J. Banasiak, Instytut Matematyki, Technical University of Lodz, Poland and
Department of Mathematics and Applied Mathematics
University of Pretoria, Pretoria, South Africa

MODELS AND METHODS IN MATHEMATICAL EPIDEMIOLOGY

Contents

1 Introduction 5

1 Principles of mathematical modelling 5

1.1 Conservation principles and constitutive relations 6

1.2 Basic unstructured continuous population models 8

1.3 Modelling interacting populations 16

2 Basic Epidemiological Models 19

1 Basic epidemiological terminology 19

2 First models 20

2.1 SIR model 20

2.2 A malaria model 22

2.3 Warm-up – analysis of a simple SIR model 24

2.4 (Mis)-matching models 26

2.5 Models reducible to one-dimensional problems 28

3 SIR models with Demography 37

1 Non-dimensionalization 37

2 Basic phase-plane analysis 38

4 SEIR model and global stability by Lyapunov function 45

1 Local stability 45

2 Global stability 47

2.1 Global stability of the disease free equilibrium 47

2.2 Global stability of the endemic equilibrium 48

5 Appendices 51

1 Stability of equilibria of autonomous differential equations 51

2 Stability by linearization 54

2.1 Solvability of linear systems 54

2.2 Stability of equilibrium solutions 67
3 The Poincaré-Bendixon Theory 76
3.1 Preliminaries 77
4 Other criteria for existence and non-existence of periodic orbit 82
5 Stability through the Lyapunov function 86

1

Introduction

1 Principles of mathematical modelling

By a **mathematical model** we understand an equation, or a set of equations, that describe some phenomenon that we observe in science, engineering, economics, or some other area, that provides a quantitative explanation and, ideally, prediction of observations.

Mathematical modelling is the process by which we formulate and analyze model equations and compare observations to the predictions that the model makes.

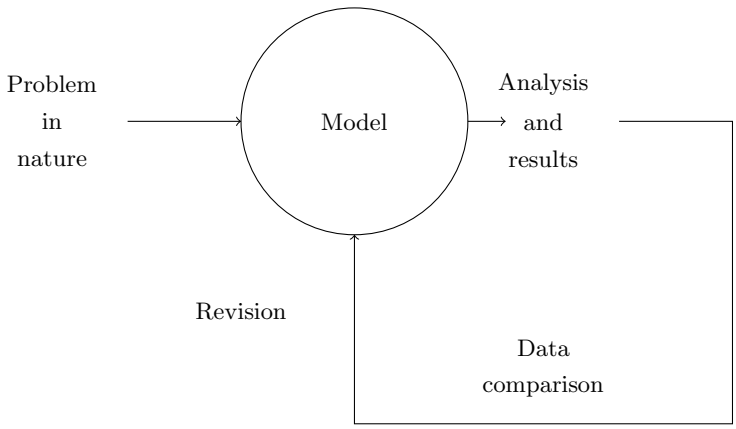


Fig. 1.1. The process of mathematical modelling.

Note:

- Modelling is not mathematics – it is impossible to prove that a model is correct;
- One counterexample disproves the model. However, this not always means that the model is useless – it may just require corrections.

A good model:

- has predictive powers – a model based on available observations gives correct answers in other cases:
  – General Theory of Relativity – light deflection, perihelion precession of Mercury, gravitational waves,
  – Dirac equations – existence of positrons;
- contains earlier working models as subcases:
  – Newton's mechanics is contained in Special/General Theory of Relativity for small velocities and away from

large masses,
– Quantum mechanics yields the same results as Newton's mechanics for large distances, large energies.

Descriptive versus explanatory models. Abundance of data often leads to statistical fitting the data with formulae. One can get a variety of statistical information such as expectations, medians, variance, correlations...
Remember: do not mistake correlations for causation!

Example: it has been observed that since the 1950s, both the atmospheric CO_2 levels and obesity levels in the US have increased sharply. Hence, obesity is caused by high levels of CO_2.

We shall focus on models which try to understand the underlying reasons for the phenomena we observe. Nevertheless, statistical analysis of the data is important as it separates their significant part from the noise.

Statistical (descriptive) models must not be mixed up with **stochastic models**. Stochastic modelling aims to explain the underlying mechanisms of the observed phenomena taking into account inherent(?) randomness of nature. Such models give probabilities of certain events and are indispensable in modeling small populations. We shall focus, however, on **deterministic models** that sometimes can be thought as stochastic models averaged over many individual trajectories (Law of Large Numbers) and giving answers in terms of the evolution of the densities of the populations. Nevertheless, stochastic models are often used explicitly to derive a deterministic model.

1.1 Conservation principles and constitutive relations

Conservation principles

Mathematical biology and epidemiology must obey laws of physics; in particular the balance law. Let Q be a quantity of interest (the number of animals, mass of a pollutant, amount of heat energy, number of infected individuals) in a fixed domain Ω. Over any fixed time interval in Ω we have

$$\begin{aligned}\text{The change of } Q = {} & \text{Inflow of Q} - \text{Outflow of Q} \\ & + \text{Creation of Q} - \text{Destruction of Q}.\end{aligned} \tag{1.1.1}$$

In probabilistic approach this is the same as saying that the probability that one of all possible events occurs equals one.

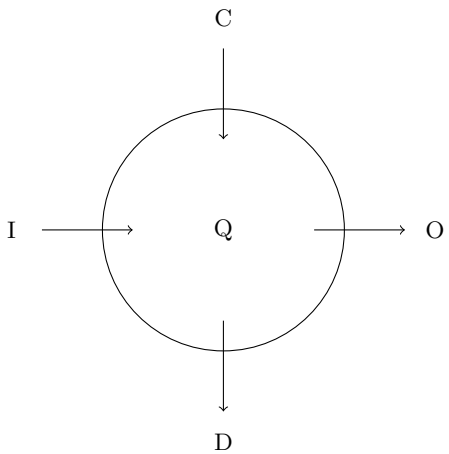


Fig. 1.2. Conservation law for the substance Q.

Continuous and discrete time

Before we proceed, we must decide whether we model with the **continuous time**, or with the **discrete time**.

We use **discrete time models** if we believe that significant changes in the system only occur during evenly spaced short time intervals, or we only can observe the system at evenly spaced time instances and have a reason to believe that essential parameters of the system remain unchanged between successive observations.

Then we use the time between the events/observations as the time unit and count time using the number of elapsed events/observations and (1.1.1) can be written as

$$Q(k+1) - Q(k) = I(k) - O(k) + C(k) - D(k). \tag{1.1.2}$$

Quantities $I(k), O(k), C(k), D(k)$ are the amounts of Q, respectively, that inflows, outflows, is created and destroyed in the time interval $[k, k+1]$.

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

(i) Monocarpic plants flower once and then die. Such plants may be annual but, for instance, bamboos grow vegetatively for 20 years and then flower and die. (ii) Animals with such a life cycle are called semelparous.
a) Insects typically die after lying eggs but their life-cycle may range from several days (e.g. house flies) to 13–17 years (cicads).
b) Similar life cycle is observed in some species of fish, such as the Pacific salmon or European eel. The latter lives 10-15 years in freshwater lakes, migrates to the Sargasso Sea, spawns and dies.
c) Some marsupials (antechinus) ovulate once per year and produce a single litter. There occurs abrupt and total mortality of males after mating. The births are synchronized to within a day or two with a predictable 'bloom' of insects.

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

If the assumptions allowing us to use discrete time modelling are not satisfied, we use **continuous time**. This, however requires some preparation, as all quantities may change at any instance of time. Thus, I, O, D, C should be considered as the **rates** of inflow, outflow, destruction or creation, respectively; in other words, the amount of Q at a given time t will be given by

$$Q(t) = Q(t_0) + \int_{t_0}^{t} I(s)ds - \int_{t_0}^{t} O(s)ds + \int_{t_0}^{t} C(s)ds - \int_{t_0}^{t} D(s)ds,$$

where $Q(t_0)$ is the initial amount of Q.

Hence, assuming that I, O, D, C are continuous functions, so that Q is differentiable, we obtain the conservation law in differential form,

$$\frac{dQ}{dt}(t) = I(t) - O(t) + C(t) - D(t). \tag{1.1.3}$$

Note 1. The meaning of I, O, C and D (and the dimension) in (1.1.3) is different than in (1.1.2).
Note 2. If we consider populations, then the value of Q always is a nonnegative integer. Such a function can never be continuous. Thus already (1.1.3) is an approximation the validity of which requires that Q be so large that it can be considered a continuum.

In epidemiology we are predominantly concerned with continuous time models.

Constitutive relations.

Real modelling consists in determining the form of Q, I, O, C and D and the relations between them – these are known as **constitutive relations**.

We try to build the functions I, O, D, C to encompass all we know about the process. However, this is usually impossible.

> There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.
> Donald Rumsfeld

Nevertheless, let us try. The functions I, O, D, C may depend on

- other unknown quantities – this leads to systems of equations that are the main topic of the lectures;
- space or other independent quantities – this leads to partial differential equations that will be discussed in the third lecture;
- explicitly on time – this results in non-autonomous equations which will be discussed later in this lecture;
- the unknown Q in
  a) a nonlinear way, such as $I(t) = I(Q(t)) = Q^2(t)$, or
  b) a linear way, such as $I(t) = I(Q(t)) = 2Q(t)$,
  in which case we talk, respectively, about autonomous nonlinear or linear equations.

Note. It is important to realize that non-autonomous equations often are derived from a larger systems of autonomous nonlinear equations in which the coefficients depend on partial solutions of this system which can be determined explicitly.

1.2 Basic unstructured continuous population models

Malthusian model.

If births and death rates are constant then, denoting the net growth rate by r we obtain

$$\frac{dP}{dt} = rP. \tag{1.1.4}$$

which has a general solution given by

$$P(t) = P(0)e^{rt}, \tag{1.1.5}$$

where $P(0)$ is the size of the population at $t = 0$. The U.S. Department of Commerce estimated that the Earth population in 1965 was 3.34 billion and that the population was increasing at an average rate of 2% per year during the decade 1960-1970. Thus $P(0) = 3.34 \times 10^9$ with $r = 0.02$, and

$$P(t) = 3.34 \times 10^9 e^{0.02t}. \tag{1.1.6}$$

Then the population will double in

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

which is in a good agreement with the estimated value of 6070 billion inhabitants of Earth in 2000. It also agrees relatively well with the observed data if we don't go too far into the past. On the other hand, if we try to extrapolate this model then in, say, 2515, the population would reach $199980 \approx 200000$ billion giving each of us area of $(86.3\,\text{cm} \times 86.3\,\text{cm})$ to live on.

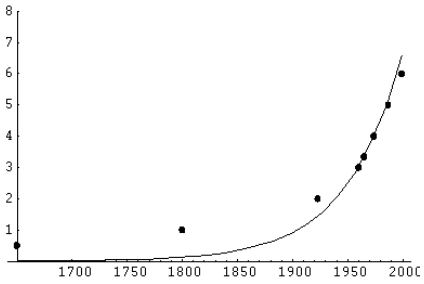


Fig 1.1. Comparison of actual population figures (points) with those obtained from equation (1.1.6).

Nevertheless, the Malthusian model has its uses for short term prediction. It also provides a useful link about the death rate and the expected life span of an individual.

Consider a population in which individuals die at a constant rate μ

$$P' = -\mu P.$$

Then the probability that an individual dies in a time interval Δt is approximately equal to $\mu \Delta t$. Let $p(t)$ be the probability that the individual is alive at time t. Then the probability $p(t + \Delta t)$ of it being alive at $t + \Delta t$ provided he/she was alive at t is $p(t + \Delta t) = (1 - \mu \Delta t)p(t)$ which, as above, yields

$$p' = -\mu p$$

with $p(0) = 1$ (expressing the fact that the individual was born, and thus alive, at $t = 0$) yielding $p(t) = e^{-\mu t}$. The average life span is given by

$$L = \int_0^\infty sm(s)ds,$$

where $m(s)$ is the probability (density) of dying exactly at age s. Since the probability of dying at the age between t and $t + \Delta t$ is

$$1 - p(t + \Delta t) - (1 - p(t)) = -\int_t^{t+\Delta t} \frac{d}{ds} p(s)ds$$

(one should be alive at t and dead at $t + \Delta t$, we have $m(s) = -\frac{d}{ds}p(s)$ and

$$L = -\int_0^\infty s\frac{d}{ds}e^{-\mu s}ds = \mu \int_0^\infty se^{-\mu s}ds = \frac{1}{\mu}. \tag{1.1.7}$$

Nonlinear models with size controlled growth

Logistic equation.

Passing to the limit in the discrete logistic equation valid between t and $t + \Delta t$,

$$P(t + \Delta t) - P(t) = r\Delta t \left(1 - \frac{P(t)}{K}\right)$$

we obtain the continuous logistic model

$$\frac{dP}{dt} = rP(t)\left(1 - \frac{P}{K}\right), \tag{1.1.8}$$

which proved to be one of the most successful models for describing a single species population. The equation has two constant solutions, $P(t) = 0$ and $P(t) = K$, with the latter being the carrying capacity of the environment. Other solutions can be obtained by separation of variables:

$$P(t) = \frac{P(0)K}{P(0) + (K - P(0))e^{-rt}}. \tag{1.1.9}$$

We have

$$\lim_{t\to\infty} P(t) = K, \quad P(0) > 0,$$

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

$$\frac{dP}{dt} > 0 \text{ if } 0 < P(0) < K,$$
$$\frac{dP}{dt} < 0 \text{ if } P(0) > K,$$

thus, if $P(0) < K$, the population monotonically increases, whereas if $P(0) > K$, then such a population will decrease until it reaches K. Also, for $0 < P(0) < K$,

$$\frac{d^2P}{dt^2} > 0 \text{ if } 0 < P(t) < K/2,$$
$$\frac{d^2P}{dt^2} < 0 \text{ if } P(0) > K/2,$$

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 that describes saturation process. On the other hand, Verhulst in 1845 predicted, on the basis of the

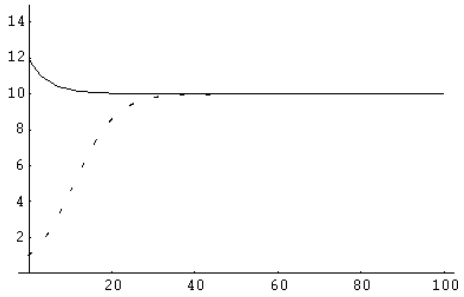


Fig. 1.3. Logistic curves with $P_0 < K$ (dashed line) and $P_0 > K$ (solid line) for $K = 10$ and $r = 0.02$.

logistic equation, that the maximum population of Belgium is 6 600 000. However, already in 1930 it was close to 8 100 000. This is attributed to the global change that happened for Belgium in the XIX century - acquisition of Congo that provided resources to support increasing population (at the cost of the African population of Congo).

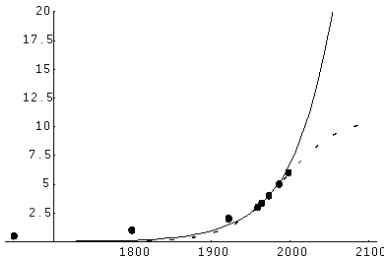


Fig. 1.4. Human population on Earth with $K = 10.76$ billion and $r = 0.029$ and $P(1965) = 3.34$ billion. Observational data (points), exponential growth (solid line) and logistic growth (dashed line).

	Actual	Predicted	Error	%
1790	3,929,000	3,929,000	0	0.0
1800	5,308,000	5,336,000	28,000	0.5
1810	7,240,000	7,228,000	− 12,000	− 0.2
1820	9,638,000	9,757,000	119,000	1.2
1830	12,866,000	13,109,000	243,000	1.9
1840	17,069,000	17,506,000	437,000	2.6
1850	23,192,000	23,192,000	0	0.0
1860	31,443,000	30,412,000	− 1,031,000	− 3.3
1870	38,558,000	39,372,000	814,000	2.1
1880	50,156,000	50,177,000	21,000	0.0
1890	62,948,000	62,769,000	− 179,000	− 0.3
1900	75,995,000	76,870,000	875,000	1.2
1910	91,972,000	91,972,000	0	0.0
1920	105,711,000	107,559,000	1,848,000	1.7
1930	122,775,000	123,124,000	349,000	0.3
1940	131,669,000	136,653,000	4,984,000	3.8
1950	150,697,000	149,053,000	− 1,644,000	− 1.1

Fig. 1.5. Comparison of actual and logistic model population in the United States

A simplified logistic model

We have considered two basic demographical models, the Malthusian model and the logistic model. The drawback of the Malthusian model is that it only can describes a very simple dynamics: the population either decays to zero, or exponentially grows to infinity. The drawback of the logistic model is that it is nonlinear and thus may create additional difficulties in analysis. For this reason an intermediate model is often used in analysis. The model takes the form

$$N' = \Lambda - \mu N, \tag{1.1.10}$$

where Λ is the total birth/recruitment rate and μ is per capita death rate. This is a linear nonhomogeneous equation in N with the solution

$$N(t) = N_0 e^{-\mu t} + \frac{\Lambda}{\mu}(1 - e^{-\mu t}). \tag{1.1.11}$$

It is easy to see that

$$N^* = \frac{\Lambda}{\mu} \tag{1.1.12}$$

is the only equilibrium (!). It is globally asymptotically stable.

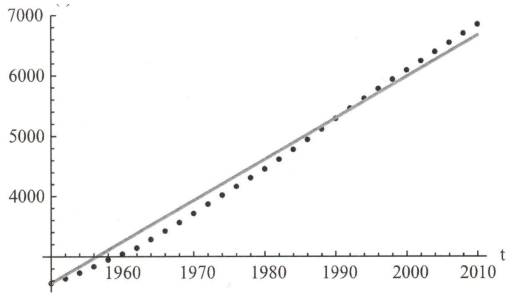


Fig. 1.6. World population alongside the simplified logistic model prediction. The value $P(1950) = 2556.5 \times 10^6$ billion people. The least square error is 703 483. However, the parameters are $\Lambda = 68.5 \times 10^6$ while $\mu = 5.54 \times 10^{-12}$. This give the average lifespan of 1.8×10^{11} years – completely unrealistic.

Holling II type argument.

Consider a sexually reproducing population. We begin with assumption, that over some time T, the number of offspring p per individual in the population of size/density P is proportional to it but we take into account the the reproduction happens only over some shorter period of time T_{as} over which individuals are sexually active:

$$p = rPT_{as}.$$

However, we have to take into account that adult individuals are not always sexually active, for instance during the gestation period. If an adult spends time T_g per offspring for gestation/rearing and, apart from that, it is ready for reproduction,

$$T = T_{as} + pT_r = T_{as} + rPT_{as}T_g$$

then

$$T_{as} = \frac{T}{1 + rT_rP}$$

and hence, the birth rate; that is, the number of offspring per unit time, is given as

$$\frac{dP}{dt} = \frac{pP}{T} = \frac{rP^2}{1 + rT_rP}.$$

A similar argument can be used to derive a Holling type death term used in Allee type models. Consider a population in which individuals must leave a refuge in order to mate. During this time they are exposed to dangers and thus the number of deaths induced by this activity in some period of time T is

$$d = \mu PT_{exp}, \tag{1.1.13}$$

where μ is the additional rate of death due to this activity. Now

$$T = T_s + T_{exp},$$

where T_s is the time spend in a shelter. We assume that the time of an individual is divided between searching for a mate and caring for the offspring. Hence, T_s can be obtained as the product of the number of successful matings in T times T_r that is the average time spend on looking after the offspring. Now, the number of successful matings is the product of the area searched in T_{exp}, the density of males and the efficiency of mating. We assume that the ratio of females and males is constant, so that the density of males is proportional to P. Further, the searched area is proportional to the search time T_{exp} so we can write

$$T_s = aPT_{exp}$$

for some constant a. Hence, as before, we have

$$\frac{dP}{dt} = -\frac{d}{T} = -\frac{\mu P}{1+aP}.$$

An analogous argument can be used in the population that is exposed to a generalist predator; that is, a predator that can eat other prey, so that its number is not affected by the presence of the prey from this particular population. In this case, (1.1.13) is replaced by

$$d = \mu N P T_{hunt}.$$

where N is the (constant) population of predators and T_{hunt} is the hunting time of the predator. Then $T = T_{hunt} + T_{handl}d/N$ where T_{handl} is the handling time of a single prey. Hence $T_{hunt} = T/(1+\mu T_{handl}P)$ and

$$\frac{dP}{dt} = -\frac{\mu N P}{1+\mu T_{handl}P}.$$

Gompertz model.

In the logistic equation we assumed $r(P) = 1 - P/K$, or

$$\frac{dr}{dP} = -\frac{1}{K}, \quad r(K) = 0.$$

A variety of models can be obtained by varying the equation for r. For instance if

$$\frac{dr}{dP} = -\frac{\alpha}{P}, \quad r(K) = 0, \tag{1.1.14}$$

then we have the so-called Gompertz model.

The above equation can be easily solved giving

$$r(P) = \alpha \ln\left(\frac{K}{P}\right)$$

and thus the population equation takes the form

$$\frac{dP}{dt} = \alpha \ln\left(\frac{K}{P}\right) P. \tag{1.1.15}$$

We see that the equation has two equilibria, with $P = K$ asymptotically stable.

One can derive another, possibly more instructive form of this equation. Using (1.1.14) and the Chain Rule, we get

$$\frac{dr}{dt} = \frac{dr}{dP}\frac{dP}{dt} = -\frac{\alpha}{P} rP = -\alpha r.$$

Hence the growth rate decays exponentially as $r(t) = r_0 e^{-\alpha t}$ and we can write (1.1.15) as

$$\frac{dP}{dt} = (r_0 e^{-\alpha t})P = r_0 \cdot (e^{-\alpha t}P). \tag{1.1.16}$$

Different places of brackets indicate different interpretations - the left one suggests that the growth rate is decreasing, while the right one suggests that the pool of fertile individuals is decreasing. That is why the model has been quite successful in modelling cancer. The equation can be solved by separation of variables, giving

$$P(t) = P(0) e^{\frac{r_0}{\alpha}} e^{-\frac{r_0}{\alpha} e^{-\alpha t}}. \tag{1.1.17}$$

Since K is the only asymptotically stable equilibrium,

$$\lim_{t\to\infty} P(t) = P(0) e^{\frac{r_0}{\alpha}} = K$$

we can rewrite the solution in terms of the carrying capacity as

$$P(t) = K e^{e^{-\alpha t} \ln \frac{P(0)}{K}}. \tag{1.1.18}$$

Allee type model

In all previous models with density dependent growth rates the bigger the population (or the higher the density), the slower was 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 $P' < 0$ for $0 < P < L$ and $P > K$ and $P' > 0$ for $L < P < K$.

The simplest equation of this type has a cubic nonlinearity:

$$\frac{dP(t)}{dt} = r(L - P(t))(P(t) - K))P(t). \tag{1.1.19}$$

A more complex model

$$\frac{dP}{dt} = \lambda P \left(1 - \frac{P}{C} - \frac{A}{1 + BP}\right), \tag{1.1.20}$$

$\lambda, C, A, B > 0$, can be obtained by adding to the logistic growth the additional mortality term $-\lambda AP/(1 + BP)$ that, as we know from modelling of Holling type effects, can be caused by exposure to danger due to search for mates, or by a presence of a generalist predator.

We have to prove that it indeed describes a behaviour required from the Allee model. Let us recall that for this, the equation must have three equilibria, 0 and, say, $0 < L < K$ such that if the size of the population P satisfies $0 < P < L$, then P decreases to 0 and if $L < P < K$, then P increases to K. In the terminology of this section, 0 and K should be asymptotically stable equilibria of (1.1.20) and L should be its unstable equilibrium.

Since (1.1.20) is difficult to solve explicitly (though it is possible as it is a separable equation) we analyse it using the 'phase-plane' argument. The equilibria are solutions to

$$f(P) := P \left(1 - \frac{P}{C} - \frac{A}{1 + BP}\right) = 0. \tag{1.1.21}$$

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

$$f'(P) = 1 - \frac{2P}{C} - \frac{A}{(1 + BP)^2} \tag{1.1.22}$$

and since $f'(0) = 1 - A$ we obtain that if $A > 1$, then $P = 0$ is an asymptotically stable equilibrium. By analysing the second derivative we can also state that if $A = 1$ and $BC < 1$, then $P = 0$ is semi-stable, that is, it attracts trajectories originating from positive initial conditions but this case is not relevant in studying the Allee type behaviour. Now we can focus on the other equilibria. For (1.1.20) to describe an Allee model first we must show that

$$1 - \frac{P}{C} - \frac{A}{1 + BP} = 0 \tag{1.1.23}$$

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

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

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

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

This gives

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

with the maximum

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

Now, the nonzero equilibria of (1.1.20) are the points at which the horizontal line $A = const$ cuts the the

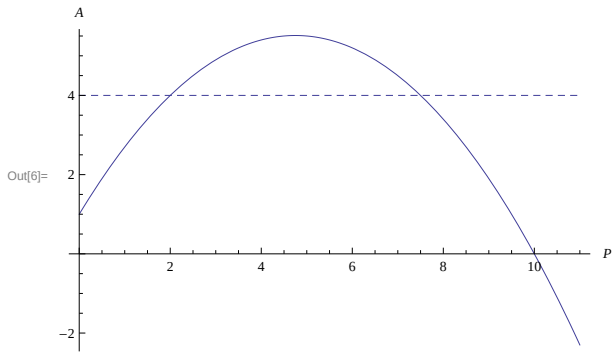


Fig. 1.7. The equilibria as a function of A.

graph of $A(P)$, see Fig. 1.7. First, we note that if $BC < 1$, then the stationary point P^* is negative and thus there is a positive and a negative solution for $0 < A < 1$, a negative and 0 solution for $A = 1$, two negative solutions if $1 < A < A^*$, one (double) negative solution if $A = A^*$ and no solutions if $A > A^*$. If $BC = 1$, then we have one positive, one negative solution for $0 < A < 1$, double 0 solution for $A = A^* = 1$ and no solutions for $A > 1$. Thus, in none case with $BK \leq 1$ we can expect the Allee type behaviour. Let us focus then on the case $BK > 1$. Since $A > 0$, we have the following cases

(a) If $0 < A < 1$, then there are two solutions to (1.1.23), but only one is positive while the other is negative;

(b) If $A = 1$, then there is one 0 and one positive solution to (1.1.23);

(c) If $1 < A < A^*$, then there are two distinct positive solutions to (1.1.23);

(d) If $A = A^*$, then there is a double positive solution to (1.1.23);

(e) If $A > A^*$, then there are no solutions to (1.1.23).

To determine the stability of the equilibria, we re-write (1.1.20) in the following form

$$\begin{aligned}\frac{dP}{dt} &= \lambda P\left(1 - \frac{P}{C} - \frac{A}{1+BP}\right) = \frac{\lambda BP}{C(1+BP)}\left(-P^2 + P\frac{BC-1}{B} + \frac{C(1-A)}{B}\right)\\ &= \frac{\lambda BP}{C(1+BP)}(P-L)(K-P). \qquad (1.1.24)\end{aligned}$$

Using the results of the first part of this section, we can describe the dynamics of (1.1.20) as follows. Let $BC > 1$. Then

(i) For $0 < A < 1$, there is one negative, L, and two nonnegative equilibria of (1.1.20), 0 and K. Zero is unstable and K is asymptotically stable;

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

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

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

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

If $BC \leq 1$, then we cannot have two positive equilibria so that the Allee effect cannot occur in this case. However, to complete analysis, we note that if $0 < BC \leq 1$ then the only case in which there is a positive equilibrium K is for $0 < A < 1$ and in this case K is asymptotically stable while 0 is unstable. For all other cases the only biologically relevant equilibrium is 0 and it is stable if $1 < A$, semistable (attracting positive trajectories) if $A = 1$ and $BC < 1$ and stable if $A = 1 = BC$. Summarizing, (1.1.20) describes the Allee effect if and only if

$$BC > 1 \quad \text{and} \quad 1 < A < \frac{(BC+1)^2}{4BC}. \tag{1.1.25}$$

In any other case with a positive equilibrium the dynamics described by (1.1.20) is similar to the dynamics described by the logistic equation.

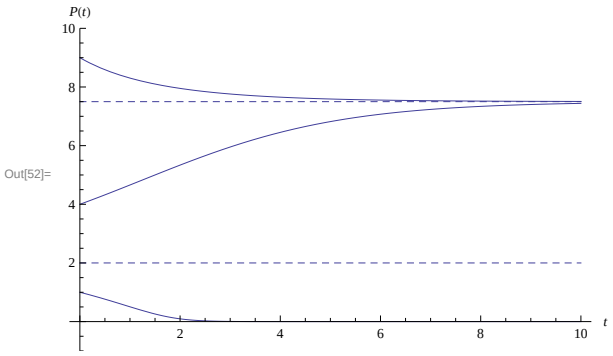


Fig. 1.8. Trajectories $P(t)$ of (1.1.20) for various initial conditions. Here $A = 4$, $C = 10$, $B = 2$, $L = 2$ (lower dashed line), $K = 7.5$ (upper dashed line).

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

1.3 Modelling interacting populations

Usually we split the system as

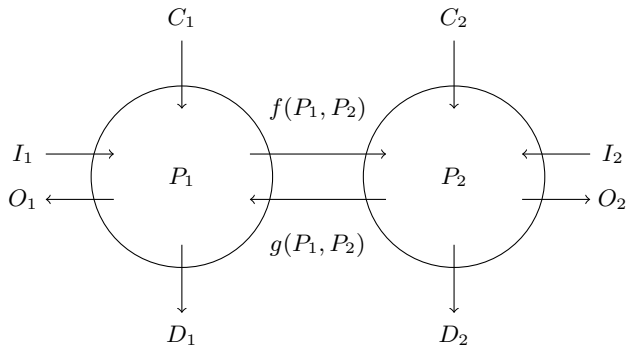


Fig. 1.9. Interactions between populations P_1 and P_2.

$$\frac{dP_1}{dt} = \text{Rate of change of} P_1 \text{ without } P_2 + \text{Impact of } P_2 \text{ on } P_1,$$
$$\frac{dP_2}{dt} = \text{Rate of change of} P_2 \text{ without } P_1 + \text{Impact of } P_1 \text{ on } P_2.$$

P_1 and P_2 typically will be densities of the populations over a certain area. For the first terms in each line we can use any suitable model from the first part.

The second terms require attention.

The simplest terms would be linear terms, e.g.

$$\frac{dP_1}{dt} = r_1 P_1 \left(1 - \frac{P_1}{K_1}\right) + a_{12} P_2$$
$$\frac{dP_2}{dt} = r_2 P_2 \left(1 - \frac{P_2}{K_2}\right) + a_{21} P_1.$$

This system could describe two populations with logistic vital dynamics each, where the additional rate of change of one population due to the other would be directly proportional to the density of the later.
Understanding the interaction terms.
We could distinguish:

- competition, when $a_{12}, a_{21} < 0$, as the presence of each species has a negative impact on the other;
- predator-prey interaction, when $a_{12} > 0$, $a_{21} < 0$, as the presence of species 2 (prey) has a positive impact on the species 1 (predator) and the as the presence of species 1 has a negative impact on the species 2;
- mutualism, when $a_{12}, a_{21} > 0$, as the presence of each species has a positive impact on the other.

In, say, the predator-prey case, the model above implies that the predator eats the same amount of prey per unit time, irrespective of the prey density. This is unrealistic so the coefficients a_{ij} should depend on the density of the ith species. Again, the simplest assumption is that a single predator will consume a proportion of the available prey, leading to the term

$$\alpha_{21} P_1 P_2, \quad \alpha_{21} < 0$$

that is called the *mass action law*, the term borrowed from chemical kinetics where it is assumed that the rate of reactions is proportional to the product of the concentrations of substrats. Hence we have a family of mass-action models of interactions between two species:

$$x' = x(\beta_1 + \mu_1 x + \alpha_{12} y),$$
$$y' = y(\beta_2 + \mu_2 y + \alpha_{21} x),$$

using which we can model various types of interactions and vital dynamics. For instance, assuming predator-prey interactions, $(\alpha_{12} > 0, \alpha_{21} < 0)$, we have $\beta_2 > 0, \mu_2 < 0$ if the prey population follows the logistic vital dynamics in the absence of predator. On the other hand, if we are to model a specialist predator, $\beta_1 < 0, \mu_1 \leq 0$ but for a generalist we can have $\beta_1 > 0, \mu_1 < 0$.

It is clear that the mass action law also is not realistic: for instance, it implies that the predator could eat an arbitrary amount of prey in a unit time (if the density of prey is large enough). To be more realistic, we must incorporate at least some saturation effect. We describe one such model, called Holling type 2 functional response (mass action is termed Holling type 1 response).

We assume that the amount P of prey consumed by a single predator in time T is proportional to the prey density and the time spent on hunting T_h

$$P = cyT_h.$$

The mass action law assumes that $T_h = T$; that is, the predator does not stop hunting while devouring the prey. While it is sometimes possible (e.g. adult salmon eating its offspring), in most cases the time T_e spent on eating the prey is positive. If P is the number of prey caught in time T, then the time used on consuming it is PT_s and thus $T_h = T - PT_s$. Thus

$$P = cy(T - PT_s)$$

and the density of prey eaten per unit time per predator is

$$\frac{P}{T} = \frac{cy}{1 + cT_s y}.$$

Adjusting the number of prey P to the density, we obtain the predation term of the form

$$-c_1 \frac{xy}{1 + c_2 y},$$

with positive constants c_1, c_2.

2

Basic Epidemiological Models

1 Basic epidemiological terminology

An infectious disease is an evident illness caused by microbial agent. The microbial agent can be:

bacteria: tuberculosis, pneumonia;

virus: HIV, influenza;

fungus: dermatomycosis;

parasite: malaria, bilharzia;

toxic protein or prions: Creutzfeldt-Jakob disease (mad cow disease).

Communicable disease are infectious disease that can be transmitted from one infectious person to another, directly or indirectly. There are infectious disease, such as tetanus, is infectious but not communicable. Transmittable diseases are infectious diseases that can be transmitted from one person to another by unnatural routes. For instance, mad cow disease can be passed from one person to another only through a surgical intervention.

For modelling purposes we distinguish the following types of transmission:

direct: when the pathogen is transmitted from one person to another by personal contact, such as sexually tramitted diseases, influenza, smallpox, measles, chickenpox, TB;

vector: when the pathogen is transmitted by a vector such as mosquito, tick or snail, that include malaria, dengue, zika, Lyme disease;

environmental: when a human is infected by a pathogen present in environment, water or food, such as cholera, salmonella, stomach flu;

vertical: mother-to-child transmission, such as HIV.

The following terminology is essential in epidemiological research:

Susceptible individuals: a member of a population who is at risk of becoming infected;

Exposed individuals: susceptible individuals that made a potentially disease-transmitting contact and may, or may not, develop the disease;

Infected and infectious individuals: if a pathogen establishes itself in an individuals, then the individual becomes infected. An infected individual who can transmit the disease is called infectious;

Latent individuals: individuals who are infected but not yet infectious;

Latent period: the time from infection to the moment the individual becomes infectious;

Incubation period: period between exposure to the pathogen to the onset of symptoms of the disease;

Incidence: the number of individuals who become ill during a specified time;

Prevalence: the number of people who have the disease at a specific time.

