one step to go to 1 and $i-1$ steps to reach from 1 to i. If $i > j$, then we need $i-j$ steps but we can add the full loop of n steps and in both cases we can reach i from j in $n-j+i$ steps. We have to show that we must be able to cycle at 1 appropriate number of steps k to eliminate the dependence on j and i. Since $1 \leq i, j \leq n$, we have $-n+1 \leq -j+i \leq n-1$. Clearly that $[\mathcal{L}^k]_{11} > 0$ for any

$$k = \alpha(n-1) + \beta n \tag{2.4.13}$$

with natural, or 0, α and β.

We observe that if $\alpha = \left[\frac{k}{n-1}\right] n - k$ and $\beta = k - \left[\frac{k}{n-1}\right](n-1)$, where $[\cdot]$ denotes the integer part of a number, then the above identity is satisfied. However, it is not clear that $\alpha > 0$. We write $k = r(n-1) + s$ with integer r and integer s satisfying $0 \leq s < n-1$ so that $\left[\frac{k}{n-1}\right] = r$. Thus, we must have

$$r \geq r\left(1 - \frac{1}{n}\right) + \frac{s}{n-1}$$

for any s as above, so that we get $r \geq n-1$ and the representation is valid for any $k \geq (n-1)^2$

Then, we write

$$n - j + i + k = n - j + i + (n-1)^2 + l = n^2 - n - j + i + l$$

for some $l \geq 0$ and observe that $n + j - i \geq 0$ so that taking $l = n + j - i$ and $k = (n-1)^2 + l$ we see that

$$n - j + i + k = n^2$$

for some k expressible through (2.4.13) Thus, one can reach any i from any j in n^2 steps: $n-j$ steps from j to n, 1 step to state 1, 'cycling' for k times around 1 and then going from 1 to i in $i-1$ steps.

Let us consider a more complicated case where the fertility is restricted to some interval $[n_1, n_2]$, that is, when $b_j > 0$ for $j \in [n_1, n_2]$. First we note that if $n_2 < n$, the matrix cannot be primitive as there is no communication between postreproductive stages and the reproductive ones. This can be also check by direct multiplication of matrices as the last column of any power will be always zero. 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 with slight modification which we shall explain 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

$$\tilde{\mathcal{L}} = \begin{pmatrix} f_1 & f_2 & \cdots & f_{n_2-1} & f_{n_2} \\ s_2 & 0 & \cdots & 0 & 0 \\ 0 & s_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & \cdots & s_{n_2} & 0 \end{pmatrix}$$

and the matrix providing (one-way) link from reproductive to postreproductive stages is given by

$$\mathcal{R} = \begin{pmatrix} 0 & \cdots & s_{n_2+1} & 0 & \cdots & 0 & 0 \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & s_n & 0 \end{pmatrix}$$

For the matrix $\tilde{\mathcal{L}}$, $f_{n_2} > 0$ and $f_{n_2-1} > 0$ and we can apply considerations of the previous part. Thus, for some $\lambda > 0$ there are sequences $\mathbf{x}^* = (x_1^*, \dots x_{n_2}^*)$ and $\mathbf{y}^* = (y_1^*, \dots y_{n_2}^*)$ such that $\tilde{\mathcal{L}}\mathbf{x}^* = \lambda \mathbf{x}^*$ and

$$\lim_{k\to\infty} \lambda^{-k}\tilde{\mathcal{L}}\mathbf{x} = \mathbf{x}^* < \mathbf{y}^*, \mathbf{x} >, \quad \mathbf{x} \in \mathbb{R}^{n_2}. \tag{2.4.14}$$

For $n_2 \leq j < n, k \geq 0$, we have $x_{j+1}(k+1) = s_{k+1}x_j(k)$. Hence, starting from $x_{n_2}(k)$ we get $x_{n_2+i}(k+i) = c_i x_{n_2}(k)$, where $c_i = s_{n_2+i}s_{n_2+1-i} \cdot \ldots \cdot s_{n_2+1}$, as long as $i \leq n - n_2$. So

$$\lim_{k\to\infty} \lambda^{-k} x_{n_2+i}(k+i) = c_i x_{n_2}^* < \mathbf{y}^*, \mathbf{x} >, \quad \mathbf{x} \in \mathbb{R}^{n_2},$$

and hence, changing $k+i$ into k

$$\lim_{k\to\infty} \lambda^{-k} x_{n_2+i}(k) = c_i \lambda^{-i} x_{n_2}^* < \mathbf{y}^*, \mathbf{x} >, \quad \mathbf{x} \in \mathbb{R}^{n_2},$$

for any $i = 1, \dots, n - n_2$.

Hence, we see that the formula (2.3.57) is satisfied if we take

$$\begin{aligned} \mathbf{x}^* &= (x_1^*, \dots x_{n_2}^*, c_1\lambda^{-1}x_{n_2}^*, \dots, c_{n-n_2}\lambda^{-(n-n_2)}x_{n_2}^*) \\ \mathbf{y}^* &= (y_1^*, \dots y_{n_2}^*, 0, \dots, 0). \end{aligned}$$

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 2.3.17.

Finally, let us consider the continuous model

$$v_i'(t) = \sum_{j=1}^{N} q_{ij} v_j(t), \quad , i = 1, \ldots, n, \tag{2.4.15}$$

with non-negative off-diagonal entries. It turns out that to get asynchronous exponential growth it is enough to assume a weaker property of $\mathcal{Q}$, namely *irreducibility*. In graph terms it mean each state communicate with any other but not necessarily in the same number of steps, as was the case for primitive matrices. Algebraically, we assume that for $i \neq j$, $q_{ij} \geq 0$ and there is a sequence of indices $i_1, \ldots, i_m$ such that

$$a_{ii_1} a_{i_1 i_2} \cdot \ldots \cdot a_{i_{m-1} i_m} a_{i_m j} > 0. \tag{2.4.16}$$

To prove this statement first we observe that $\tilde{\mathcal{Q}} = \mathcal{Q} + r\mathcal{I}$, where $r > \max_{1 \leq i \leq N}\{|q_{ii}|\}$, has all its diagonal entries positive. Arguing as in the discrete case, we see that $\tilde{\mathcal{Q}}$ is primitive. Indeed, we can go from j to 1 in m_{j1} steps and from m_{1i} steps. Let $M = \max_{i,j}\{m_{j1}, m_{1i}\}$. If for some (j', i') we have $m_{j'1} + m_{1i'} < M$, then we can wait in state 1 for any number of steps as $q_{11} + r > 0$. Thus, $\tilde{\mathcal{Q}}^M > 0$. Hence, $\tilde{\mathcal{Q}}$ has a simple dominant eigenvalue $\tilde{\lambda}$ and the corresponding positive eigenvector $\mathbf{x}^*$ as well as positive eigenvector $\mathbf{y}^*$ of the transposed matrix. Therefore, as in (2.3.58), we have

$$\lim_{t \to \infty} e^{-\tilde{\lambda} t} e^{t\tilde{\mathcal{Q}}} \mathbf{x} = \mathbf{x}^* < \mathbf{y}^*, \mathbf{x} >,$$

but $\tilde{\lambda} = \lambda + r$ and $e^{t\tilde{\mathcal{Q}}} = e^{rt} e^{t\mathcal{Q}}$, where λ is simple eigenvalue larger then real part of any other eigenvalue. Indeed, $\tilde{\lambda}$ is real and positive with all other eigenvalues $\tilde{\lambda}'$ of $\tilde{\mathcal{Q}}$ satisfying $\tilde{\lambda} \geq |\tilde{\lambda}'|$. But then $\tilde{\lambda} > \Re\tilde{\lambda}'$ as the line $\Re z = \tilde{\lambda}$ is tangent to the circle $|z| \leq \tilde{\lambda}$ and $\tilde{\lambda}$ is simple. Thus, the same is true for $\lambda = \tilde{\lambda} - r$. Hence, by (2.4.1),

$$\mathbf{x}^* < \mathbf{y}^*, \mathbf{x} >= \lim_{t \to \infty} e^{-\tilde{\lambda} t} e^{t\tilde{\mathcal{Q}}} \mathbf{x} = \lim_{t \to \infty} e^{-\lambda t} e^{-tr} e^{tr} e^{t\mathcal{Q}} \mathbf{x} = \lim_{t \to \infty} e^{-\lambda} e^{t\mathcal{Q}} \mathbf{x} \tag{2.4.17}$$

2.5 Other ways of modelling age-structured populations

2.5.1 Continuous time - Lotka integral equation

We track female births in time (though we can also track males or pairs). Let $B(t)$ denotes the birth rate; that is,

> $B(t)\Delta t$ is approximately the number of female births that occur in the time interval $[t, t + \Delta t)$

We shall need several parameters of the population to derive the model. First, let

> $n(a, t)$ - the density of females of age a at time t; that is, $n(a, t)\Delta a$ is the number of females of age $[a, a + \Delta a)$.

Next, we must have information about the age-specific survivorship and fertility. We let

> $l(a) =$ fraction of newborn females surviving to age a.

Here we assume that $l(a)$ is continuous and piecewise differentiable. Furthermore, l must be non-increasing and there is some maximum age of survivorship ω. Finally, we denote

> $m(a)\Delta a =$ number of females born, on average, to a female of age between a and $a + \Delta a$.

The rate $m(a)$ is referred to as a *maternity function.* We assume that m is continuous and piecewise smooth and that there is a minimum age of reproduction α (menarche) and a maximum age of reproduction β (menopause).