2 First models

2.1 SIR model

We begin with a simple model of a nonlethal disease in a homogeneous population divided into three classes: susceptible S, infective I and recovered R. Let us denote

λ = the force of infection; that is the rate at which susceptibles become infected,

μ = the death rate,

ν = the recovery rate,

γ = the rate of immunity loss,

B = the birth rate of the population.

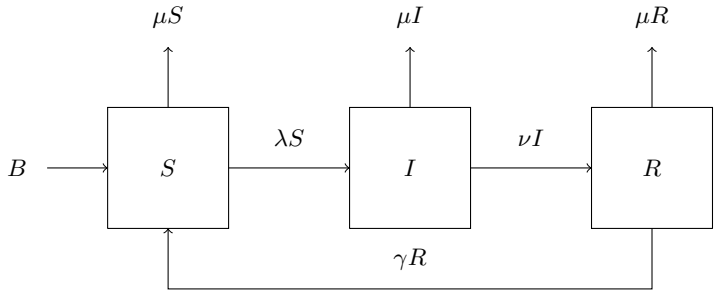


Fig. 2.1. Compartments in the SIRS model

On the basis of the above diagram, we build the following system of equations

$$\begin{aligned} S' &= B(N) - \lambda S + \gamma R - \mu S, \\ I' &= \lambda S - \nu I - \mu I, \\ R' &= \nu I - \gamma R - \mu R. \end{aligned} \tag{2.2.1}$$

S, I and R typically denote the number densities of, respectively, susceptibles, infectives and recovered, B maybe any function describing the vital dynamics of a healthy population (here we tacitly assumed that there is no vertical transmission of the disease or immunity).

Parameter interpretation

We showed that $1/\mu$ gives the expected lifespan of an individual; in the same way $1/\nu$ is the average duration of the disease and $1/\gamma$ is the average period of the acquired immunity.

For a directly transmitted pathogen the force of infection λ is the product of

1. the contact rate;

2. the proportion of these contacts that are with infective;
3. the proportion of such contacts that actually result in infection.

How should we model λ?

A simple assumption would be the mass action law - a single infective meets a fraction $c_1 S$ of susceptibles in a unit time and infects a fraction c_2 of those met:

$$c_2 c_1 SI.$$

As before, this may be a reasonable assumption if the densities are low but for large densities we need to take into account that contacts take time so that there can be only a finite time of contacts in a unit time interval. Also, saturation may be caused by satiation which playes a role in sexual transmission, but also in blood meals taken by mosquitoes.

In a population of size/density N we define $C(N)$ to be the fraction of the population engaged in a contact at any given time. Then $NC(N)$ (precisely $0.5NC(N)$) is the number of pairs in the population at any given time. Since the probability of choosing at random a pair consisting of a susceptible and an infective is

$$\frac{S}{N}\frac{I}{N},$$

the density of pairing that potentially can lead to infection is

$$C(N)\frac{SI}{N}.$$

The function $C(N)$ should be nonnegative, nondecreasing, linear in N for small N and having a limit (≤ 1) as $N \to \infty$. Let us try to derive such a function using a Holling type argument. First, in a population of size N we introduce the number of singles X and pairs P so that

$$N = X + 2P \tag{2.2.2}$$

so that

$$S' = -\beta 2P\frac{SI}{N^2}.$$

Assume that an individual can be either an available single, or form a pair, and that the contact lasts some time T_h. Denote by Z the total number contacts over some time T and let $Y = Z/T$, the number of contacts per unit time. As in the Holling derivation, we have

$$Z = \rho X T_s = \rho X(T - ZT_h), \tag{2.2.3}$$

where ρ is a constant. However, contrary to the predator-prey, where prey was unlimited, here the number of available singles is limited by time - singles are available only when they are not engaged in another contact. For a given single, in T it is available for $T - ZT_h$, and thus for the fraction $1 - YT_h$ of time. In other words, any given single at any given time is available for contact with probability $p = 1 - YT_h$. Thus the expected number of available singles at any given time is given by

$$\begin{aligned} X &= \sum_{k=0}^{N} k\binom{N}{k}p^k(1-p)^{N-k} = Np\sum_{l=0}^{N-1}\binom{N-1}{l}p^l(1-p)^{N-1-l} \\ &= Np = N(1 - YT_h). \end{aligned}$$

Hence, from (2.2.3),

$$\frac{Z}{T} = Y = \rho N(1 - YT_h)^2.$$

Using again (2.2.3), we find that the average number of pairs is

$$P = \frac{1}{2}NYT_h = \frac{\rho T_h}{2}N^2(1 - YT_h)^2 = \frac{\nu}{2}X^2.$$

By (2.2.2),

$$P = \frac{\nu}{2}(N - 2P)^2,$$

or

$$P = \frac{2\nu N + 1 \pm \sqrt{4\nu N + 1}}{4\nu}$$

and we have to select the negative sign,

$$P = \frac{2\nu N + 1 - \sqrt{4\nu N + 1}}{4\nu},$$

to keep $2P < N$. We re-write this as

$$P = \frac{\nu N^2}{2\nu N + 1 + \sqrt{4\nu N + 1}}.$$

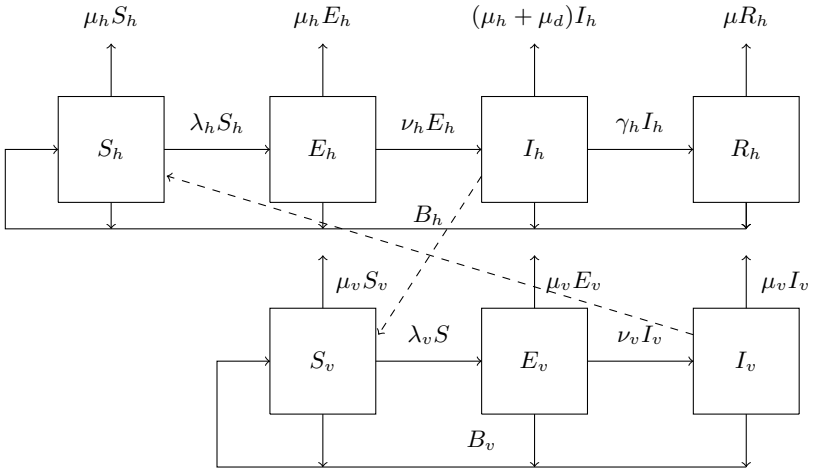
Finally,

$$S' = -C(N)\frac{SI}{N} = -\frac{2\nu\beta N}{2\nu N + 1 + \sqrt{4\nu N + 1}}\frac{SI}{N}.$$

However, most often we use one of the following simplifications:

- low density approximation $C(N) \sim N$ that leads to the mass action transmission rate βIS;
- constant approximation $C(N) \sim 1$ that gives frequency dependent transmission rate $\beta IS/N$.

2.2 A malaria model



$$\begin{aligned}
S_h' &= B_h(N_h) - \lambda_h S_h + \rho_h R_h - \mu_h(N_h)S_h,\\
E_h' &= \lambda_h S_h - \nu_h E_h - \mu_h(N_h)E_h,\\
I_h' &= \nu_h E_h - \gamma_h I - \mu_h(N_h)I_h - \delta_h I_h,\\
R_h' &= \gamma_h I_h - \rho_h R_h - \mu_h(N_h)R_h,\\
S_v' &= B_v(N_v) - \lambda_v S_v - \mu_v(N_v)S_v,\\
E_v' &= \lambda_v S_v - \nu_v E_v - \mu_v(N_v)E_v,\\
I_v' &= \nu_v E_v - \mu_v(N_v)I_v.
\end{aligned} \tag{2.2.4}$$

Here, we have the state variables

S_h: number of susceptible humans,

E_h: number of exposed humans,

I_h: number of infectious humans,

R_h: number of recovered (immune and asymptomatic, but slightly infectious) humans,

S_v: number of susceptible mosquitoes,

E_v: number of exposed mosquitoes,

I_v: number of infectious mosquitoes,

N_h: total number of humans,

N_v: total number of mosquitoes,

and parameters

σ_v: number one mosquito could bite a human per unit time, if humans were freely available. This is a function of the mosquito gonotropic cycle, its preference for human blood and time used for feeding. Time^{-1}.

σ_h: the maximum of mosquito bites a human can have per unit time. This is a function of the human's exposed area, awareness, etc. Time^{-1}.

β_{hv}: probability of infection from an infectious mosquito to a susceptible human, given that a contact between the two occurs. Dimensionless.

β_{vh}: probability of infection from an infectious human to a susceptible mosquito, given that a contact between the two occurs. Dimensionless.

$\tilde{\beta}_{hv}$: probability of infection from a recovered human to a susceptible mosquito, given that a contact between the two occurs. Dimensionless.

ν_h: per capita rate of progression of humans from the exposed state to the infectious state. $1/\nu_h$ is the average duration of the latent period. Time^{-1}.

ν_v: per capita rate of progression of mosquitoes from the exposed state to the infectious state. $1/\nu_v$ is the average duration of the latent period. Time^{-1}.

γ_h: per capita recovery rate of. $1/\gamma_h$ is the average duration of the infectious period. Time^{-1}.

ρ_h: per capita rate of the immunity loss of humans. $1/\rho_h$ is the average duration of the immune period. Time^{-1}.

δ_h: per capita disease induced death rate for humans. Time^{-1}.

Modelling the infection rates

The infection rates are given by

$$\lambda_h = b_h(N_h, N_v)\beta_{hv}\frac{I_v}{N_v}, \quad \text{and} \quad \lambda_v = b_v(N_v, N_h)\left(\beta_{vh}\frac{I_h}{N_h} + \tilde{\beta}_{vh}\frac{R_h}{N_h}\right). \tag{2.2.5}$$

In other words, λ_h is the product of the number of mosquito bites a human can have per unit time, b_h, the probability of the transmission of the infection, $\beta_h v$ and the probability that the bite comes from an infected mosquito, I_v/N_v. Similarly, λ_v is the product the number of human bites a mosquito has per unit times and the sum of probabilities that the bite comes from an infectious human and the transmission occurs. To

model the numbers of bites we first define the total number of bites that occur per unit time, $b(N_h, N_v)$ so that

$$b(N_h, N_v) = b_h(N_h, N_v)N_h = b_v(N_h, N_v)N_v.$$

To derive the formula, we use Holling type argument. In time T the total number of bites received by humans can be written as

$$b(N_h, N_v)T = \sigma_h T_{av} N_h$$

where T_{av} is the time available for mosquitoes to bite. Thus

$$T = T_{av} + T_{nav}.$$

Now, a mosquito cannot bite if it had a meal, and in time T a mosquito has $\sigma_h N_h T_{av}/N_v$ meals and the mosquito is not available for $1/\sigma_v$ after each meal. Thus

$$T = T_{av} + \frac{\sigma_h N_h T_{av}}{\sigma_v N_v}$$

and hence

$$b(N_h, N_v) = \frac{\sigma_v N_v \sigma_h N_h}{\sigma_v N_v + \sigma_h N_h}. \tag{2.2.6}$$

This gives

$$\lambda_h = \frac{\sigma_v \sigma_h}{\sigma_v N_v + \sigma_h N_h}\beta_{hv} I_v, \quad \text{and} \quad \lambda_v = \frac{\sigma_v \sigma_h}{\sigma_v N_v + \sigma_h N_h}\left(\beta_{vh} I_h + \tilde{\beta}_{vh} R_h\right). \tag{2.2.7}$$

2.3 Warm-up – analysis of a simple SIR model

For short lasting diseases, such as flu or common cold, it is customary to discard demographical processes. If the diseases induces immunity, at least in the time covered by the model, one of the simplest models is the SIR Kermack-McKendrick model

$$\begin{aligned} S' &= -\beta SI, \\ I' &= \beta SI - \nu I, \\ R' &= \nu I. \end{aligned} \tag{2.2.8}$$

As we see, we use the mass action transmission rate. The total population at time t is given by $N(t) = S(t) + I(t) + R(t)$ and, by adding the equations in (2.2.8), we obtain

$$N' = 0$$

hence $N(t) = N(0)$, reflecting the assumption that there is no demographic processes included in the model.

The dynamics of the model can be fully analysed without any sophisticated tools.

Step 1. The model is well-posed; that is, for every $(S(0), I(0), R(0)) = (S_0, I_0, R_0)$ there is exactly one solution defined at least on some interval $(-\tau, \tau)$, $\tau > 0$. This follows from the Picard theorem. We shall be interested in $t \geq 0$. Once we know that there is a solution $(S(t), I(t), R(t)), t \in (0, \tau)$, we can prove that it is positive provided $S(0), I(0)$ and $R(0)$ are positive. Indeed, for instance for S, we see that it satisfies the linear equation

$$S'(t) = -\beta I(t) S(t),$$

and hence $S(t) = S(0)e^{-\int_0^t I(s)ds}$, where I is a known function. Hence, $S(t) \geq 0$ as long as $I(t)$ is defined.

Thus, we have $0 \leq S(t), I(t), R(t) \leq N(0)$ for t in any interval on which the solution is defined and hence the solution is defined globally for $t \geq 0$.

Step 2. We see that $S' < 0$; that is, S is decreasing and bounded from below. Since it is defined for all $t \geq 0$, we have

$$\lim_{t\to\infty} S(t) = S_\infty.$$

Similarly, R is growing and satisfies

$$\lim_{t\to\infty} R(t) = R_\infty.$$

Further, since $S(t) + I(t) + R(t) = N(0)$, we must have

$$\lim_{t\to\infty} I(t) = I_\infty.$$

However, the number of infected individuals can increase or decrease depending on the sign of $\beta S(t) - \nu$. In particular, if $\beta S(0) - \nu > 0$, or

$$\frac{\beta S(0)}{\nu} > 1,$$

then the number of infectives initially will increase. Then we say that we have an outbreak or epidemic.

The number $\beta S(0)/\nu$ has an important interpretation. The coefficient β gives the number of infections per unit time induced by one infective, whereas $1/\nu$ is the average time an infective remains infectious. The number of susceptibles at the beginning is $S(0)$. Thus, we have arrived at the common interpretation of $\mathcal{R}_0$

Definition 2.1. *The basic reproduction number $\mathcal{R}_0$ is the number of infections that one infectious individual will introduce in a population consisting only of susceptible individuals.*

Next we estimate the limits. First, observe that $I(t) \neq 0$ for any finite t_0. Otherwise $(S(t_0), 0, R(t_0))$ would be a constant solution to the problem taking the same value as $(S(t), I(t), R(t))$ at $t = t_0$, contradicting the uniqueness of solutions. Hence $t \to R(t)$ is strictly increasing and we can consider $t = t(R)$ on $[R_0, R_\infty)$. Thus, we can consider $S(R) = S(t(R))$ and, using the Chain Rule and the derivative of the inverse function formula, we get

$$\frac{dS}{dR} = \frac{dS}{dt}\frac{dt}{dR} = -\frac{\beta SI}{\nu I} = -\frac{\beta}{\nu}S.$$

Hence

$$S(R) = S(R_0)e^{-\frac{\beta}{\nu}(R-R_0)} \geq S(R_0)e^{\frac{\beta}{\nu}R_0}e^{-\frac{\beta}{\nu}N(0)} > 0$$

for any R. Therefore we must have $S_\infty > 0$ which shows that no epidemic can infect all susceptibles.

Let us consider I_∞. Integrating the first equation in (2.2.8) we obtain

$$S_\infty - S_0 = \int_0^\infty S'(t)dt = -\beta \int_0^\infty S(t)I(t)dt \leq -\beta S_\infty \int_0^\infty I(t)dt.$$

In other words

$$\int_0^\infty I(t)dt \leq \frac{S_0 - S_\infty}{\beta S_\infty} < \infty.$$

Since we know that I_∞ exists and is nonnegative, we must have $I_\infty = 0$.

Step 3. We note that (2.2.8) is really a two-dimensional system and we can find orbits in the (S, I) plane. The two first equations are independent of R and can be solved separately yielding $R = N - S - I$. So, let us focus on

$$\begin{aligned} S' &= -\beta SI, \\ I' &= \beta SI - \nu I. \end{aligned} \tag{2.2.9}$$

From the above discussion, we know that S is a monotonic function of t for $S, I > 0$ and hence it can be inverted $t = t(S)$ allowing for writing $I = I(S)$

$$\frac{dI}{dS} = \frac{dI}{dt}\frac{dt}{dS} = \frac{\beta SI - \nu I}{-\beta SI} = -1 + \frac{\nu}{\beta S}.$$

Separation of variables and integration yields

$$I - I_0 = S_0 - S + \frac{\nu}{\beta} \ln \frac{S}{S_0}.$$

In particular, using that fact that $\lim_{t\to\infty} I(t) = 0$ and $\lim_{t\to\infty} S(t) = S_\infty$ we obtain

$$-I_0 = S_0 - S_\infty + \frac{\nu}{\beta} \ln \frac{S_\infty}{S_0}$$

or

$$\frac{\beta}{\nu} = \frac{\ln \frac{S_0}{S_\infty}}{S_0 + I_0 - S_\infty}. \tag{2.2.10}$$

Let us draw a few conclusions. First, since we know that S is a decreasing function, $S(t) \geq S_\infty$ for any $t \geq 0$. Thus we obtain

$$S_\infty \leq S_0 + I_0.$$

An important information is the maximum number of infectives. This occurs for $I' = 0$ or at $I(0) = I_0$. $I' = 0$ if $S = \nu/\beta$ (and hence this can occur if $S(0) > \nu/\beta$ since S is decreasing). Thus

$$I_{max} = I_0 + S_0 - \frac{\nu}{\beta} + \frac{\nu}{\beta} \ln \frac{\nu}{\beta S_0}. \tag{2.2.11}$$

2.4 (Mis)-matching models

In 1978 there was a report with detailed statistics of a flu epidemic in a boys boarding school with a total of 763 boys. Of these, 512 were confined to bed during the epidemic, which lasted from 22nd January to 4th February 1978. It seems that one infected boy initiated the epidemic. When a boy was infected he was put to bed and so we have $I(t)$ directly from the data.

A best fit numerical technique was used directly on the equations

$$\begin{aligned} S' &= -\beta IS, \\ I' &= \beta IS - \nu I, \\ R' &= \nu I. \end{aligned} \tag{2.2.12}$$

for comparison of the data. These gave $\beta = 2.1810^{-3}$/day, $\nu = 0.44$; that is, infectious period of 2.27 days, and

$$\mathcal{R}_0 = 2.18 \cdot 10^{-3} \cdot 762 \cdot 2.27 \approx 3.77.$$

However, the above approach ignores that flu, like most other diseases, has a latent period – there is a delay of 1 to 4 days in an infected becoming infective. The simplest way of incorporating the delay is to introduce the exposed class(es). In the case discussed here

$$\begin{aligned} S' &= -\beta IS, \\ E' &= \beta IS - \sigma E, \\ I' &= \sigma E - \nu I, \\ R' &= \nu I. \end{aligned} \tag{2.2.13}$$

How different can be SIR and SEIR models resulting from fitting the same data? We compare the estimated $\mathcal{R}_0$ numbers. First, observe that for the SEIR model, $\mathcal{R}_0^{SIR}$ is given by the same formula

$$\mathcal{R}_0^{SEIR} = \frac{\beta S(0)}{\nu}.$$

We use A. Lloyd approach. If SIR and SEIR models give the same data, their initial growth rate of I should be the same. For SIR initially

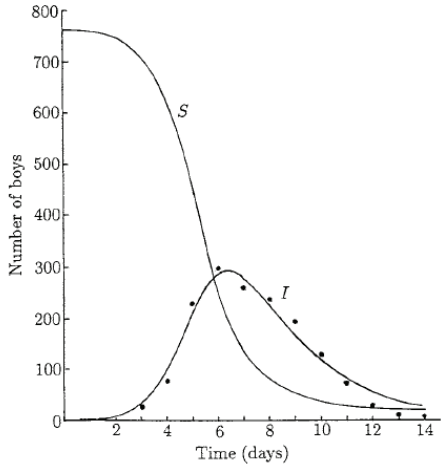


Fig. 2.2. Parameter values are $N = 763$, $S_0 = 762$, $I_0 = 1$ and, fitted, $\beta = 2.1810^{-3}$/day, $\nu = 0.44$.

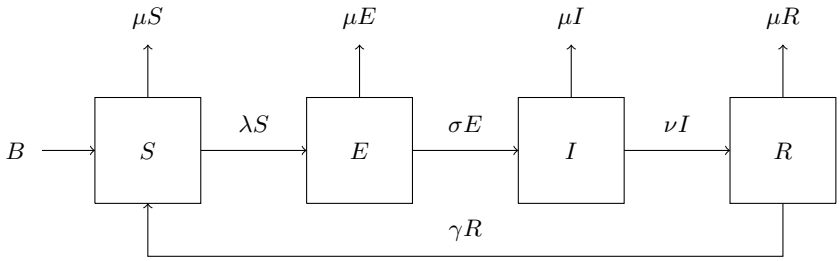


Fig. 2.3. Compartments in SEIRS model

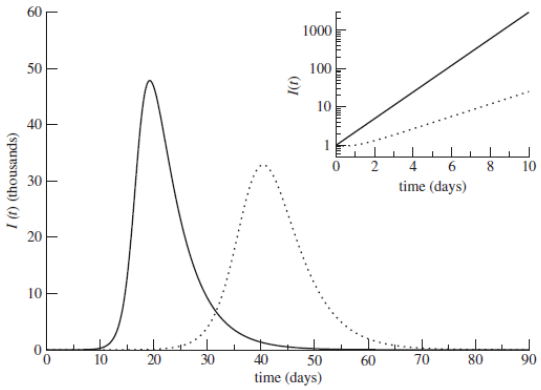


Fig. 2.4. Solid curve: SIR model, dotted curve: SEIR model. The inset compares the initial behavior of the two outbreaks. $1/\nu = 5$ days, $\mathcal{R}_0 = 5$, $1/\sigma = 2$ days, $N_0 = S_0 = 10^6$.

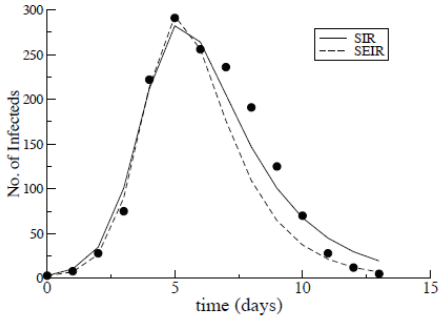


Fig. 2.5. SIR and SEIR models fitted to the available data (first half).

$$I' = \beta S(0)I - \nu I = \nu(\mathcal{R}_0^{SIR} - 1)I$$

initially $I(t) = I_0 e^{rt}$ with $r = \nu(\mathcal{R}_0^{SIR} - 1)$ or $\mathcal{R}_0^{SIR} = 1 + \nu^{-1} r$. On the other hand, for SEIR

$$E' = \beta S(0)I - \sigma E,$$
$$I' = \sigma E - \nu I.$$

and the initial growth rate r is the biggest root of the eigenvalue equation

$$r^2 + (\nu + \sigma)r - \sigma\nu(R_0^{SEIR} - 1) = 0$$

or

$$\mathcal{R}_0^{SEIR} = (1 + \nu^{-1} r)(1 + \sigma^{-1} r) = \mathcal{R}_0^{SIR}(1 + \sigma^{-1} r).$$

If we observe the same data and try to use SIR and SEIR models, the observable r

$$r = r^{SIR} = r^{SEIR} = 1.22.$$

If we add the latency period of 1 day

$$\mathcal{R}_0^{SEIR} = \mathcal{R}_0^{SIR}(1 + r) = 3.77 \cdot 2.22 = 8.37.$$

2.5 Models reducible to one-dimensional problems

The SIS model

If the disease does not induce immunity but, instead, after recovery the infected individuals become again susceptible, then the SIR model turns into the SIS model

$$S' = -\beta SI + \alpha I,$$
$$I' = \beta SI - \alpha I, \tag{2.2.14}$$

where α is the rate of recovery. Here, again, if we add the equations, we will find that the total population $N = S + I$ is constant in time. Thus, we can write

$$S = N - I$$

and thus (5.5.81) reduces to

$$I' = \beta I(N - I) - \alpha I = (\beta N - \alpha) I \left(1 - \frac{I}{\frac{\beta N - \alpha}{\beta}}\right) = rI\left(1 - \frac{I}{K}\right). \tag{2.2.15}$$

This is the logistic equation that was analysed earlier. In particular, we have the following cases

a) $r = \beta N - \alpha < 0$, or $\frac{\beta N}{\alpha} < 1$, then the solution has only one nonnegative equilibrium, 0, that is attractive. It can be easily seen as then $K < 0$ and thus

$$I' \leq rI;$$

that is

$$I(t) \leq I(0)e^{rt}.$$

Hence $I(t) \to 0$ faster than e^{rt} and thus the disease will die out.

b) if $r > 0$, then the properties of the logistic equation shows that

$$\lim_{t\to\infty} I(t) = K = \frac{\beta N - \alpha}{\beta}.$$

Hence, the disease will permanently stay in the population.

Remark 2.2. In any epidemiological model the equilibrium $I = 0$, that always exists, is called the disease free equilibrium. A positive equilibrium, if it exists, is called an endemic equilibrium.

Remark 2.3. In both models there is a parameter $\mathcal{R}_0$ that determines the progression of the disease: if $\mathcal{R}_0 < 1$, the disease will die out and if $\mathcal{R}_0 > 1$ it will spread. In the SIR model we have

$$\mathcal{R}_0 = \frac{\beta S(0)}{\nu}$$

while in the SIS model

$$\mathcal{R}_0 = \frac{\beta N}{\alpha}.$$

Seemingly these two constants are unrelated. However, let us look at their biological meaning. The coefficient β gives the number of infections per unit time induced by one infective whereas $1/\nu$ (respectively $1/\alpha$ is the average time an infective remains infectious. Finally, if we assume that consider a population that at time $t = 0$ had no infective individuals, then the number of susceptibles at the beginning is $S(0)$ in the first case and $N = N(0)$ in the second. Thus, we have arrived at the common interpretation of $\mathcal{R}_0$

SIS model with treatment

In many cases the return of an infective to the susceptible class is due to a treatment. In the simplest case we can assume that the constant α in (5.5.81) represents the efficacy of the treatment. A more realistic model takes into account that the treatment of a single patient takes some time and thus the rate of recovery should be rather modelled by the Holling type functional response. As before, let the number of treated infectives in time T by one nurse be given by

$$C = \nu\gamma I T_a,$$

where the constant γ is the rate at which the infectives are treated (number per unit time), ν is the efficacy of the treatment and T_a is the time available for administering the treatment. Since

$$T = T_a + \gamma I T_a T_t = T_a(1 + \gamma I T_t),$$

where T_t is the average time of treatment,

$$C = \frac{\nu\gamma}{1 + \gamma T_t I} I.$$

However, $\gamma = 1/T_t$, hence we obtain the SIS model with saturated treatment as

$$\begin{aligned} S' &= -\beta SI + \frac{\nu\gamma M}{1+I} I, \\ I' &= \beta SI - \frac{\nu\gamma M}{1+I} I, \end{aligned} \tag{2.2.16}$$

where M is the number of the available medical personnel. By defining $\alpha = \nu\gamma M$, we have

$$S' = -\beta SI + \frac{\alpha}{1+I}I,$$
$$I' = \beta SI - \frac{\alpha}{1+I}I. \tag{2.2.17}$$

As in the previous subsection, $N(t) = S(t) + I(t) = N = S_0 + I_0$ is constant. Hence, substituting $S(t) = N - I(t)$ we obtain the single equation

$$I'(t) = \beta I(N-I) - \frac{\alpha I}{1+I}. \tag{2.2.18}$$

Eqn (5.5.85) is in the form of the Allee model (1.1.20). It is a separable equation that, in principle, can be solved. This, however, one one hand would produce a messy and difficult to analyse formula and, on the other, would hide a general structure that can be utilised in cases when an explicit solution is not available.

We use general one dimensional 'phase-plane' analysis to find the properties of equilibria. Denote

$$F(I) = \beta I(N-I) - \frac{\alpha I}{1+I} = I\left(\beta(N-I) - \frac{\alpha}{1+I}\right) = If(I).$$

Clearly, $I = 0$ is an equilibrium so, in particular, any solution originating from $I(0) = I_0 > 0$ satisfies $I(t) > 0$. We see that

$$F'(I) = f(I) + If'(I) \tag{2.2.19}$$

and hence $F'(0) = f(0) = \beta N - \alpha$ we obtain that if $\beta N/\alpha > 1$, then $I = 0$ is a repelling equilibrium and if $\beta N/\alpha < 1$, it is an asymptotically stable equilibrium. In the expression

$$\mathcal{R}_0 = \frac{\beta N}{\alpha}$$

we recognize the basic reproduction number. Here it requires some explanation as the average duration of the disease is I dependent. However, the definition requires the basic reproduction number to be calculated in a population consisting only of susceptible individuals; that is, whenever in calculations of $\mathcal{R}_0$ we have an I dependent term, we put $I = 0$.

To find stability for $N = \alpha/\beta$, we use the geometrical argument. In this case

$$F(I) = -\frac{\beta I^2}{1+I}(I + N - 1)$$

so $F(I) < 0$ for $I < 0$ (provided $N > 1$) and $F(I) > 0$ for $I > 0$. Hence, $I = 0$ is repelling. Conversely, if $0 < N < 1$ (it is possible if N gives the density of the population and not the total population), then $I = 0$ becomes asymptotically stable. Finally, if $N = 1$, then $F(I)$ behaves at $I = 0$ as the negative cubic and $I = 0$ stays asymptotically stable.

Consider now endemic equilibria. These are the solutions to the quadratic equation

$$g(I) := (N-I)(1+I) = \frac{\alpha}{\beta}. \tag{2.2.20}$$

The graph of g is the downward parabola with roots at $I = -1$ and $I = N$. The maximum of g is taken at

$$I_{\max} = \frac{N-1}{2}$$

and equals

$$g(I_{\max}) = \frac{(N+1)^2}{4}.$$

Clearly, if $N < 1$, then $I_{\max} < 0$. Then, if $N > \alpha/\beta$ (that is, $\mathcal{R}_0 > 1$), then there is a unique positive equilibrium and if $\mathcal{R}_0 < 1$, there is none. If $N = 1$, $I_{\max} = 0$ and again, we have a unique positive equilibrium when $\mathcal{R}_0 > 1$ and none for $\mathcal{R}_0 \leq 1$.

A different picture emerges when $N > 1$. Then, as before, for $\mathcal{R}_0 \geq 1$ there is a unique positive solution to (5.5.87), see Fig. 5.8. If, however, $\mathcal{R}_0 < 1$, (5.5.87) may have two, one, or no solutions. The first case occurs if

$$N < \frac{\alpha}{\beta} < \frac{(N+1)^2}{4}, \tag{2.2.21}$$

see Fig. 5.9. Equivalently, in terms of $\mathcal{R}_0$,

$$\frac{4N}{(N+1)^2} < \mathcal{R}_0 < 1. \tag{2.2.22}$$

Then, if

$$\frac{\alpha}{\beta} = \frac{(N+1)^2}{4}, \tag{2.2.23}$$

then again we have one positive equilibrium and, finally, for

$$\frac{\alpha}{\beta} > \frac{(N+1)^2}{4} \tag{2.2.24}$$

there is no positive equilibrium, see Fig. 5.10.

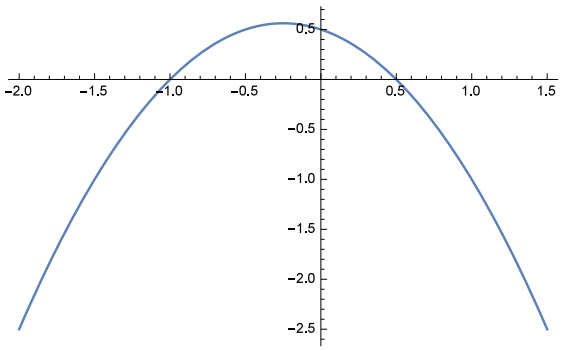


Fig. 2.6. The graph of $g(I)$ for $N = 0.5 < 1$

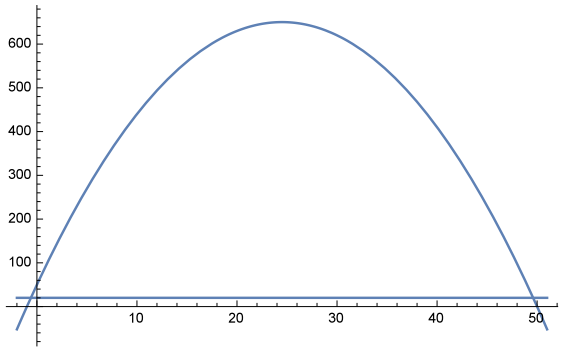


Fig. 2.7. The unique endemic equilibrium for $\mathcal{R}_0 > 1$ ($N = 50 > 1$ and $\alpha/\beta = 20$.)

To find the stability of the equilibria, we write

$$F(I) = \frac{\beta I}{1+I}\left((N-I)(1+I) - \frac{\alpha}{\beta}\right) = \frac{\beta I}{1+I}\left(g(I) - \frac{\alpha}{\beta}\right).$$

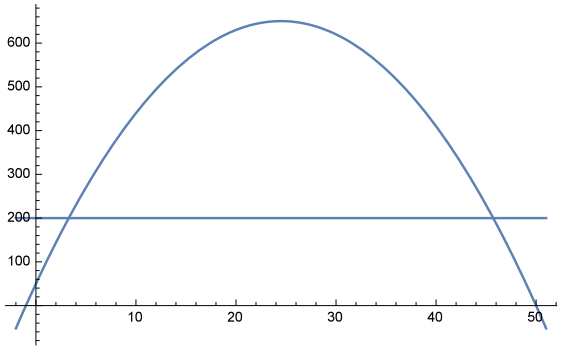


Fig. 2.8. Two endemic equilibria in the case (5.5.88).

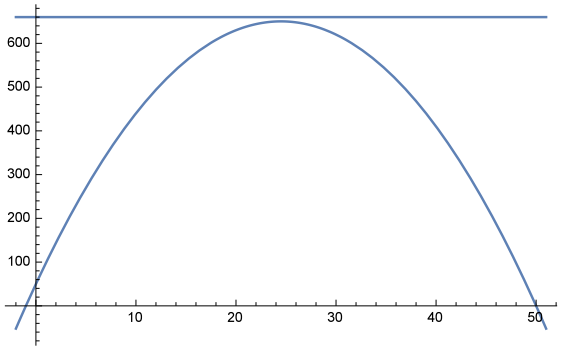


Fig. 2.9. No endemic equilibria in the case (5.5.91).

Let us denote by I_2^* the equilibrium larger than $I_{\max}$, by I_1^* the one smaller than $I_{\max}$ and by I^* the equilibrium equal to $I_{\max}$.

Case $N \leq 1$.

$\mathcal{R}_0 \leq 1$. There is only the disease free equilibrium. In this case F changes sign from positive to negative, hence 0 is asymptotically stable. Since there is no other positive equilibrium, it is globally asymptotically stable, see Figs. 2.10 and 5.15.

$\mathcal{R}_0 > 1$. There are a disease free and endemic equilibria. Here F changes sign at $I = 0$ from negative to positive, hence 0 is repelling. At the endemic equilibrium I_2^*, the function F changes sign from positive to negative and hence I_2^* is asymptotically stable (it is globally asymptotically stable for positive initial conditions), see Fig. 2.12.

Case $N > 1$.

$\mathcal{R}_0 < 1$ and $\alpha/\beta > (N+1)^2/4$. There is only the disease free equilibrium that, as above, is globally asymptotically stable, see Fig. 5.14.

$\mathcal{R}_0 < 1$ and $\alpha/\beta = (N+1)^2/4$. There is a disease free equilibrium and an endemic equilibrium I^*. The disease free equilibrium is asymptotically stable, as above, but not globally asymptotically stable. The endemic equilibrium is unstable (precisely, semi-stable – it repels solutions smaller than I^* and attracts solutions bigger than I^*, see Fig. 5.13.

$\mathcal{R}_0 < 1$ and $\alpha/\beta < (N+1)^2/4$. There is a disease free equilibrium and two endemic equilibria I_1^*, I_2^*. The disease free equilibrium is asymptotically stable, as above, but not globally asymptotically stable. The endemic equilibrium I_1^* is unstable and I_2^* is asymptotically stable. Neither stable equilibrium is globally

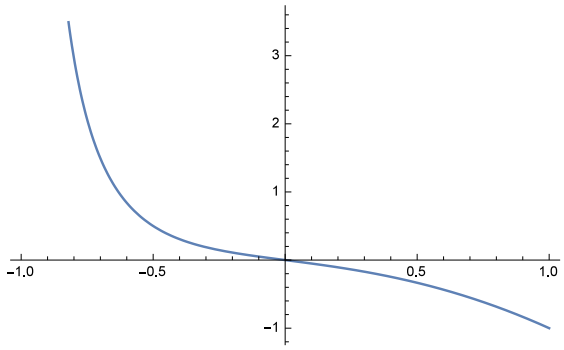


Fig. 2.10. The graph of $F(I)$ for $N = 0.5 < 1, \beta = 1, \alpha = 1, \mathcal{R}_0 = 0.5$

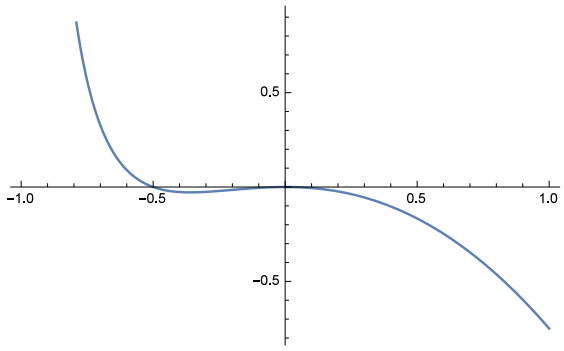


Fig. 2.11. The graph of $F(I)$ for $N = 0.5 < 1, \beta = 1, \alpha = 0.5, \mathcal{R}_0 = 1$

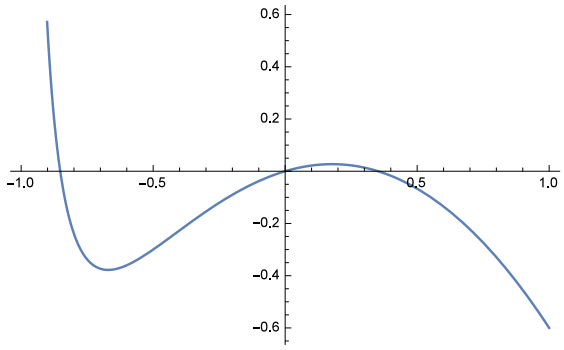


Fig. 2.12. The graph of $F(I)$ for $N = 0.5 < 1, \beta = 1, \alpha = 0.2, \mathcal{R}_0 = 2.5$

asymptotically stable: $I = 0$ attracts solutions in $[0, I_1^*)$ while I_2^* attracts solutions from (I_1^*, ∞). The intervals $[0, I_1^*)$ and (I_1^*, ∞) are called basins of attraction of respective equilibria, see Fig. 5.12.

$\mathcal{R}_0 \geq 1$. There is a disease free equilibrium and an endemic equilibrium I_2^*. The disease free equilibrium is unstable. The endemic equilibrium is asymptotically stable (and globally asymptotically stable in $(0, \infty)$), see Fig. 5.11.

Remark 2.4. It is a common (mis)perception that to control a disease if is sufficient to bring $\mathcal{R}_0$ below 1. We have seen that, indeed, the disease free equilibrium is asymptotically stable in this case but, nevertheless, the disease can persist – if the population of infectives is sufficiently large, then it will be attracted to the endemic equilibrium and the disease will not be eradicated. Only by bringing $\mathcal{R}_0$ down below $4N/(N+1)^2$ we will make the disease free equilibrium globally asymptotically stable and thus the disease will be eradicated.

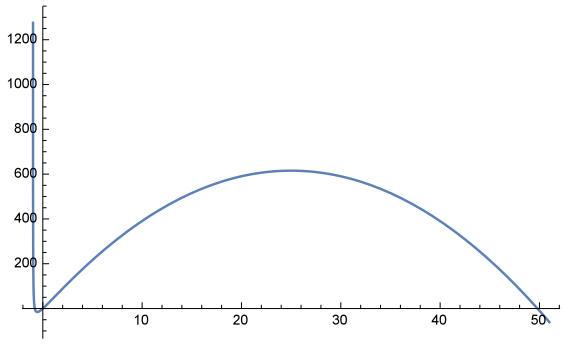


Fig. 2.13. The graph of $F(I)$ for $N = 50 > 1, \beta = 1, \alpha = 10, \mathcal{R}_0 = 5$

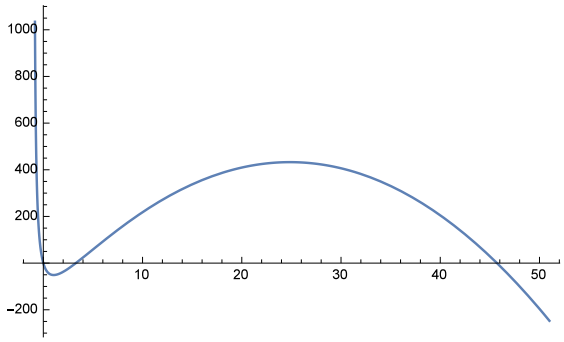


Fig. 2.14. The graph of $F(I)$ for $N = 50 > 1, \beta = 1, \alpha = 200, \mathcal{R}_0 = 0.25$

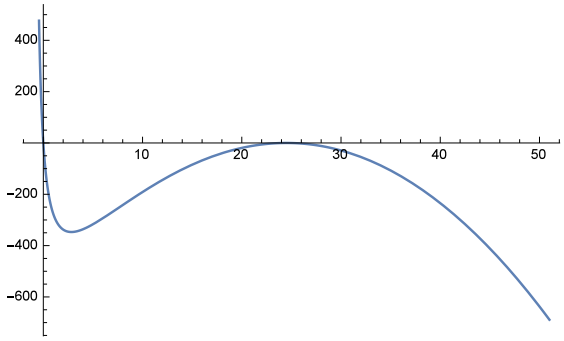


Fig. 2.15. The graph of $F(I)$ for $N = 50 > 1, \beta = 1, \alpha = 650.25 = (N+1)^2/4, \mathcal{R}_0 = 0.077$

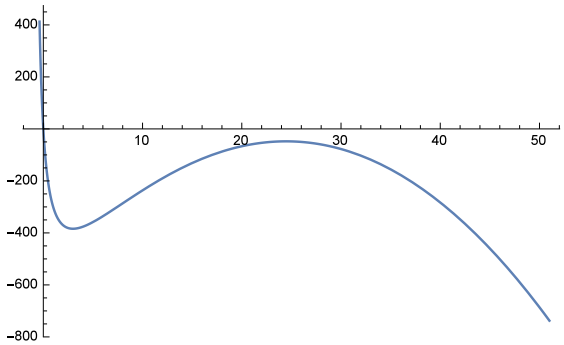


Fig. 2.16. The graph of $F(I)$ for $N = 50 > 1, \beta = 1, \alpha = 700, \mathcal{R}_0 = 0.0714$

Let us consider the implications of this observation. Assume that we have a disease that is spreading. We found the basic reproduction number

$$\mathcal{R}_0 = \frac{\beta N}{\alpha} = \frac{\beta N}{\nu\gamma M} = \frac{\beta N T_t}{\nu M} > 1.$$

To reduce $\mathcal{R}_0$ we can:

1. reduce β by e.g. using protective clothes (lower probability of transmission through contact);
2. reduce N by e.g. culling (mad cow disease, foot-and-mouth disease) or quarantine;
3. shortening the treatment time T_t;
4. improving the efficacy of the treatment ν;
5. increasing the number of medical personnel M.

Clearly, if we manage to bring $\mathcal{R}_0$ below $4N/(N+1)^2$, then the disease will be eradicated as the disease free equilibrium is then globally asymptotically stable. This is, however, often impossible or too costly. We observe that the model suggests the following alternative. If we manage to make $\mathcal{R}_0 < 1$, then the progress of the disease depends on the size of the infective population. If it is below I_1^*, that this alone will suffice to eradicate the disease. If not, then without intervention the disease will settle at I_2^*. So one needs to bring the number of infectives below I_1^*, e.g. by culling in animal diseases or the quarantine in the case of humans – this may prove less costly than further reduction of $\mathcal{R}_0$.

3

SIR models with Demography

We combine SIR model with the demography described by (1.1.10). We assume that there is no vertical transmission; that is, all individuals are born susceptible, and we assume law of mass action for the transmission of the disease. Then we have

$$\begin{aligned} S' &= \Lambda - \beta SI - \mu S, \\ I' &= \beta SI - \nu I - \mu I, \\ R' &= \nu I - \mu R. \end{aligned} \tag{3.0.1}$$

The evolution of the total population is given by (1.1.10); that is, $N(t)$ is given by (1.1.11). Contrary to the previous cases, it is not constant.

1 Non-dimensionalization

As before, the first two equations are independent of R and hence we shall work with

$$\begin{aligned} S' &= \Lambda - \beta SI - \mu S, \\ I' &= \beta SI - \nu I - \mu I, \end{aligned} \tag{3.1.2}$$

and $R(t) = N(t) - S(t) - I(t)$. To find the basic reproduction number, we must realize that here we deal with a changing population so the definition requires some modification. We are interested in stability of the disease free equilibrium so the fully susceptible population is given by $N(0) = S(0) = \Lambda/\mu$. Hence

$$\mathcal{R}_0 = \frac{\Lambda\beta}{\mu(\nu+\mu)} \tag{3.1.3}$$

The next typical step is nondimensionalization of the system that also reduces the number of parameters. First we observe that both sides must have dimension [number of people]/time. Usually to nondimensionalize, we choose new variables as ratio of the old variables to some typical quantities of the same dimension. Here, we have the average length of stay of an individual in the infective compartment $1/(\nu+\mu)$ for time and Λ/μ for the size of the population. Thus, we define new time τ by

$$\tau = (\nu+\mu)t$$

and $\hat{S}(\tau) = S(t), \hat{I}(\tau) = I(t)$. This gives

$$\hat{S}'_\tau = \frac{1}{\nu+\mu} S'_t, \quad \hat{I}'_\tau = \frac{1}{\nu+\mu} I'_t.$$

Then we introduce

$$x(\tau) = \frac{\mu \hat{S}(\tau)}{\Lambda}, \quad y(\tau) = \frac{\mu \hat{I}(\tau)}{\Lambda}$$

which allows for writing (3.1.2) in the form

$$\begin{aligned} x' &= \rho(1-x) - \mathcal{R}_0 xy, \\ y' &= (\mathcal{R}_0 x - 1)y, \end{aligned} \tag{3.1.4}$$

where

$$\rho = \frac{\mu}{\nu + \mu}.$$

We consider (3.1.4) writing t instead of τ as rescaling of time does not change the long term behaviour.

2 Basic phase-plane analysis

Since the LHS of the system is polynomial, it is locally (but not globally) Lipschitz continuous. Thus, we have the local solvability.

Further, since the equation for y is in the form

$$y' = \phi(t)y$$

where ϕ is a given function (as we know that the solution to (3.1.4) exists), $y(t) = y_0 e^{\int_0^t \psi(s)ds} \geq 0$. The equation for x is not of this form, but we can argue as follows. We have

$$x' = -x\psi(t) + \rho,$$

where $\psi(t) = \rho + \mathcal{R}_0 y(t)$ is a known function. So, the variation of constant formula gives

$$x(t) = x_0 e^{-\int_0^t \psi(s)ds} + \rho e^{-\int_0^t \psi(s)ds} \int_0^t e^{\int_0^r \psi(s)ds} dr \geq 0.$$

Hence $(x(t), y(t)) \geq 0$ as long as they exist. But

$$(x+y)' = x' + y' = \rho(1-x) - y$$

or, putting $z = x + y$

$$z' = -\rho z - (1-\rho)y + \rho.$$

Since $y(t) \geq 0$ and $\rho \leq 1$, we can write

$$z' \leq -\rho z + \rho;$$

that is

$$(ze^{\rho t})' \leq \rho e^{\rho t}$$

that, upon integration, yields

$$z(t) \leq e^{-\rho t} z_0 + (1 - e^{-\rho t}) = 1 + (z_0 - 1)e^{-\rho t}.$$

From this we see that if $z_0 \leq 1$ (equivalently, $S(0) + I(0) \leq \Lambda/\mu$), then $z(t) \leq 1$ (equivalently, $N(t) \leq \Lambda/\mu$). On the other hand, if $z_0 > 1$, then

$$z(t) \leq 1 + (z_0 - 1)e^{-\rho t} \leq z_0$$

as the second term is decreasing with maximum attained at $t = 0$. Moreover, as the RHS of the inequality converges to 1, we obtain

$$\limsup_{t \to \infty} z(t) \leq 1. \tag{3.2.5}$$

We say that the region

$$V = \{(x, y);\ x \geq 0, y \geq 0, x + y \leq 1\}$$

is positively invariant under the flow generated by (3.1.4). Also, due to (3.2.5), $\mathbb{R}^2_+$ is a basin of attraction for V.

From these considerations we obtain that the solution $(x(t), y(t))$ exists globally.

Next we consider the equilibria of (3.1.4):

$$\begin{aligned} 0 &= \rho(1 - x) - \mathcal{R}_0 xy, \\ 0 &= (\mathcal{R}_0 x - 1)y. \end{aligned} \tag{3.2.6}$$

Solving, gives $(x_0^*, y_0^*) = (1, 0)$, or

$$(S_0^*, I_0^*) = \left(\frac{\Lambda}{\mu}, 0\right)$$

which is the disease free equilibrium. We say that this is a boundary equilibrium as it is placed on the boundary of the feasible region $\mathbb{R}_+$. Further, if $y \neq 0$, then we obtain $x = 1/\mathcal{R}_0$ and $y = \rho(1 - 1/\mathcal{R}_0)$. Thus, we have the endemic equilibrium

$$(x^*, y^*) = \left(\frac{1}{\mathcal{R}_0}, \rho\left(1 - \frac{1}{\mathcal{R}_0}\right)\right). \tag{3.2.7}$$

We start with basic phase-plane analysis of (3.1.4) to get a better understanding of its dynamics.

We begin with sketching the nullclines of (3.1.4).

1. **x-nullcline.** $x' = 0$ if and only if
$$\rho(1 - x) - \mathcal{R}_0 xy = 0;$$
that is,
$$y = \frac{\rho}{\mathcal{R}_0}\frac{1 - x}{x}. \tag{3.2.8}$$

2. **y-nullcline.** $y' = 0$ if and only if
$$y = 0, \text{or } x = \frac{1}{\mathcal{R}_0}.$$

Clearly, the equilibria are given by the intersections of the nullclines. The x-nullcline intersects the y-nullcline $y = 0$ at $x = 1$, as expected. Also, the x-nullcline intersects the y-nullcline $x = 1/\mathcal{R}_0$ at $y = \rho(1 - 1/\mathcal{R}_0)$ and this equilibrium is positive only if $\mathcal{R}_0 > 0$.

There are two cases to consider.

$\mathcal{R}_0 < 1$

In this case, the x-nullcline and the boundary of $x + y = 1$ of the feasible region divide V into two subsets

$$V_1 = \left\{(x, y);\ x \geq 0, y \geq 0, y \leq \min\left\{1 - x, \frac{\rho}{\mathcal{R}_0}\frac{1 - x}{x}\right\}\right\}$$

and

$$V_2 = \left\{(x, y);\ \frac{\rho}{\mathcal{R}_0}\frac{1 - x}{x} \leq y \leq 1 - x\right\}.$$

We observe that $x' > 0$ and $y' < 0$ in the interior of V_1 and $x' < 0$ and $y' < 0$ in the interior of V_2.

Let us assume that we have a trajectory that is in the interior of V_1. Since we know that the trajectory cannot escape through $y = 0, x = 0$, or the piece of the boundary given by $y = 1 - x$, it can only escape through the

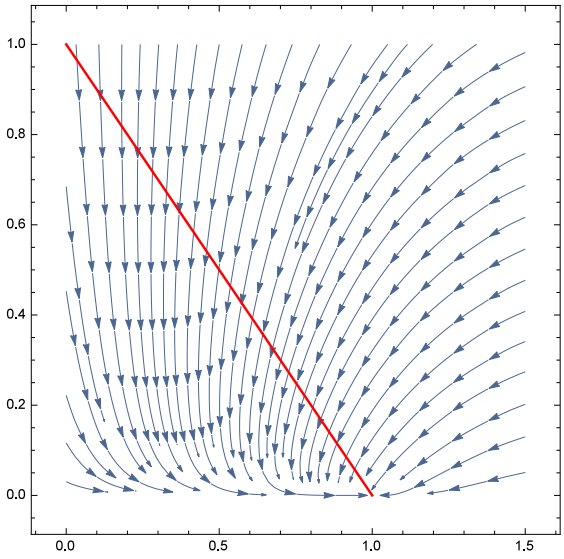


Fig. 3.1. Streamlines for $\mathcal{R}_0 = 0.5$.

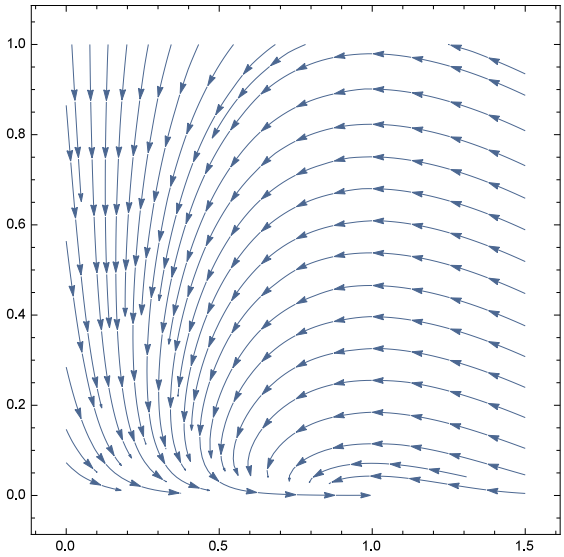


Fig. 3.2. Streamlines for $\mathcal{R}_0 = 1$.

isocline. The direction of the vector field along the x-nullcline; that is, the tangent to the trajectory, is given by $\mathbf{t} = (x', y') = (0, (\mathcal{R}_0 x - 1)y(x))$, where $y(x)$ is given by (3.2.8). At the same time, the normal vector at this part of the boundary, pointing inward V_1, is given by $\mathbf{n} = (-\mathcal{R}_0 x^2/\rho, -1)$. We see that

$$\mathbf{t} \cdot \mathbf{n} = (1 - \mathcal{R}_0 x)y(x).$$

However, $y > 0$ along the nullcline (apart from the value at DFE and $x < 1/\mathcal{R}_0$ since $\mathcal{R}_0 < 1$. Thus $\mathbf{t} \cdot \mathbf{n} > 0$ along the nullcline and thus the trajectory cannot escape V_1.

Since $x' > 0$ and $y' < 0$ in V_1, $x(t)$ is increasing and bounded, while $y(t)$ is decreasing and bounded. Hence, there is $(x_0, y_0) \in V_1$ such that

$$\lim_{t\to\infty} x(t) = x_0, \quad \lim_{t\to\infty} y(t) = y_0.$$

However, we have the following result.

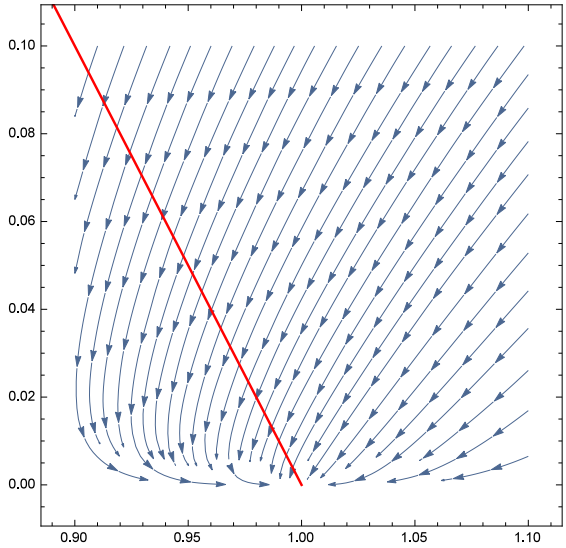


Fig. 3.3. Close up of the streamlines for close to the DFE for $\mathcal{R}_0 = 1$. Note that there are trajectories with $x(t) > 1$ for all t.

Lemma 3.1. *Assume that* $\mathbf{x} : [0, \infty) \to \mathbb{R}^n$ *is differentiable with* $\mathbf{x}'$ *uniformly continuous on* $[0, \infty)$. *If* $\mathbf{x}$ *satisfies* $\lim_{t\to\infty} \mathbf{x}(t) = \mathbf{x}_0$, *then* $\lim_{t\to\infty} \mathbf{x}'(t) = 0$.

Proof. Let us take an arbitrary $\epsilon > 0$. From uniform continuity of $\mathbf{f}$ we know that there is $h > 0$ such that

$$\|\mathbf{x}'(t_1) - \mathbf{x}'(t_2)\| < \epsilon$$

if only $|t_1 - t_2| < h$. Then from the Mean Value Theorem, for each $i = 1, \ldots, n$, there is $0 \le \tau_i(t) < h$ such that

$$x_i(t + h) - x_i(t) = x_i'(t + \tau(t))h = x_i'(t)h + (x_i'(t + \tau(t)) - x_i'(t))h.$$

Also, for that ϵ and h selected above, we can find t_0 such that for all $t \ge t_0$ we have

$$|x_i(t + h) - x_i(t)| \le |x_i(t + h) - x_{i,0}| + |x_i(t) - x_{i,0}| < 2\epsilon h.$$

Hence

$$|x_i'(t)| \le h^{-1}|x_i(t + h) - x_i(t)| + |(x_i'(t + \tau(t)) - x_i'(t))| < 3\epsilon.$$

Thus $\lim_{t\to\infty} \mathbf{x}'(t) = 0$. □

To use this lemma we observe that

$$\begin{aligned} x'' &= -\rho x' - \mathcal{R}_0 x' y - \mathcal{R}_0 x y' = -(\rho + \mathcal{R}_0)(\rho(1 - x) - \mathcal{R}_0 x y) - \mathcal{R}_0 x y(\mathcal{R}_0 x - 1) \\ y'' &= \mathcal{R}_0 x' y + \mathcal{R}_0 x y' - y' = \mathcal{R}_0(\rho(1 - x) - \mathcal{R}_0 x y) + y(\mathcal{R}_0 x - 1)^2 \end{aligned}$$

and since the solutions are bounded

$$\|(x'(t + h), y'(t + h)) - (x'(t), y'(t))\| \le \|(x''(\tau), y''(\tau)\|h \le Kh$$

for some constant K, the derivative (x', y') is uniformly continuous. Passing now in

$$\begin{aligned} x'(t) &= \rho(1 - x(t)) - \mathcal{R}_0 x(t) y(t), \\ y'(t) &= (\mathcal{R}_0 x(t) - 1) y(t), \end{aligned}$$

with t to infinity and using the lemma, we obtain

$$0 = \rho(1 - x_0) - \mathcal{R}_0 x_0 y_0,$$
$$0 = (\mathcal{R}_0 x_0 - 1) y_0,$$

hence (x_0, y_0) is an equilibrium. Since, however, $(1, 0)$ is the only equilibrium in this case, we have $(x_0, y_0) = 0$. Next, if we have a solution originating in V_2, then it cannot escape through $x + y = 1$ so it will either stay in V_2 or enter V_1. If it stays in V_2 then, as before, the derivatives x' and y' are of constants sign and, as the solutions are bounded, they must converge within V_2. By the same argument, the solution must converge to the DFE. On the other hand, if the solution enters V_1, then we can apply the argument of the first part.

Finally, assume that $z(0) > 1$ so that the trajectory starts outside V. By (3.2.5), $\limsup_{t\to\infty} z(t) \leq 1$. There are two cases to consider. If $\liminf_{t\to\infty} z(t) < 1$, then for some t_0 we must have $z(t_0) < 1$ and thus the trajectory is inside V and the previous argument applies (by the flow property). On the other hand, if $\liminf_{t\to\infty} z(t) = 1$, then $\lim_{t\to\infty} z(t) = 1$. The argument below takes into account the fact that we may have $\mathcal{R}_0 = 1$ and thus the y isocline could pass through the DFE. First we observe that along the line $x = 1$ the normal pointing to the left is given by $\mathbf{n} = (-1, 0)$ while the field is given by $\mathbf{t} = (-\mathcal{R}_0 y, (\mathcal{R}_0 - 1)y)$ so

$$\mathbf{t} \cdot \mathbf{n} = \mathcal{R}_0 y > 0,$$

provided $y > 0$. However, the point $(1, 0)$ is the equilibrium, so no trajectory can move from the region $\{x \leq 1\}$ to $\{x > 1\}$. Hence, if $z(0) > 1$ with $x(0) \leq 1$, we have $y'(t) < 0$ for $t > 0$ and thus $y(t)$ converges. Since $z(t)$ converges and so must $x(t)$. So, the last case to consider is when $x(0) > 1$ (this argument is important only if $x(0) > 1/\mathcal{R}_0$, as then initially $y' > 0$). If $x(t)$ converges, then $y(t)$ converges and the limit must be the DFE, hence $y(t)$ must converge to 0. Assume then that $y(t)$ does not converge to 0. Then $y(t_n) \geq c > 0$ for some $t_n \to \infty$. But since $x(t) + y(t) = z(t) \to 1$, then $x(t_n) < 1$ for sufficiently large t_n and hence $(x(t), y(t))$ enters the region $\{x \leq 1\}$, to which the previous argument applies.

$\mathcal{R}_0 > 1$

In this case we have two equilibria – the disease free equilibrium (DFE)

$$(x_0^*, y_0^*) = (1, 0)$$

and the endemic equilibrium (EE)

$$(x_1^*, y_1^*) = \left(\frac{1}{\mathcal{R}_0}, \rho\left(1 - \frac{1}{\mathcal{R}_0}\right)\right).$$

Clearly, DFE cannot be globally stable (but it can be globally stable in the interior of the feasible domain).

Local stability.

We proceed by analysing the local stability of these equilibria. The Jacobian of the nondimensional model (3.1.4) at an equilibrium (x^*, y^*) is

$$\mathcal{J}(x^*, y^*) = \begin{pmatrix} -\rho - \mathcal{R}_0 y^* & -\mathcal{R}_0 x^* \\ \mathcal{R}_0 y^* & \mathcal{R}_0 x^* - 1 \end{pmatrix}. \tag{3.2.9}$$

Hence, at DFE $(1, 0)$ we have

$$\mathcal{J}(1, 0) = \begin{pmatrix} -\rho & -\mathcal{R}_0 \\ 0 & \mathcal{R}_0 - 1 \end{pmatrix} \tag{3.2.10}$$

and, since the matrix is upper triangular, we immediately get eigenvalues

$$\lambda_1 = -\rho, \quad \lambda_2 = \mathcal{R}_0 - 1.$$

The first eigenvalue is negative, while the second is negative if $\mathcal{R}_0 < 1$. As we observed earlier, in this case there is no other equilibrium (in the biologically feasible domain) and we proved that DFE is globally

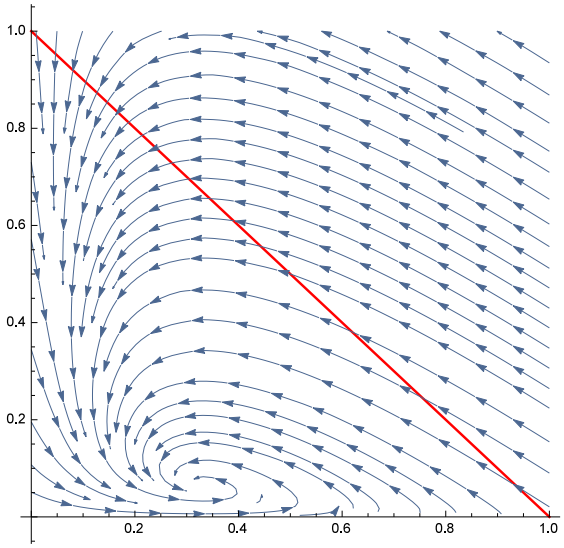


Fig. 3.4. Streamlines and V as the basin of attraction for (3.1.4) with $\mathcal{R}_0 > 1$.

asymptotically stable. The same was proved to be true if $\mathcal{R}_0 = 1$ so we only need to to consider the case $\mathcal{R}_0 > 1$. In this case DFE is a saddle and we move to investigating the endemic equilibrium. We have

$$\mathcal{J}(x_1^*, y_1^*) = \begin{pmatrix} -\rho - \mathcal{R}_0 y_1^* & -\mathcal{R}_0 x_1^* \\ \mathcal{R}_0 y_1^* & \mathcal{R}_0 x_1^* - 1 \end{pmatrix}. \tag{3.2.11}$$

From the equation for EE we have $\mathcal{R}_0 x_1^* - 1 = 0$ so

$$\mathcal{J}(x_1^*, y_1^*) = \begin{pmatrix} -\rho - \mathcal{R}_0 y_1^* & -1 \\ \mathcal{R}_0 y_1^* & 0 \end{pmatrix} \tag{3.2.12}$$

and we obtain the characteristic equation as

$$\lambda^2 + (\rho + \mathcal{R}_0 y_1^*)\lambda + \mathcal{R}_0 y_1^* = 0 \tag{3.2.13}$$

Using $y_1^* = \rho\left(1 - \frac{1}{\mathcal{R}_0}\right)$, we find

$$\rho + \mathcal{R}_0 y_1^* = \rho + \mathcal{R}_0 \rho \left(1 - \frac{1}{\mathcal{R}_0}\right) = \rho \mathcal{R}_0,$$
$$\mathcal{R}_0 y_1^* = \mathcal{R}_0 \rho \left(1 - \frac{1}{\mathcal{R}_0}\right) = \rho(\mathcal{R}_0 - 1),$$

so that (3.2.13) becomes

$$\lambda^2 + \rho \mathcal{R}_0 \lambda + \rho(\mathcal{R}_0 - 1) = 0. \tag{3.2.14}$$

Thus, the roots of the characteristic equation are

$$\lambda_{1,2} = \frac{-\rho \mathcal{R}_0 \pm \sqrt{\Delta}}{2}, \quad \Delta = (\rho \mathcal{R}_0)^2 - 4\rho(\mathcal{R}_0 - 1).$$

Hence, if $\Delta > 0$, the characteristic equation has two negative real roots and therefore the endemic equilibrium is a stable node. If $\Delta < 0$, then the characteristic equation has two complex conjugate roots with negative real parts and therefore the endemic equilibrium is a stable focus.

'Global' stability of the endemic equilibrium

We prove the following theorem.

Theorem 3.2. *Assume $\mathcal{R}_0 > 1$. If $y_0 > 0$, then*

$$\lim_{t\to\infty} \phi(t,(x_0,y_0)) = (x_1^*, y_1^*) = \left(\frac{1}{\mathcal{R}_0}, \rho\left(1-\frac{1}{\mathcal{R}_0}\right)\right).$$

In other words

$$\omega(\Gamma_{(x_0,y_0)}) = (x_1^*, y_1^*)$$

provided $y_0 > 0$.

Proof. We know from the previous considerations that

$$\omega(\Gamma_{(x_0,y_0)}) \subset V = \{(x,y);\ x \geq 0, y \geq 0, x+y \leq 1\}.$$

Moreover, for the initial condition of the form $(x_0,y_0) = (x_0,0)$, system (3.1.4) can be reduced to the single equation

$$x' = \rho(1-x), \quad x(0) = x_0$$

with the solution $x(t) = 1 + (x_0-1)e^{-\rho t}$ and hence the solution to (3.1.4) is $(1+(x_0-1)e^{-\rho t}, 0)$. Therefore the whole line $y = 0$ is a trajectory and $\omega(\Gamma_{(x_0,0)}) = (1,0)$. First, we rule out periodic trajectories in V using the Dulac–Bendixon criterion, Proposition 5.41. We observe that taking $g = 1/y$ simplifies the coordinates of the vector field and we gave

$$\frac{\partial}{\partial x}\frac{\rho(1-x)-\mathcal{R}_0 xy}{y} + \frac{\partial}{\partial y}\frac{y(R_0x-1}{y} = -\frac{\rho}{y} - \mathcal{R}_0 < 0.$$

Hence, there are no periodic orbits for $y > 0$. However, if there was a periodic orbit with $y = 0$, then this orbit and the orbit $\{y = 0\}$ would intersect which is impossible.