Let us derive the Lotka equation for B. The number of birth in the time interval $[t, t+\Delta t)$ is, as we know, $B(t)\Delta t$. On the other hand, the births can be divided into two classes: one class attributed to females born between time 0 and t and the other due to females which were alive at time 0. Females that are of age a at time t were born at time $t - a$. The number of females born around $t - a$ is given by $B(t - a)\Delta t$. The number of them that survive till the age a (that is, till time t) is $l(a)B(t - a)\Delta t$ and thus the number of births by females of age circa a is $l(a)B(t - a)\Delta t m(a)\Delta a$. Summing up, we obtain

$$\Delta t \int\limits_0^t B(t-a)l(a)m(a)da.$$

To find the contribution of the females who were present at time $t = 0$ we begin with taking the number of females of age c. a present at $t = 0$; that is, $n(a, 0)\Delta a$. Now, these females must live till the age $t + a$, that is we must take survival rate till $t+a$, $l(t+a)$ conditioned upon the female having survived till a. Since

$$l(a + t) = l(a) \cdot \{\text{fraction of age } a \text{ females surviving till } t + a\}$$

we see that $n(a, 0)l(a + t)/l(a)\Delta a$ females survived till time t. These gave birth to $m(a+t)n(a, 0)l(a+t)/l(a)\Delta a$ new females. To find the number of all births due to females older then t we again integrate over all ages. However, no individual survives beyond ω so the integration terminates at $\omega - t$ (no female older that $\omega - t$ at time $t = 0$ will survive till t. Combining these two formulae and dropping Δt we obtain the basic Lotka *renewal* equation

$$B(t) = \int\limits_0^t B(t - a)l(a)m(a)da + G(t) \tag{2.5.1}$$

where

$$G(t) = \int\limits_0^{\omega - t} m(a + t)n(a, 0)\frac{l(a + t)}{l(a)}da, \tag{2.5.2}$$

is a known function.

Note, that the 'age' a in both integrals is measured from 0. To be able to think about it in a continuous way, we have to change $a + t \to a$ and in this way we obtain

$$G(t) = \int\limits_t^{\omega} m(a)n(a - t, 0)\frac{l(a)}{l(a - t)}da. \tag{2.5.3}$$

Previous considerations allow to express the population density as

$$n(a, t) = \begin{cases} n(a - t, 0)\frac{l(a)}{l(a-t)} & \text{for} \quad a > t, \\ B(t - a)l(a) & \text{for} \quad a < t, \end{cases} \tag{2.5.4}$$

and thus we also can express B in terms of n by combining both integrals in (2.6.14)

$$B(t) = \int\limits_0^{\omega} n(a, t)m(a)da. \tag{2.5.5}$$

2.5.2 Discrete time - Lotka difference equation

We can obtain similar equation in discrete time. In analogy with the previous subsection

$$\begin{aligned} B_t &= \text{number of female births at time } t \\ n_{a,t} &= \text{number of females of age a at time } t \\ l_a &= \text{fraction of females surviving from birth to age } a \\ m_a &= \text{number of females born on average to a female of age } a. \end{aligned}$$

Here, a and t are integer valued. Arguing as before, we obtain the discrete renewal equation

$$B_t = \sum_{a=1}^{t} B_{t-a} l_a m_a + G_t, \tag{2.5.6}$$

where

$$G_t = \sum_{a=1}^{\omega-t} n_{a,0} \frac{l_{a+t}}{l_a} m_{a+t} = \sum_{a=t+1}^{\omega} n_{a-t,0} \frac{l_a}{l_{a-t}} m_a.$$

Using these notation, we can also write the difference equation for the age distribution $n_{a,t}$. We assume that births occur at $t+1$ if females (of one age class or more) survive from time t and reproduce. Starting from $n_{a,t}$ we project to $n_{a,t+1}$ as follows. For neonates we have

$$n_{0,t+1} \quad = \quad n_{0,t} l_1 m_1 + n_{1,t} \frac{l_2}{l_1} m_2 + \ldots + n_{\omega-1,t} \frac{l_\omega}{l_{\omega-1}} m_\omega,$$

whereas older females ($a > 0$) will be found at time $t+1$ only if females of age $a-1$ survived; that is

$$n_{a,t+1} = n_{a-1,t} \frac{l_a}{l_{a-1}}, \quad a = 1, \ldots, \omega - 1.$$

Denoting $\mathbf{n}_t = (n_{0,t}, \ldots, n_{\omega-1,t})$, we obtain an evolution governed by Leslie matrix

$$\mathbf{n}_{t+1} = \mathcal{L} \mathbf{n}_t$$

where

$$\mathcal{L} := \left(\begin{array}{ccccc} l_1 m_1 & m_2 l_2/l_1 & \cdots & m_{\omega-1} l_{\omega-1}/l_{\omega-2} & m_\omega l_\omega / l_{\omega-1} \\ l_1 & 0 & \cdots & 0 & 0 \\ 0 & l_2/l_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & l_{\omega-1}/l_{\omega-2} & 0 \end{array} \right). \tag{2.5.7}$$

2.5.3 McKendrick-von Vorester partial differential equation

Yet another way of modelling age-structured population is to look at the population as if it was 'transported' through stages of life. Taking into account that $n(a,t)\Delta a$ is the number of females in the age group $[a, a+\Delta a)$ at time t, we may write that the rate of change of this number

$$\frac{\partial}{\partial t}[n(a,t)\Delta a]$$

equals rate of entry at a minus rate of exit at $a + \Delta a$ -deaths. Denoting per capita mortality rate for individuals by $\mu(a,t)$, the last term is simply $-\mu(a,t)n(a,t)\Delta t$. The first two terms require introduction of the 'flux' of individuals J describing 'speed' of ageing. Thus, passing to the limit $\Delta a \to 0$, we get

$$\frac{\partial n(a,t)}{\partial t} + \frac{\partial J(a,t)}{\partial a} = -\mu(a,t)n(a,t).$$

Let us determine the flux in the simplest case when ageing is just the passage of time measured in the same units. Then, if the number of individuals in the age bracket $[a, a+\Delta a)$ is $n(a,t)\Delta a$, then after Δt we will have $n(a, t+\Delta t)\Delta a$. On the other hand, $u(a-\Delta t, t)\Delta t$ individuals moved in while $u(a+\Delta a - \Delta t)\Delta t$ moved out, where we assumed, for simplicity, $\Delta t < \Delta a$. Thus

$$n(a,t+\Delta t)\Delta a - n(a,t)\Delta a = u(a-\Delta t, t)\Delta t - u(a+\Delta a - \Delta t)\Delta t - \mu(a,t)n(a,t)\Delta a\Delta t$$

or, using the Mean Value Theorem $(0 \leq \theta, \theta' \leq 1)$

$$n_t(a, t+\theta\Delta t)\Delta a\Delta t = -n_a(a+\theta'\Delta a, t)\Delta a\Delta t - \mu(a,t)n(a,t)\Delta a\Delta t$$

and, passing to the limit with $\Delta t, \Delta a \to 0$, we obtain

$$\frac{\partial n(a,t)}{\partial t} + \frac{\partial n(a,t)}{\partial a} = -\mu(a,t)n(a,t). \tag{2.5.8}$$

This equation is defined for $a > 0$ and the flow is to the right hence we need a boundary condition. In this model the birth rate enters here: the number of neonates $(a = 0)$ is the number of births across the whole age range:

$$n(0,t) = \int\limits_0^{\omega} n(a,t)m(a,t)da,$$

where m is the maternity function. Eq. (2.5.8) also must be supplemented by the initial condition

$$n(a,0) = n_0(a)$$

describing the initial age distribution.

2.6 Rudimentary mathematical analysis of the models

2.6.1 Discrete Lotka equation

In this subsection we analyse the equation

$$B_t = \sum_{a=1}^{t} B_{t-a} l_a m_a + G_t, \tag{2.6.1}$$

where

$$G_t = \sum_{a=1}^{\omega-t} n_{a,0} \frac{l_{a+t}}{l_a} m_{a+t} = \sum_{a=t+1}^{\omega} n_{a-t,0} \frac{l_a}{l_{a-t}} m_a.$$

We assume that there is largest survival rate ω and the minimum and maximum reproduction ages (α and β, respectively). Looking at (2.6.1), we see that it splits into two classes $t \geq \beta$ and $t < \beta$. For $t > \omega \geq \beta$, the inhomogeneity satisfies $G_t = 0$ (in fact, summation goes as $\beta+1, \beta+2, \ldots$ and $m_{1+\beta} = m_{2+\beta} = \ldots = 0$. Hence the problem reduces to

$$B_t = \sum_{a=\alpha}^{\beta} B_{t-a} l_a m_a + G_t, \tag{2.6.2}$$

which is just a homogeneous linear difference equation of order $n = \beta - \alpha + 1$.

Long time behaviour

If we are only interested in long time behaviour of the solution, we can focus on (2.6.2). In this case, we know that the solution is spanned by n linearly independent solutions determined by solutions of the characteristic equation which, in the above notation, takes the form

$$\sum_{a=\alpha}^{\beta} \lambda^{-a} l_a m_a = 1. \tag{2.6.3}$$

Remark 2.6.1. We observe that the characteristic equation of the Leslie matrix (2.5.7), associated with this model, is

$$\lambda^{\omega} - \sum_{a=1}^{\beta} \lambda^{\omega-a} l_a m_a = 0$$

which is exactly (2.6.3) if we take into account menarche and menopause and divide by λ^{ω}. Thus, analysis can be done by Frobenius-Perron theory but in this case it is instructive do the work from scratch. Also, as we will see in the forthcoming example, validity of Eq. (2.6.3) can be stretched to include more general situations.

Example 2.6.2. The Northern Spotted Owl has the following characteristics:

$$m_a = \begin{cases} 0 & \text{for} \quad a < 3, \\ 0.24 & \text{for} \quad a \geq 3, \end{cases}$$

$l_3 = 0.0722$ and $P = l_{a+1}/l_a = 0.942$ for $a \geq 3$ hence, in principle, we have infinitely many reproductive classes. nevertheless, it turns out that (2.6.3) still makes sense: denoting $m_3 = m$ for $a \geq 3$, we have

$$\begin{aligned} 1 &= \sum_{a=3}^{\infty} \lambda^{-a} l_a m_a = \frac{l_3 m}{\lambda^3} + \frac{l_3 P m}{\lambda^4} + \frac{l_3 P^2 m}{\lambda^5} + \ldots \\ &= \frac{l_3 m}{\lambda^3} \sum_{a=0}^{\infty} \left(\frac{P}{\lambda}\right)^a \end{aligned}$$

$$= \frac{l_3 m}{\lambda^3} \frac{1}{1 - P/\lambda},$$

which can be re-written as

$$\lambda^3 - P\lambda^2 - l_3 m = 0.$$

Using, say, *Mathematica*, we get $\{\lambda \to 0.960772\}, \{\lambda \to -0.00938594 + 0.133968i\}$ and $\{\lambda \to -0.00938594 - 0.133968i\}$. We note the the complex roots are not the roots of the original equation as for them $P/|\lambda| > 1$ and the series would be divergent. So, $\lambda = 0.960772$ is the only root outside $|\lambda| = P$.

Sensitivity analysis The dominant real eigenvalue gives an indication of the rate of growth for large times of the population. The characteristic equation can be used to find out how sensitive this parameter is with respect to environmental changes. The dominant eigenvalue can be thought of as a function of the parameters P, l_3, m determined implicitly through the equation

$$\lambda^3(P, l_3, m) - P\lambda^2(P, l_3, m) - l_3 m = 0.$$

Sensitivity of a function with respect to a parameter is given by the value of the partial derivative of the function with respect to this parameter. In this case we can find the derivatives differentiating the above equation implicitly. Thus,

$$3\lambda^2 \frac{\partial \lambda}{\partial P} - \lambda^2 - P 2\lambda \frac{\partial \lambda}{\partial P} = 0$$

or

$$\frac{\partial \lambda}{\partial P} = \frac{\lambda^2}{3\lambda^2 - 2P\lambda}$$

which, evaluated at $P = 0.942$ and $\lambda = 0.96$, gives

$$\frac{\partial \lambda}{\partial P} = 0.962.$$

In the same way,

$$\frac{\partial \lambda}{\partial l_3} = 0.254$$

and

$$\frac{\partial \lambda}{\partial m} = 0.075.$$

We see that the growth rate of the population is most sensitive to changes in adult annual survival, less to so to survival to breeding stage, and only lastly to average reproductive rate.

Let us turn our attention to question what we can say in general about the roots of the equation (2.6.3). We prove the following result.

Proposition 2.6.3. *Equation (2.6.3) has exactly one positive real root, $\lambda = \lambda_0$, of algebraic multiplicity 1.*

Proof. We define

$$\psi(\lambda) = \sum_{a=\alpha}^{\beta} \lambda^{-a} l_a m_a \tag{2.6.4}$$

so that (2.6.3) reads

$$\psi(\lambda) = 1.$$

Since all coefficients are non-negative we have

$$\lim_{\lambda \to 0} \psi(\lambda) = \infty$$

and

$$\lim_{\lambda \to \infty} \psi(\lambda) = 0.$$

Furthermore, ψ is strictly decreasing:

$$\frac{d\psi(\lambda)}{d\lambda} = -\sum_{a=\alpha}^{\beta} a\lambda^{-a-1} l_a m_a < 0.$$

So, ψ is continuous, strictly decreasing and thus it can cross any horizontal line in the upper half-plane, including $\psi(\lambda) = 1$, only once. In particular, there is only one positive real solution $\lambda = \lambda_0$ of $\psi(\lambda) = 1$. The proposition is proved. □

All other roots λ_j are either complex or negative and if, for instance, they are simple, we can write

$$B_t = c_0 \lambda_0^t + \sum_{j=1} c_j \lambda_j^t, \quad t \geq \beta.$$

However, as we noted, solutions to $\psi(\lambda) = 1$ are the eigenvalues of the corresponding Leslie matrix $\mathcal{L}$ and, in general, if $\mathcal{L}$ is not primitive, there are other eigenvalues λ satisfying $|\lambda| = \lambda_0$. In the current context, we can improve the statement of the Frobenius-Perron theorem.

Consider the maternity function (sequence) $m(a)$ ad assume that for some as, $m(a) = 0$. Consider the set of all ages a for which $m(a) > 0$ and let d be the greatest common divisor of these ages. If $d > 1$, we say that the maternity function is *periodic* with period d. Otherwise, m_a is *aperiodic*.

Proposition 2.6.4. *No other root λ_j can be greater than λ_0 in modulus. If the maternity function is periodic with period d, then there are $d-1$ other eigenvalues with the same modulus as λ_0. Otherwise, λ_0 is strictly dominant eigenvalue.*

Proof. Let us suppose that

$$\lambda_j = |\lambda_j|e^{i\theta}, \quad \theta \neq 2\pi n,$$

is a negative or complex root to $\psi(\lambda) = 1$. Then,

$$\sum_{a=\alpha}^{\beta} |\lambda_j|^{-a} e^{-ia\theta} l_a m_a = 1 \tag{2.6.5}$$

or, taking real and imaginary part,

$$\sum_{a=\alpha}^{\beta} |\lambda_j|^{-a} \cos(a\theta) l_a m_a \quad = \quad 1 \tag{2.6.6}$$

$$\sum_{a=\alpha}^{\beta} |\lambda_j|^{-a} \sin(a\theta) l_a m_a \quad = \quad 0 \tag{2.6.7}$$

Now, if $m(a)$ is periodic, then the only nonzero terms correspond to multiples of d. Taking $\theta_n = 2\pi n/d$, $n = 0, 1, \ldots, d-1$, we see $\cos a\theta_n = 1$, $\sin a\theta_n = 0$ and so, if the above equations are satisfied by λ_0, they are also satisfied for any $\lambda_n = \lambda_0 e^{\theta_n}$, so that the second statement if proved. If $m(a)$ is aperiodic, then for some as we have $\cos a\theta < 1$. But then, if (2.6.6) is satisfied, we must have

$$\sum_{a=\alpha}^{\beta} |\lambda_j|^{-a} l_a m_a > 1,$$

On the other hand, since

$$\sum_{a=\alpha}^{\beta} \lambda_0^{-a} l_a m_a = 1,$$

we obtain $|\lambda_j| < \lambda_0$, thus proving the first and third assertion of the proposition. □

Example 2.6.5. Consider semelparous reproduction defined by

$$m_a = \begin{cases} 0 & \text{for} \quad a = 1, \\ 0 & \text{for} \quad a = 2, \\ 6 & \text{for} \quad a = 3, \end{cases}$$

and

$$l_a = \begin{cases} 1 & \text{for} \quad a = 1, \\ 1/2 & \text{for} \quad a = 2, \\ 1/6 & \text{for} \quad a = 3. \end{cases}$$

The greatest common divisor is 3. The Euler-Lotka equation reduces to

$$\frac{1}{\lambda^3} = 1$$

so that

$$\lambda_0 = 1, \lambda_{1,2} = -\frac{1}{2} \pm \frac{\sqrt{3}}{2} i,$$

with all roots of modulus 1. However, if we introduce immature reproduction

$$m_a = \begin{cases} 1/4 & \text{for} \quad a = 1, \\ 0 & \text{for} \quad a = 2, \\ 6 & \text{for} \quad a = 3, \end{cases}$$

and

$$l_a = \begin{cases} 1 & \text{for} \quad a = 1, \\ 1/2 & \text{for} \quad a = 2, \\ 1/6 & \text{for} \quad a = 3, \end{cases}$$

we have aperiodic maternity function and the Euler-Lotka equation is

$$\frac{1}{4}\frac{1}{\lambda} + \frac{1}{\lambda^3} = 0,$$

yielding $\lambda_0 = 1.09$, $\lambda_{1,2} = -0.42 \pm 0.86i$ with $|\lambda_{1,2}| = 0.957$. Hence, the single positive root is dominant.

For aperiodic maternity schedule (and simple roots)

$$B_t = c_0\lambda_0^t + \sum c_j \lambda_j^t = c_0\lambda_0^t \left(1 + \sum \frac{c_j}{c_0} \left(\frac{\lambda_j}{\lambda_0}\right)^t\right),$$

where the sum is finite. If λ_j are not simple, then solutions are of the form $\sum \lambda_j^{t-k} t^k$ and also are of lower order than λ_0^t for large t. Thus,

$$B_t \approx c_0 \lambda_0^t$$

as $t \to \infty$. After the death of founder females, the number of females in each age class is given by

$$n_{a,t} = B_{t-a} l_a$$

so that

$$n_{a,t} \approx c_0\lambda_0^t \left(\frac{l_a}{\lambda_0^a}\right). \tag{2.6.8}$$

Now, it can be proved (Tutorial problems) that the stable age distribution for a Leslie matrix is

$$\left(\frac{l_1}{\lambda_0}, \frac{l_2}{\lambda_0^2}, \ldots, \frac{l_\omega}{\lambda_0^\omega}\right)$$

so that the fraction in (2.6.8) is the relative proportion of a-years-old to newborns for the stable age distribution.