Thus, we only have to prove that $(1,0) \notin \omega(\Gamma_{(x_0,y_0)})$ with $y_0 > 0$. Assume to the contrary. Then there is a sequence $(t_n)_{n\in\mathbb{N}}$ with $t_n \to \infty$ as $n \to \infty$ such that $(x(t_n), y(t_n)) \to (1,0)$. We can assume that $x(t_n) > 1/\mathcal{R}_0$ for all n. Then we have $y'(t_n) > 0$ for any n. Let us take arbitrary n. Since $y(t_n) > 0$ and $y(t_n) \to 0$ as $n \to \infty$, we cannot have $y'(t) \geq 0$ for $t \geq t_n$. Thus, there is $t_n' > t_n$ with $y'(t_n') < 0$. Then there must be an index $n' > n$ such that $y'(t_{n'}) > 0$. To avoid working with subsequences, we select from $(t_n)_{n\in\mathbb{N}}$ a subsequence (without changing notation) in such a way that $t_n < t_n' < t_{n+1}$ (that is, we throw away from $(t_n)_{n\in\mathbb{N}}$ all elements with indices between n and n' and rename n' to be $n+1$). Then there must be $\hat{t}_n \in (t_n', t_{n+1})$ where $y'(\hat{t}_n) = 0$. Again, we select $\hat{t}_n$ to be the largest t in (t_n', t_{n+1}) with this property. Then we have

$$0 \leq y(\hat{t}_n) \leq y(t_n), \quad x(\hat{t}_n) = 1/\mathcal{R}_0$$

as $(x(\hat{t}_n), y(\hat{t}_n))$ is on the y' isocline and $y'(t) > 0$ for $t \in (\hat{t}_n, t_{n+1})$. Thus, by the Sandwich Theorem, $y(\hat{t}_n) \to 0$ as $n \to \infty$ and thus $(1/\mathcal{R}_0, 0) \in \omega(\Gamma_{(x_0,y_0)}$. But then, by Lemma 5.25, p. 3, $\{y = 0\} \in \omega(\Gamma_{(x_0,y_0)}$ as $\{y = 0\}$ is an orbit. On the other hand, as before,

$$(x+y)' = x' + y' = \rho(1-x) - y$$

hence

$$(x+y)' \geq -(x+y) + \rho;$$

that is

$$(x+y)(t) \geq \rho + (x_0+y_0-1)e^{-t}.$$

Thus, if $(x,y) \in \omega(\Gamma_{(x_0,y_0)})$, then $x + y \geq \rho$ only points $(x,0)$ with $x \geq \rho$ could potentially be in $\omega(\Gamma_{(x_0,y_0)})$. This contradicts the above result that $\{y = 0\} \subset \omega(\Gamma_{(x_0,y_0)})$. Hence DFE is not in $\omega(\Gamma_{(x_0,y_0)})$ and the theorem is proved. □

4

SEIR model and global stability by Lyapunov function

1 Local stability

Let us consider an SEIR model in which an infected person does not become infective for some time. Such a person, infected but not infective, is called exposed; the class of exposed individuals accordingly is denoted by E. Again, the simplest assumption is that exposed individuals move to the infective class at a constant rate. Using the standard compartmental modelling argument with mass action infection force and constant influx rate (as in (3.0.1), we arrive at the system

$$\begin{aligned} S' &= \Lambda - \beta SI - \mu S, \\ E' &= \beta SI - (\mu + \gamma)E, \\ I' &= \gamma E - (\mu + \nu)I, \\ R' &= \nu I - \mu R, \end{aligned} \tag{4.1.1}$$

where, as before, Λ is a constant recruitment rate, β is the transmission coefficient, μ is the constant death rate and ν is the recovery rate (that is, $1/\nu$ is the average infectious time of an individual); the new parameter γ is the rate at which the exposed individuals become infective.

The global solvability of (4.2.10), and positivity and boundedness of solutions can be done as in the SIR model. Our main task is to determine long time behaviour of solutions. As usual, we begin with the equilibria. For this we solve

$$\begin{aligned} 0 &= \Lambda - \beta SI - \mu S, \\ 0 &= \beta SI - (\mu + \gamma)E, \\ 0 &= \gamma E - (\mu + \nu)I, \\ 0 &= \nu I - \mu R. \end{aligned} \tag{4.1.2}$$

Clearly, we have the disease free equilibrium

$$E_0 = (\Lambda/\mu, 0, 0, 0). \tag{4.1.3}$$

Now, we can define the basic reproduction number $\mathcal{R}_0$. The totally susceptible population at equilibrium is Λ/μ and one infective will infect $\beta\Lambda/\mu$ susceptibles in the unit time. Now, the probability that an infected person will become infective (that is, it will survive the exposed class) is $\gamma/(\mu + \gamma)$ and it is infective for $1/(\mu + \nu)$ units of time. Thus

$$\mathcal{R}_0 = \frac{\Lambda\beta\gamma}{(\mu + \gamma)(\mu + \nu)\mu}. \tag{4.1.4}$$

To find the endemic equilibria, we see that

$$E = \frac{\mu + \nu}{\gamma} I$$

and then, taking into account that in an endemic equilibrium $I \neq 0$

$$S = \frac{(\mu+\gamma)(\mu+\nu)}{\gamma\beta}.$$

Then, from the first equation

$$I = \frac{\Lambda}{\beta S} - \frac{\mu}{\beta} = \frac{\mu}{\beta}(\mathcal{R}_0 - 1).$$

Re-writing the equilibria in terms of $\mathcal{R}_0$, we can formulate the first result.

Proposition 4.1. *The SEIR system (4.2.10) has a unique disease free equilibrium*

$$E_0 = \left(\frac{\Lambda}{\mu}, 0, 0, 0\right).$$

If $\mathcal{R}_0 > 1$, it has a unique endemic equilibrium

$$\mathcal{E}^* = (S^*, E^*, I^*, R^*), \tag{4.1.5}$$

where

$$\begin{aligned} S^* &= \frac{(\mu+\gamma)(\mu+\nu)}{\gamma\beta}, \\ E^* &= \frac{\mu+\nu}{\gamma}\frac{\mu}{\beta}(\mathcal{R}_0 - 1), \\ I^* &= \frac{\mu}{\beta}(\mathcal{R}_0 - 1), \\ R^* &= \frac{\nu}{\beta}(\mathcal{R}_0 - 1). \end{aligned}$$

To investigate local stability, we find the Jacobian of (4.2.10). We get

$$J = \begin{pmatrix} -\beta I - \mu & 0 & -\beta S & 0 \\ \beta I & -(\mu+\gamma) & \beta S & 0 \\ 0 & \gamma & -(\mu+\nu) & 0 \\ 0 & 0 & \nu & -\mu \end{pmatrix}. \tag{4.1.6}$$

At the DFE we have

$$J(E_0) = \begin{pmatrix} -\mu & 0 & -\beta\frac{\Lambda}{\mu} & 0 \\ 0 & -(\mu+\gamma) & \beta\frac{\Lambda}{\mu} & 0 \\ 0 & \gamma & -(\mu+\nu) & 0 \\ 0 & 0 & \nu & -\mu \end{pmatrix}. \tag{4.1.7}$$

Expanding the characteristic equation $\det(J(E_0) - \lambda I) = 0$ we have

$$(\lambda+\mu)^2 \begin{vmatrix} -(\mu+\gamma+\lambda) & \beta\frac{\Lambda}{\mu} \\ \gamma & -(\mu+\nu+\lambda) \end{vmatrix} = (\lambda+\mu)^2 \left((\mu+\gamma+\lambda)(\mu+\nu+\lambda) - \frac{\gamma\beta\Lambda}{\mu}\right) = 0.$$

The quadratic equation in brackets can be written as

$$\lambda^2 + (2\mu+\nu+\gamma)\lambda + (\mu+\gamma)(\mu+\nu)(1-\mathcal{R}_0) = 0$$

from which it follows that if $\mathcal{R}_0 < 0$, then the equation has either two negative real roots or two complex conjugate roots with negative real parts. If $\mathcal{R}_0 > 1$, then we have real roots of opposite sign. Hence we have

Proposition 4.2. *If $\mathcal{R}_0 < 1$, then the disease free equilibrium is locally asymptotically stable. If $\mathcal{R}_0 > 1$, then the disease free equilibrium is unstable.*

Let us turn our attention to the endemic equilibrium E^*. The characteristic equation is given by

$$\det(J(E^*)-\lambda I)=\begin{vmatrix} -\beta I^*-\mu-\lambda & 0 & -\beta S^* & 0\\ \beta I^* & -(\mu+\gamma+\lambda) & \beta S^* & 0\\ 0 & \gamma & -(\mu+\nu+\lambda) & 0\\ 0 & 0 & \nu & -(\mu+\lambda)\end{vmatrix}=0. \tag{4.1.8}$$

Adding the first row to the second, we obtain

$$\begin{vmatrix} -\beta I^*-\mu-\lambda & 0 & -\beta S^* & 0\\ -(\mu+\lambda) & -(\mu+\gamma+\lambda) & 0 & 0\\ 0 & \gamma & -(\mu+\nu+\lambda) & 0\\ 0 & 0 & \nu & -(\mu+\lambda)\end{vmatrix}=0.$$

Expanding by the last column yields

$$(\mu+\lambda)\begin{vmatrix} -\beta I^*-\mu-\lambda & 0 & -\beta S^*\\ -(\mu+\lambda) & -(\mu+\gamma+\lambda) & 0\\ 0 & \gamma & -(\mu+\nu+\lambda)\end{vmatrix}=0.$$

Hence, we obtain an eigenvalue $\lambda=\mu$ and, from the determinant, the cubic equation

$$(\beta I^*+\mu+\lambda)(\mu+\gamma+\lambda)(\mu+\nu+\lambda)=\beta\gamma S^*=(\mu+\gamma)(\mu+\nu)(\mu+\lambda).$$

The sign of the real parts of the eigenvalues can be determined from Routh–Hurwitz criterion but often it can be done in a simpler way, by some smart observation. By inspection we see that $\lambda=-\mu$ is not a root of the above hence, dividing by the right hand side, we get

$$\left(1+\frac{\beta I^*}{\mu+\lambda}\right)\left(1+\frac{\lambda}{\mu+\gamma}\right)\left(1+\frac{\lambda}{\mu+\nu}\right)=1. \tag{4.1.9}$$

Now, if we have complex number $1+z$, then $|1+z|=\sqrt{(1+\Re z)^2+\Im z^2}$ and $|1+z|\geq 1$ provided $\Re z\geq 0$ with strict equality if $\Re z>0$. Now,

$$\Re\frac{\beta I^*}{\mu+\lambda}=\frac{\beta I^*(\mu+\Re\lambda)}{|\mu+\lambda|^2}>0$$

and hence

$$\left|\left(1+\frac{\beta I^*}{\mu+\lambda}\right)\left(1+\frac{\lambda}{\mu+\gamma}\right)\left(1+\frac{\lambda}{\mu+\nu}\right)\right|>1$$

as long as $\Re\lambda\geq 0$. Thus (4.1.9) cannot have solutions with $\Re\lambda\geq 0$. Thus, we can formulate

Proposition 4.3. *Let $\mathcal{R}_0>1$. Then the endemic equilibrium is locally asymptotically stable.*

2 Global stability

2.1 Global stability of the disease free equilibrium

Proposition 4.4. *If $\mathcal{R}_0<1$, then the disease free equilibrium is globally asymptotically stable (in the admissible domain).*

Proof. Since in (4.2.10) the first three equations are independent of the last one, if $(0,0,0)$ is attracting in $\mathbb{R}^3_+$, then also $R(t)\to 0$ as $t\to\infty$ for any nonnegative $R(0)$. This follows from

$$R(t) = e^{-\mu t}R(0) + \nu e^{-\mu t}\int_0^t e^{\mu s} I(s)ds$$

and the l'Hôspital rule. Hence, consider

$$\begin{aligned} S' &= \Lambda - \beta SI - \mu S, \\ E' &= \beta SI - (\mu+\gamma)E, \\ I' &= \gamma E - (\mu+\nu)I. \end{aligned} \tag{4.2.10}$$

A candidate for the Lyapunov function for such problems is

$$V(S,E,I) = \kappa\left(S - S^* - S^*\ln\frac{S}{S^*}\right) + \frac{1}{\mu+\gamma}E + \frac{1}{\gamma}, \tag{4.2.11}$$

where κ is to be determined and $S^* = \frac{\Lambda}{\mu}$.

To establish that $V > 0$ for $(S,E,I) \neq (\Lambda/\mu, 0, 0)$ we see that it is sufficient to establish

$$S - S^* - S^*\ln\frac{S}{S^*} > 0, \quad S \neq S^*. \tag{4.2.12}$$

For this we consider

$$g(x) = x - 1 - \ln x.$$

We have $g(x) \to \infty$ as $x \to 0^+$ and $x \to +\infty$. Further, $g'(x) = 1 - 1/x$, $g''(x) = 1/x^2$, hence the global minimum is attained at $x = 1$ and it is equal to $g(1) = 0$. Using $x = S/S^*$, we see that (4.2.12) is satisfied.

Next,

$$\begin{aligned} \frac{d}{dt}V &= \kappa\left(1 - \frac{S}{S^*}\right)S' + \frac{1}{\mu+\gamma}E' + \frac{1}{\gamma}I' \\ &= \kappa\left(1 - \frac{S}{S^*}\right)(\Lambda - \beta SI - \mu S) + \frac{\beta SI - (\mu+\gamma)E}{\mu+\gamma} + \frac{\gamma E - (\mu+\nu)I}{\gamma} \\ &= 2\kappa\Lambda - \beta\kappa SI - \kappa\mu S - \frac{\Lambda^2\kappa}{\mu S} + \frac{\Lambda\beta\kappa}{\mu}I + \frac{\beta}{\mu+\gamma}SI - \frac{\mu+\nu}{\gamma}I. \end{aligned}$$

Choosing $\kappa = 1/(\mu+\gamma)$ eliminates the cross term SI, giving in the last line

$$2\kappa\Lambda - \kappa\mu S - \frac{\Lambda^2\kappa}{\mu S} + \frac{\Lambda\beta\kappa}{\mu}I - \frac{\mu+\nu}{\gamma}I = -\kappa\Lambda\left(\frac{\Lambda}{\mu S} + \frac{\mu S}{\Lambda} - 2\right) + \frac{\mu+\nu}{\gamma}(\mathcal{R}_0 - 1)I.$$

Since the last term is negative due to $\mathcal{R}_0 < 1$, we focus on the first term. The term inside the bracket can be written as

$$x + \frac{1}{x} - 2 = \frac{(x-1)^2}{x} > 0, \quad x \neq 1, x > 0.$$

Thus, $V' < 0$ if $(S,E,I) \neq (S^*,0,0)$. Since V is radially unbounded in the admissible domain, we find that $(S^*,0,0)$ is globally stable. □

2.2 Global stability of the endemic equilibrium

Proposition 4.5. *If $\mathcal{R}_0 > 1$, then the endemic equilibrium*

$$\mathcal{E}^* = (S^*, E^*, I^*, R^*) = \left(\frac{(\mu+\gamma)(\mu+\nu)}{\gamma\beta}, \frac{\mu+\nu}{\gamma}\frac{\mu}{\beta}(\mathcal{R}_0 - 1), \frac{\mu}{\beta}(\mathcal{R}_0 - 1), \frac{\nu}{\beta}(\mathcal{R}_0 - 1)\right) \tag{4.2.13}$$

is globally asymptotically stable (in $\mathbb{R}^n_+$).

Proof. As before, it is sufficient to consider the first three components, (S, E, I). We assume that they stay in the positive orthant $\mathbb{R}^3_+$. We consider the candidate Lyapunov function

$$V(S,E,I) = \kappa_1 \left(S - S^* - S^* \ln \frac{S}{S^*} \right) + \kappa_2 \left(E - E^* - E^* \ln \frac{E}{E^*} \right) + \kappa_3 \left(I - I^* - I^* \ln \frac{I}{I^*} \right), \qquad (4.2.14)$$

where $\kappa_i, i = 1, 2, 3$ are constants to be determined. As in the previous proof, $V(S^*, E^*, I^*) = 0$ and $V > 0$ otherwise. Moreover, V is radially unbounded in $\mathbb{R}^3_+$. So, it remains to determine the sign of V'. We have

$$\begin{aligned}\frac{dV}{dt} &= \kappa_1 \left(1 - \frac{S^*}{S}\right) S' + \kappa_2 \left(1 - \frac{E^*}{E}\right) E' + \kappa_3 \left(1 - \frac{I^*}{I}\right) I' \\ &= \kappa_1 \left(1 - \frac{S^*}{S}\right) (\Lambda - \beta SI - \mu S) + \kappa_2 \left(1 - \frac{E^*}{E}\right) (\beta SI - (\mu + \gamma)E) + \kappa_3 \left(1 - \frac{I^*}{I}\right) (\gamma E - (\mu + \nu)I).\end{aligned}$$

First we observe that

$$\Lambda = \beta S^* I^* + \mu S^*$$

and thus

$$\kappa_1 \left(1 - \frac{S^*}{S}\right) (\Lambda - \beta SI - \mu S) = -\kappa_1 \mu \frac{(S - S^*)^2}{S} + \kappa_1 \beta S^* I^* - \kappa_1 \beta SI - \kappa_1 \beta \frac{S^{*2} I^*}{S} + \kappa_1 \beta S^* I.$$

Multiplying out the other brackets, we obtain

$$\begin{aligned}\frac{dV}{dt} = &-\kappa_1 \mu \frac{(S - S^*)^2}{S} + \kappa_1 \beta S^* I^* - \kappa_1 \beta SI - \kappa_1 \beta \frac{S^{*2} I^*}{S} + \kappa_1 \beta S^* I \\ &+ \kappa_2 \beta SI - \kappa_2 (\mu + \gamma) E - \kappa_2 \beta \frac{E^*}{E} SI + \kappa_2 E^* (\mu + \gamma) \\ &+ \kappa_3 \gamma E - \kappa_3 (\mu + \nu) I - \kappa_3 \gamma \frac{I^* E}{I} + \kappa_3 I^* (\mu + \nu).\end{aligned}$$

First, we observe that taking $\kappa_1 = \kappa_2 = 1$, we cancel the terms containing SI. Next, we group together terms containing linear terms. We have

$$\begin{aligned}&E(\gamma \kappa_3 - (\mu + \gamma)) \\ &I(\beta S^* - \kappa_3 (\mu + \nu)) = I(\mu + \nu) \left(\frac{\mu + \gamma}{\gamma} - \kappa_3 \right),\end{aligned}$$

where we used (4.2.13) to express S^*. Hence, putting

$$\kappa_3 = \frac{\mu + \gamma}{\gamma}$$

eliminates the linear terms leaving

$$\begin{aligned}\frac{dV}{dt} = &-\mu \frac{(S - S^*)^2}{S} + \beta S^* I^* - \beta \frac{S^{*2} I^*}{S} \\ &- \beta \frac{E^*}{E} SI + E^* (\mu + \gamma) - (\mu + \gamma) \frac{I^* E}{I} + \frac{\mu + \gamma}{\gamma} I^* (\mu + \nu).\end{aligned}$$

Now, we use

$$\beta S^* I^* = (\mu + \gamma) E^*$$

and again the formula for S^* to get

$$\frac{dV}{dt} = -\mu \frac{(S - S^*)^2}{S} + \beta S^* I^* \left(3 - \frac{S^*}{S} - \frac{E^* SI}{E S^* I^*} - \frac{I^* E}{I E^*} \right),$$

where again we used

$$\mu + \gamma = \frac{\beta S^* I^*}{E^*}.$$

Let us denote

$$x_1 = \frac{S^*}{S}, \quad x_2 = \frac{E^* SI}{ES^* I^*}, \quad x_3 = \frac{I^* E}{IE^*}$$

and observe that

$$x_1 x_2 x_3 = 1.$$

What is the maximum of $x_1 + x_2 + x_3$ under this constraint? We can use the relation between the harmonic and geometric means:

$$\frac{x_1 + \ldots + x_n}{n} \geq \sqrt[n]{x_1 \ldots x_n}, \quad x_1, \ldots, x_n \geq 0$$

to ascertain that $x_1 + x_2 + x_3 \geq 3$ with the equality attained if $x_1 = x_2 = x_3 = 1$. Hence we obtain

$$\frac{dV}{dt} \leq 0$$

with

$$\frac{dV}{dt} = 0$$

on the set

$$\left\{ (S, E, I);\ S = S^*, I = \frac{I^*}{E^*} E \right\}.$$

Now, along the line $I = I^* E / E^*$ the first two components of the field are given by $\Lambda - \beta I S^* - \mu S^*$ and 0 whereas the direction along the line is $(0, I^*, E^*)$. Thus, the only invariant set on the line is the equilibrium (S^*, E^*, I^*) and, by the LaSalle principle, the ω limit set of any trajectory originating from initial conditions not on the S axis consists of the endemic equilibrium. □

5

Appendices

1 Stability of equilibria of autonomous differential equations

Consider the problem

$$x' = f(x), \qquad x(t_0) = x_0. \tag{5.1.1}$$

We recall that the word autonomous refers to the fact that f in (5.1.1) does not explicitly depend on time. To have anything to talk about, we must ensure that (5.1.1) has solutions different from the equilibrium solutions. This is settled by the Picard–Lindelöf theorem which asserts that if f is sufficiently regular, for instance, differentiable on $\mathbb{R}$, then the initial value problem has exactly one solution defined for t on some interval (t_{min}, t_{max}) containing t_0. Furthermore, if the solution is bounded at the endpoints, then it can be extended to a larger interval. It is possible that the solution only is defined on a finite interval. However, if we can show that the solution is bounded on each finite interval of its existence, then it is defined on $\mathbb{R}$. In other words, if a solution to (5.1.1) with differentiable f is defined only on an interval with a finite endpoint, then it must be unbounded at this endpoint.

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

In many problems it is important to emphasize the dependence of the solution on the initial conditions. Thus we introduce the notion of the flow $x(t, t_0, x_0)$ of (5.1.1), which is the solution of the Cauchy problem (5.1.1). Here we only use $t_0 = 0$ and then we write $x(t, 0, x_0) = x(t, x_0)$.

If (5.1.1) has a stationary solution $x(t) \equiv x^*$ that, by definition, is constant in time, then such a solution satisfies $x'(t) \equiv 0$ and consequently

$$f(x^*) = 0. \tag{5.1.2}$$

Conversely, if the equation $f(x) = 0$ has a solution, which we call an *equilibrium point* then, since f is independent of time, such a solution is a number, say x^*. If we now consider a function defined by $x(t) \equiv x^*$, then $x'(t) \equiv 0$. Consequently,

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

and such a function is a stationary solution. Summarizing, equilibrium points are solutions to the algebraic equation (5.1.2) and, treated as constant functions, they are (the only) stationary, or equilibrium, solutions to (5.1.1). Therefore usually we will not differentiate between these terms.

Next we give a definition of stability of an equilibrium.

Definition 5.1. *1. The equilibrium x^* is stable if for given $\epsilon > 0$ there is $\delta > 0$ such that for any x_0 $|x_0 - x^*| < \delta$ implies $|x(t, x_0) - x^*| < \epsilon$ for all $t > 0$. If x^* is not stable, then it is called unstable.*

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

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

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

Lemma 5.2. *If x_0 is not an equilibrium point of (5.1.1), then $x(t, x_0)$ is never equal to an equilibrium point. In other words, $f(x(t, x_0)) \neq 0$ for any t for which the solution exists.*

Proof. An equilibrium point x^* generates a stationary solution, given by $x(t) \equiv x^*$. Thus, if $x(t_1, x_0) = x^*$ for some t_1, then (t_1, x_0) belongs to two different solutions, which contradicts the Picard theorem. □

From the above lemma it follows that if f has several equilibrium points, then the stationary solutions corresponding to these points divide the (t, x) plane into horizontal strips having the property that any solution always remains confined to one of them. We shall formulate and prove a theorem that strengthens this observation.

Theorem 5.3. *Let $x(t, x_0)$ be a non-stationary solution of (5.1.1) with $x_0 \in \mathbb{R}$ and let $I_{max} = (t_-, t_+)$ be its maximal interval of existence. Then $x(t, x_0)$ is either a strictly decreasing or a strictly increasing function of t. Moreover, $x(t, x_0)$ either diverges to $+\infty$ or to $-\infty$, or converges to an equilibrium point, as $t \to t_\pm$. In the latter case $t_\pm = \pm\infty$.*

Proof. Assume that for some $t_* \in I_{max}$ the solution $x(t) := x(t, x_0)$ has a local maximum or minimum $x_* = x(t_*)$. Since $x(t)$ is differentiable, we must have $x'(t_*) = 0$ but then $f(x_*) = 0$ which makes x_* an equilibrium point of f. This means that a non-stationary solution $x(t, x_0)$ reaches an equilibrium in finite time, which contradicts Lemma 5.2. Thus, if $x(t, x_0)$ is not a stationary solution, then it cannot attain local maxima or minima and thus must be either strictly increasing or strictly decreasing.

Since the solution is monotonic, it either diverges to $\pm\infty$ (depending on whether it decreases or increases) or converges to finite limits as $t \to t_\pm$. Let us focus on the right end point t_+ of I_{max}. If $x(t, x_0)$ converges as $t \to t_+$, then $t_+ = \infty$, by the property of the maximal interval of existence. Thus

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

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

$$t(x) - t(x_0) = \int_{x_0}^{x} \frac{ds}{f(s)}. \tag{5.1.3}$$

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

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

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

Remark 5.4. We note that Eq. (5.1.3) is of independent interest as it gives a formula for the blow up time of the solution $x(t, x_0)$. To wit, let the interval $[x_0, \infty)$ be free of equilibria and let $x(t, x_0)$ be increasing for $t > 0$. Then $\lim_{t\to t_+} x(t, x_0) = \infty$ so that, by (5.1.3),

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

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

Remark 5.5. Theorem 5.3 shows that for scalar differential equations with regular right hand sides, the distinction between different properties of an equilibrium made in Definition 5.1 is superfluous. Indeed, if an equilibrium x^* is stable, then solutions originating close to it stay close to it. However, by Theorem 5.3, these solutions are monotonic. Hence, the solutions are closer to x^* than their initial conditions are. In particular, they must be bounded and, being monotonic, they must converge as $t \to \infty$. From the proof of Theorem 5.3 it follows that the limit point must be the equilibrium x^*. This implies that x^* is attracting and hence asymptotically stable. Also, by monotonicity of solutions, any attracting equilibrium must be stable and thus asymptotically stable.

Remark 5.6. Theorem 5.3 usually is used in the following weaker form. Let f be continuously differentiable function. Then the equilibrium x^* is stable provided $f'(x^*) < 0$ and unstable provided $f'(x^*) > 0$. The proof is obvious–if $f'(x^*) < 0$, then $f'(x) < 0$ in some neighbourhood of x^*, by continuity of f'. Thus, $f > 0$ to the left and $f < 0$ to the right of x^* and any solution originating in such a left neighbourhood of x^* is increasing and must converge to x^*. Similarly, any solution originating in such a right neighbourhood of x^* is decreasing and also must converge to x^*. Thus x^* is asymptotically stable. An analogous argument shows that $f'(x^*) > 0$ means that x^* is unstable. However, Theorem 5.3 is much more general and allows to ascertain stability or instability in the so called nonhyperbolic cases when $f'(x^*)$ by considering the sign of f to the left and to the right of x^*.

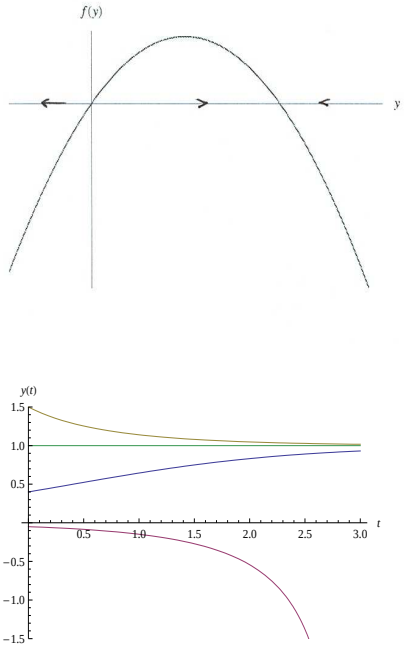


Fig. 5.1. Monotonic behaviour of solutions to (5.1.1) depends on the right hand side f of the equation.

Application to the logistic equation

Consider the Cauchy problem for the logistic equation

$$y' = y(1-y), \qquad y(0) = y_0. \tag{5.1.4}$$

Let us get as much information as possible about the solutions to this problem without actually solving it. First, we observe that the right hand side is given by $f(y) = y(1-y)$, which is a polynomial, and therefore at each point of $\mathbb{R}^2$ the assumptions of Picard's theorem are satisfied, that is, only one solution of (5.1.4) passes through each point (t_0, y_0). However, f is not a globally Lipschitz function, so that this solution may be defined only on a finite time interval.

The equilibrium points are found solving $y(1-y) = 0$, hence $y \equiv 0$ and $y \equiv 1$ are the only stationary solutions. Moreover, $f(y) < 0$ for $y < 0$ and $y > 1$ and $f(y) > 0$ for $0 < y < 1$. Hence, from Lemma 5.2, it follows that the solutions starting from $y_0 < 0$ will stay strictly negative, those starting from $0 < y_0 < 1$ will stay in this interval and those with $y_0 > 1$ will be larger than 1, for all times of their respective existence, as they cannot cross the equilibrium solutions. Then, from Theorem 5.3, we see that the solutions with negative initial condition are decreasing and therefore tend to $-\infty$ if time increases. In fact, they blow up in finite time since, by integrating the equation, we obtain

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

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

Next, solutions with $0 < y_0 < 1$ are bounded and thus they are defined for all times. They are increasing and thus they must converge to the larger equilibrium, that is, $\lim_{t\to\infty} y(t, y_0) = 1$. Finally, if we start with $y_0 > 1$, then $y(t, y_0)$ decreases and thus is bounded from below, satisfying again $\lim_{t\to\infty} y(t, y_0) = 1$. The shape of the solution curves can be determined as follows. By differentiating Eq. 5.1.4 with respect to time, we obtain

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

Since for each solution (apart from the stationary ones), y' has a fixed sign, we see that an inflection point can exist only for solutions starting at $y_0 \in (0,1)$ and it occurs at $y = 1/2$, where the solution changes from being convex downward to being convex upward. In the two other cases, the second derivative is of constant sign, giving the solution convex upward for negative solutions and convex downward for solutions larger than 1.

We see that we got the same picture as when solving the equation but with much less work.

2 Stability by linearization

2.1 Solvability of linear systems

We shall consider only linear systems of first order differential equations.

$$\begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 + \ldots + a_{1n}y_n + g_1(t), \\ &\vdots\ \vdots\ \vdots, \\ y_n' &= a_{n1}y_1 + a_{n2}y_2 + \ldots + a_{nn}y_n + g_n(t), \end{aligned} \tag{5.2.5}$$

where $y_1, \ldots, y_n$ are unknown functions, $a_{11}, \ldots a_{nn}$ are constant coefficients and $g_1(t) \ldots, g_n(t)$ are known continuous functions. If $g_1 = \ldots = g_n = 0$, then the corresponding system (5.2.5) is called the associated

homogeneous system. The structure of (5.2.5) suggest that a more economical way of writing is is to use the vector-matrix notation. Denoting $\mathbf{y} = (y_1, \ldots, y_n)$, $\mathbf{g} = (g_1, \ldots, g_n)$ and $\mathcal{A} = \{a_{ij}\}_{1 \le i,j \le n}$, that is

$$\mathcal{A} = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \ldots & a_{nn} \end{pmatrix},$$

we can write (5.2.5) is a more concise notation as

$$\mathbf{y}' = \mathcal{A}\mathbf{y} + \mathbf{g}. \tag{5.2.6}$$

Here we have n unknown functions and the system involves first derivative of each of them so that it is natural to consider (5.2.6) in conjunction with the following initial conditions

$$\mathbf{y}(t_0) = \mathbf{y^0}, \tag{5.2.7}$$

or, in the expanded form,

$$y_1(t_0) = y_1^0, \ldots, y_n(t_0) = y_n^0, \tag{5.2.8}$$

where t_0 is a given argument and $\mathbf{y^0} = (y_1^0, \ldots, y_n^0)$ is a given vector.

Let us denote by $\mathbf{X}$ the set of all solutions to the homogeneous system (5.2.5). Due to linearity of differentiation and multiplication by $\mathcal{A}$, it is easy to see that $\mathbf{X}$ is a vector space. We have two fundamental results.

Theorem 5.7. *The dimension of* $\mathbf{X}$ *is equal to* n.

Theorem 5.8. *Let* $\mathbf{y_1}, \ldots, \mathbf{y_k}$ *be* k *linearly independent solutions of* $\mathbf{y}' = \mathcal{A}\mathbf{y}$ *and let* $t_0 \in \mathbb{R}$ *be an arbitrary number. Then,* $\{\mathbf{y_1}(t), \ldots, \mathbf{y_k}(t)\}$ *for 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* $\mathbb{R}$.

These two results show that if we construct solutions emanating from n linearly independent initial vectors, then these solutions are linearly independent and therefore they span the space of all solutions to the homogeneous system (5.2.5).

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

$$\mathcal{A}\mathbf{v} = \lambda\mathbf{v}. \tag{5.2.9}$$

Such a solution is called an *eigenvector* of $\mathcal{A}$. The set of eigenvectors corresponding to a given eigenvalue is a vector subspace. Eq. (5.2.9) 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

$$det(\mathcal{A} - \lambda\mathcal{I}) = \begin{vmatrix} a_{11} - \lambda & \ldots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \ldots & a_{nn} - \lambda \end{vmatrix} = 0. \tag{5.2.10}$$

Evaluating the determinant we obtain a polynomial in λ of degree n. This polynomial is also called the characteristic polynomial of the system (5.2.5) (if (5.2.5) arises from a second order equation, then this is the same polynomial as the characteristic polynomial of the equation). We shall denote this polynomial by $p(\lambda)$. From algebra we know that there are exactly n, possibly complex, root of $p(\lambda)$. Some of them may be multiple, so that in general $p(\lambda)$ factorizes into

$$p(\lambda) = (\lambda_1 - \lambda)^{n_1} \cdot \ldots \cdot (\lambda_k - \lambda)^{n_k}, \tag{5.2.11}$$

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 roots of the characteristic polynomial of $\mathcal{A}$. The exponent n_i appearing in the factorization (5.2.11) is called the *algebraic multiplicity* of λ_i. For each eigenvalue λ_i there corresponds an eigenvector $\mathbf{v_i}$ and eigenvectors corresponding to distinct eigenvalues are linearly independent. The set of all eigenvectors corresponding to λ_i spans a subspace, called the *eigenspace* corresponding to λ_i which we will denote by E_{λ_i}. The dimension of E_{λ_i} is called the *geometric multiplicity* of λ_i. In general, algebraic and geometric multiplicities are different with geometric multiplicity being at most equal to the algebraic one. Thus, in particular, if λ_i is a single root of the characteristic polynomial, then the eigenspace corresponding to λ_1 is one-dimensional.

If the geometric multiplicities of eigenvalues add up to n, that is, if we have n linearly independent eigenvectors, then these eigenvectors form a basis for $\mathbb{R}^n$. In particular, this happens if all eigenvalues are single roots of the characteristic polynomial. If this is not the case, then we do not have sufficiently many eigenvectors to span $\mathbb{R}^n$ and if we need a basis for $\mathbb{R}^n$, then we have to find additional linearly independent vectors. A procedure that can be employed here and that will be very useful in our treatment of systems of differential equations is to find solutions to equations of the form $(\mathcal{A} - \lambda_i\mathcal{I})^k\mathbf{v} = 0$ for $1 < k \leq n_i$, where n_i is the algebraic multiplicity of λ_i. Precisely speaking, if λ_i has algebraic multiplicity n_i and if

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

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

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

It follows that all the solutions of the preceding equation solve this equation but there is at least one more independent solution so that we have at least $\nu_i + 1$ independent vectors (note that these new vectors are no longer eigenvectors). If the number of independent solutions is still less than n_1, 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 $(\mathcal{A} - \lambda_i\mathcal{I})^{j-1}\mathbf{v} \neq 0$.