General solution

Let us consider the full equation

$$B_t = \sum_{a=1}^{t} B_{t-a}l_a m_a + G_t, \tag{2.6.9}$$

for $t < \beta$. One of the method used to solve (2.6.9), which utilizes its convolution structure, is the Z-transform.

Mathematical interlude-the Z-transform For a sequence $(f_t)_{t\in\mathbb{N}}$ we define

$$\hat{f}(\lambda) = Z(f_t) = \sum_{t=0}^{\infty} f_t\lambda^{-t} \tag{2.6.10}$$

$Z(\mathbf{f})$ is a Laurent series convergent for $|\lambda| > R$ where

$$R = \overline{\lim_{t\to\infty}} \sqrt[t]{|f_t|}.$$

Z transform is a linear operation. A crucial property for applications in difference equations is

$$Z(f_{t+k})(\lambda) = \lambda^k\hat{f}(\lambda) - \sum_{r=0}^{k-1} f_r\lambda^{k-r} \tag{2.6.11}$$

Indeed, taking $k = 1$, we have

$$\begin{aligned} Z(f_{t+1})(\lambda) &= f_1 + f_2\lambda^{-1} + \ldots = \lambda(f_0 + f_1\lambda^{-1} + f_2\lambda^{-2} + \ldots) - \lambda f_0 \\ &= \lambda Z(f_t) - \lambda f_0 \end{aligned}$$

and the formula for larger values of k can be derived by induction.

The most important property is the Z-transform of a convolution of two sequences, defined by

$$(f * g)(t) = \sum_{a=0}^{t} f_{t-a}g_a = \sum_{a=0}^{t} f_a g_{t-a}.$$

We have

$$\begin{aligned}
Z(f * g)(\lambda) &= \sum_{t=0}^{\infty}\left(\sum_{a=0}^{t} f_{t-a}g_a\right)\lambda^{-t} \\
&= \sum_{t=0}^{\infty}\left(\sum_{a=0}^{t} f_{t-a}\lambda_{-(t-a)}g_a\right)\lambda^{-a} \\
&= \sum_{a=0}^{\infty} g_a\lambda^{-a}\left(\sum_{t=a}^{\infty} f_{t-a}\lambda^{-(t-a)}\right) \\
&= \sum_{a=0}^{\infty} g_a\lambda^{-a}\sum_{b=0}^{\infty} f_b\lambda^{-b} \\
&= Z(f_t)Z(g_t)
\end{aligned}$$

Suppose we know that a given function $f(\lambda)$ is the Z-transform of a sequence $(f_t)_{t\in\mathbb{N}}$. How we can find this sequence. First, we observe that by uniqueness of Laurent expansion, no two different sequences can have the same Z transform. To find $(f_t)_{t\in\mathbb{N}}$ we can use one of the following method:

a) power series method;

b) partial fraction method;

c) inversion integral method.

In the power series method, we find the Laurent expansion of $f(\lambda)$; then the coefficients of the expansion give the required series. Clearly, this is a long and not very satisfactory method. In many cases, $f(\lambda)$ can be written in terms of partial fractions. Then we can use the fact that

$$Z(a^t) = \sum_{t=0}^{\infty}\left(\frac{a}{\lambda}\right)^t = \frac{\lambda}{\lambda - a}, \quad |\lambda| > a \tag{2.6.12}$$

to invert the Z-transform.

Example 2.6.6. Solve the difference equation

$$f_{t+2} + 3f_{t+1} + 2f_t = 0, \quad f_0 = 1, f_1 = -4$$

Applying the Z transform, we obtain

$$\lambda^2\hat{f}(\lambda) - f(0)\lambda^2 - f(1)\lambda + 3\lambda\hat{f}(\lambda) - 3f(0)\lambda + 2\hat{f}(\lambda) = \hat{f}(\lambda)(\lambda^2+3\lambda+2) - \lambda^2 + \lambda = 0$$

and

$$\frac{\hat{f}(\lambda)}{\lambda} = \frac{\lambda - 1}{(\lambda+1)(\lambda+2)}.$$

The partial fraction representation is

$$\hat{f}(\lambda) = \frac{-2\lambda}{\lambda+1} + \frac{3\lambda}{\lambda+2},$$

so that, by (2.6.12),

$$f_t = -2(-1)^t + 3(-2)^t.$$

The general way of finding $(f_t)_{t\in\mathbb{N}}$ is through the complex integration. Writing

$$\hat{f}(\lambda)\lambda^{t-1} = f_0\lambda^{t-1} + f_1\lambda^{t-2} + \ldots + f_t\lambda^{-1} + f_{t+1}\lambda^{-2} + \ldots$$

and, by the Cauchy theorem,

$$f_t = \frac{1}{2\pi i}\int\limits_C \hat{f}(\lambda)\lambda^{t-1}d\lambda \tag{2.6.13}$$

where C is a circle centred at the origin and enclosing all singularities of $\hat{f}(\lambda)\lambda^{t-1}$. If $\hat{f}(\lambda)\lambda^{t-1}$ has only poles inside the circle of integration, f_t can be evaluated by summing up all residues of $\hat{f}(\lambda)\lambda^{t-1}$.

Let us turn back our attention to (2.6.9) and take the Z-transform of both side. We get

$$\hat{B}(\lambda) = \hat{B}(\lambda)\hat{f}(\lambda) + \hat{G}(\lambda)$$

where f is the net maternity function, defined as $f_a = l_a m_a$. Thus

$$\hat{B}(\lambda) = \frac{\hat{G}(\lambda)}{1-\hat{f}(\lambda)}$$

where, as we easily note $\hat{f}(\lambda) = \sum\limits_{a=\alpha}^{\beta}\lambda^{-a}f_a$ so that zeroes of the denominator are exactly the solutions of the Euler-Lotka equation.

To simplify further presentation, we assume that the solutions of the Euler-Lotka equation are simple $(\lambda_0, \lambda_1, \ldots, \lambda_{\beta-1})$ and therefore, by (2.6.13) and the comment below it,

$$B_t = \sum_{j=0}^{\beta-1} res_{\lambda=\lambda_j}\frac{\lambda^{t-1}\hat{G}(\lambda)}{1-\hat{f}(\lambda)}.$$

Using the assumption that the poles are simple, we get

$$res_{\lambda=\lambda_j}\frac{\lambda^{t-1}\hat{G}(\lambda)}{1-\hat{f}(\lambda)} = \frac{\lambda_j^{t-1}\hat{G}(\lambda_j)}{-\hat{f}'(\lambda_j)} = \lambda_j^t\frac{\sum\limits_{r=0}^{\beta-1}\lambda_j^{-r}G_r}{\sum\limits_{a=\alpha}^{\beta}a\lambda^{-a}l_a m_a}$$

and thus we have

$$B_t = \sum_{j=0}^{\beta-1}c_j\lambda_j^t,$$

where

$$c_j = \frac{\sum\limits_{r=0}^{\beta-1}\lambda_j^{-r}G_r}{\sum\limits_{a=\alpha}^{\beta}a\lambda^{-a}l_a m_a}.$$

To conclude, we note that $\lambda_0 > 1$ if and only if

$$R_0 = \psi(1) = \sum_{a=\alpha}^{\beta} l_a m_a > 1.$$

This follows since ψ is a strictly decreasing function and if at $\lambda = 1$ is bigger than 1, then it will be 1 for some $\lambda > 1$. Conversely, if $\psi(\lambda) = 1$ for some $\lambda < 1$, then $\psi(1) < 1$. The number $\psi(1)$ is easily seen to be the average number of (female) offspring produced by a female during her lifetime. Thus, R_0 is the basic reproductive ratio and the population will grow if and only if $R_0 > 1$.

2.6.2 Continuous Lotka equation

Let us recall the integral renewal equation

$$B(t) = \int\limits_0^t B(t-a)l(a)m(a)da + G(t) \tag{2.6.14}$$

where

$$G(t) = \int\limits_0^{\omega-t} m(a+t)n(a,0)\frac{l(a+t)}{l(a)}da, \tag{2.6.15}$$

is a known function. The existence of solutions to (2.6.14) can be proved, under mild assumptions, using Picard iterations. Under our assumptions l and m are only non-zero on finite intervals as, assuming they are piecewise continuous, they are bounded. Using this fact, we obtain, by Gronwall's inequality, that B is exponentially bounded.

Therefore we can apply the Laplace transform to solve (2.6.14). The Laplace transform 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 the right half-plane (determined by the rate of growth of f. We also note that if the f is non-zero only over a finite interval $[a,b]$, then its Laplace transform is defined and analytic everywhere in $\mathbb{C}$. Such functions are called entire.