Matrix exponentials

The above theory can be used to provide a unified framework for solving systems of differential equations.

Recall that for a single equation $y' = ay$, where a is a constant, the general solution is given by $y(t) = e^{at}C$, where C is a constant. In a similar way, we would like to say that the general solution to

$$\mathbf{y}' = \mathcal{A}\mathbf{y},$$

where $\mathcal{A}$ is an $n \times n$ matrix, is $\mathbf{y} = e^{\mathcal{A}t}\mathbf{v}$, where $\mathbf{v}$ is any constant vector in $\mathbb{R}^n$. The problem is that we do not know what it means to evaluate an 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

$$e^{\mathcal{A}} = \mathcal{I} + \mathcal{A} + \frac{1}{2}\mathcal{A}^2 + \frac{1}{3!}\mathcal{A}^3 + \ldots + \frac{1}{k!}\mathcal{A}^k + \ldots. \tag{5.2.12}$$

It can be shown that if $\mathcal{A}$ is a matrix, then the above series always converges and the sum is a matrix. For example, if we take

$$\mathcal{A} = \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix} = \lambda\mathcal{I},$$

then

$$\mathcal{A}^k = \lambda^k \mathcal{I}^k = \lambda^k \mathcal{I},$$

and

$$\begin{aligned} e^{\mathcal{A}} &= \mathcal{I} + \lambda\mathcal{I} + \frac{\lambda^2}{2}\mathcal{I} + \frac{\lambda^3}{3!}\mathcal{I} + \ldots + \frac{\lambda^k}{k!} + \ldots \\ &= \left(1 + \lambda + \frac{\lambda^2}{2} + \frac{\lambda^3}{3!} + \ldots + \frac{\lambda^k}{k!} + \ldots\right)\mathcal{I} \\ &= e^{\lambda}\mathcal{I}. \end{aligned} \tag{5.2.13}$$

Unfortunately, in most cases finding the explicit form for $e^{\mathcal{A}}$ directly is impossible.

Matrix exponentials have the following algebraic properties

$$\left(e^{\mathcal{A}}\right)^{-1} = e^{-\mathcal{A}}$$

and

$$e^{\mathcal{A}+\mathcal{B}} = e^{\mathcal{A}} e^{\mathcal{B}} \tag{5.2.14}$$

provided the matrices $\mathcal{A}$ and $\mathcal{B}$ commute: $\mathcal{A}\mathcal{B} = \mathcal{B}\mathcal{A}$.

Let us define a function of t by

$$e^{t\mathcal{A}} = \mathcal{I} + t\mathcal{A} + \frac{t^2}{2}\mathcal{A}^2 + \frac{t^3}{3!}\mathcal{A}^3 + \ldots + \frac{t^k}{k!}\mathcal{A}^k + \ldots\ . \tag{5.2.15}$$

It follows that this function can be differentiated with respect to t by termwise differentiation of the series, as in the scalar case, that is,

$$\begin{aligned} \frac{d}{dt}e^{\mathcal{A}t} &= \mathcal{A} + t\mathcal{A}^2 + \frac{t^2}{2!}\mathcal{A}^3 + \ldots + \frac{t^{k-1}}{(k-1)!}\mathcal{A}^k + \ldots \\ &= \mathcal{A}\left(\mathcal{I} + t\mathcal{A} + \frac{t^2}{2!}\mathcal{A}^2 + \ldots + \frac{t^{k-1}}{(k-1)!}\mathcal{A}^{k-1} + \ldots\right) \\ &= \mathcal{A}e^{t\mathcal{A}} = e^{t\mathcal{A}}\mathcal{A}, \end{aligned}$$

proving thus that $\mathbf{y}(t) = e^{t\mathcal{A}}\mathbf{v}$ is a solution to our system of equations for any constant vector $\mathbf{v}$. Since $\mathbf{y}(0) = e^{0\mathcal{A}}\mathbf{v} = \mathbf{v}$, from Picard's theorem $\mathbf{y}(t)$ is a unique solution to the Cauchy problem

$$\mathbf{y}' = \mathcal{A}\mathbf{y}, \qquad \mathbf{y}(0) = \mathbf{v}.$$

As we mentioned earlier, in general it is difficult to find directly the explicit form of $e^{t\mathcal{A}}$. However, we can always find n linearly independent vectors $\mathbf{v}$ for which the series $e^{t\mathcal{A}}\mathbf{v}$ can be summed exactly. This is based on the following two observations. Firstly, since $\lambda\mathcal{I}$ and $\mathcal{A} - \lambda\mathcal{I}$ commute, we have by (5.2.13) and (5.2.14)

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

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

$$(\mathcal{A} - \lambda\mathcal{I})^r\mathbf{v} = \mathbf{0}, \tag{5.2.16}$$

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

$$e^{t\mathcal{A}}\mathbf{v} = e^{\lambda t}e^{t(\mathcal{A}-\lambda\mathcal{I})} = e^{\lambda t}\left(\mathbf{v} + t(\mathcal{A}-\lambda\mathcal{I})\mathbf{v} + \ldots + \frac{t^{m-1}}{(m-1)!}(\mathcal{A}-\lambda\mathcal{I})^{m-1}\mathbf{v}\right). \tag{5.2.17}$$

Thus, to find all solutions to $\mathbf{y}' = \mathcal{A}\mathbf{y}$ it is sufficient to find n independent vectors $\mathbf{v}$ satisfying (5.2.16) for some scalars λ. But these are precisely the eigenvectors or associated eigenvectors and we know that it is possible to find exactly n of them.

Thus, for example, if $\lambda = \lambda_1$ is a simple eigenvalue of $\mathcal{A}$ with a corresponding eigenvector $\mathbf{v^1}$, then $(\mathcal{A} - \lambda_1\mathcal{I})\mathbf{v^1} = 1$, thus m of (5.2.16) is equal to 1. Consequently, the sum in (5.2.17) terminates after the first term and we obtain

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

From our discussion of eigenvalues and eigenvectors it follows that if λ_i is a multiple eigenvalue of $\mathcal{A}$ of algebraic multiplicity n_i and the geometric multiplicity is less 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. Thus, we have the following algorithm for finding n linearly independent solutions to $\mathbf{y}' = \mathcal{A}\mathbf{y}$:

1. Find all eigenvalues of $\mathcal{A}$;
2. If λ is a single real eigenvalue, then there is an eigenvector $\mathbf{v}$ so that the solution is given by

$$\mathbf{y}(t) = e^{\lambda t}\mathbf{v} \tag{5.2.18}$$

3. If λ is a single complex eigenvalue $\lambda = \xi + i\omega$, then there is a complex eigenvector $\mathbf{v} = \Re\mathbf{v} + i\Im\mathbf{v}$ such that two solutions corresponding to λ (and $\bar{\lambda}$) are given by

$$\begin{aligned}\mathbf{y^1}(t) &= e^{\xi t}(\cos\omega t\,\Re\mathbf{v} - \sin\omega t\,\Im\mathbf{v})\\ \mathbf{y^2}(t) &= e^{\xi t}(\cos\omega t\,\Im\mathbf{v} + \sin\omega t\Re\mathbf{v})\end{aligned} \tag{5.2.19}$$

4. If λ is a multiple eigenvalue with algebraic multiplicity k (that is, λ is a multiple root of the characteristic equation, of multiplicity k), then we first find eigenvectors by solving $(\mathcal{A} - \lambda\mathcal{I})\mathbf{v} = \mathbf{0}$. For these eigenvectors the solution is again given by (5.2.18) (or (5.2.19), if λ is complex). If we found k independent eigenvectors, then our work with this eigenvalue is finished. If not, then we look for vectors that satisfy $(\mathcal{A} - \lambda\mathcal{I})^2\mathbf{v} = \mathbf{0}$ but $(\mathcal{A} - \lambda\mathcal{I})\mathbf{v} \neq \mathbf{0}$. For these vectors we have the solutions

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

If we still do not have k independent solutions, then we find vectors for which $(\mathcal{A} - \lambda\mathcal{I})^3\mathbf{v} = \mathbf{0}$ and $(\mathcal{A} - \lambda\mathcal{I})^2\mathbf{v} \neq \mathbf{0}$, and for such vectors we construct solutions

$$e^{t\mathcal{A}}\mathbf{v} = e^{\lambda t}\left(\mathbf{v} + t(\mathcal{A} - \lambda\mathcal{I})\mathbf{v} + \frac{t^2}{2}(\mathcal{A} - \lambda\mathcal{I})^2\mathbf{v}\right).$$

This procedure is continued till we have k solutions (by the properties of eigenvalues we have to repeat this procedure at most k times).

If λ is a complex eigenvalue of multiplicity k, then also $\bar{\lambda}$ is an eigenvalue of multiplicity k and we obtain pairs of real solutions by taking real and imaginary parts of the formulae presented above.

Fundamental solutions and nonhomogeneous problems

Let us suppose that we have n linearly independent solutions $\mathbf{y^1}(t), \ldots, \mathbf{y^n}(t)$ of the system $\mathbf{y}' = \mathcal{A}\mathbf{y}$, where $\mathcal{A}$ is an $n \times n$ matrix, like the ones constructed in the previous paragraphs. Let us denote by $\mathcal{Y}(t)$ the matrix

$$\mathcal{Y}(t) = \begin{pmatrix} y_1^1(t) \dots y_1^n(t) \\ \vdots \qquad \vdots \\ y_n^1(t) \dots y_n^n(t) \end{pmatrix},$$

that is, the columns of $\mathcal{Y}(t)$ are the vectors $\mathbf{y^i}$, $i = 1, \dots, n$. Any such matrix is called a *fundamental matrix* of the system $\mathbf{y}' = \mathcal{A}\mathbf{y}$.

We know that for a given initial vector $\mathbf{y^0}$ the solution is given by

$$\mathbf{y}(t) = e^{t\mathcal{A}}\mathbf{y^0}$$

on one hand, and, by Theorem 5.7, by

$$\mathbf{y}(t) = C_1\mathbf{y^1}(t) + \dots + C_n\mathbf{y^n}(t) = \mathcal{Y}(t)\mathbf{C},$$

on the other, where $\mathbf{C} = (C_1, \dots, C_n)$ is a vector of constants to be determined. By putting $t = 0$ above we obtain the equation for $\mathbf{C}$

$$\mathbf{y^0} = \mathcal{Y}(0)\mathbf{C}$$

Since $\mathcal{Y}$ has independent vectors as its columns, it is invertible, so that

$$\mathbf{C} = \mathcal{Y}^{-1}(0)\mathbf{y^0}.$$

Thus, the solution of the initial value problem

$$\mathbf{y}' = \mathcal{A}\mathbf{y}, \qquad \mathbf{y}(0) = \mathbf{y^0}$$

is given by

$$\mathbf{y}(t) = \mathcal{Y}(t)\mathcal{Y}^{-1}(0)\mathbf{y^0}.$$

Since $e^{t\mathbf{A}}\mathbf{y^0}$ is also a solution, by the uniqueness theorem we obtain explicit representation of the exponential function of a matrix

$$e^{t\mathcal{A}} = \mathcal{Y}(t)\mathcal{Y}^{-1}(0). \tag{5.2.20}$$

Let us turn our attention to the non-homogeneous system of equations

$$\mathbf{y}' = \mathcal{A}\mathbf{y} + \mathbf{g}(t). \tag{5.2.21}$$

The general solution to the homogeneous equation ($\mathbf{g}(t) \equiv 0$) is given by

$$\mathbf{y_h}(t) = \mathcal{Y}(t)\mathbf{C},$$

where $\mathcal{Y}(t)$ is a fundamental matrix and $\mathbf{C}$ is an arbitrary vector. Using the technique of variation of parameters, we will be looking for the solution in the form

$$\mathbf{y}(t) = \mathcal{Y}(t)\mathbf{u}(t) = u_1(t)\mathbf{y^1}(t) + \dots + u_n(t)\mathbf{y^n}(t) \tag{5.2.22}$$

where $\mathbf{u}(t) = (u_1(t), \dots, u_n(t))$ is a vector-function to be determined so that (5.2.22) satisfies (5.2.21). Thus, substituting (5.2.22) into (5.2.21), we obtain

$$\mathcal{Y}'(t)\mathbf{u}(t) + \mathcal{Y}(t)\mathbf{u}'(t) = \mathcal{A}\mathcal{Y}(t)\mathbf{u}(t) + \mathbf{g}(t).$$

Since $\mathcal{Y}(t)$ is a fundamental matrix, $\mathcal{Y}'(t) = \mathcal{A}\mathcal{Y}(t)$ and we find

$$\mathcal{Y}(t)\mathbf{u}'(t) = \mathbf{g}(t).$$

As we observed earlier, $\mathcal{Y}(t)$ is invertible, hence

$$\mathbf{u}'(t) = \mathcal{Y}^{-1}(t)\mathbf{g}(t)$$

and

$$\mathbf{u}(t) = \int^t \mathcal{Y}^{-1}(s)\mathbf{g}(s)ds + \mathbf{C}.$$

Finally, we obtain

$$\mathbf{y}(t) = \mathcal{Y}(t)\mathbf{C} + \mathcal{Y}(t)\int^t \mathcal{Y}^{-1}(s)\mathbf{g}(s)ds \tag{5.2.23}$$

This equation becomes much simpler if we take $e^{t\mathcal{A}}$ as a fundamental matrix because in such a case $\mathcal{Y}^{-1}(t) = \left(e^{t\mathcal{A}}\right)^{-1} = e^{-t\mathcal{A}}$, that is, to calculate the inverse of $e^{t\mathcal{A}}$ it is enough to replace t by $-t$. The solution (5.2.23) takes then the form

$$\mathbf{y}(t) = e^{t\mathcal{A}}\mathbf{C} + \int e^{(t-s)\mathcal{A}}\mathbf{g}(s)ds. \tag{5.2.24}$$

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 $e^{t\mathcal{A}}$; that is, if $\mathbf{y} \in \tilde{E}_\lambda$, then $\mathcal{A}\mathbf{y} \in \tilde{E}_\lambda$ (and $e^{t\mathcal{A}}\mathbf{y} \in \tilde{E}_\lambda$ for all $t > 0$). 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, $e^{t\mathcal{A}}\mathbf{y} = e^{\lambda t}\alpha\mathbf{v} \in \tilde{E}_\lambda$. 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 5.9. *Let $E_{\lambda_i} = Span\{\mathbf{v}^1, \ldots, \mathbf{v}^{n_i}\}$ be the generalized eigenspace corresponding to an eigenvalue λ_i and let $\mathbf{v}^r$ satisfy*

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

for some $1 < k < n_i$, while $(\mathcal{A} - \lambda_i\mathcal{I})^{k-1}\mathbf{v}^r = 0$. Then $\mathbf{v}^r$ satisfies

$$(\mathcal{A} - \lambda_i\mathcal{I})\mathbf{v}^r = \mathbf{v}^{r'}, \tag{5.2.25}$$

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

$$(\mathcal{A} - \lambda_i\mathcal{I})^{k-1}\mathbf{v}^r = \mathbf{v}^{r'}, \tag{5.2.26}$$

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, $\{\mathbf{v}^\rho\}_{\nu_i+1 \le \rho \le r'}$ satisfy $(\mathcal{A} - \lambda\mathcal{I})^2\mathbf{v}^\rho = 0$, etc. Then $\mathbf{v}^\rho$, $\nu_i + 1 \le \rho \le r'$, 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 $\nu_i + 1 \le \rho \le r'$ satisfies (after possibly multiplication by a scalar)

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

for some eigenvector $\mathbf{v}^j$, $j \le \nu_i$. If $r' < n_i$, then the elements from the next group, $\{\mathbf{v}^\rho\}_{r'+1 \le \rho \le r''}$ satisfy

$$0 = (\mathcal{A} - \lambda\mathcal{I})^3\mathbf{v}^\rho = (\mathcal{A} - \lambda\mathcal{I})(\mathcal{A} - \lambda\mathcal{I})^2\mathbf{v}^\rho \tag{5.2.27}$$

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

$$(\mathcal{A}-\lambda\mathcal{I})^2\mathbf{v}^\rho = \mathbf{v}^j \tag{5.2.28}$$

for some $1 \le j \le \nu_i$; that is, for some eigenvector $\mathbf{v}^j$. Alternatively, we can write (5.2.27) 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,

$$(\mathcal{A}-\lambda\mathcal{I})\mathbf{v}^\rho = \mathbf{v}^{\rho'} \tag{5.2.29}$$

for some ρ' between $\nu_i + 1$ and r'. By induction, we obtain a basis of E_λ consisting of vectors satisfying (5.2.28) 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 5.10. *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 $\mathcal{A}^k, k = 1, 2, \dots$ 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 $1 \le j \le \nu_i$ (eigenvectors). Thus

$$\mathcal{A}\mathbf{x} = \lambda\mathbf{x} - \sum_{j'>\nu_i} a_{j'}\mathbf{v}^{j'} \in E_\lambda,$$

which ends the proof of the first part.

From the first part, by induction, we obtain that $(\mathcal{A}-\lambda_i\mathcal{I})^k E_{\lambda_i} \subset E_{\lambda_i}$. In fact, let $\mathbf{x} \in E_{\lambda_i}$ and assume $(\mathcal{A}-\lambda_i\mathcal{I})^{k-1}\mathbf{x} \in E_{\lambda_i}$. Then $(\mathcal{A}-\lambda_i\mathcal{I})^k\mathbf{x} = (\mathcal{A}-\lambda_i\mathcal{I})(\mathcal{A}-\lambda_i\mathcal{I})^{k-1}\mathbf{x} \in E_{\lambda_i}$ by the induction assumption and the first part.

For $\mathcal{A}^k$ we have

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

where the inner sum must terminate at at most $n_i - 1$ term since $\mathbf{v}^j$ are determined by solving $(\mathcal{A}-\lambda\mathcal{I})^\nu\mathbf{v} = 0$ with ν being at most equal to n_i. From the previous part of the proof we see that $(\mathcal{A}-\lambda_i\mathcal{I})^r\mathbf{v}^j \in E_{\lambda_i}$ and thus $\mathcal{A}^k\mathbf{x}$.

The same argument works for $e^{t\mathcal{A}}$. Indeed, for $\mathbf{x} \in E_{\lambda_i}$ and using (5.2.17) we obtain

$$e^{t\mathcal{A}}\mathbf{x} = e^{\lambda_i t}\sum_{j=1}^{n_i} a_j e^{t(\mathcal{A}-\lambda\mathcal{I})}\mathbf{v}^j = e^{\lambda_i t}\sum_{j=1}^{n_i} a_j \sum_{r=0}^{r_j}\frac{t^{r-1}}{(r-1)!}(\mathcal{A}-\lambda\mathcal{I})^{r-1}\mathbf{v}^j. \tag{5.2.30}$$

with $r_j \le n_i$ 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 representation

$$e^{t\mathcal{A}} \overset{\circ}{\mathbf{x}} = \begin{pmatrix} | & \dots & | \\ e^{t\mathcal{A}}\mathbf{v^1} & \dots & e^{t\mathcal{A}}\mathbf{v^n} \\ | & \dots & | \end{pmatrix} \mathcal{V}^{-1} \overset{\circ}{\mathbf{x}}, \tag{5.2.31}$$

where

$$\mathcal{V} = \begin{pmatrix} | & \dots & | \\ \mathbf{v^1} & \dots & \mathbf{v^n} \\ | & \dots & | \end{pmatrix}. \tag{5.2.32}$$

Following our considerations, we select the vectors $\mathbf{v^1}, \dots, \mathbf{v^n}$ to be eigenvectors and generalized eigenvectors of $\mathcal{A}$ as then the entries of the solution matrices can be evaluated explicitly with relative ease. We want to split these expressions into generalized eigenspaces.

Let us introduce the matrix

$$\mathcal{P}_i = \begin{pmatrix} 0 \dots & | & \dots 0 \\ 0 \dots & \mathbf{v^i} & \dots 0 \\ 0 \dots & | & \dots 0 \end{pmatrix} \begin{pmatrix} | & \dots & | \\ \mathbf{v^1} & \dots & \mathbf{v^n} \\ | & \dots & | \end{pmatrix}^{-1}. \tag{5.2.33}$$

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

$$\mathcal{P}_i^2 = \mathcal{P}_i, \qquad \mathcal{P}_i\mathcal{P}_j = 0, \tag{5.2.34}$$

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

$$\mathcal{P}_{\lambda_i} = \sum_{j;\, \mathbf{v^j} \in E_{\lambda_i}} \mathcal{P}_j, \tag{5.2.35}$$

where $\mathcal{P}_i$, then such operators again will be projections. This follows from (5.2.34) by termwise multiplication. They are called *spectral projections*. Let $\sigma(\mathcal{A})$ denote the set of all eigenvalues of $\mathcal{A}$, called the *spectrum* of $\mathcal{A}$. The decomposition

$$\mathcal{I} = \sum_{\lambda \in \sigma(\mathcal{A})} \mathcal{P}_\lambda, \tag{5.2.36}$$

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

$$e^{t\mathcal{A}} = \sum_{\lambda \in \sigma(\mathcal{A})} e^{\lambda t}\mathcal{P}_\lambda. \tag{5.2.37}$$

In general case, we use (5.2.36) to write

$$\mathcal{A}\mathbf{x} = \sum_{\lambda \in \sigma(A)} \mathcal{A}\mathcal{P}_\lambda\mathbf{x}, \tag{5.2.38}$$

where, by Corollary 5.10, we have $\mathcal{A}\mathcal{P}_\lambda \mathbf{x} \in E_\lambda$. Thus, using (5.2.34), we get $\mathcal{P}_{\lambda_i}\mathcal{A}\mathcal{P}_{\lambda_j} = 0$ for $i \neq j$. Using (5.2.35) 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, (5.2.38) 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 $e^{t\mathcal{A}}$, generalizing (5.2.37), we observe that, by Corollary 5.10, 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

$$e^{t\mathcal{A}}\mathbf{x} = \sum_{\lambda \in \sigma(\mathcal{A})} e^{\lambda t}\mathcal{P}_\lambda \mathbf{x} = \sum_{\lambda \in \sigma(\mathcal{A})} e^{\lambda t}\mathbf{q}_\lambda(t)\mathbf{x}, \tag{5.2.39}$$

where $\mathbf{q}_\lambda$ are polynomials 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.

Planar linear systems

In this section we shall present a complete description of all orbits of the linear differential system

$$\mathbf{y}' = \mathcal{A}\mathbf{y} \tag{5.2.40}$$

where $\mathbf{y}(t) = (y_1(t), y_2(t))$ and

$$\mathcal{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

We shall assume that $\mathcal{A}$ is invertible, that is, $ad - bc \neq 0$. In such a case $\mathbf{y} = (0, 0)$ is the only equilibrium point of (5.2.40).

The phase portrait is fully determined by the eigenvalues of the matrix $\mathcal{A}$. Let us briefly describe all possible cases, as determined by the theory of the preceding section. The general solution can be obtained as a linear combination of two linearly independent solutions. To find them, we have to find first the eigenvalues of $\mathcal{A}$, that is, solutions to

$$(\lambda - \lambda_1)(\lambda - \lambda_2) = (a - \lambda)(d - \lambda) - bc = \lambda^2 - \lambda(d + a) + ad - bc.$$

Note that by the assumption on invertibility, $\lambda = 0$ is not an eigenvalue of $\mathcal{A}$. We have the following possibilities:

a) $\lambda_1 \neq \lambda_2$. In this case each eigenvalue must be simple and therefore we have two linearly independent eigenvectors $\mathbf{v}^1$, $\mathbf{v}^2$. The expansion $e^{t\mathcal{A}}\mathbf{v}^i$ for $i = 1, 2$ terminates after the first term. We distinguish two cases.

   ◇ If λ_1, λ_2 are real numbers, then the general solution is given simply by

$$\mathbf{y}(t) = c_1 e^{\lambda_1 t}\mathbf{v}^1 + c_2 e^{\lambda_2 t}\mathbf{v}^2. \tag{5.2.41}$$

   ◇ If λ_1, λ_2 are complex numbers, then the general solution is still given by the above formula but the functions above are complex and we would rather prefer solution to be real. To achieve this, we note that λ_1, λ_2 must be necessarily complex conjugate $\lambda_1 = \xi + i\omega, \lambda_2 = \xi - i\omega$, where ξ and ω are real. It can be also proved that the associated eigenvectors $\mathbf{v}^1$ and $\mathbf{v}^2$ are also complex conjugate. Let $\mathbf{v}^1 = \mathbf{u} + i\mathbf{v}$; then the real-valued general solution is given by

$$\mathbf{y}(t) = c_1 e^{\xi t}(\mathbf{u}\cos\omega t - \mathbf{v}\sin\omega t) + c_2 e^{\xi t}(\mathbf{u}\sin\omega t + \mathbf{v}\cos\omega t). \tag{5.2.42}$$

   This solution can be written in a more compact form

$$\mathbf{y}(t) = e^{\xi t}\left(A_1\cos(\omega t - \phi_1), A_2\cos(\omega t - \phi_2)\right), \tag{5.2.43}$$

   for some choice of constants $A_1, A_2 > 0$ and ϕ_1, ϕ_2.

b) $\lambda_1 = \lambda_2 = \lambda$. There are two cases to distinguish.

$\diamond$ There are two linearly independent eigenvectors $\mathbf{v}^1$ and $\mathbf{v}^2$ corresponding to λ. In this case the general solution is given by

$$\mathbf{y}(t) = e^{\lambda t}(c_1\mathbf{v}^1 + c_2\mathbf{v}^2). \tag{5.2.44}$$

$\diamond$ If there is only one eigenvector, then following the discussion above, we must find a vector $\mathbf{v}^2$ satisfying $(\lambda I - \mathcal{A})\mathbf{v}^2 \neq 0$ and $(\lambda I - \mathcal{A})^2\mathbf{v}^2 = 0$. However, since we are in the two-dimensional space, the latter is satisfied by any vector $\mathbf{v}^2$ and, since the eigenspace is one dimensional, from

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

it follows that $(\lambda I - \mathcal{A})\mathbf{v}^2 = k\mathbf{v}^1$. Thus, the formula for $e^{\mathcal{A}t}\mathbf{v}^2$ simplifies as

$$e^{t\mathcal{A}}\mathbf{v}^2 = e^{\lambda t}\left(\mathbf{v}^2 + t(\lambda I - \mathcal{A})\mathbf{v}^2\right) = e^{\lambda t}\left(\mathbf{v}^2 + kt\mathbf{v}^1\right).$$

Thus, the general solution in this case can be written as

$$\mathbf{y}(t) = e^{\lambda t}\left((c_1 + c_2kt)\mathbf{v}^1 + c_2\mathbf{v}^2\right). \tag{5.2.45}$$

Remark 5.11. Before we embark on describing phase portraits, let us observe that if we change the direction of time in (5.2.40): $\tau = -t$ and $\mathbf{z}(\tau) = \mathbf{y}(-\tau) = \mathbf{y}(t)$, then we obtain

$$\mathbf{z}'_\tau = -\mathcal{A}\mathbf{z}$$

and the eigenvalues of $-\mathcal{A}$ are precisely the negatives of the eigenvalues of $\mathcal{A}$. Thus, the orbits of solutions corresponding to systems governed by $\mathcal{A}$ and $-\mathcal{A}$ or, equivalently, with eigenvalues that differ only by sign, are the same with only difference being the direction in which they are traversed.

We are now in a position to describe all possible phase portraits of (5.2.40). Again we have to go through several cases.

i) $\lambda_2 < \lambda_1 < 0$. Let $\mathbf{v}^1$ and $\mathbf{v}^2$ be eigenvectors of $\mathcal{A}$ with eigenvalues λ_1 and λ_2, respectively. In the $y_1 - y_2$ plane we draw four half-lines l_1, l'_1, l_2, l'_2 parallel to $\mathbf{v}^1$, $-\mathbf{v}^1$, $\mathbf{v}^2$ and $-\mathbf{v}^2$, respectively, and emanating from the origin, as shown in Fig 2.1. Observe first that $\mathbf{y}(t) = ce^{\lambda_i t}\mathbf{v}^i$, $i = 1, 2$, are the solutions to (5.2.40) for any choice of a non-zero constant c and, as they are parallel to $\mathbf{v}^i$, the orbits are the half-lines l_1, l'_1, l_2, l_2 (depending on the sign of the constant c) and all these orbits are traced towards the origin as $t \to \infty$. Since every solution $\mathbf{y}(t)$ of (5.2.40) can be written as

$$\mathbf{y}(t) = c_1e^{\lambda_1 t}\mathbf{v}^1 + c_2e^{\lambda_2 t}\mathbf{v}^2$$

for some choice of constants c_1 and c_2. Since $\lambda_1, \lambda_2 < 0$, every solution tends to $(0, 0)$ as $t \to \infty$, and so every orbit approaches the origin for $t \to \infty$. We can prove an even stronger fact – as $\lambda_2 < \lambda_1$, the second term becomes negligible for large t and therefore the tangent of the orbit of $\mathbf{y}(t)$ approaches the direction of l_1 if $c_1 > 0$ and of l'_1 if $c_1 < 0$. Thus, every orbit except that with $c_1 = 0$ approaches the origin along the same fixed line. Such a type of an equilibrium point is called a *stable node*. If we have $0 < \lambda_1 < \lambda_2$, then by Remark 5.11, the orbits of (5.2.40) will have the same shape as in case i) but the arrows will be reversed so that the origin will repel all the orbits and the orbits will be unbounded as $t \to \infty$. Such an equilibrium point is called an *unstable node*.

ii) $\lambda_1 = \lambda_2 = \lambda < 0$. In this case the phase portrait of (5.2.40) depends on whether $\mathcal{A}$ has one or two linearly independent eigenvectors. In the latter case, the general solution in given (see b) above) by

$$\mathbf{y}(t) = e^{\lambda t}(c_1\mathbf{v}^1 + c_2\mathbf{v}^2),$$

so that orbits are half-lines parallel to $c_1\mathbf{v}^1 + c_2\mathbf{v}^2$. These half-lines cover every direction of the $y_1 - y_2$ plane and, since $\lambda < 0$, each solution will converge to $(0, 0)$ along the respective line. Thus, the phase

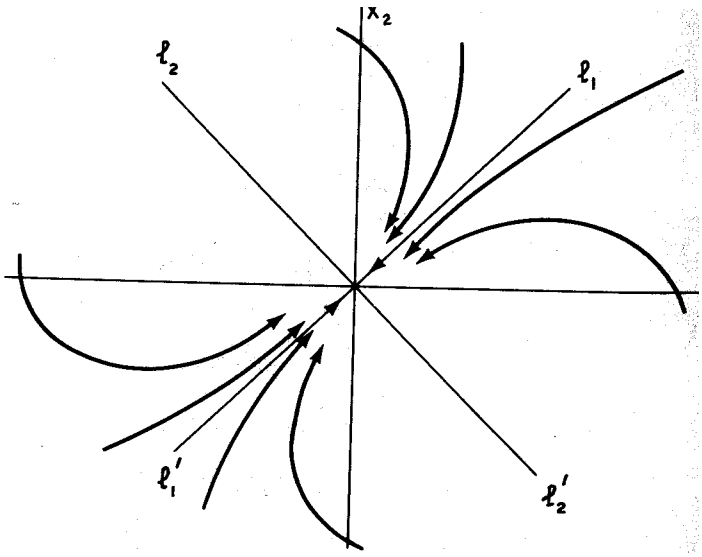
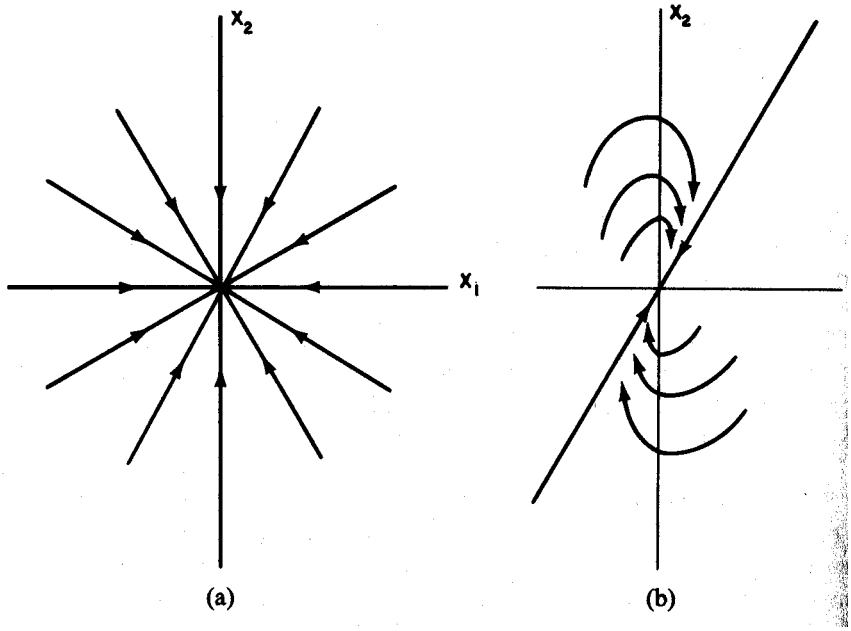


Fig. 2.1 Stable node

portrait looks like in Fig. 2.2a. If there is only one independent eigenvector corresponding to λ then, by (5.2.45),

$$\mathbf{y}(t) = e^{\lambda t}\left((c_1 + c_2 kt)\mathbf{v}^1 + c_2\mathbf{v}^2\right)$$

for some choice of constants c_1, c_2, k. Obviously, every solution approaches $(0,0)$ as $t \to \infty$. Putting $c_2 = 0$, we obtain two half-line orbits $c_1 e^{\lambda t}\mathbf{v}^1$ but, contrary to the case i), there are no other half-line orbits. In addition, the term $c_1\mathbf{v}^1 + c_2\mathbf{v}^2$ becomes small in comparison with $c_2 kt\mathbf{v}^1$ as $t \to \infty$ so that the orbits approach the origin in the direction of $\pm\mathbf{v}^1$. The phase portrait is presented in Fig. 2.2b. The equilibrium in both cases is called the *stable degenerate node*. If $\lambda_1 = \lambda_2 > 0$, then again by Remark



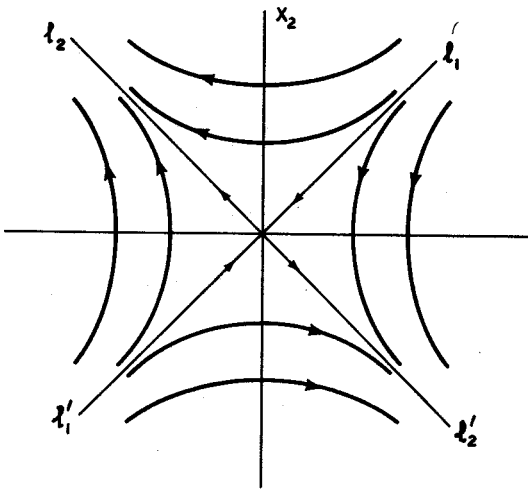
2.2 Stable degenerate node

5.11, the picture in this case will be the same as in Fig. 2.a-b but with the direction of arrows reversed. Such equilibrium point is called an *unstable degenerate node.*

iii) $\lambda_1 < 0 < \lambda_2$. As in case i), in the $y_1 - y_2$ plane we draw four half-lines l_1, l_1', l_2, l_2' that emanate from the origin and are parallel to $\mathbf{v}^1$, $-\mathbf{v}^1$, $\mathbf{v}^2$ and $-\mathbf{v}^2$, respectively, as shown in Fig 2.3. Any solution is given by

$$\mathbf{y}(t) = c_1 e^{\lambda_1 t}\mathbf{v}^1 + c_2 e^{\lambda_2 t}\mathbf{v}^2$$

for some choice of c_1 and c_2. Again, the half-lines are the orbits of the solutions: l_1, l_1' for $c_1 e^{\lambda_1 t}\mathbf{v}^1$ with $c_1 > 0$ and $c_1 < 0$, and l_2, l_2' for $c_2 e^{\lambda_2 t}\mathbf{v}^2$ with $c_2 > 0$ and $c_2 < 0$, respectively. However, the direction of arrows is different on each pair of half-lines: while the solution $c_1 e^{\lambda_1 t}\mathbf{v}^1$ converges towards $(0,0)$ along l_1 or l_1' as $t \to \infty$, the solution $c_2 e^{\lambda_2 t}\mathbf{v}^2$ becomes unbounded moving along l_2 or l_2', as $t \to \infty$. Next, we observe that if $c_1 \neq 0$, then for large t the second term $c_2 e^{\lambda_2 t}\mathbf{v}^2$ becomes negligible and so the solution becomes unbounded as $t \to \infty$ with asymptotes given by the half-lines l_2, l_2', respectively. Similarly, for $t \to -\infty$ the term $c_1 e^{\lambda_1 t}\mathbf{v}^1$ becomes negligible and the solution again escapes to infinity, but this time with asymptotes l_1, l_1', respectively. Thus, the phase portrait, given in Fig. 2.3, resembles a saddle near $y_1 = y_2 = 0$ and, not surprisingly, such an equilibrium point is called a *saddle.* The case $\lambda_2 < 0 < \lambda_1$ is



A saddle point

of course symmetric.

iv) $\lambda_1 = \xi + i\omega, \lambda_2 = \xi - i\omega$. In (5.2.43) we derived the solution in the form

$$\mathbf{y}(t) = e^{\xi t}\left(A_1 \cos(\omega t - \phi_1), A_2 \cos(\omega t - \phi_2)\right).$$

We have to distinguish three cases:

α) If $\xi = 0$, then

$$y_1(t) = A_1 \cos(\omega t - \phi_1), \qquad y_2(t) = A_2 \cos(\omega t - \phi_2),$$

both are periodic functions with period $2\pi/\omega$ and y_1 varies between $-A_1$ and A_1 while y_2 varies between $-A_2$ and A_2. Consequently, the orbit of any solution $\mathbf{y}(t)$ is a closed curve containing the origin inside and the phase portrait has the form presented in Fig. 3.4a. For this reason we say that the equilibrium point of (5.2.40) is a *center* when the eigenvalues of $\mathcal{A}$ are purely imaginary. The direction of arrows

must be determined from the equation. The simplest way of doing this is to check the sign of y_2' when $y_2 = 0$. If at $y_2 = 0$ and $y_1 > 0$ we have $y_2' > 0$, then all the orbits are traversed in the anticlockwise direction, and conversely.

β) If $\xi < 0$, then the factor $e^{\xi t}$ forces the solution to come closer to zero at every turn so that the solution spirals into the origin giving the picture presented in Fig. 2.4b. The orientation of the spiral must be again determined directly from the equation. Such an equilibrium point is called a *stable focus.*

γ) If $\xi > 0$, then the factor $e^{\xi t}$ forces the solution to spiral outwards creating the picture shown in Fig. 4c. Such an equilibrium point is called an *unstable focus.*

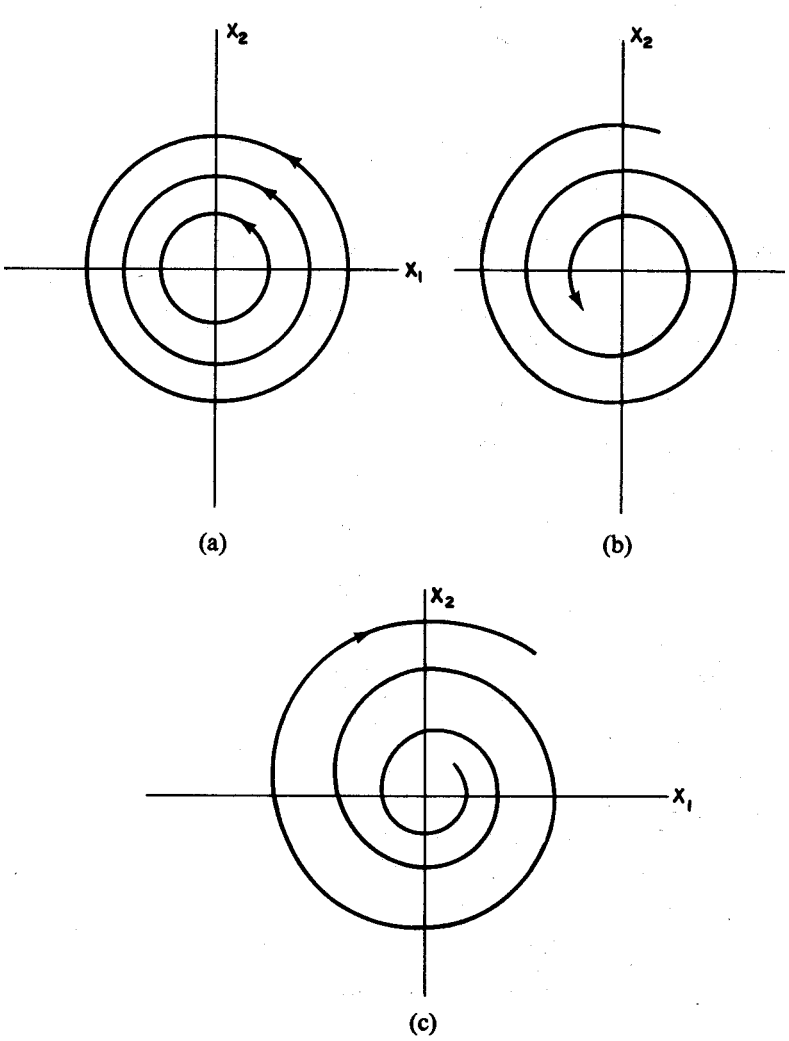


Fig.4 Center, stable and unstable foci

2.2 Stability of equilibrium solutions

Linear systems

The discussion of phase-portraits for two-dimensional linear, given in the previous section allows to determine easily under which conditions $(0, 0)$ is stable. Clearly, the only stable cases are when real parts of both

eigenvalues are nonnegative with asymptotic stability offered by eigenvalues with strictly negative ones (the case of the centre is an example of a stable but not asymptotically stable equilibrium point).

Analogous results can be formulated for linear systems in higher dimensions. By considering formulae for solutions we ascertain that the equilibrium point is (asymptotically stable) if all the eigenvalues have negative real parts and is unstable if at least one eigenvalue has positive real part. The case of eigenvalues with zero real part is more complicated as in higher dimension we can have multiple complex eigenvalues. Here, again from the formula for solutions, we can see that if for each eigenvalue with zero real part of algebraic multiplicity k there is k linearly independent eigenvectors, the solution is stable. However, if geometric and algebraic multiplicities of at least such eigenvalue are different, then in the solution corresponding to this eigenvalue there will appear a polynomial in t which will cause the solution to be unstable.

Nonlinear systems–stability by linearization

The above considerations can be used to determine stability of equilibrium points of arbitrary differential equations

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}). \tag{5.2.46}$$

Let us first note the following result.

Lemma 5.12. *If* $\mathbf{f}$ *has continuous partial derivatives of the first order in some neighbourhood of* $\mathbf{y}^0$*, then*

$$\mathbf{f}(\mathbf{x}+\mathbf{y}^0) = \mathbf{f}(\mathbf{y}^0) + \mathcal{A}\mathbf{x} + \mathbf{g}(\mathbf{x}) \tag{5.2.47}$$

where

$$\mathcal{A} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{y^0}) & \dots & \frac{\partial f_1}{\partial x_n}(\mathbf{y^0}) \\ \vdots & & \vdots \\ \frac{\partial f_1}{\partial x_n}(\mathbf{y^0}) & \dots & \frac{\partial f_n}{\partial x_n}(\mathbf{y^0}) \end{pmatrix},$$

and $\mathbf{g}(\mathbf{x})/\|\mathbf{x}\|$ *is continuous in some neighbourhood of* $\mathbf{y}^0$ *and vanishes at* $\mathbf{x} = \mathbf{y}^0$*.*

Proof. The matrix $\mathcal{A}$ has constant entries so that $\mathbf{g}$ defined by

$$\mathbf{g}(\mathbf{x}) = \mathbf{f}(\mathbf{x}+\mathbf{y}^0) - \mathbf{f}(\mathbf{y}^0) - \mathcal{A}\mathbf{x}$$

is a continuous function of $\mathbf{x}$. Hence, $\mathbf{g}(\mathbf{x})/\|\mathbf{x}\|$ is also continuous for $\mathbf{x} \neq \mathbf{0}$. Using now Taylor's formula for each component of $\mathbf{f}$ we obtain

$$f_i(\mathbf{x}+\mathbf{y}^0) = f_i(\mathbf{y}^0) + \frac{\partial f_i}{\partial x_1}(\mathbf{y}^0)x_1 + \ldots + \frac{\partial f_i}{\partial x_n}x_n(\mathbf{y}^0) + R_i(\mathbf{x}), \quad i = 1, \ldots, n,$$

where, for each i, the remainder R_i satisfies

$$|R_i(x)| \leq M(\|\mathbf{x}\|)\|\mathbf{x}\|$$

and M tends to zero is $\|\mathbf{x}\| \to 0$. Thus,

$$\mathbf{g}(\mathbf{x}) = (R_1(\mathbf{x}), \ldots, R_n(\mathbf{x}))$$

and

$$\frac{\|\mathbf{g}(\mathbf{x})\|}{\|\mathbf{x}\|} \leq M(\|\mathbf{x}\|) \to 0 \tag{5.2.48}$$

as $\|\mathbf{x}\| \to 0$ and, $\mathbf{f}(\mathbf{y}^0) = 0$, the lemma is proved. ■

The linear system

$$\mathbf{x}' = \mathcal{A}\mathbf{x}$$

is called the linearization of (5.5.78) around the equilibrium point $\mathbf{y}^0$.

Theorem 5.13. *Suppose that* $\mathbf{f}$ *is a differentiable function in some neighbourhood of the equilibrium point* $\mathbf{y}^0$*. Then,*

1. *The equilibrium point* $\mathbf{y}^0$ *is asymptotically stable if all the eigenvalues of the matrix* $\mathcal{A}$ *have negative real parts, that is, if the equilibrium solution* $\mathbf{x}(t) = \mathbf{0}$ *of the linearized system is asymptotically stable. In particular, for sufficiently small initial conditions the solutions are defined for all* t*.*
2. *The equilibrium point* $\mathbf{y}^0$ *is unstable if at least one eigenvalue has a positive real part.*
3. *If all the eigenvalues of* $\mathcal{A}$ *have non-negative real part but at least one of them has real part equal to* 0*, then the stability of the equilibrium point* $\mathbf{y}^0$ *of the nonlinear system (5.5.78) cannot be determined from the stability of its linearization.*

Proof. To prove 1) we use the variation of constants formula applied to (5.5.78) written in the form of Lemma 5.12 for $\mathbf{y}(t) = \mathbf{x}(t) + \mathbf{y}^0$:

$$\mathbf{x}' = \mathbf{y}' = \mathbf{f}(\mathbf{y}) = \mathbf{f}(\mathbf{x} + \mathbf{y}^0) = \mathcal{A}\mathbf{x} + \mathbf{g}(\mathbf{x}). \tag{5.2.49}$$

Thus

$$\mathbf{x}(t) = e^{t\mathcal{A}}\mathbf{x}(0) + \int_0^t e^{(t-s)\mathcal{A}}\mathbf{g}(\mathbf{x}(s))ds.$$

Denoting by α' the maximum of real parts of eigenvalues of $\mathcal{A}$ we observe that for any $\alpha > \alpha'$

$$\|e^{t\mathcal{A}}\mathbf{x}(0)\| \le Ke^{\alpha t}\|\mathbf{x}(0)\|, \quad t \ge 0,$$

for some constant $K \ge 1$. Note that in general we have to take $\alpha > \alpha'$ to account for possible polynomial entries in $e^{t\mathcal{A}}$. Thus, since $\alpha' < 0$, then we can take also $\alpha < 0$ keeping the above estimate satisfied. From the assumption on $\mathbf{g}$, for any ϵ we find $\delta > 0$ such that if $\|\mathbf{x}\| \le \delta$, then

$$\|\mathbf{g}(\mathbf{x})\| \le \epsilon\|\mathbf{x}\|. \tag{5.2.50}$$

Assuming for a moment that for $0 \le s \le t$ we can keep $\|\mathbf{x}(s)\| \le \delta$, we can write

$$\begin{aligned}\|\mathbf{x}(t)\| &\le \|e^{\mathcal{A}t}\mathbf{x}(0)\| + \int_0^t \|e^{\mathcal{A}(t-s)}\mathbf{g}(\mathbf{x}(s))\|ds \\ &\le Ke^{\alpha t}\mathbf{x}(0) + K\epsilon\int_0^t e^{\alpha(t-s)}\|\mathbf{x}(s)\|ds\end{aligned}$$

or, multiplying both sides by $e^{-\alpha t}$ and setting $z(t) = e^{-\alpha t}\|\mathbf{x}(t)\|$,

$$z(t) \le K\|\mathbf{x}(0)\| + K\epsilon\int_0^t z(s)ds. \tag{5.2.51}$$

Using Gronwall's lemma we obtain thus

$$\|\mathbf{x}(t)\| = e^{\alpha t}z(t) \le K\|\mathbf{x}(0)\|e^{(K\epsilon+\alpha)t},$$

providing $\|\mathbf{x}(s)\| \le \delta$ for all $0 \le s \le t$. Let us take $\epsilon \le -\alpha/2K$, then the above can be written as

$$\|\mathbf{x}(t)\| \le K\|\mathbf{x}(0)\|e^{\frac{\alpha t}{2}}. \tag{5.2.52}$$

Assume now that $\|\mathbf{x}(0)\| < \delta/K \le \delta$ where δ was fixed for $\epsilon \le \alpha/2K$. Then $\|\mathbf{x}(0)\| < \delta$ and, by continuity, $\|\mathbf{x}(t)\| \le \delta$ for some time. Let $\mathbf{x}(t)$ be defined on some interval I and $t_1 \in I$ be the first time for which $\|\mathbf{x}(t)\| = \delta$. Then for $t \le t_1$ we have $\|\mathbf{x}(t)\| \le \delta$ so that for all $t \le t_1$ we can use (5.2.52) getting, in particular,

$$\|\mathbf{x}(t_1)\| \leq \delta e^{\frac{\alpha t_1}{2}} < \delta;$$

that is, a contradiction. Thus $\|\mathbf{x}(t)\| < \delta$ if $\|\mathbf{x}(0)\| < \delta_1$ in the whole interval of existence but then, if the interval was finite, then we could extend the solution to a larger interval as the solution is bounded at the endpoint and the same procedure would ensure that the solution remains bounded by δ on the larger interval. Thus, the extension can be carried out for all the values of $t \geq 0$ and the solution exists for all t and satisfies $\|\mathbf{x}(t)\| \leq \delta$ for all $t \geq 0$. Consequently, (5.2.52) holds for all t and the solution $\mathbf{x}(t)$ converges exponentially to 0 as $t \to \infty$ proving the asymptotic stability of the stationary solution $\mathbf{y}^0$.

We shall prove Statement 2 only in the two dimensional case to avoid some algebraic technicalities. For a given system of differential equations, let $\phi(t, \mathbf{x}_0)$ be the solution $\mathbf{x}(t)$ of this system satisfying $\mathbf{x}(0) = \mathbf{x}_0$. Let us recall that to prove that $\mathbf{x} = 0$ is unstable for (5.2.49) it is enough to show that there is $\epsilon >$ such that for any $\delta > 0$ there is $\mathbf{x}_0$ and $t_0 < \infty$ such that $\|\mathbf{x}_0\| \leq \delta$ and $\|\phi(t_0, \mathbf{x}(t)\| \geq \epsilon$.

Thus, let λ_1 and λ_2 be the eigenvalues of $\mathcal{A}$. There are two possible cases.

(i) Both $\Re\lambda_1, \Re\lambda_2 > 0$. If we replace t by $\tau = -t$, then (5.2.49) takes the form

$$\frac{d\mathbf{z}}{d\tau} = -\mathcal{A}\mathbf{z} - \mathbf{g}(\mathbf{z}). \tag{5.2.53}$$

Since the eigenvalues of $-\mathcal{A}$ both have negative real parts, the equilibrium $\mathbf{z} = 0$ is locally asymptotically stable. This means, in particular, that there is $\epsilon > 0$ such that if $\|\mathbf{z}_0\| = \epsilon$, then the solution $\mathbf{z}(\tau) = \phi(\tau, \mathbf{z}_0)$ approaches $\mathbf{z} = 0$ as $\tau \to \infty$.

Let us fix arbitrary $\delta > 0$ and $\mathbf{z}_0$ with $\|\mathbf{z}_0\| = \epsilon$. Then there is $\hat{\tau}$ such that $\|\phi(\tau, \mathbf{z}_0)\| = \epsilon$. Consider now (5.2.49) with $\mathbf{x}_0 = \phi(\hat{\tau}, \mathbf{z}_0)$. Then $\mathbf{x}(t) = \phi(-\tau, \phi(\hat{\tau}, \mathbf{z}_0))$ satisfies $\mathbf{x}(\hat{t}) = \phi(-\hat{\tau}, \phi(\hat{\tau}, \mathbf{z}_0)) = \mathbf{z}_0$. Thus, for any $\delta > 0$ we can find an initial condition $\mathbf{x}_0$ such that the solution $\mathbf{x}(t)$ emanating from this initial condition will be at a fixed distance ϵ from the origin and hence the equilibrium $\mathbf{x} = 0$ is unstable.

(ii) Assume now that only one eigenvalue has a positive real part. This implies that the eigenvalues are real and thus the second one must be non-positive. In other words, we have $\lambda_1 \leq 0 < \lambda_2$. Then, by a linear change of coordinates, (5.2.49) can be written in the form

$$\begin{aligned} y_1' &= \lambda_1 y_1 + g_1(y_1, y_2), \\ y_2' &= \lambda_2 y_2 + g_2(y_1, y_2), \end{aligned} \tag{5.2.54}$$

where the function $\mathbf{g} = (g_1, g_2)$ has the required property (5.2.48). We observe that, in principle, the solution can contract along the y_1 axis and so we have to somehow enhance the expansion along the y_2-axis. We use an argument similar to the classical Lyapunov approach. Consider the function

$$V(y_1, y_2) = \frac{1}{2}(y_2^2 - y_1^2). \tag{5.2.55}$$

This is a hyperboloid with level curves

$$\frac{1}{2}(y_2^2 - y_1^2) = c, \quad c \geq 0, \tag{5.2.56}$$

being hyperbolas in the sector $|y_2| \geq |y_1|$ (with two intersecting lines $y_2 = \pm y_1$ for $c = 0$. Notice that the larger the value of c in (5.2.56), the higher is the level curve. Let us consider how the function V behaves along the solutions to (5.2.49). From (5.2.48) we see that for any $m > 0$ there is ϵ such that

$$\|\mathbf{g}(\mathbf{y})\| \leq m\|\mathbf{y}\|, \tag{5.2.57}$$

provided $\|\mathbf{y}(t)\| < \epsilon$. Let us assume that $\|\mathbf{y}\| < \epsilon$. Then, using the Cauchy-Schwarz inequality,

$$-\|\mathbf{y}\|\|\mathbf{g}\| \leq -\sqrt{y_1^2 + y_2^2}\sqrt{g_1^2 + g_2^2} \leq y_2 g_2 - y_1 g_1 \leq \sqrt{y_1^2 + y_2^2}\sqrt{g_1 + g_2^2} = \|\mathbf{y}\|\|\mathbf{g}\|,$$

we have

$$\begin{aligned}\frac{dV(\mathbf{y}(t))}{dt} &= y_2y_2' - y_1y_1' = \lambda_2y_2^2 + y_2g_2(y_1,y_2) - \lambda_1y_1^2 - y_1g_1(y_1,y_2)\\ &\geq \lambda_2y_2^2 - \|\mathbf{g}\|\|\mathbf{y}\| - \lambda_1y_1^2\\ &\geq \lambda_2y_2^2 - m\|\mathbf{y}\|^2 - \lambda_1y_1^2\\ &= (\lambda_2 - m)y_2^2 + (\lambda_1 + m)y_1^2.\end{aligned}$$

Consider now the set

$$\Omega = \{(y_1, y_2);\ y_2 > |y_2|\}$$

that is the wedge above lines $y_2 = \pm y_1$. Now, chose $m < \lambda_2$ and corresponding ϵ so that (5.2.57) is satisfied. Further, consider the ball $B_\epsilon = \{\mathbf{y};\ \|y\| < \epsilon\}$. Then in $\Omega \cap B_\epsilon$ we have $V(\mathbf{y}) > 0$ and

$$\frac{dV(\mathbf{y}(t))}{dt} \geq (\lambda_2 - m)y_2^2 > 0.$$

Let us consider arbitrary $0 < \delta < \epsilon$ and $\mathbf{y}_0 \in \Omega$ with $\|\mathbf{y}_0\| = \delta$. Then $y_{0,2} > \delta/\sqrt{2}$. Further, $\mathbf{y}_0 \in V^{-1}(c_0)$ for some $c_0 > 0$. Since $V(\mathbf{y}(t))$ is strictly increasing as long as $\mathbf{y}(t) \in \Omega \cap B_\epsilon$, we see that with increasing t the solution $\mathbf{y}(t)$ will move upward, through level curves $V^{-1}(c)$ with an increasing c. This shows, in particular, that $\mathbf{y}(t)$ cannot reach $y_2 = \pm y_1$ within $\Omega \cap B_\epsilon$. Hence $y_2(t) > \delta/\sqrt{2}$ implying $dV(\mathbf{y}(t))/dt \geq (\lambda_1 - m)\delta/\sqrt{2} > 0$. In other words, $c(t) = V(\mathbf{y}(t))$ is a strictly increasing function such that $\mathbf{y}(t) \in V^{-1}(c(t))$ for $t \in \mathbf{y}^{-1}(B_\epsilon)$. If we assume that $\|\mathbf{y}(t)\| < \epsilon$ for all t, then $c(t)$ diverges to infinity as $t \to \infty$ since $c'(t)$ is bounded away from zero by a constant. But then $\|\mathbf{y}(t)\| \to \infty$ since $\|\mathbf{y}(t)\| \geq \inf \|V^{-1}(c(t))\| = \sqrt{2c(t)}$. We arrived at a contradiction, and thus $\|\mathbf{y}(t)\| = \epsilon$ in finite time.

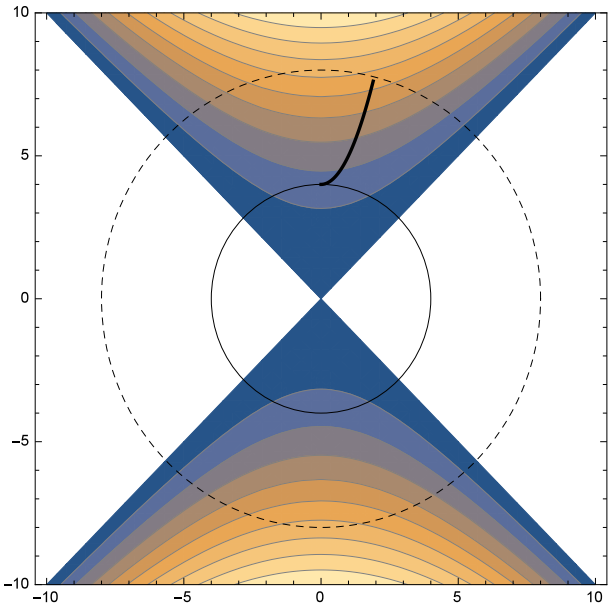


Fig. 5.2. Level sets of V, the circle $\|\mathbf{y}\| = \delta$ and $\|\mathbf{y}\| = \epsilon$, and the solution $\mathbf{y}(t)$.

To prove 3., it is enough to display two systems with the same linear part and different behaviour of solutions. Let us consider

$$\begin{aligned}y_1' &= y_2 - y_1(y_1^2 + y_2^2)\\ y_2' &= -y_1 - y_2(y_1^2 + y_2^2)\end{aligned}$$

with the linearized system given by

$$y_1' = y_2$$
$$y_2' = -y_1$$

The eigenvalues of the linearized system are $\pm i$. To analyze the behaviour of the solutions to the non-linear system, let us multiply the first equation by y_1 and the second by y_2 and add them together to get

$$\frac{1}{2}\frac{d}{dt}(y_1^2 + y_2^2) = -(y_1^2 + y_2^2)^2.$$

Solving this equation we obtain

$$y_1^2 + y_2^2 = \frac{c}{1 + 2ct}$$

where $c = y_1^2(0) + y_2^2(0)$. Thus $y_1^2(t) + y_2^2(t)$ approaches $\mathbf{0}$ as $t \to \infty$ and $y_1^2(t) + y_2^2(t) < y_1^2(0) + y_2^2(0)$ for any $t > 0$ and we can conclude that the equilibrium point $\mathbf{0}$ is asymptotically stable.

Consider now the system

$$y_1' = y_2 + y_1(y_1^2 + y_2^2)$$
$$y_2' = -y_1 + y_2(y_1^2 + y_2^2)$$

with the same linear part and thus with the same eigenvalues. As above we obtain that

$$y_1^2 + y_2^2 = \frac{c}{1 - 2ct}$$

with the same meaning for c. Thus, any solution with non-zero initial condition blows up at the time $t = 1/2c$ and therefore the equilibrium solution $\mathbf{0}$ is unstable. ■

Example 5.14. Find all equilibrium solutions of the system of differential equations

$$y_1' = 1 - y_1 y_2,$$
$$y_2' = y_1 - y_2^3,$$

and determine, if possible, their stability.

Solving equation for equilibrium points $1 - y_1 y_2 = 0, y_1 - y_2^3 =$ *we find two equilibria:* $y_1 = y_2 = 1$ *and* $y_1 = y_2 = -1$. *To determine their stability we have to reduce each case to the equilibrium at* $\mathbf{0}$. *For the first case we put* $u(t) = y_1(t) - 1$ *and* $v(t) = y_2 - 1$ *so that*

$$u' = -u - v - uv,$$
$$v' = u - 3v - 3v^2 - v^3,$$

so that the linearized system has the form

$$u' = -u - v,$$
$$v' = u - 3v,$$

and the perturbing term is given by $\mathbf{g}(u, v) = (-uv, -3v^2 + v^3)$ *and, as the right-hand side of the original system is infinitely differentiable at* $(0, 0)$ *the assumptions of the stability theorem are satisfied. The eigenvalues of the linearized system are given by* $\lambda_{1,2} = -2$ *and therefore the equilibrium solution* $\mathbf{y}(t) \equiv (1, 1)$ *is asymptotically stable.*

For the other case we set $u(t) = y_1(t) + 1$ *and* $v(t) = y_2 + 1$ *so that*

$$u' = u + v - uv,$$
$$v' = u - 3v + 3v^2 - v^3,$$

so that the linearized system has the form

$$u' = u + v,$$
$$v' = u - 3v,$$

and the perturbing term is given by $\mathbf{g}(u, v) = (-uv, 3v^2 - v^3)$. *The eigenvalues of the linearized system are given by* $\lambda_1 = -1 - \sqrt{5}$ *and* $\lambda_2 = -1 + \sqrt{5}$*and therefore the equilibrium solution* $\mathbf{y}(t) \equiv (-1, -1)$ *is unstable.*

Flows, orbits and limit sets

From now on our interest lies with the Cauchy problem for the autonomous system of equations in $\mathbb{R}^n$

$$\mathbf{x}' = \mathbf{f}(\mathbf{x}), \tag{5.2.58}$$
$$\mathbf{x}(0) = \mathbf{x}_0 \tag{5.2.59}$$

To simplify considerations, we assume that $\mathbf{f}$ satisfies the assumptions of the Picard theorem on $\mathbb{R}^n$ so that the solutions exist for all $-\infty < t < \infty$. The *flow* of (5.2.58) is the map

$$\mathbb{R} \times \mathbb{R}^n \ni (t, \mathbf{x}_0) \to \phi(t, \mathbf{x_0}) \in \mathbb{R}^n$$

where $\mathbf{x}(t) = \phi(t, \mathbf{x}_0)$ is the solution to (5.2.58) satisfying $\mathbf{x}(0) = \mathbf{x}_0$. We note the following important properties of the flow:

$$\phi(0, \mathbf{x}) = \mathbf{x}, \tag{5.2.60}$$
$$\phi(t_1, \phi(t_2, \mathbf{x})) = \phi(t_1 + t_2, \mathbf{x}) \tag{5.2.61}$$

for any $t_1, t_2 \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}$. Property (5.2.61) follows from the simple lemma which w note for further reference

Lemma 5.15. *If $\mathbf{x}(t)$ is a solution to*

$$\mathbf{x}' = \mathbf{f}(\mathbf{x}),$$

then for any c the function $\hat{\mathbf{x}}(t) = \mathbf{x}(t + c)$ also satisfies this equation.

Proof. Define $\tau = t + c$ and use the chain rule for $\hat{x}$. We get

$$\frac{d\hat{\mathbf{x}}(t)}{dt} = \frac{d\mathbf{x}(t+c)}{dt} = \frac{d\mathbf{x}(\tau)}{d\tau}\frac{d\tau}{dt} = \frac{d\mathbf{x}(\tau)}{d\tau} = \mathbf{f}(\mathbf{x}(\tau)) = \mathbf{f}(\mathbf{x}(t+c)) = \mathbf{f}(\hat{\mathbf{x}}(t)).$$

In terms of the flow we can rephrase the lemma by noting that $\mathbf{u}(t) := \phi(t + t_2, \mathbf{x})$ is the solution of (5.2.58), with $\mathbf{u}(0) = \phi(t_2, \mathbf{x})$, and thus, by Picard's theorem and the definition of the flow, must coincide with $\phi(t, \phi(t_2, \mathbf{x}))$.

Remark 5.16. Occasionally we could need solutions satisfying the initial condition at $t = t_0 \neq 0$. Then we should use the notation $\phi(t, t_0, \mathbf{x})(= \phi(t + t_0, \mathbf{x}))$ so that the first property above is $\phi(t_0, t_0, \mathbf{x}) = \mathbf{x}$.

Definition 5.17. *The set*

$$\Gamma_{\mathbf{x}_0} = \{\mathbf{x} \in \mathbb{R}^n;\ \mathbf{x} = \phi(t, \mathbf{x}_0), t \in \mathbb{R}\}$$

is called the trajectory, or orbit, of the flow through $\mathbf{x}_0$. If $\mathbf{x}_0$ plays no role in the considerations, we shall drop it from the notation. By positive (negative) half-orbit we understand the curve

$$\Gamma^{\pm}_{\mathbf{x}_0} = \{\mathbf{x} \in \mathbb{R}^n;\ \mathbf{x} = \phi(t, \mathbf{x}_0), t \gtreqless 0\}.$$

The n-dimensional $\mathbf{y}$-space, in which the orbits are situated, is called the phase space *of the solutions of (5.2.58).*

Theorem 5.18. *Assume that the assumptions of the Picard theorem are satisfied on $\mathbb{R}^n$. Then*

(i) there exists one and only one orbit through every point $\mathbf{x^0} \in \mathbb{R}^2$. In particular, if the orbits of two solutions $\mathbf{x}(t)$ and $\mathbf{y}(t)$ have one point in common, then they must be identical.

(ii) Let $\mathbf{x}(t)$ be a solution to (5.2.58). If for some $T > 0$ and some t_0 we have $\mathbf{x}(t_0 + T) = \mathbf{x}(t_0)$, then $\mathbf{x}(t + T) = \mathbf{x}(t)$ for all t. In other words, if a solution $\mathbf{x}(t)$ returns to its starting value after a time $T > 0$, then in must be periodic (that is, it must repeat itself over every time interval of length T).

Proof. ad (i) Let $\mathbf{x^0}$ be any point in $\mathbb{R}^2$. Then, from the Picard theorem, we know that there is a solution of the problem $\mathbf{x}' = \mathbf{f}(\mathbf{x}), \mathbf{x}(0) = \mathbf{x^0}$ and the orbit of this solution passes through $\mathbf{x^0}$ from the definition of the orbit. Assume now that there is another orbit passing through $\mathbf{x^0}$, that is, there is a solution $\mathbf{y}(t)$ satisfying $\mathbf{y}(t_0) = \mathbf{x^0}$ for some t_0. From Lemma 5.15 we know that $\hat{\mathbf{y}}(t) = \mathbf{y}(t + t_0)$ is also a solution. However, this solution satisfies $\hat{\mathbf{y}}(0) = \mathbf{y}(t_0) = \mathbf{x^0}$, that is, the same initial condition as $\mathbf{x}(t)$. By the uniqueness part of Picard theorem we must then have $\mathbf{x}(t) = \hat{\mathbf{y}}(t) = \mathbf{y}(t + t_0)$ for all t for which the solutions are defined. This implies that the orbits are identical. In fact, if ξ is an element of the orbit of $\mathbf{x}$, then for some t' we have $\mathbf{x}(t') = \xi$. However, we have also $\xi = \mathbf{y}(t' + t_0)$ so that ξ belongs to the orbit of $\mathbf{y}(t)$. Conversely, if ξ belongs to the orbit of $\mathbf{y}$ so that $\xi = \mathbf{y}(t'')$ for some t'', then by $\xi = \mathbf{y}(t'') = \mathbf{x}(t'' - t_0)$, we see that ξ belongs to the orbit of $\mathbf{x}$.

ad (ii) Assume that for some numbers t_0 and $T > 0$ we have $\mathbf{x}(t_0) = \mathbf{x}(t_0 + T)$. The function $\mathbf{y}(t) = \mathbf{x}(t + T)$ is again a solution satisfying $\mathbf{y}(t_0) = \mathbf{x}(t_0 + T) = \mathbf{x}(t_0)$, thus from Picard Theorem, $\mathbf{x}(t) = \mathbf{y}(t)$ for all t for which they are defined and therefore $\mathbf{x}(t) = \mathbf{x}(t + T)$ for all such t. ■

Example 5.19. A curve in the shape of a figure 8 cannot be an orbit. In fact, suppose that the solution passes through the intersection point at some time t_0, then completing the first loop returns after time T, that is, we have $\mathbf{x}(t_0) = \mathbf{x}(t_0 + T)$. From (ii) it follows then that this solution is periodic, that is, it must follow the same loop again and cannot switch to the other loop.