We use the property of the Laplace transform which is similar to that of the Z-transform: convolution is transformed into algebraic product. In this context, the convolution of two functions is defined by

$$(f * g)(t) = \int_0^t f(t-s)g(s)ds = \int_0^t f(s)g(t-s)ds$$

Using the definition of the Laplace transform and changing the order of integration in a similar way we changed the order of summation in the discrete case of the Z-transform, we obtain

$$[\mathcal{L}(f * g)](\lambda) = (\mathcal{L}f)(\lambda) \cdot (\mathcal{L}g)(\lambda) \tag{2.6.16}$$

Using this result, we obtain from (2.6.14)

$$\hat{B}(\lambda) = \hat{B}(\lambda)\hat{f}(\lambda) + \hat{G}(\lambda) \tag{2.6.17}$$

where f is, as before the net maternity rate $f(a) = m(a)g(a)$. Hence,

$$\hat{B}(\lambda) = \frac{\hat{G}(\lambda)}{1 - \hat{f}(\lambda)} \tag{2.6.18}$$

As we noted above, $\hat{G}$ is an entire function so the only singularities of $\hat{B}$ are due to zeroes of $1 - \hat{f}$. Since $\hat{f}$ 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 $\hat{g}$ is the Laplace transform of a continuous function g, then

$$g(t) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} e^{\lambda t}\hat{g}(\lambda)d\lambda$$

where integration is carried along any vertical line in the domain of analyticity of $\hat{g}$. In the case of isolated poles in λ_j, we can write

$$\frac{1}{2\pi i}\int_{c-iR}^{c+iR} e^{\lambda t}\hat{g}(\lambda)d\lambda + \frac{1}{2\pi i}\int_{\Gamma_R} e^{\lambda t}\hat{g}(\lambda)d\lambda = \sum_j res_{\lambda=\lambda_j}\hat{f}(\lambda),$$

where Γ_R is the left semicircle of radius R with centre at c, which does not pass through any of the poles, and the summation is carried over all poles enclosed by this circle and the segment of the line $c \pm i\infty$. If the sum of residues converges (or equivalently, the integrals over Γ_R converge to zero, then we can write

$$g(t) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} e^{\lambda t}\hat{g}(\lambda)d\lambda = \sum_{j=1}^{\infty} res_{\lambda=\lambda_j}\hat{f}(\lambda).$$

If we assume that all the zeroes of $1-\hat{f}$ are of first order; that is, the poles are simple, giving

$$res_{\lambda=\lambda_j}\frac{e^{\lambda t}\hat{G}(\lambda)}{1-\hat{f}(\lambda)} = \lim_{\lambda\to\lambda_j}(\lambda-\lambda_j)\frac{e^{\lambda t}\hat{G}(\lambda)}{1-\hat{f}(\lambda)} = C_j e^{\lambda_j t}$$

where

$$C_j = \frac{\int\limits_0^\beta e^{-\lambda_j s}G(s)ds}{-\hat{f}'(\lambda_j)} = \frac{\int\limits_0^\beta e^{-\lambda_j s}G(s)ds}{\int\limits_\alpha^\beta ae^{-\lambda_j a}l(a)m(a)ds}$$

and the above assumptions are satisfied, we can write

$$B(t) = \sum_{j=1}^{\infty} C_j e^{\lambda_j t}. \tag{2.6.19}$$

If the poles are not simple, then instead of purely exponential terms, we have combinations of exponents and powers of t but general picture remains the same.

As usual, we ask the standard question of population dynamics: is there a dominant trend in the evolution of the population. If we assume that the above expression for B is valid, this question boils down to the existence of dominant real solution to the equation

$$1 = \int\limits_\alpha^\beta e^{-\lambda a}m(a)l(a)da \tag{2.6.20}$$

As in the discrete case, we introduce

$$\psi(\lambda) = \int\limits_\alpha^\beta e^{-\lambda a}m(a)l(a)da$$

and we note

$$\begin{aligned} \lim_{\lambda\to-\infty}\psi(\lambda) &= \infty, \\ \lim_{\lambda\to\infty}\psi(\lambda) &= 0. \end{aligned}$$

Moreover,

$$\begin{aligned} \psi'(\lambda) &= -\int\limits_\alpha^\beta ae^{-\lambda a}l(a)m(a)da < 0, \\ \psi''(\lambda) &= \int\limits_\alpha^\beta a^2e^{-\lambda a}l(a)m(a)da > 0, \end{aligned}$$

so that ψ is strictly decreasing and concave up function. Since it is continuous, it takes on every positive value exactly once. Thus, we proved the counterpart of the discrete case result

Proposition 2.6.7. *Equation (2.6.20) has exactly one real root, $\lambda = \lambda_0$, of algebraic multiplicity 1.*

Remark 2.6.8. The continuity of ψ is a consequence of the boundeness of the domain of integration. If we allow $\beta = \infty$ and consider, say, $l(a) = (1+a^2)^{-1}$ and $m(a) = const < 2/\pi$

$$\psi(\lambda) = \int\limits_{\alpha}^{\infty} \frac{e^{-\lambda a} m(a)}{1+a^2} dt$$

then $\psi(\lambda)$ is finite for $\lambda \geq 0$ but $\psi(\lambda) = \infty$ for $\lambda < 0$ and $\psi(\lambda) < 1$ for all $\lambda \geq 0$ and Eq. (2.6.20) has no real solution.

The function ψ crosses the ordinate at

$$R_0 := \psi(0) = \int\limits_{\alpha}^{\beta} l(a)m(a)da \tag{2.6.21}$$

As before, R_0 is called the net reproductive rate. It is the lifetime reproductive potential of a female corrected for mortality. R_0 must exceed 1 for λ_0 to be positive, If $R_0 = 1$ if and only if $\lambda_0 = 0$ and, finally, $R_0 < 1$ if and only if $\lambda_0 < 0$.

The next step is, however, different from the discrete case. Namely, we have

Proposition 2.6.9. *All other roots λ_j of (2.6.20) occur as complex conjugates (real root is its own conjugate). Moreover, $\Re\lambda_j < \lambda_0$ for any j.*

Proof. Suppose $\lambda_j = u + iv$ is a root of (2.6.20). Then

$$1 = \int\limits_{\alpha}^{\beta} e^{-va}(\cos(-ua) + i\sin(-ua))m(a)l(a)da$$

and, taking real and imaginary part

$$\begin{aligned} \int\limits_{\alpha}^{\beta} e^{-va}\cos(ua)m(a)l(a)da &= 1, \\ \int\limits_{\alpha}^{\beta} e^{-va}\sin(va)m(a)l(a)da &= 0 \end{aligned}$$

and these two equation are invariant under the change $v \to -v$ so that $\bar{\lambda_j} = u - iv$ also satisfies (2.6.20).

To prove the second part, we note that, since the variable a is continuous, there must be a range of a between $[\alpha, \beta]$ over which $\cos va < 1$. Thus, from nonnegativity of the integrand

$$\int\limits_{\alpha}^{\beta} e^{-va} m(a) l(a) da > 1.$$

However

$$\int\limits_{\alpha}^{\beta} e^{-\lambda_0 a} m(a) l(a) da = 1,$$

and direct comparison of these two integrals yields $\lambda_0 > v = \Re\lambda_j$. □

Thus, we re-write the formula (2.6.19) for B as

$$B(t) = C_0 e^{\lambda_0 t} \left(1 + \sum_{j=1}^{\infty} \frac{C_j}{C_0} e^{(\lambda_j - \lambda_0)t} \right)$$

where $|e^{\lambda_j - \lambda_0}| = e^{\Re\lambda_j - \lambda_0} < 1$. Hence, each term of the series tends to zero as $t \to \infty$ which, unfortunately, does not yield that the whole series converges to zero. This can be proved under some mild assumptions and so we accept here that this is the case; that is, indeed

$$B(t) \approx C_0 e^{\lambda_0 t}, \quad t \to \infty \tag{2.6.22}$$

with

$$C_0 = \frac{\int\limits_0^{\beta} e^{-\lambda_0 s} G(s) ds}{\int\limits_{\alpha}^{\beta} a e^{-\lambda_0 a} l(a) m(a) ds}$$

After the death of founder females, the whole population grows according to

$$N(t) = \int\limits_0^{\omega} B(t-a) l(a) da$$

thus, for large times, we have

$$N(t) \approx C_0 e^{\lambda_0} \int\limits_0^{\omega} e^{-\lambda_0 a} l(a) da, \quad t \to \infty,$$

so that the population will eventually grow with exponential rate.

Similarly, the number of females in some small age range Δa is given by

$$n(a,t)\Delta a = B(t-a)l(a)\Delta a \tag{2.6.23}$$

(using $\Delta a = \Delta t$) so that

$$n(a,t)\Delta a \approx C_0 e^{\lambda_0}[e^{-\lambda_0 a}l(a)]\Delta a \quad t \to \infty.$$

If we denote the fraction of females in the age range Δa by

$$c(a,t) = \frac{n(a,t)\Delta a}{N(t)},$$

then

$$c(a,t) \approx c^*(a) := \frac{[e^{-\lambda_0 a}l(a)]\Delta a}{\int\limits_0^\omega e^{-\lambda_0 a}l(a)da}, \quad t \to \infty$$

is asymptotically independent of t for large t. In other words, the population tens towards a stable age distribution.