Corollary 5.20. *A solution $\mathbf{x}(t)$ of (5.2.58) is periodic if and only if its orbit is a closed curve in $\mathbb{R}$.*

Proof. Assume that $\mathbf{x}(t)$ is a periodic solution of (5.2.58) of period T, that is $\mathbf{x}(t) = t + T$. If we fix t_0, then, as t runs from t_0 to $t_0 + T$, the point $\mathbf{x}(t) = (x_1(t), x_2(t))$ traces a curve, say C, from $\xi = \mathbf{x}(t)$ back to the same point ξ without intersections and, if t runs from $-\infty$ to ∞, the curve C is traced infinitely many times.

Conversely, suppose that the orbit C is a closed curve (containing no equilibrium points). The orbit is parametrically described by $\mathbf{x}(t)$, $-\infty < t < \infty$ in a one-to-one way (as otherwise we would have $\mathbf{x}(t') = \mathbf{x}(t'')$ for some $t' \neq t''$ and, by the previous theorem, the solution would be periodic). Consider a sequence $(t_n)_{n\in\mathbb{N}}$ with $t_n \to \infty$. Since the sequence $\mathbf{x}(t)$ is bounded, we find a subsequence $t'_n \to \infty$ such that $\mathbf{x}(t'_n) \to \mathbf{x} \in C$. Then, however, $\mathbf{x} = \mathbf{x}(t_0)$ for some finite t_0 since C does not contain equilibria. Consider a neighbourhood of $\mathbf{x}(t_0)$ which is the image of some interval $(t_0 - \epsilon, t_0 + \epsilon)$. Since $\mathbf{x}(t'_n) \to \mathbf{x}(t_0)$, $t'_n \in (t_0 - \epsilon, t_0 + \epsilon)$ for sufficiently large n which contradicts $t'_n \to \infty$. ■

Definition 5.21. *The ω-limit set of the trajectory $\Gamma_{\mathbf{x}_0}$ is the set of all points $\mathbf{p} \in \mathbb{R}^n$ such that there is a sequence $(t_n)_{n\in\mathbb{N}}$ such that $t_n \to \infty$ as $n \to \infty$ for which*

$$\lim_{n\to\infty} \phi(t_n, \mathbf{x}_0) = \mathbf{p}.$$

Similarly, the α-limit set of the trajectory $\Gamma_{\mathbf{x}_0}$ is the set of all points $\mathbf{q} \in \mathbb{R}^n$ such that there is a sequence $(t_n)_{n\in\mathbb{N}}$ such that $t_n \to -\infty$ as $n \to \infty$ for which

$$\lim_{n\to\infty} \phi(t_n, \mathbf{x}_0) = \mathbf{q}.$$

Since for a given equation (5.2.58)) any trajectory is uniquely determined by any point on it (and conversely), sometimes we shall use the notation $\omega(\mathbf{x}_0)$ instead of $\omega(\Gamma_{\mathbf{x}_0})$ (and the same for α-limit sets).

Example 5.22. If $\mathbf{v}_0$ is an equilibrium, then $\Gamma_{\mathbf{v}_0} = \{\mathbf{v}_0\} = \omega(\Gamma_{\mathbf{v}_0}) = \alpha(\Gamma_{\mathbf{v}_0})$. The only ω and α limit sets of scalar equations are equilibria.

Example 5.23. Consider the system in polar coordinates

$$r' = r(1 - r^2),$$
$$\theta' = 1.$$

Since $r' > 0$ if $r \in (0,1)$ and $r' < 0$ if $r > 1$, trajectories which start with $0 < r < 1$ and $r > 1$ tend to $r = 1$ which, since $\theta' \neq 0$, is a periodic orbit. The origin $r = 0$ is a stationary point and so $\omega(\{r = 0\}) = \alpha(\{r = 0\}) = (0,0)$. If $r \neq 0$, then

$$\omega(\Gamma_{(r,\theta)}) = \{(r,\theta);\ r = 1\},$$

and

$$\omega(\Gamma_{(r,\theta)}) = \begin{cases} \{(r,\theta);\ r = 0\} \text{ for } r < 1, \\ \text{does not exist for } r > 1 \end{cases}$$

We start with two observations often used in the sequel. They are seemingly obvious but require some reflection.

Remark 5.24. 1. How do we prove that $\phi(t, \mathbf{x}_0) \to \mathbf{x}$ as $t \to \infty$? We take arbitrary sequence $(t_n)_{n\in\mathbb{N}}$ and show that it contains a subsequence with $\phi(t_{n_k}, \mathbf{x}_0) \to \mathbf{x}$. In fact, assume that the above holds but $\phi(t, \mathbf{x}_0) \nrightarrow \mathbf{x}$. Then there must be a sequence $(t_n)_{n\in\mathbb{N}}$ for which $\|\phi(t_n, \mathbf{x}_0) - \mathbf{x}\| \geq r$ for some r. But such a sequence cannot contain a subsequence converging to $\mathbf{x}$, so we proved the thesis.

2. Consider a bounded sequence $\mathbf{y}_n = \phi(t_n, \mathbf{x}_0)$ with $t_n \to \infty$. Then, as we know, we have a subsequence $\mathbf{y}_{n_k}$ converging to, say, $\mathbf{y}$. Can we claim that $\mathbf{y} \in \omega(\Gamma_{\mathbf{x}_0})$; that is, there is a sequence $(t'_n)_{n\in\mathbb{N}}$ converging to ∞ such that $\phi(t'_n, \mathbf{x}_0)) \to \mathbf{y}$? The reason it is not obvious is that by selecting $\mathbf{y}_{n_k} = \phi(t_{n_k}, \mathbf{x}_0)$ we could create a sequence $\{t_{n_1}, \ldots, t_{n_k} \ldots\}$ which does not diverge to ∞ as we do not have any control on how the latter is ordered with respect to the former. First, we note that we can assume that $t_n - t_{n-1} \geq 1$ for any n. Indeed, $(t_n)_{n\in\mathbb{N}}$ must contain a subsequence with this property so, if necessary, we can select such a subsequence for further analysis. Next, we note that we can assume that $\mathbf{y}_n \neq \mathbf{y}_m$ for $n \neq m$ as otherwise, by Theorem 5.18, the orbit would be periodic since $t_n \neq t_m$ and in the case of periodic orbit with period T we can take $t_n = nT$. Thus, $\mathbf{y}_{n_k} \neq \mathbf{y}_{n_l}$ and hence $t_{n_k} \neq t_{n_l}$ for $k \neq l$. Now, $(n_k)_{k\in\mathbb{N}}$ is an infinite sequence of mutually different natural numbers and thus we can select a monotonic subsequence $(n_{k_l})_{l\in\mathbb{N}}$.

Consider a nested sequence of balls $B(\mathbf{y}, 1/N)$. For each N there is n_N such that $\mathbf{y}_{n_k} = \phi(t_{n_k}, \mathbf{x}_0)$ for all $n_k \geq n_N$. In particular, each set $\{t_{n_k};\ n_k \geq n_N\}$ is infinite. Now, the crucial observation is that an infinite subset of a sequence $(t_n)_{n\in\mathbb{N}}$ converging to ∞ must be unbounded and thus must contain a subsequence converging to ∞. Indeed, from the definition of convergence to ∞, for each M only finite number of elements of the sequence is smaller that M. If so, having selected $\phi(t_N, \mathbf{x}_0)$ in $B(\mathbf{y}, N^{-1})$, we select $\phi(t_{N+1}, \mathbf{x}_0)$ in $B(\mathbf{y}, (N+1)^{-1})$ with $t_{N+1} > t_N + 1$ as an infinite collection of t_{n_k}s corresponding to $\mathbf{y}_{n_k} \in B(\mathbf{y}, (N+1)^{-1})$ must contain arbitrary large t_{n_k}s. This shows that we have $\phi(t_N, \mathbf{x}_0) \to \mathbf{y}_0$ with $t_N \to \infty$ as $N \to \infty$.

Lemma 5.25. *ω-limit sets have the following properties:*

1. *If Γ is bounded, then $\omega(\Gamma)$ is non-empty;*
2. *An ω-limit set is closed;*
3. *If $\mathbf{y} \in \omega(\Gamma)$, then $\Gamma_{\mathbf{y}} \subset \omega(\Gamma)$; that is, $\omega(\Gamma)$ is invariant.*
4. *If Γ is bounded, then $\omega(\Gamma)$ is connected;*
5. *If $\mathbf{z} \in \omega(\Gamma)$ and $\Gamma \cap \omega(\Gamma') \neq \emptyset$, then $\mathbf{z} \in \omega(\Gamma')$.*
6. *If $\Gamma_{\mathbf{y}}$ is bounded, then $\phi(t, \mathbf{y}) \to \omega(\Gamma_{\mathbf{y}})$ as $t \to \infty$ in the sense that for each $\epsilon > 0$ there is $t > 0$ such that for every $t' > t$ there is $\mathbf{p} \in \omega(\Gamma_{\mathbf{y}})$ (possibly depending on t) which satisfies $\|\phi(t, \mathbf{y}) - \mathbf{p}\| < \epsilon$.*

The same properties are valid for α-limit sets.

Proof. ad 1) Let $\mathbf{x} \in \Gamma$. Taking e.g. $\phi(n, \mathbf{x})$ we obtain a bounded sequence of points which must have a converging subsequence with the limit in $\omega(\Gamma)$ by Remark 5.24.

ad 2) Let $(\mathbf{p}_n)_{n\in\mathbb{N}}$ be a sequence of points of $\omega(\Gamma_{\mathbf{x}})$ converging to $\mathbf{p} \in \mathbb{R}^n$. We must prove that there is a sequence $t_n \to \infty$ such that $\phi(t_n, \mathbf{x}) \to \mathbf{p}$. We proceed as follows. For $\epsilon = 1/2$ we can find $\mathbf{p}_{n_1}$ and t_{n_1} such that

$$\|\mathbf{p} - \phi(t_{n_1}, \mathbf{x})\| \leq \|\mathbf{p} - \mathbf{p}_{n_1}\| + \|\mathbf{p}_{n_1} - \phi(t_{n_1}, \mathbf{x})\| \leq 1/2 + 1/2 = 1$$

Similarly, for $\epsilon = 1/4$ we can find $\mathbf{p}_{n_2}$ and $t_{n_2} > t_{n_1} + 1$ such that

$$\|\mathbf{p} - \phi(t_{n_2}, \mathbf{x})\| \leq \|\mathbf{p} - \mathbf{p}_{n_2}\| + \|\mathbf{p}_{n_2} - \phi(t_{n_2}, \mathbf{x})\| \leq 1/4 + 1/4 = 1/2,$$

and, by induction, for any given k we can find $\mathbf{p}_{n_k}$ and $t_{n_k} > t_{n_{k-1}} + 1$ such that

$$\|\mathbf{p} - \phi(t_{n_k}, \mathbf{x})\| \leq \|\mathbf{p} - \mathbf{p}_{n_k}\| + \|\mathbf{p}_{n_k} - \phi(t_{n_k}, \mathbf{x})\| \leq 1/2k + 1/2k = 1/k.$$

Since the sequence $(t_{n_k})_{k\in\mathbb{N}}$ is infinite and increasing by 1 at each step, it must diverge to infinity. Thus, $\mathbf{p} \in \omega(\Gamma_{\mathbf{x}})$.

ad 3) Let $\mathbf{y} \in \omega(\Gamma_{\mathbf{x}})$ and consider arbitrary $\mathbf{q} \in \Gamma_{\mathbf{y}}$. Thus $\mathbf{q} = \phi(t', \mathbf{y})$. Since $\mathbf{y}$ is in the ω-limit set,

$$\mathbf{y} = \lim_{t_n \to \infty} \phi(t_n, \mathbf{x}).$$

But

$$\phi(t_n + t', \mathbf{x}) = \phi(t', \phi(t_n, \mathbf{x})) \to \phi(t', \mathbf{y}) = \mathbf{q}$$

by continuity of the flow with respect to the initial value. Since $t' + t_n \to \infty$, we obtain that $\mathbf{q} \in \omega(\Gamma_{\mathbf{y}})$ and since $\mathbf{q} \in \omega(\Gamma_{\mathbf{x}})$ was arbitrary, the thesis follows.

ad 4.) By 1.), the set $\omega(\Gamma)$ is bounded and closed. Thus, if it was not connected, then $\omega(\Gamma) = A \cup B$ with A, B closed bounded and a distance $d > 0$ apart. Fix $\mathbf{x} \in \Gamma$ and chose $\mathbf{v} \in A$. There is $t_n \to \infty$ such that $\mathbf{x_n} := \phi(t_n, \mathbf{x}) \to \mathbf{v}$. Thus, for sufficiently large n, the distance between $\mathbf{x}_n$ and B is not less than $3d/4$. Let $\mathbf{w} \in B$. For each $\mathbf{x}_n$ we can find $t'_n > 0$ such that the distance between $\mathbf{y}_n = \phi(t'_n, \mathbf{x}_n) = \phi(t_n + t'_n, \mathbf{x})$ and $\mathbf{w}$ is not greater than $d/4$, thus the distance between A and $\mathbf{y}_n$ is at least $3d/4$. On the other hand, from the continuity of the flow, for each n there is a point $\mathbf{z}_n = \phi(t''_n, \mathbf{x})$ whose distance from A is $d/2$. The set of such points is bounded and thus have a convergent subsequence whose limit $\mathbf{z} \in \omega(\Gamma)$ is at the distance $d/2$ from A, which contradicts the assumption that the components of $\omega(\Gamma)$ are d apart.

ad 5.) Let $\mathbf{y} \in \omega(\Gamma)$ and $\mathbf{y} = \lim_{t'_n \to \infty} \phi(t'_n, \mathbf{x})$ with $\mathbf{x} \in \Gamma'$. Let us fix $\epsilon > 0$. From continuity of the flow with respect to the initial condition, we know that for a given ϵ and T, there is $\delta_{T,\epsilon}$ such that $\|\phi(t, \mathbf{x}_1) - \phi(t, \mathbf{x}_2)\| < \epsilon$ provided $\|\mathbf{x}_1 - \mathbf{x}_2\| < \delta_{T,\epsilon}$ for all $0 \leq t \leq T$.

For this given $\epsilon > 0$ we find t_n such that $\|\mathbf{z} - \phi(t_n, \mathbf{y})\| < \epsilon$ and also t'_n such that $\|\mathbf{y} - \phi(t'_n, \mathbf{x}_0)\| < \delta_{t_n,\epsilon}$ (since $\mathbf{y} \in \omega(\Gamma')$). Hence

$$\|\mathbf{z} - \phi(t_n, \phi(t'_n, \mathbf{x}_0))\| \leq \|\mathbf{z} - \phi(t_n, \mathbf{y})\| + \|\phi(t_n, \mathbf{y}) - \phi(t_n, \phi(t'_n, \mathbf{x}_0))\| < 2\epsilon$$

but since $\phi(t_n, \phi(t'_n, \mathbf{x}_0)) = \phi(t_n + t'_n, \mathbf{x}_0)$, we see that $\mathbf{z} \in \omega(\Gamma')$ (we note that the sequence $\tau_n := t_n + t'_n$ can be made increasing with n and thus convergent to ∞ as t_n and t'_n are).

ad 6.) Suppose that the statement is false; that is, there is $\epsilon > 0$ and a sequence $t_n \to \infty$ with $\|\phi(t_n, \mathbf{y}) - \mathbf{p}\| > \epsilon$ for all $\mathbf{p} \in \omega(\Gamma_{\mathbf{y}})$. This means that the sequence $\phi(t_n, \mathbf{y})$ stays at least ϵ away from $\omega(\Gamma_{\mathbf{y}})$. But the orbit is bounded so, by Remark 5.24, there is a converging subsequence $\phi(t_{n_k}, \mathbf{y})$ of this sequence which, by the definition of $\omega(\Gamma_{\mathbf{y}})$ must belong to it. ■

3 The Poincaré-Bendixon Theory

The linearization theorem, Theorem 5.13, may suggest that locally nonlinear dynamics is the same as linear. This is, however, false even in 2 dimensions. The Poincaré-Bendixon theory provides a complete description of two-dimensional dynamics by giving full classification of possible limit sets for planar autonomous systems.

3.1 Preliminaries

Let $\phi(t, \mathbf{x})$ be the flow generated by the system

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}), \tag{5.3.62}$$

where $\mathbf{y} = (y_1, y_2) \in \mathbb{R}^2$. Throughout this chapter we assume that $\mathbf{f}$ is a differentiable function. By a *local transversal* to ϕ we mean a line segment L which all trajectories cross from one side. In other words, the field $\mathbf{f}$ always forms either an acute or an obtuse angle with the normal to L.

Lemma 5.26. *If $\mathbf{x}_0$ is not a stationary point of (5.3.62), then there is always possible to construct a local transversal in a neighbourhood of $\mathbf{x}_0$.*

Proof. If $\mathbf{x}_0$ is not stationary point, then $\mathbf{v} = \mathbf{f}(\mathbf{x}_0) \neq 0$. Take coordinate system with origin at $\mathbf{x}_0$ and axes parallel to $\mathbf{v}$ and $\mathbf{w} \perp \mathbf{v}$. Let (v, w) be the coordinates of a point in this system. Then (5.3.62) can be written as

$$\begin{aligned} v' &= a + O(\|(v, w) - \mathbf{x}_0\|), \\ w' &= O(\|(v, w) - \mathbf{x}_0\|). \end{aligned}$$

Here $a \neq 0$ is the norm of $\mathbf{v}$. Now we can chose line L through $\mathbf{x}_0$ along $\mathbf{w}$. As long as we are close to $\mathbf{x}_0$, $v' \neq 0$ and the trajectories cross L in the same direction (positive if $a > 0$ and negative otherwise). □

Having found a local transversal L at $\mathbf{x}$ we can construct a *flow box* around $\mathbf{x}$ by taking a collection of trajectories with starting points at L and t running from $-\delta$ to δ with δ small enough that all trajectories exist over this time interval; that is

$$V_{\mathbf{x}} = \{\mathbf{y};\ \mathbf{y} = \phi(t, \mathbf{z}),\ \mathbf{z} \in L, t \in (-\delta, \delta)\}.$$

An important property of flow box is that if a trajectory starts close enough to $\mathbf{x}$, it crosses L (either forward or backward). Precisely speaking, we have

Lemma 5.27. *Let L be a transversal at $\mathbf{x}$. There is a neighbourhood D of $\mathbf{x}$ such that for any $\mathbf{y} \in D$ there are $t_1 < t_2$ such that the trajectory segment $\Gamma_{t_1,t_2} = \{\mathbf{z};\ \mathbf{z} = \phi(t, \mathbf{y}), t \in (t_1, t_2)\} \subset D$ and $\Gamma_{t_1,t_2} \cap L \neq \emptyset$.*

Proof. We chose coordinates (z_1, z_2) as in the proof of Lemma 5.26 so that $\mathbf{x} = (0, 0)$ and $\mathbf{f}(\mathbf{x}) = (a, 0)$ where we select $a > 0$. Then L is on the x_2 axis. Let D_δ be the ball $\|(z_1, z_2)\| = |z_1| + |z_2| < \delta$ (that is a square with diagonals on L and $\mathbf{f}(\mathbf{x})$. We can chose δ small enough for the following condition to hold:

a) The slope of $\mathbf{f}(\mathbf{z})$ is strictly between -1 and 1; that is, $|f_2(\mathbf{z})/f_1(\mathbf{z})| < 1$,

b) $f_1(\mathbf{z}) > a/2$,

for $\mathbf{z} \in D_\delta$. This is possible by continuity of $\mathbf{f}$ in a neighbourhood of $\mathbf{x}$. Notice, in particular, that $\mathbf{f}$ is always transversal to the sides of D_δ and pointing outward at the righ-hand side and inward at the left-hand side of it. Hence, a trajectory can only leave D_δ at the right and enter at the left. Now, clearly, for any solution $\mathbf{z}(t)$

$$z_1(t) - z_1(0) = \int_0^t f_1(\mathbf{z}(s))ds > \frac{a}{2}t.$$

Since the maximum value of $z_1(t) - z_1(0)$ is 2δ, the solution starting from any point $\mathbf{z}_0 \in D_\delta$ must leave it in time shorter than $4\delta/a$. Let t_2 be the smallest value at which $\phi(t, \mathbf{z}_0)$ intersects the right-hand side of D_δ. Similarly, $\phi(t, \mathbf{z}_0)$ intersects the left-hand side of D_δ at some time $t \in (-4\delta/a, 0)$ nd hence there is $t_1 < 0$ at which this happens for the first time (going backward). Hence the segment $\Gamma_{t_1,t_2} = \{\mathbf{z};\ \mathbf{z} = \phi(t, \mathbf{y}), t \in (t_1, t_2)\} \in D$ and also $\Gamma_{t_1,t_2} \cap L \neq \emptyset$. □

Lemma 5.28. *If a trajectory $\Gamma_{\mathbf{x}}$ intersects a local transversal several times, the successive crossings points move monotonically along the transversal.*

Proof. Consider two successive crossings $\mathbf{y}_1 = \phi(t_1, \mathbf{x})$ and $\mathbf{y}_2 = \phi(t_2, \mathbf{x})$ with, say, $t_1 < t_2$ and a closed curve S composed of the piece Γ' of the trajectory between $\mathbf{y}_1$ and $\mathbf{y}_2$ and the piece L' of the transversal between these two points. Using Jordan's theorem, S divides $\mathbb{R}^2$ into two disjoint open sets with one, say D_1, bounded the other, say D_2, unbounded. We can assume that the flow through L is from D_1 into D_2. Consider $\mathbf{y}_3 = \phi(t_3, \mathbf{x}) \in L$ with $t_3 > t_2$ and first assume $\mathbf{y}_3 \in D_2$. Taking $t' = t_2 + \epsilon$ with ϵ sufficiently small we can be sure that $\phi(t', \mathbf{x})$ is outside D_1. If we assume that $\mathbf{y}_3 \in L'$, then there is ϵ' such that $\phi(t_3 - \epsilon', \mathbf{x}) \in D_1$. Hence, the trajectory between t' and $t_3 - \epsilon'$ joins points of D_1 and D_2. However, it cannot cross Γ' as trajectories cannot cross; also it cannot enter D_1 through L' by its definition. By similar argument, $\mathbf{y}_3$ cannot belong to the sub-segment of L which is inside D_1 (we note that this sub-segment cannot stick outside D_1 as this would require that a point moves along a piece of trajectory in two directions at once). Thus, it must belong to subsegment with $\mathbf{y}_2$ as the end-point. □

An important corollary is:

Corollary 5.29. *If $\mathbf{x} \in \omega(\Gamma_{\mathbf{x}_0})$ is not a stationary point and $\mathbf{x} \in \Gamma_{\mathbf{x}_0}$, then $\Gamma_{\mathbf{x}} (= \Gamma_{\mathbf{x}_0})$ is a closed curve.*

Proof. Since $\mathbf{x} \in \Gamma_{\mathbf{x}_0}$, $\omega(\Gamma_{\mathbf{x}_0}) = \omega(\Gamma_{\mathbf{x}})$ (the limit set depends on the trajectory and not on the initial point). Hence, $\mathbf{x} \in \omega(\Gamma_{\mathbf{x}})$. Chose L to be a local transversal at $\mathbf{x}$. By Lemma 5.27, for sufficiently large T we have an increasing sequence $t_i > t_{i-1} \geq T$ such that $\phi(t_i, \mathbf{x}) \to \mathbf{x}$ as $t_i \to \infty$ and $\phi(t_i, \mathbf{x}) \in L$ (see Lemma 5.27). Also, $\phi(0, \mathbf{x}) = \mathbf{x}$. Suppose $\phi(t_i, \mathbf{x}) \neq \mathbf{x}$, $t_i > T$, then successive intercepts are bounded away from $\mathbf{x}$ which contradicts the fact that $\mathbf{x} \in \omega(\Gamma_{\mathbf{x}})$. Hence $\phi(t_1, \mathbf{x}) = \mathbf{x}$ for some t and, by Theorem 5.18, the solution is periodic. □

Theorem 5.30. *If an orbit $\Gamma_{\mathbf{x}_0}$ enters and does not leave a closed bounded domain Ω which contains no equilibrium points, then $\omega(\Gamma_{\mathbf{x}_0})$ is a closed orbit.*

Proof. First we prove that $\omega(\Gamma_{\mathbf{x}_0})$ contains a closed orbit. Let $\mathbf{x} \in \omega(\Gamma_{\mathbf{x}_0})$. There are two possibilities.

(i) If $\mathbf{x} \in \Gamma_{\mathbf{x}_0}$, then by Corollary 5.29, $\Gamma_{\mathbf{x}_0}$ is a closed orbit.

(ii) If $\mathbf{x} \notin \Gamma_{\mathbf{x}_0}$, then since $\mathbf{x} \in \omega(\Gamma_{\mathbf{x}_0})$, the orbit $\Gamma_{\mathbf{x}} \subset \omega(\Gamma_{\mathbf{x}_0})$ by Lemma 5.25 (3) and, because $\omega(\Gamma_{\mathbf{x}_0})$ is closed, $\omega(\Gamma_{\mathbf{x}}) \subset \omega(\Gamma_{\mathbf{x}_0})$. Let $\mathbf{x}^* \in \omega(\Gamma_{\mathbf{x}}) \subset \omega(\Gamma_{\mathbf{x}_0})$. If $\mathbf{x}^* \in \Gamma_{\mathbf{x}}$ then, again by Corollary 5.29, $\Gamma_{\mathbf{x}} \subset \omega(\Gamma_{\mathbf{x}_0})$ is a closed orbit. This leaves the only possibility that $\mathbf{x}^* \notin \Gamma_{\mathbf{x}}$. Consider a local transversal L at $\mathbf{x}^*$. Arguing as in the proof of Corollary 5.29 we have a sequence $(\mathbf{p}_i)_{i \in \mathbb{N}}$, $\mathbf{p}_i = \phi(t_i, \mathbf{x}) \in L$ with $\mathbf{p}_i \to \mathbf{x}^*$ as $t_i \to \infty$ in a monotonic way. On the other hand, $p_i \in \omega(\Gamma_{\mathbf{x}_0})$ and L is also a local transversal at each p_i. This means that there are sequences on the trajectory $\Gamma_{\mathbf{x}_0}$ converging monotonically to, say, $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$. Assume $\|\mathbf{p}_i - \mathbf{p}_{i+1}\| < \epsilon$ We have, say, $\|\phi(t_1, \mathbf{x}_0) - \mathbf{p}_i\| < \epsilon/4$, then there must be $t_2 > t_1$ such that $\|\phi(t_2, \mathbf{x}_0) - \mathbf{p}_{i+1}\| < \epsilon/4$ but then the next t_3 at which $\Gamma_{\mathbf{x}_0}$ intersects L must be closer than $\epsilon/4$ to both $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$, by Lemma 5.28. This is a contradiction, which shows that $\mathbf{x}^* \in \Gamma_{\mathbf{x}}$ and thus $\omega(\Gamma_{\mathbf{x}_0})$ contains a closed orbit.

The final step of the proof is to show that this orbit, say, Γ is equal to $\omega(\Gamma_{\mathbf{x}_0})$. To do this, we must show that $\phi(t, \mathbf{x}_0) \to \Gamma$ as $t \to \infty$ in the sense of Lemma 5.25(6); that is, that for each $\epsilon > 0$ there is $t > 0$ such that for every $t' > t$ there is $\mathbf{p} \in \Gamma$ (possibly depending on t) which satisfies $\|\phi(t, \mathbf{x}_0) - \mathbf{p}\| < \epsilon$. The argument again uses the properties of a local transversal. Let $\mathbf{z} \in \Gamma \subset \omega(\Gamma_{\mathbf{x}_0})$ and consider a local transversal L at $\mathbf{z}$. Using Lemmas 5.27 and 5.28 we find a sequence $t_0 < t_1 < \ldots t_n \to \infty$ such that $\mathbf{x}_n = \phi(t_k, \mathbf{x}_0) \in L$, $\mathbf{x}_n \to \mathbf{z}$ (in a monotonic way along L). Moreover, we select the sequence $(t_n)_{n \in \mathbb{N}}$ to be subsequent intercepts of $\Gamma_{\mathbf{x}_0}$ with L so that $\phi(t, \mathbf{x}_0) \notin L$ for $t_n < t < t_{n+1}$. Thus, we must estimate what happens to $\phi(t, \mathbf{x}_0)$ for $t \neq t_n$. For any δ there is n_0 such that for $n \geq n_0$ $\|\phi(t_n, \mathbf{x}_0) - \mathbf{z}\| < \delta$. Hence, We by continuity of the flow with respect to the initial condition, for any fixed T an any ϵ there is n_0 such that for any $n \geq n_0$ and $0 \leq t' \leq T$

$$\|\phi(t', \mathbf{x}_n) - \phi(t', \mathbf{z})\| = \|\phi(t' + t_n, \mathbf{x}_0) - \phi(t', \mathbf{z})\| < \epsilon.$$

Now, we observe that if t' changes from 0 to $t_{n+1} - t_n$, the point $\mathbf{x}_n = \phi(t' + t_n, \mathbf{x}_0)$ moves from $\mathbf{x}_n$ to $\mathbf{x}_{n+1}$ which is even closer to $\mathbf{z}$ than $\mathbf{x}_n$ and the trajectory after $\mathbf{x}_{n+1}$ will again stay close to Γ for some finite

time T. However, at each cycle the trajectory may wander off when $t' > T$ So, the problem is to determine whether it is possible for $t_{n+1} - t_n$ to become arbitrary large. We know that Γ is closed, hence it is an orbit of a periodic solution, say, $\phi(\lambda, \mathbf{z}) = \mathbf{z}$. Using again continuity of the flow with respect to the initial condition, for any $\mathbf{x}_n$ sufficiently close to $\mathbf{z}$ (that is, for all sufficiently large n), $\phi(\lambda, \mathbf{x}_n)$ will be in some some neigbourhood D_δ of $\mathbf{z}$ (described in Lemma 5.27). But then for all such n, there is $t'_n \in (-\delta, \delta)$ such that $\phi(\lambda + t'_n, \mathbf{x}_0) \in L$. This means that $t_{n+1} - t_n \leq \lambda + \delta$ and the time interval T can be chosen independently of n. Precisely, let us fix $\epsilon > 0$, then there is η such that for all $\|\mathbf{x}_n - \mathbf{z}\| \leq \eta$ and $|t| < \lambda + \delta$ we have

$$\|\phi(t', \mathbf{x}_n) - \phi(t', \mathbf{z})\| = \|\phi(t' + t_n, \mathbf{x}_0) - \phi(t', \mathbf{z})\| < \epsilon.$$

For given η and δ there is n_0 such that for all $n \geq n_0$ we have both $\|\mathbf{x}_n - \mathbf{z}\| < \eta$ and $t_{n+1} - t_n \leq \lambda + \delta$. Hence, taking $t > t_{n_0}$ and selecting n with $t_n \leq t \leq t_{n+1}$ we have, using $t = t' + t_n$ with $0 < t' < \lambda + \delta$

$$\|\phi(t, \mathbf{x}_0) - \phi(t - t_n, \mathbf{z}\| = \|\phi(t - t_n, \mathbf{x}_n) - \phi(t - t_n, \mathbf{z}\| < \epsilon$$

and the proof of the theorem is complete. □

Remark 5.31. The proof of Theorem 5.30 actually shows that a stronger result is valid. Namely, if $\mathbf{x} \in \omega(\Gamma_{\mathbf{x}_0})$ is such that $\omega(\Gamma_{\mathbf{x}})$ contains a non-stationary point, then $\omega(\Gamma_{\mathbf{x}})$ is a closed orbit. This fact will be used in the proof of the following corollary.

Corollary 5.32. *Let $\Gamma_{\mathbf{x}}$ be a bounded trajectory of the system $\mathbf{y}' = \mathbf{f}(\mathbf{y})$ with C^1 planar field for which equilibria are isolated. Then the following three possibilities hold:*

1. *$\omega(\Gamma_{\mathbf{x}})$ is an equilibrium,*
2. *$\omega(\Gamma_{\mathbf{x}})$ is a periodic orbit,*
3. *$\omega(\Gamma_{\mathbf{x}})$ consists of a finite number of equilibria $\mathbf{p}_1, \ldots, \mathbf{p}_k$ connected by trajectories Γ with $\alpha(\Gamma) = \mathbf{p}_i$ and $\omega(\Gamma) = \mathbf{p}_j$.*

Proof. By Lemma 5.25(1), the set $\omega(\Gamma_{\mathbf{x}})$ is bounded an closed and thus can contain only finite number of equilibria. If it contains only equilibria, then it contains only one of them, since $\omega(\Gamma_{\mathbf{x}})$ is connected by Lemma 5.25(4), which covers the first possibility. If this is not the case, then $\omega(\Gamma_{\mathbf{x}})$ contains non-equilibria and, by Lemma 5.25(3) (invariance), it contains trajectories through these points. Let $\mathbf{u} \in \omega(\Gamma_{\mathbf{x}})$ be an arbitrary non-equilibrium. If $\omega(\Gamma_{\mathbf{u}})$ and $\alpha(\Gamma_{\mathbf{u}})$ are equilibria, then case 3. holds. Assume then that $\omega(\Gamma_{\mathbf{u}})$ contains a non-equilibrium point, say, $\mathbf{z}$. Then, arguing as in the proof of Theorem 5.30 (which is possible as $\mathbf{z}$ is a non-equilibrium, $\Gamma_{\mathbf{u}}$ is a periodic orbit, which means that $\omega(\Gamma_{\mathbf{x}})$ contains a periodic orbit. But then, by the second part of the proof of Theorem 5.30, the whole $\omega(\Gamma_{\mathbf{x}})$ is a periodic orbit. The same argument applies if $\alpha(\Gamma_{\mathbf{u}})$ contains a non-equilibrium point. □

Remark 5.33. Case 3 of the previous corollary can be fine-tuned. Namely, if $\mathbf{p}_1$ and $\mathbf{p}_2$ are two equilibria in $\omega(\Gamma_{\mathbf{x}})$, then there is at most one trajectory $\Gamma \subset \omega(\Gamma_{\mathbf{x}})$ with $\alpha(\Gamma) = \mathbf{p}_1$ and $\omega(\Gamma) = \mathbf{p}_2$. Indeed, assume to the contrary that we have two trajectories Γ_1, Γ_2 with this property and take points $\mathbf{q}_1 \in \Gamma_1$ close to $\mathbf{p}_1$ and $\mathbf{q}_2 \in \Gamma_2$ close to $\mathbf{p}_2$. Since $\mathbf{q}_i$ are not equilibria, there are local transversals L_1 and L_2 at these points. Since Γ_i, $i = 1, 2$ are in $\omega(\Gamma_{\mathbf{x}})$ the trajectory $\Gamma_{\mathbf{x}}$ crosses L_1 in the same direction as Γ_1 and L_2 in the direction of Γ_2. Since the direction of the field along the transversals is the same as of the above trajectories, the region bounded by L_1, the segment of $\Gamma_{\mathbf{x}}$ between L_1 and L_2, L_2, and corresponding segments of Γ_2 and Γ_1, is positively invariant (that is the trajectory $\Gamma_{\mathbf{x}}$ must stay inside). But this is a contradiction as the segments of trajectories Γ_1 and Γ_2 outside this region form a part of the limit set for $\Gamma_{\mathbf{x}}$.