2.6.3 McKendrick-van Foerster equation

We analayse the initial-boundary value problem

$$\begin{aligned}\frac{\partial n(a,t)}{\partial t} + \frac{\partial n(a,t)}{\partial a} &= -\mu(a)n(a,t)\\ n(0,t) &= \int\limits_0^\omega n(a,t)m(a,t)da = B(t),\\ n(a,0) &= n_0(a).\end{aligned}$$

We use the method of characteristics; that is, we introduce new variables one of which will be running along characteristics while the other will label particular characteristic. In our case, characteristics are straight lines $\eta = t - a$. As the complementary variable we take $a = \xi$ which gives a nonsingular change of variables. Denoting by $\tilde{n}(\xi,\eta) = n(a,t)$ we obtain

$$\begin{aligned} n_a &= \tilde{n}_\xi \xi_a + \tilde{n}_\eta \eta_a = \tilde{n}_\xi - \tilde{n}_\eta,\\ n_t &= \tilde{n}_\xi \xi_t + \tilde{n}_\eta \eta_t = \tilde{n}_\eta\end{aligned}$$

so that (2.6.24) turns into

$$\tilde{n}_\xi(\xi,\eta) = -\mu(\xi)\tilde{n}(\xi,\eta)$$

whose general solution is given by

$$\tilde{n}(\xi,\eta) = C(\eta)e^{-\int\limits_0^\xi \mu(s)ds}$$

where C is an arbitrary function to be determined by initial and boundary conditions. Returning to the original variables, we have

$$n(a,t) = C(t-a)e^{-\int\limits_0^a \mu(s)ds} \tag{2.6.24}$$

Now, using the initial condition at $t = 0$ and $a > 0$ (so that $\eta = t - a < 0$), we get

$$n_0(a) = n(a,0) = C(-a)e^{-\int\limits_0^a \mu(s)ds}$$

that is, introducing dummy variable r

$$C(r) = n_0(-r)e^{\int\limits_0^{-r} \mu(s)ds}$$

so that, returning to (2.6.24),

$$n(a,t) = n_0(a-t)e^{\int\limits_0^{a-t} \mu(s)ds}e^{-\int\limits_0^a \mu(s)ds} = n_0(a-t)e^{-\int\limits_{a-t}^a \mu(s)ds} \tag{2.6.25}$$

for $t - a < 0$. On the other hand, for $a = 0$ and $t > 0$; that is $\eta = t - a > 0$, we have

$$B(t) = n(0,t) = C(t)$$

so that, by (2.6.23),

$$n(a,t) = B(t-a)e^{-\int\limits_0^a \mu(s)ds} = n(0,t-a)e^{-\int\limits_0^a \mu(s)ds},$$

(the number of births at time t is equal to the number of neonates at time t). Let us relate the integrals above with the definitions introduced earlier. If μ is the per capita death rate, then the age a population satisfies

$$N'(a) = -\mu(a)N(a)$$

so that

$$N(a) = N(0)e^{-\int\limits_0^a \mu(s)ds}$$

so that the fraction of newborns surviving till the age a is precisely

$$l(a) = \frac{N(a)}{N(0)}e^{-\int\limits_0^a \mu(s)ds}.$$

Thus, we can write down our solution as

$$n(a,t) = \begin{cases} n(0,t-a)l(a) & \text{for} \quad t > a, \\ n(a-t,0)\frac{l(a)}{l(a-t)} & \text{for} \quad a > t, \end{cases} \tag{2.6.26}$$

in accordance with (2.5.4).

Of course, this is not a complete solution as we do not know $n(0, t-a) = B(t-a)$. However, using again the boundary condition, we get

$$\begin{aligned} B(t) &= n(0,t) = \int\limits_0^\omega n(a,t)m(a)da \\ &= \int\limits_0^t B(t-a)l(a)m(a)da + \int\limits_t^\omega n(a-t,0)\frac{l(a)}{l(a-t)}m(a)da \\ &= \int\limits_0^t B(t-a)l(a)m(a)da + \int\limits_0^{\omega-t} n(a,0)\frac{l(a+t)}{l(a)}m(a+t)da \end{aligned}$$

which is precisely the integral Lotka equation. Hence, all conclusions derived earlier are also valid for solutions of the McKendrick-van Foerster model

2.7 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

$$p_n(t) = P[N(t) = n], \quad n = 1, 2, \ldots \tag{2.7.1}$$

the probability that the population has n individuals at t.

2.7.1 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

$$\begin{aligned} P\{1 \text{ birth in } (t, t+\Delta t]|N(t) = 1\} &= \beta\Delta t + o(\Delta), & (2.7.2) \\ P\{\text{more than 1 birth in } (t, t+\Delta t]|N(t) = 1\} &= o(\Delta t), & (2.7.3) \\ P\{0 \text{ births in } (t, t+\Delta t]|N(t) = 1\} &= 1 - \beta\Delta t + o(\Delta t). & (2.7.4) \end{aligned}$$

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

$$\begin{aligned} P\{1 \text{ birth in } (t, t+\Delta t]|N(t) = n\} &= n(\beta\Delta t + o(\Delta t))(1 - \beta\Delta t + o(\Delta t))^{n-1} \\ &= n\beta\Delta t + o(\Delta t). \end{aligned} \tag{2.7.5}$$

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

$$P\{\text{more than 1 birth in } (t, t+\Delta t]|N(t)=n\} = o(\Delta t). \tag{2.7.6}$$

Finally, no birth occurs if none individual produces an offspring; that is

$$\begin{aligned} P\{\text{0 births in } (t, t+\Delta t]|N(t)=n\} &= (1-\beta\Delta t + o(\Delta t))^n \\ &= 1 - n\beta\Delta t + o(\Delta t). \end{aligned} \tag{2.7.7}$$

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

$$\begin{aligned} p_n(t+\Delta t) &= p_{n-1}(t)P\{\text{1 birth in } (t, t+\Delta t]|N(t)=n-1\} \\ &\quad + p_n(t)P\{\text{0 births in } (t, t+\Delta t]|N(t)=n\} \end{aligned} \tag{2.7.8}$$

that is, using the formulae

$$p_n(t) = (n-1)\beta\Delta t p_{n-1} + (1-n\beta\Delta t)p_n(t) + o(\Delta) + o(\Delta t). \tag{2.7.9}$$

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

$$\frac{dp_n(t)}{dt} = -n\beta p_n(t) + (n-1)\beta p_{n-1}(t). \tag{2.7.10}$$

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,

$$p_n(0) = \begin{cases} 1 & \text{for} \quad n = n_0, \\ 0 & \text{for} \quad n \neq n_0. \end{cases} \tag{2.7.11}$$

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

$$\frac{dp_{n_0}(t)}{dt} = -n_0\beta p_{n_0}(t), \tag{2.7.12}$$

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

$$p_{n_0+m}(t) = \left(\begin{array}{c} n_0+m-1 \\ n_0-1 \end{array}\right) e^{-\beta n_0 t}(1-e^{-\beta t})^m.$$

Indeed, we proved the validity of the formula for $m = 1$. Next

$$\frac{dp_{n_0+m+1}(t)}{dt} = -(n_0+m+1)\beta p_{n_0+m+1}(t) + \beta \left(\begin{array}{c} n_0+m-1 \\ n_0-1 \end{array}\right) e^{-\beta n_0 t}(1-e^{-\beta t})^m.$$

and, as before

$$\left(p_{n_0+m+1}(t)e^{\beta(n_0+m+1)t}\right)' = \beta \left(\begin{array}{c} n_0+m-1 \\ n_0-1 \end{array}\right) e^{\beta t}(1-e^{-\beta t})^m.$$

and, integrating

$$\begin{aligned} & p_{n_0+m+1}(t)e^{\beta(n_0+m+1)t} \\ = \;& C + \beta(n_0+m)\left(\begin{array}{c} n_0+m-1 \\ n_0-1 \end{array}\right)\int e^{\beta(m+1)t}(1-e^{-\beta t})^m dt \\ = \;& C + \beta(n_0+m)\left(\begin{array}{c} n_0+m-1 \\ n_0-1 \end{array}\right)\int e^{\beta t}(e^{\beta t}-1)^m dt \\ = \;& C + (n_0+m)\left(\begin{array}{c} n_0+m-1 \\ n_0-1 \end{array}\right)\int u^m du \\ = \;& C + \frac{n_0+m}{m+1}\left(\begin{array}{c} n_0+m-1 \\ n_0-1 \end{array}\right)(e^{\beta t}-1)^{m+1} \\ = \;& C + \left(\begin{array}{c} n_0+m \\ n_0-1 \end{array}\right)(e^{\beta t}-1)^{m+1} \end{aligned}$$

Using the initial condition $p_{n_0+m+1}(0) = 0$ we find $C = 0$ and so

$$p_{n_0+m+1}(t) = \left(\begin{array}{c} n_0+m \\ n_0-1 \end{array}\right) e^{-\beta n_0 t}(1-e^{-\beta t})^{m+1}.$$

2.7.2 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,

$$\begin{aligned}
&P\{1 \text{ birth in } (t, t+\Delta t]|N(t) = 1\} = \beta\Delta t + o(\Delta), &(2.7.13)\\
&P\{1 \text{ death in } (t, t+\Delta t]|N(t) = 1\} = \delta\Delta t + o(\Delta), &(2.7.14)\\
&P\{\text{no change in } (t, t+\Delta t]|N(t) = 1\} = 1 - (\beta+\delta)\Delta t + o(\Delta t). &(2.7.15)
\end{aligned}$$

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

$$\begin{aligned}
& P\{1 \text{ birth in } (t, t+\Delta t]|N(t) = n\}\\
= \;& n(\beta\Delta t + o(\Delta t))(1 - (\beta+\delta)\Delta t + o(\Delta t))^{n-1}\\
= \;& n\beta\Delta t + o(\Delta t). \qquad (2.7.16)
\end{aligned}$$

Similarly, probability of 1 (net) death in the population

$$\begin{aligned}
& P\{1 \text{ birth in } (t, t+\Delta t]|N(t) = n\}\\
= \;& n(\delta\Delta t + o(\Delta t))(1 - (\beta+\delta)\Delta t + o(\Delta t))^{n-1}\\
= \;& n\delta\Delta t + o(\Delta t). \qquad (2.7.17)
\end{aligned}$$

and, finally,

$$\begin{aligned}
P\{\text{no change in } (t, t+\Delta t]|N(t) = n\} &= (1 - (\beta+\delta)\Delta t + o(\Delta t))^{n}\\
&= 1 - n(\beta+\delta)\Delta t + o(\Delta t). \qquad (2.7.18)
\end{aligned}$$

We can set up the equation describing evolution of $p_n(t)$. Arguing as before

$$p_n(t) = (n-1)\beta\Delta t p_{n-1} + (n+1)\delta\Delta t p_{n+1} + (1 - n(\beta+\delta)\Delta t)p_n(t) + o(\Delta t) \qquad (2.7.19)$$

and, finally

$$\frac{dp_n(t)}{dt} = -n(\beta+\delta)p_n(t) + (n-1)\beta p_{n-1}(t) + (n+1)\delta p_{n+1}(t). \qquad (2.7.20)$$

This system has to be supplemented by the initial condition

$$p_n(0) = \left\{\begin{array}{lll} 1 & \text{for} & n = n_0,\\ 0 & \text{for} & n \neq n_0. \end{array}\right. \qquad (2.7.21)$$

Remark 2.7.1. Equations similar to (2.7.20) 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 (2.7.10), the system (2.7.20) 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

$$\sum_{n=0}^{\infty} p_n(t) = 1, \quad t \geq 0. \tag{2.7.22}$$

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 (2.7.22), 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

$$p_0(t) = F(t, 0). \tag{2.7.23}$$

(2) The probabilities $p_n(t)$ are given by

$$p_n(t) = \frac{1}{n!}\frac{\partial^n F}{\partial x^n}\bigg|_{x=0} \tag{2.7.24}$$

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

$$E[N(t)] = \sum_{n=0}^{\infty} np_n(t) = \left.\frac{\partial F}{\partial x}\right|_{x=1} \tag{2.7.25}$$

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

$$Var[N(t)] = \left.\left(\frac{\partial^2 F}{\partial x^2} + \frac{\partial F}{\partial x} - \left(\frac{\partial F}{\partial x}\right)^2\right)\right|_{x=1} \tag{2.7.26}$$

Let us find the equation satisfied by F. Using (2.7.20) and remembering that $p_{-1} = 0$, we have

$$\begin{aligned}\frac{\partial F}{\partial t}(t,x) &= \sum_{n=0}^{\infty} n\frac{dp_n}{dt}(t) = -(\beta+\delta)\sum_{n=0}^{\infty} np_n(t)x^n \\ &\quad +\beta\sum_{n=0}^{\infty}(n-1)p_{n-1}(t)x^n + \delta\sum_{n=0}^{\infty}(n+1)p_{n+1}(t)x^n \\ &= -(\beta+\delta)x\frac{\partial F}{\partial x}(t,x) + \beta x^2\frac{\partial F}{\partial x}(t,x) + \delta\frac{\partial F}{\partial x}(t,x).\end{aligned}$$

That is, to find F we have to solve the equation

$$\frac{\partial F}{\partial t} = \left(\beta x^2 - (\beta+\delta)x + \delta\right)\frac{\partial F}{\partial x}. \tag{2.7.27}$$

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

$$F(x,t) = \left(\frac{e^{-rt}\frac{\beta x-\delta}{x-1} - \delta}{e^{-rt}\frac{\beta x-\delta}{x-1} - \beta}\right)^{n_0} = \left(\frac{e^{rt}\delta(1-x) + (\beta x-\delta}{e^{rt}\beta(1-x) + (\beta x-\delta)}\right)^{n_0} \tag{2.7.28}$$

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 (2.7.27) 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,

$$F(t,x) = \begin{cases} \left(\frac{e^{rt}\delta(1-x)+(\beta x-\delta)}{e^{rt}\beta(1-x)+(\beta x-\delta)}\right)^{n_0} & \text{if} \quad \beta \neq \delta \\ \left(\frac{\beta t+(1-\beta t)x}{1-x\beta t+\beta t}\right)^{n_0} & \text{if} \quad \beta = \delta \end{cases} \tag{2.7.29}$$

Let us complete this section by evaluating some essential parameters. The probability of extinction at time t is given by

$$p_0(t) = F(t,0) = \begin{cases} \left(\frac{\delta(e^{rt}-1}{e^{rt}\beta-\delta}\right)^{n_0} & \text{if} \quad \beta \neq \delta \\ \left(\frac{\beta t}{1+\beta t}\right)^{n_0} & \text{if} \quad \beta = \delta. \end{cases} \tag{2.7.30}$$

Hence, the asymptotic probability of extinction is given by

$$\lim_{t\to\infty} p_0(t) = \begin{cases} \left(\frac{\delta}{\beta}\right)^{n_0} & \text{if} \quad \beta > \delta \\ 1 & \text{if} \quad \beta \leq \delta. \end{cases} \tag{2.7.31}$$

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 (2.7.25). We have

$$\begin{aligned} E[N(t)] &= \left.\frac{\partial F}{\partial x}\right|_{x=1} \\ &= n_0 \left(\frac{e^{rt}\delta(1-x)+(\beta x-\delta)}{e^{rt}\beta(1-x)+(\beta x-\delta)}\right)^{n_0-1} \\ &\left.\frac{(-e^{rt}\delta+\beta)(e^{rt}\beta(1-x)+(\beta x-\delta))+\beta(e^{rt}-1)(e^{rt}\delta(1-x)+(\beta x-\delta))}{(e^{rt}\beta(1-x)+(\beta x-\delta))^2}\right|_{x=1} \\ &= n_0 \frac{(-e^{rt}\delta+\beta)(\beta-\delta)+\beta(e^{rt}-1)(\beta-\delta)}{(\beta-\delta)^2} \\ &= n_0 e^{rt} \end{aligned}$$

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

$$V(t) = 2n_0\beta t.$$

Chapter 3

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.

3.1 Models

3.1.1 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:

$$\begin{aligned} N_{t+1} &= \lambda N_t f(N_t, P_t), \\ P_{t+1} &= cN_t(1 - f(N_t, P_t). \end{aligned} \tag{3.1.1}$$

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

$$b(n, p; r) = \left(\begin{array}{c} n \\ r \end{array} \right) p^r (1-p)^{n-r}.$$

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.,

$$\mu = aP_t.$$

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

$$\begin{aligned} N_{t+1} &= \lambda N_t e^{-aP_t}, \\ P_{t+1} &= cN_\lambda(1-e^{-aP_t}) \end{aligned} \tag{3.1.2}$$

3.1.2 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

$$\begin{aligned} X_{t+1} &= bY_t, \\ X_{t+1} &= cX_t + Y_t(s-DY_t). \end{aligned} \tag{3.1.3}$$

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 (3.1.3) into

$$\begin{aligned} \hat{X}_{t+1} &= \hat{Y}_t, \\ \hat{X}_{t+1} &= a\hat{X}_t + \hat{Y}_t(s-\hat{Y}_t), \end{aligned} \tag{3.1.4}$$

where $a = cb > 0$.

3.1.3 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

$$\begin{aligned} S(k+1) &= S(k) - \frac{\alpha}{N}I(k)S(k) + \beta(I(k)+R(k)) \\ I(k+1) &= I(k) + \frac{\alpha}{N}I(k)S(k) - \gamma I(k) - \beta I(k) \\ R(k+1) &= R(k) - \beta R(k) + \gamma I(k) \end{aligned} \tag{3.1.5}$$

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 (3.1.5) to a two dimensional system

$$\begin{aligned} S(k+1) &= S(k) - \frac{\alpha}{N}I(k)S(k) + \beta(N - S(k)) \\ I(k+1) &= I(k)(1-\gamma-\beta) + \frac{\alpha}{N}I(k)S(k). \end{aligned} \tag{3.1.6}$$

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

3.2 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

$$\mathbf{x}(n+1) = \mathbf{f}(\mathbf{x}(n)), \tag{3.2.1}$$

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 (3.2.1) if

$$\mathbf{x}^* = \mathbf{f}(\mathbf{x}^*) \tag{3.2.2}$$

The definition of stability is analogous to the scalar case.

Definition 3.2.1. (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.

3.2.1 Stability of linear systems

We consider the linear autonomous system

$$\mathbf{x}(n+1) = \mathcal{A}\mathbf{x}(n), \quad \mathbf{x}(0) = \overset{\circ}{\mathbf{x}}, \tag{3.2.3}$$

We assume that $\mathcal{A}$ is non-singular. The origin $\mathbf{0}$ is always an equilibrium point of (3.2.3). We have the following result:

Theorem 3.2.2. *The following statements hold:*

1. *The zero solution of (3.2.3) 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

$$\mathcal{A}^n \mathbf{x}_0 = \sum_{r=1}^{k} \left(\sum_{j=1}^{n_r} c_r^j \mathcal{A}^n \mathbf{v}_r^j \right) \tag{3.2.4}$$

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

$$\begin{aligned}\mathcal{A}^n \mathbf{v}_r^j &= (\lambda_r \mathcal{I} + \mathcal{A} - \lambda_r \mathcal{I})^n \mathbf{v}_r^j = \sum_{l=0}^{n} \lambda_r^{n-l} \begin{pmatrix} n \\ l \end{pmatrix} (\mathcal{A} - \lambda_r \mathcal{I})^l \mathbf{v}_r^j \\ &= \Big(\lambda_r^n \mathcal{I} + n\lambda_r^{n-1}(\mathcal{A} - \lambda_r \mathcal{I}) + \ldots \\ &\quad + \frac{n!}{(m_j^r - 1)!(n - m_j^r + 1)!} \lambda_r^{n - m_j^r + 1} (\mathcal{A} - \lambda_r \mathcal{I})^{m_j^r - 1} \Big) \mathbf{v}_r^j, \end{aligned} \tag{3.2.5}$$

It is important to note that (3.2.5) 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

$$\begin{aligned}
&\|\mathcal{A}^n\mathbf{v}_r^j\| \qquad\qquad (3.2.6)\\
&\leq |\lambda_r|^n \left(1 + n|\lambda_r|^{-1}(\|\mathcal{A}\| + |\lambda_r|) + \dots\right.\\
&\qquad \left. + P_{m_j^r-1}(n)|\lambda_r|^{-m_j^r+1}(|\mathcal{A}| + |\lambda_r|)^{m_j^r-1}\right) \|\mathbf{v}_r^j\|\\
&= |\lambda_r|^n n^{m_j^r-1}\left(n^{-m_j^r+1} + n^{m_j^r-2}|\lambda_r|^{-1}(\|\mathcal{A}\| + |\lambda_r|) + \dots\right.\\
&\qquad \left. + \frac{P_{m_j^r-1}(n)}{n^{m_j^r-1}}|\lambda_r|^{-m_j^r+1}(\|\mathcal{A}\| + |\lambda_r|)^{m_j^r-1}\right) \|\mathbf{v}_r^j\|\\
&\leq C_j^r|\lambda_r|^n n^{m_j^r-1} \leq C_j^r|\lambda_r|^n n^{n_r-1},
\end{aligned}$$

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 $\rfloor$ consisting of constants c_r^j is given by

$$\mathbf{c} = \begin{pmatrix} | & \dots & | \\ \mathbf{v}_1^1 & \dots & \mathbf{v}_k^{n_k} \\ | & \dots & | \end{pmatrix}^{-1} \mathbf{x}_0$$

and thus, for some constant M

$$\|\mathbf{c}\| \leq M\|\mathbf{x}_0\|. \tag{3.2.7}$$

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 (3.2.7) 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

$$\|\mathcal{A}^n\mathbf{x_0}\| \leq K\|\mathbf{x}_0\|\eta^n \tag{3.2.8}$$

and the zero solution is asymptotically stable.

If there are eigenvalues of unit modulus but they are semi-simple, then the expansions (3.2.7) 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 (3.2.5). 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. □

3.2.2 Stability by linearisation

Let us first note the following result.

Lemma 3.2.3. *If* $\mathbf{f}$ *has continuous partial derivatives of the first order in some neighbourhood of* $\mathbf{y}^0$*, then*

$$\mathbf{f}(\mathbf{x} + \mathbf{y}^0) = \mathbf{f}(\mathbf{y}^0) + \mathcal{A}\mathbf{x} + \mathbf{g}(\mathbf{x}) \tag{3.2.9}$$

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

$$\mathbf{y}_{t+1} = \mathcal{A}\mathbf{y}_t \tag{3.2.10}$$

is the linearization of (3.2.1) around an equilibrium $\mathbf{x}^*$.

We also note that to solve the nonhomogeneous system of equations

$$\mathbf{x}(n+1) = \mathcal{A}\mathbf{x}(n) + \mathbf{g}(n), \tag{3.2.11}$$

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 1.3.1 we find that the solution to (3.2.11) satisfying the initial condition $\mathbf{x}(0) = \mathbf{x}^0$ is given by the formula

$$\mathbf{x}(n) = \mathcal{A}^n\mathbf{x}^0 + \sum_{r=0}^{n-1}\mathcal{A}^{n-r-1}\mathbf{g}(r). \tag{3.2.12}$$

We shall need a discrete version of Gronwall's lemma.

Lemma 3.2.4. *Let $z(n)$ and $h(n)$ be two sequences of real numbers, $n \geq n_0 > 0$ and $h(n) \geq 0$.. If*

$$z(n) \leq M\left(z(n_0) + \sum_{j=n_0}^{n-1} h(j)z(j)\right) \tag{3.2.13}$$

for some $M > 0$, then

$$z(n) \leq z(n_0)\prod_{j=n_0}^{n-1}(1 + Mh(j)) \tag{3.2.14}$$

$$z(n) \leq z(n_0)\exp\sum_{j=n_0}^{n-1} Mh(j) \tag{3.2.15}$$

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 (3.2.14). The second follows from the formula $1 + Mh(j) \leq \exp(Mh(j))$. □

Theorem 3.2.5. *Assume that* $\mathbf{f}$ *is a* C^1 *function and* $\mathbf{x}^*$ *is an equilibrium point. If the zero solution of the linearised system (3.2.10) 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 3.2.3 we obtain

$$\mathbf{y}(n+1) = \mathcal{A}\mathbf{y}(n) + \mathbf{g}(\mathbf{y}n),$$

so that, by (3.2.12),

$$\mathbf{y}(n) = \mathcal{A}^n\mathbf{y}(0) + \sum_{r=0}^{n-1}\mathcal{A}^{n-r-1}\mathbf{g}(\mathbf{y}(r)).$$

Since the condition (3.2.8) 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.

3.3 Stability analysis of models

3.3.1 SIR model

Let us start with finding the equilibria of (3.1.6). These are solutions of

$$\begin{aligned} S &= F_1(S,I) = S - \frac{\alpha}{N}IS + \beta(N-S) \\ I &= F_21(S,I) = I(1-\gamma-\beta) + \frac{\alpha}{N}IS. \end{aligned} \tag{3.3.1}$$

$I = 0$ 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

$$\begin{aligned} F_{1,S}(S,I) = 1 - \frac{\alpha}{N}I - \beta, &\qquad F_{1,I}(S,I) = -\frac{\alpha}{N} \\ F_{2,S}(S,I) = \frac{\alpha}{N}I, &\qquad F_{2,I}(S,I) = 1 - \delta + \frac{\alpha}{N}S, \end{aligned}$$

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

$$|B| < 1 + A < 2 \tag{3.3.2}$$

Conversely, from (3.3.2) 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, (3.3.2) 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

$$1 - |B| + A > 0. \tag{3.3.3}$$

But in this case, $|B| < 2\sqrt{A}$ so that if

$$1 - 2\sqrt{A} + A > 0$$

holds on $(0,1)$, than (3.3.3) holds as well. But the former is nothing but $(1-\sqrt{A})^2 > 0$ on this open interval. Hence, (3.3.3) is proved.

Conversely, assume (3.3.2) 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

$$|tr\,\mathcal{A}| < 1 + det\,\mathcal{A} < 2 \tag{3.3.4}$$

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 (3.3.4) 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 (3.3.4) 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.

3.3.2 Nicholson-Bailey model

Recall that the model is given by

$$\begin{aligned} N_{t+1} &= \lambda N_t e^{-aP_t}, \\ P_{t+1} &= cN_t(1 - e^{-aP_t}). \end{aligned} \tag{3.3.5}$$

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,

$$\bar{P} = \frac{\ln\lambda}{a}, \tag{3.3.6}$$

and hence

$$\bar{N} = \frac{\lambda\ln\lambda}{(\lambda-1)ac}. \tag{3.3.7}$$

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 (3.3.4). 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 (3.3.5) would be amended as follows:

$$\begin{aligned} N_{t+1} &= \lambda(N_t)N_t e^{-aP_t}, \\ P_{t+1} &= cN_t(1-e^{-aP_t}), \end{aligned} \tag{3.3.8}$$

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

$$\begin{aligned} N_{t+1} &= N_t \exp\left(r\left(1-\frac{N_t}{K}\right) - aP_t\right), \\ P_{t+1} &= cN_t(1-\exp(-aP_t)), \end{aligned} \tag{3.3.9}$$

We simplify this system by introducing $n_t = N_t/K$ and $p_t = aP_t$. This converts (3.3.9) into

$$\begin{aligned} n_{t+1} &= n_t \exp\left(r(1-n_t) - p_t\right), \\ p_{t+1} &= Kcan_t(1-\exp(-p_t)), \end{aligned} \tag{3.3.10}$$

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

$$1 = \exp\left(r\left(1-n\right)-p\right), \tag{3.3.11}$$

$$p = Cn(1-\exp(-p)). \tag{3.3.12}$$

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

$$\bar{p} = r\left(1-\bar{n}\right) = r(1-q), \tag{3.3.13}$$

$$C\bar{n} = \frac{\bar{p}}{1-\exp(-\bar{p})}. \tag{3.3.14}$$

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 (3.3.14) 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

$$\begin{array}{rcl} f_1(n) & = & \dfrac{r\left(1-n\right)}{Cn} = \dfrac{r}{C}\left(\dfrac{1}{n} - 1\right), \\ f_2(n) & = & 1 - \exp\left(-r\left(1-n\right)\right) \end{array}$$

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 (3.3.9) and define

$$F(n,p) = n\exp\left(r\left(1-n\right) - p\right), G(n,p) = Cn(1 - \exp(-p)),$$

and thus

$$\begin{array}{rcl} F_n(n,p) & = & (1 - rn)\exp\left(r\left(1-n\right) - p\right), \\ F_p(n,p) & = & -n\exp\left(r\left(1-n\right) - p\right), \\ G_n(n,p) & = & C(1 - \exp(-p)), \\ G_p(n,p) & = & Cn\exp(-p). \end{array}$$

At $(\bar{n}_2, \bar{p}_2)$ we find, by (3.3.11),

$$\begin{aligned} F_n(\bar{n}_2, \bar{p}_2) &= (1 - rn_2) = 1 - rq \\ F_p(\bar{n}_2, \bar{p}_2) &= -\bar{n}_2 = -q, \end{aligned}$$

For the other two derivatives we find, by (3.3.12) and (3.3.11), that

$$1 - e^{-\bar{p}_2} = \frac{\bar{p}_2}{C\bar{n}_2} = \frac{r\,(1 - \bar{n}_2)}{C\bar{n}_2} = \frac{r\,(1-q)}{Cq}$$

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.