Hence, if $\omega(\Gamma_{\mathbf{x}})$ contains two equilibria, then they must be joined either by a single trajectory, or two trajectories running in opposite directions (and thus forming, together with the equilibria, a loop).

We say that a closed orbit Γ is a *limit cycle* if for some $\mathbf{x} \notin \Gamma$ we have $\Gamma \subset \omega(\Gamma_{\mathbf{x}})$ (ω-*limit cycle*) or $\Gamma \subset \alpha(\Gamma_{\mathbf{x}})$ (α-*limit cycle*).

In the proof of Theorem 5.30 we showed that a limit cycle Γ enjoys the following property: there is $\mathbf{x} \notin \Gamma$ such that

$$\phi(t, \mathbf{x}) \to \Gamma, \qquad t \to \infty.$$

Geometrically it means that some trajectory spirals towards Γ as $t \to \infty$ (or $t \to -\infty$). It follows that limit cycles have certain (one-sided) stability property.

Proposition 5.34. *Let Γ be an ω-limit cycle such that $\Gamma \subset \omega(\Gamma_{\mathbf{x}})$, $\mathbf{x} \notin \Gamma$. Then there is a neighourhood V of $\mathbf{x}$ such that for any $\mathbf{y} \in V$ we have $\Gamma \in \omega(\Gamma_{\mathbf{y}})$.*

Proof. Let Γ be an ω-limit cycle and let $\phi(t, \mathbf{x})$ spirals toward Γ as $t \to \infty$. Take $\mathbf{z} \in \Gamma$ and a transversal L at $\mathbf{z}$. Take $t_0 < t_1$ such that $\mathbf{x}_0 = \phi(t_0, \mathbf{x}), \mathbf{x}_1 = \phi(t_1, \mathbf{x}) \in L$ with $\phi(t, \mathbf{x}) \notin L$ for $t_0 < t < t_1$. For sufficiently large t_0 the segment $L_{\mathbf{x}_0, \mathbf{x}_1}$ of L between $\mathbf{x}_0$ and $\mathbf{x}_1$ does not intersect Γ. Then the region A bounded by Γ, the part of $\Gamma_{\mathbf{x}}$ between $\mathbf{x}_0$ and $\mathbf{x}_1$ and $L_{\mathbf{x}_0, \mathbf{x}_1}$ is forward invariant, as is the set $B = A \setminus \Gamma$. For sufficiently large $t > 0$, $\phi(t, \mathbf{x})$ is in the interior of A. But than, the same is true for $\phi(t, \mathbf{y})$ for $\mathbf{y}$ sufficiently close to $\mathbf{x}$. Such $\phi(t, \mathbf{y})$ stays in A and by the Poincareé-Bendixon theorem, must spiral towards Γ. □

Corollary 5.35. *Let Γ be a closed orbit enclosing an open set Ω (contained in the domain of $\mathbf{f}$). Then Ω contains an equilibrium.*

Proof. Assume Ω contains no equilibrium. Our first step is to show that there must be the 'smallest' closed orbit. We know tat orbits don't intersect so if we have a collection of closed orbits Γ_n, then the regions Ω_n enclosed by them form a nested sequence with decreasing areas A_n. We also note that if we have a sequence $\mathbf{x}_n$ of points on Γ_n converging to $\mathbf{x} \in \Omega$, then $\mathbf{x}$ also is on a closed orbit. Otherwise, by Poincaré-Bendixon theorem, $\phi(t, \mathbf{x})$ would spiral towards a limit cycle but then this would be true for $\phi(t, \mathbf{x}_n)$ for sufficiently large n, by virtue of the previous result, which contradict the assumption that $\mathbf{x}_n$ is on a closed orbit.

Let $0 \leq A$ be an infimum of all areas of regions enclosed by closed orbits. Then $A = \lim_{n \to \infty} A_n$ for some sequence of Γ_n. Let $\mathbf{x}_n \in \Gamma_n$; since $\Omega \cup \Gamma$ is compact, we may assume $\mathbf{x}_n \to \mathbf{x}$ and, by the previous part, $\mathbf{x}$ belongs to a closed orbit Γ_0. By using the standard argument with transversal at $\mathbf{x}$ we find that Γ_n get arbitrarily close to Γ_0 so A is the area enclosed by Γ_0. Since Γ_0 does not reduce to a point, $A > 0$. But then the region enclosed by Γ_0 contains neither equilibria nor closed orbit, which contradicts the Poincaré-Bendixon theorem. □

A difficult part in applying the Poincareé-Bendixon theorem is to find the *trapping region*; that is, the closed and bounded region which contains no equilibria and which trajectories cannot leave. Quite often this will be an annular region with the equilibrium (source) in the hole, so that the trajectories can enter through the inner boundary and the outer boundary chosen in such a way that the vector field there always points inside the region. We illustrate this in the following example:

Example 5.36. Show that the second order equation

$$z'' + (z^2 + 2(z')^2 - 1)z' + z = 0$$

has a nontrivial periodic solution. We start with writing this equation as the system

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

Clearly, $(0, 0)$ is an equilibrium and by linearization we see that this is an unstable equilibrium hence there is a chance that the flow in a small neighbourhood of the origin will be outward. Thus, let us write the equation in polar coordinates. We get

$$\frac{d}{dt}(x^2 + y^2) = 2xx' + 2yy' = 2y^2(1 - x^2 - 2y^2).$$

We observe that $1 - x^2 - 2y^2 > 0$ for $x^2 + y^2 < 1/2$ and $1 - x^2 - 2y^2 < 0$ for $x^2 + y^2 > 1$. Hence, any solution which starts in the annulus $1/2 < x^2 + y^2 < 1$ must stay there. Since the annulus does not contain an equilibrium, there must be a periodic orbit inside it.

Let us consider a more sophisticated example.

Example 5.37. Glycolysis is a fundamental biochemical process in which living cells obtain energy by breaking down sugar. In many cells, such as yeast cells, glycolysis can proceed in an oscillatory fashion.

A simple model of glycolysis presented by Sel'kov reads

$$\begin{aligned} x' &= -x + ay + x^2 y, \\ y' &= b - ay - x^2 y, \end{aligned} \tag{5.3.63}$$

where x and y are, respectively, concentrations of ADP (adenosine diposphate) and F6P (fructose-6-phosphate), and $a, b > 0$ are kinetic parameters. We shall show that under certain assumptions there are, indeed, periodic solutions to the system.

First we find nullclines *(that is, curves along which one or the other time derivative is zero). Note that an equilibria are the intersection points of nullclines. We obtain that $x' = 0$ along the curve $y = x/(a + x^2)$ and $y' = 0$ along $y = b/(a + x^2)$. The nullclines are sketched on Figure 5.3, along with some representative vectors.*

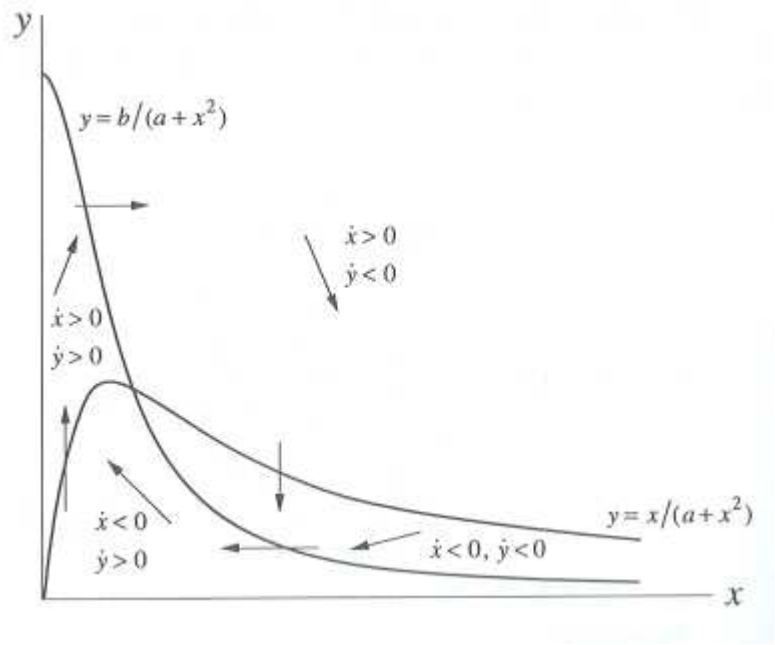


Fig. 5.3. Isoclines of (5.3.63)

Our first step is to construct a trapping region. First we observe that the trajectory cannot escape across any coordinate axis and also, since $y' < 0$ above both nullclines, no trajectory can cross a horizontal line in this region going upward. This leaves only the question of closing this region from the right. We note that $x' + y' = b - x < 0$ as long as $x > b$. Thus, in this region $y' < -x'$ and, since $x' > 0$, we have $y'/x' < -1$ which geometrically means that the slope is steeper than any straight line with slope -1. In other words, any trajectory in this region will cross a line with the slope -1 from above-right to below-left. Hence, if we draw such a line crossing the horizontal line $y = b/a$ at $x = b$, then trajectories from the left will not be able to escape through it. Finally, we continue the line till it intersects with the nullcline $y = x/(a + x^2)$ and note that below this nullcline $x' < 0$ so that we can close the region by a vertical segment to obtain a trapping region, see Fig 5.4: Can we conclude that there is a periodic trajectory? Well, no as there is an equilibrium $(b, b/(a + b^2))$ inside the region. However, this is not necessarily a bad news. If the equilibrium is a repeller (real parts of all eigenvalues are negative), then there is a neighbourhood V of the equilibrium such that any trajectory starting from V will eventually reach a sphere of a fixed radius (see the proof of point 2 of Theorem

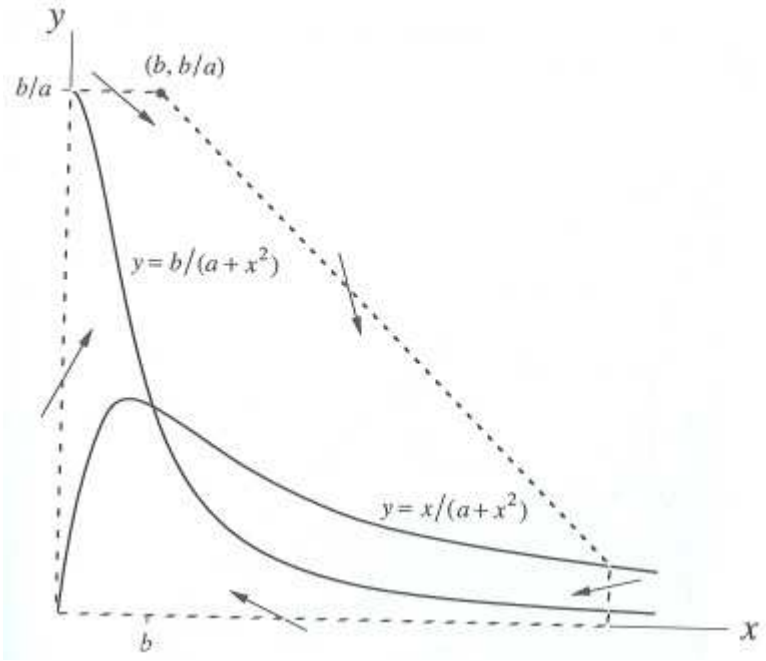


Fig. 5.4. Trapping region for (5.3.63).

5.13. Thus, the boundary of V can be taken as the inner boundary of the required set. So, we need to prove that $(b, b/(a+b^2))$ is a repeller. The Jacobi matrix at for (5.3.63) at (x_0, y_0) is

$$J = \begin{pmatrix} -1 + 2x_0y_0 & a + x_0 \\ -2x_0y_0 & -(a + x_0^2) \end{pmatrix}.$$

The Jacobian at the equilibrium is $a + b^2$ and the trace τ is

$$\tau = -\frac{b^2 + (2a-1)b^2 + (a + a^2)}{a + b^2}.$$

Hence, the equilibrium is unstable $\tau > 0$, and stable $\tau > 0$. The dividing line $\tau = 0$ occurs when

$$b^2 = \frac{1}{2}(1 - 2a \pm \sqrt{1 - 4a})$$

and this curve is shown on Fig. 5.6. For parameters in the region corresponding to $\tau > 0$ we are sure that there is a closed orbit of the system.

4 Other criteria for existence and non-existence of periodic orbit

When we were discussing the Liapunov function, we noted the cases such that if V was constant on trajectories, then the trajectories are closed so that the solutions are periodic. It turns out that this is a general situation for *conservative systems*. To be more precise, let us consider the system

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}). \tag{5.4.64}$$

We say that there is a *conservative quantity* for this system if there exists a scalar continuous function $E(\mathbf{x})$ which is constant along trajectories; that is $\frac{d}{dt}E(\phi(t, \mathbf{x}_0)) = 0$. To avoid trivial examples, we also require $E(\mathbf{x})$ to be nonconstant on every open set. We have the following result.

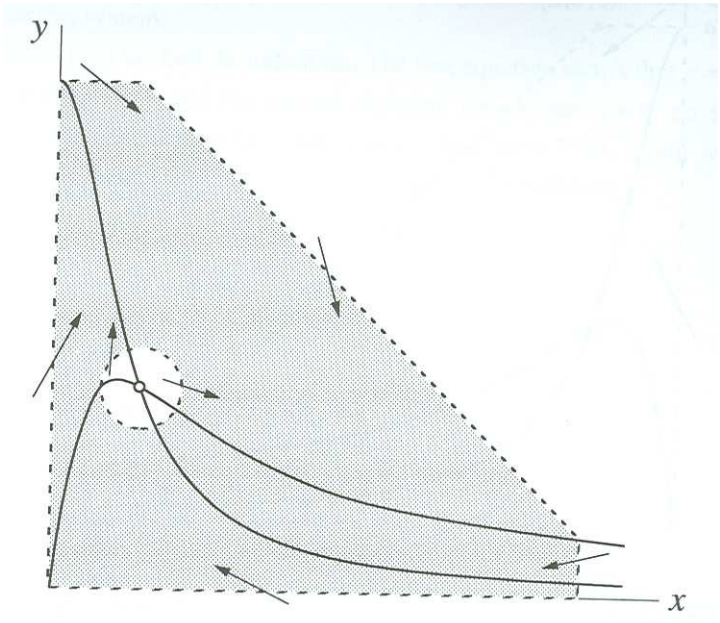


Fig. 5.5. Repelling character of equilibrium (5.3.63)

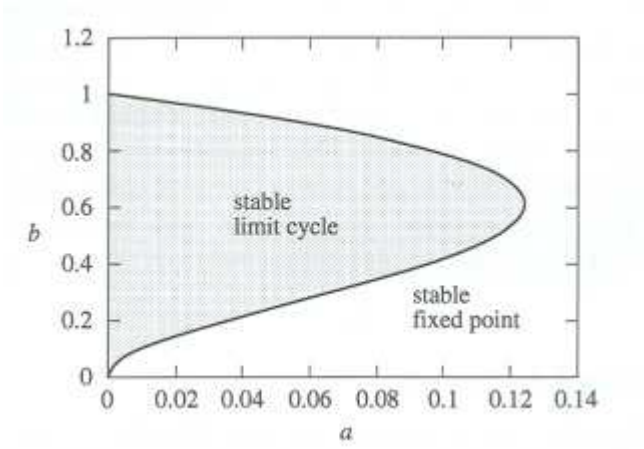


Fig. 5.6. Bifurcation curve for (5.3.63)

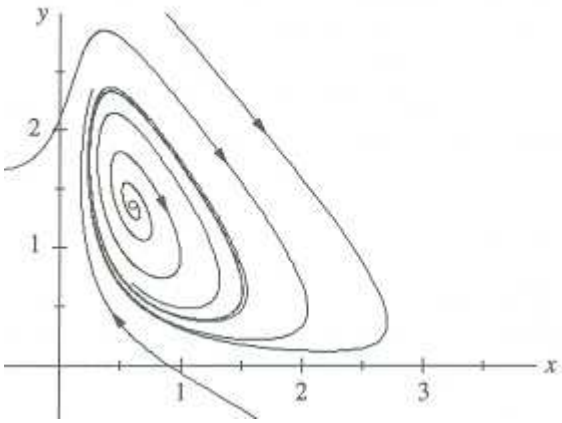


Fig. 5.7. Closed orbit for (5.3.63)

Proposition 5.38. *Assume that there is a conserved quantity of the system (5.4.64), where* $\mathbf{y} \in \mathbb{R}^2$*, and that* $\mathbf{y}^*$ *is an isolated fixed point of (5.4.64). If* $\mathbf{y}^*$ *is a local minimum of* E*, then there is a neighbourhood of* $\mathbf{y}^*$ *in which all trajectories are closed.*

Proof. Let $B_\delta = B(\mathbf{y}^*, \delta)$ be a closed neighbourhood of $\mathbf{y}^*$ in which $\mathbf{y}^*$ is absolute minimum. We can assume $E(\mathbf{y}^*) = 0$. Since E is constant on trajectories, each trajectory is contained in a level set of E:

$$E_c = \{(y_1, y_2);\ E(y_1, y_2) = c\} \cap B_\delta.$$

E_c is a nonempty bounded closed set for all sufficiently small $0 < c \leq c_0$. We use Theorem 5.30. For each c there are two possibilities: either the trajectory stays in E_c in which case also its ω-limit set is in E_c and thus it is a closed orbit or the trajectory leaves E_c so that there is a point $\mathbf{y}$ on the trajectory with $E(\mathbf{y}) = c$ and $\mathbf{y} = \delta$.

If the second case happens for some $c_n \to 0$, then we have a sequence $\mathbf{y}_n$ with $\|\mathbf{y}_n\| = \delta$ for which $E(\mathbf{y}_n)c_n \to 0$. But then, by continuity of E we find a point $\mathbf{y}_0$ with $E(\mathbf{y}_0) = 0$, contradicting the assumption that $\mathbf{y}^*$ is an absolute minimum in B_δ. □

Another case is concerned with 'time reversible' systems. We say that the system

$$\begin{aligned} y_1' &= f_1(y_1, y_2), \\ y_2' &= f_2(y_1, y_2), \end{aligned} \tag{5.4.65}$$

is time reversible, if it is invariant under $t \to -t$ and $y \to -y$. If f is odd in y_2 and f_2 is even in y_2, then such system is 'time-reversible'.

Proposition 5.39. *Suppose* $\mathbf{y}^* = (0, 0)$ *is the center for the linearization of a reversible system (5.4.65). Then sufficiently close to the origin all trajectories are closed curves.*

Proof. Change coordinates we can write (5.4.65) as

$$\begin{aligned} x_1' &= -x_2 + \omega_1(x_1, x_2), \\ x_2' &= x_1 + \omega_2(x_1, x_2), \end{aligned} \tag{5.4.66}$$

Let A denotes the matrix of the linearization. Then

$$\mathbf{x}(t) = e^{tA}\mathbf{x}_0 + e^{tA}\int\limits_0^t e^{-sA}\omega(\mathbf{x}(s))ds,$$

where $\omega = (\omega_1, \omega_2)$. The trajectories of the linear part are circles traversed anticlockwise and so $\|e^{tA}\mathbf{x}_0\| = \|\mathbf{x}_0\|$ for any t. Hence

$$\|\mathbf{x}(t)\| \le \|\mathbf{x}_0\| + \int\limits_0^t \|\omega(\mathbf{x}(s))\|ds,$$

Our aim is to show that the trajectory of the nonlinear system follows closely the trajectory of the linear system and thus starting from $\mathbf{x}_0$ on the positive horizontal semiaxis, it will cross the negative positive semiaxis. First, let us fix $\xi > 1$ satisfying $\ln\xi < 1/\xi$. Next, from the properties of linearization, for any ϵ we find δ such that $\|\omega(\mathbf{x})\| \le \epsilon\|\mathbf{x}\|$ as long as $\|\mathbf{x}\| \le \delta$. Next, select $\mathbf{x}_0$ with $\|\mathbf{x}_0\| = \delta/\xi$. Then, by the Gronwall inequality

$$\|\mathbf{x}(t)\| \le \|\mathbf{x}_0\|e^{\epsilon t} \le \delta$$

as long as $t \le \ln\xi/\epsilon > 0$. Next, calculate the norm of the difference $\mathbf{z}(t)$ between the nonlinear and linear solution starting from $\mathbf{x}_0$ over the time interval $[0, \ln\xi/\epsilon]$. We obtain

$$\|z(t)\| \le \int\limits_0^t \|\omega(\mathbf{y}(s)\|ds \le \epsilon t\delta \le \delta\ln\xi < \delta/\xi < \delta.$$

Hence, $\mathbf{x}(t)$ stays in the annulus surrounding the circle of radius δ/ϵ with the inner radius $\delta(1/\xi - \ln\xi)$. In particular, taking ϵ sufficiently small, we can take $t = 3\pi/2$ in which case $\mathbf{x}(t)$ must be necessarily below the horizontal axis. Thus, there mast be a point $t' > 0$ with $x_1(t') < 0$ and $x_2(t') = 0$. Now, trajectories are invariant with respect to the change $t \to -t$ and $y \to -y$ and hence we can complete the constructed trajectory to a closed one.

Since δ can be taken arbitrarily small, we see that all trajectories close to $(0, 0)$ are closed. □

Sometimes it is equally important to show that there are no periodic orbits in certain regions. We have already seen one such criterion related to Lyapunov functions: if there is a strict Lyapunov function in certain region surrounding an equilibrium, then there are no periodic orbits in this region. Here we consider two more 'negative' criteria.

Consider a system $\mathbf{y}' = \mathbf{f}(\mathbf{y})$ where $\mathbf{f} = -\nabla V$ in some region Ω for a scalar function V. Such systems are called *gradient systems* with *potential function* V. We have

Proposition 5.40. *Closed orbits are impossible in gradient systems with $V \neq const$.*

Proof. Suppose that there is a closed orbit Γ corresponding to a periodic solution with period T. Define ΔV to be the change of V along Γ. Clearly, $\Delta V = 0$ as the the orbit is closed and V is continuous. On the other hand

$$\Delta V = \int\limits_0^T \frac{dV}{dt}dt = \int\limits_0^T \nabla V \cdot \mathbf{y}'dt = -\int\limits_0^T \|\mathbf{y}'\|^2 dt < 0$$

as the trajectory is not an equilibrium due to $V \neq const$. This contradiction proves the statement. □

To illustrate this result, consider the system

$$\begin{aligned} y_1' &= \sin y_2, \\ y_2' &= y_1 \cos y_2. \end{aligned}$$

This is a gradient system with $V(y_1, y_2) = -y_1 \sin y_2$, so there are no periodic solutions.

Note. The above criterion works in arbitrary dimension.

Next we consider the so-called *Dulac criterion.*

Proposition 5.41. *Consider a planar system* $\mathbf{y}' = \mathbf{f}(\mathbf{y})$ *with continuously differentiable* $\mathbf{f}$*. Suppose that there is a scalar differentiable function* $g : \mathbb{R}^2 \to \mathbb{R}$ *such that* $\operatorname{div}(g\mathbf{f}) \neq 0$ *on some simply connected domain* Ω*. Then there are no periodic orbits of this system which lie in* Ω*.*

Proof. Suppose $\Gamma \subset \Omega$ is a periodic orbit. Using Green's theorem and the assumption

$$0 \neq \int\int_A \operatorname{div}(g\mathbf{f})dx_1dx_2 = \oint_\Gamma g\mathbf{f}\mathbf{n}d\sigma$$

where $\mathbf{n}$ is the normal to Γ and $d\sigma$ the line element along Γ. However, since $\mathbf{n}$ is normal to Γ, we have $\mathbf{n} \cdot \mathbf{f} = 0$ as $\mathbf{f}$ is tangent to any trajectory. Thus the left hand side is zero and we obtain a contradiction. □

We can illustrate this criterion on a variant of Lotka-Volterra model

$$\begin{aligned} y_1' &= y_1(A - a_1y_1 + b_1y_2), \\ y_2' &= y_2(B - a_2y_2 + b_2y_1). \end{aligned}$$

with $a_i > 0$ (to model the effect of overcrowding. Using the function $g = 1/y_1y_2$ we can prove that there are no periodic orbits in the first quadrant.

5 Stability through the Lyapunov function

Consider again the system (5.2.58) in $\mathbb{R}^n$. Suppose that it has an isolated equilibrium at $\mathbf{x}_0$. Then, by writing (5.2.58) as

$$\mathbf{y}' = (\mathbf{y} - \mathbf{x}_0)' = \mathbf{x}' = \mathbf{f}(\mathbf{x} + \mathbf{x}_0) = \tilde{\mathbf{f}}(\mathbf{x}),$$

we obtain an equivalent system for which $\mathbf{x} = 0$ becomes an isolated equilibrium. Thus there is no loss of generality to consider (5.2.58) with $\mathbf{x} = 0$ as its equilibrium.

Let Ω be an open neighbourhood of 0 and let $V : \Omega \to \mathbf{R}$ be a continuously differentiable function. We define the derivative of V along trajectories of (5.2.58) by the chain rule

$$V' = \frac{dV}{dt} = \mathbf{x}' \cdot \nabla V = \mathbf{f} \cdot \nabla V = \sum_{i=1}^{n} f_i \frac{\partial V}{\partial x_i} \tag{5.5.67}$$

Example 5.42. In this example we provide another point of view at the result formulated in Proposition 5.40. Let us consider a system

$$\mathbf{y}' = \mathbf{f}(\mathbf{y})$$

with $\mathbf{f}$ being a potential field; that is, there is a scalar function V satisfying

$$\mathbf{f}(\mathbf{x}) = -\operatorname{grad}V(x).$$

In general, not every vector field has a potential. An exception is offered by one dimensional fields when we have

$$V(x) = -\int_a^x f(z)dz$$

for some fixed a (the potential of a field is determined up to a constant). We note that since $dV/dx = -f$, the stationary points of V correspond to equilibria of f. Furthermore, if $t \to x(t)$ is any solution of the equation $x' = f(x)$ then we have

$$V'(x(t)) = \frac{dV(x(t))}{dt} = \frac{dV}{dx}(x(t))\frac{dx}{dt} = -f(x(t))f(x(t)) < 0,$$

so that $V(x(t))$ strictly decreases along trajectories. In other words, the point $x(t)$ moves always in the direction of decreasing V and thus equilibria corresponding to minima of V are asymptotically stable and corresponding to maxima are unstable.

In this section we shall discuss a generalization of the above concept.

Definition 5.43. *A continuously differentiable function V on $\Omega \ni 0$ is called a Lyapunov function for (5.2.58) if*

1. *$V(0) = 0$ and $V(\mathbf{x}) > 0$ on Ω;*
2. *$V' \leq 0$ on Ω.*

Theorem 5.44. *Assume that there exists a Lyapunov function defined on a neighbourhood Ω of an equilibrium $\mathbf{x} = 0$ of system (5.2.58). Then the solutions originating from Ω are globally defined (for all $t \geq 0$) and the equilibrium $\mathbf{x} = 0$ is stable.*

Proof. There is a ball $B(0, r) \subset \Omega$ (centred at 0 with radius r) such that $0 < V(\mathbf{x})$ on $B(0, r) \setminus 0$ and $V' \geq 0$ on $B(0, r)$. Let us take $0 \neq \mathbf{x}_0 \in B(0, r)$ and consider the flow $\phi(t, \mathbf{x}_0)$. Let $[0, t_{max})$ be the maximal interval of existence of the trajectory. We do not know whether t_{max} is finite or not. Since V is decreasing along trajectories, we have

$$0 < V(\phi(t, \mathbf{x}_0)) \leq V(\mathbf{x}_0), \quad t \in [0, t_{max}),$$

where the left-hand side inequality follows from the fact that $\phi(t, \mathbf{x}_0) \neq 0$ (by Theorem 5.18(i)) and strict positivity of V away from 0). Let $\mu = \min_{\|\mathbf{y}\|=r} V(\mathbf{y})$. Since $V(\mathbf{x}) \to 0$ as $\|\mathbf{x}\| \to 0$, we can find ball $B(0, \delta)$ with $\delta < r$ such that $V(\mathbf{x}) < \mu$ for $\mathbf{x} \in B(0, \delta)$. Then, for $\|\mathbf{y}_0\| < \delta$ we have

$$0 < V(\phi(t, \mathbf{y}_0)) \leq V(\mathbf{y}_0) < \mu, \quad t \in [0, t_{max}),$$

(with t_{max} is not necessarily the same as above). By the definition of μ and continuity of the flow, $\|\phi(t, \mathbf{y}_0)\| \leq r$ for $[0, t_{max})$. Indeed, otherwise there would be $t' > 0$ with $\|\phi(t', \mathbf{y}_0)\| > r$ and, by continuity, for some t'' we would have $\|\phi(t'', \mathbf{y}_0)\| = r$ so that $V(\phi(t'', \mathbf{y}_0)) \geq \mu$.

This means that the maximal interval of existence is infinite and, at the same time, yields stability, as r was arbitrary. ∎

Example 5.45. Consider the equation

$$u'' + g(u) = 0$$

where g is a continuously differentiable function for $|u| < k$, with some constant $k > 0$, and $ug(u) > 0$ for $u \neq 0$. Thus, by continuity, $g(0) = 0$. Particular examples include $g(u) = \omega^2 u$ which gives harmonic oscillator of frequency ω, or $g(u) = \sin u$: the undamped simple pendulum. Writing the equation as a system, we get

$$\begin{aligned} x_1' &= x_2, \\ x_2' &= -g(x_1). \end{aligned} \tag{5.5.68}$$

It is clear that $(0, 0)$ is an isolated equilibrium point. To construct Lyapunov function we employ mechanical interpretation of the model to find the energy of the system. If we think of g as the restoring force of a spring, the potential energy of the particle at a displacement $u = x_1$ from equilibrium is given by

$$\int_0^{x_1} g(\sigma)d\sigma.$$

On the other hand, the kinetic energy is

$$\frac{1}{2}x_2^2$$

as $x_2 = u'$ *which is the velocity of the particle. This suggests to take the total energy of the system as a Lyapunov function*

$$V(x_1, x_2) = \frac{1}{2}x_2^2 + \int_0^{x_1} g(\sigma)d\sigma.$$

This function is defined on the region

$$\Omega = \{(x_1, x_2);\ |x_1| < k, x_2 \in \mathbb{R}\}.$$

Clearly, V *is positive definite on* Ω*. Let us calculate the derivative of* V *along trajectories. We have*

$$V'(x_1, x_2) = x_2x_2' + g(x_1)x_1' = -g(x_1)x_2 + g(x_1)x_2 = 0.$$

Thus, V *is a Lyapunov function for (5.5.68) and the equilibrium at* $(0, 0)$ *is stable.*

Actually, we have proved more. For any $\mathbf{x}_0 = (x_{1,0}, x_{2,0}) \in \Omega$*, we obtain*

$$V(\phi(t, \mathbf{x}_0)) = V(\mathbf{x}_0)$$

for any t*. Thus, the orbits are given by implicit equation*

$$\frac{1}{2}x_2^2 + \int_0^{x_1} g(\sigma)d\sigma = V(\mathbf{x}_0).$$

Because of the hypotheses on g *the integral is positive for both* $x_1 > 0$ *and* $x_1 < 0$*; moreover it is an increasing function of* $|x_1|$*, which is zero at* $x_1 = 0$*. On the other hand,* $V(\mathbf{x}_0) \to 0$ *as* $\|\mathbf{x}_0\| \to 0$*. This means that, for sufficiently small* $\|\mathbf{x}_0\|$ *(*$V(\mathbf{x}_0) < \sup_{|x_1|<k} \int_0^{x_1} g(\sigma)d\sigma$*) the orbits are closed orbits symmetric with respect to the* y_2 *axis and thus solutions are periodic. Hence,* $(0, 0)$ *is stable but not asymptotically stable.*

It is rare to be able to find a Lyapunov function in one go. For left hand side of polynomial type, the method of undetermined coefficients is often employed.

Example 5.46. Consider the system

$$\begin{aligned} x_1' &= x_2, \\ x_2' &= -cx_2 - ax_1 - bx_1^3, \end{aligned} \tag{5.5.69}$$

where a, b, c *are positive constants. We are looking for a Lyapunov function as a polynomial in two variables. Let us try*

$$V(x_1, x_2) = \alpha x_1^2 + \beta x_1^4 + \gamma x_1^3,$$

with $\alpha, \beta, \gamma > 0$*. Clearly,* $V(\mathbf{x}) > 0$ *for* $\mathbf{x} \neq 0$*. Differentiating* V *along trajectories, we have*

$$\begin{aligned} V'(\mathbf{x}) &= (2\alpha x_1 + 4\beta x_1^3)x_1' + 2\gamma x_2 x_2' \\ &= (2\alpha x_1 + 4\beta x_1^3)x_2 + 2\gamma x_2(-cx_2 - ax_1 - bx_1^3) = (2\alpha - 2\gamma a)x_1x_2 + (4\beta - 2\gamma b)x_1^3x_2 - 2\gamma cx_2^2. \end{aligned}$$

Since $c, \gamma > 0$ *the last term is non-positive. The first two terms are more difficult, but we have freedom to chose free parameters* α *and* β*. Fixing* $\gamma > 0$ *and setting*

$$\alpha = a\gamma, \quad \beta = \frac{\gamma b}{2} = \frac{\alpha b}{2\alpha}$$

we obtain

$$V'(\mathbf{x}) = -2\gamma cx_2^2 \leq 0.$$

Hence $V(\mathbf{x})$ *is a Lyapunov function on any open bounded set of* $\mathbb{R}^2$ *which contains* $(0, 0)$ *and hence* $(0, 0)$ *is a stable equilibrium point.*

The first Liapunov theorem, Theorem 5.47, ensures stability but not asymptotic stability. From Example 5.45 we see that a possible problem is created by trajectories along which V is constant as these could give rise to periodic orbits which prevent asymptotic stability. The next theorem shows that indeed, preventing the possibility of V being constant in a neighborhood of zero solves the problem, at least partially.

Theorem 5.47. *Assume that there exists a Lyapunov function defined on a neighbourhood Ω of an equilibrium $\mathbf{x} = 0$ of system (5.2.58), which additionally satisfies*

$$V' < 0 \quad in \quad \Omega \setminus \{0\}. \tag{5.5.70}$$

Then $\mathbf{x} = 0$ is asymptotically stable.

Definition 5.48. *A Lyapunov function satisfying assumptions of this theorem is called* strict Lyapunov function.

Proof. Using the notation of the previous proof, we consider $\phi(t, \mathbf{y}_0)$ with $\|\mathbf{y}_0\| < \delta$ so that the solution stays in the closed ball $\overline{B(0,r)}$. Since this set is compact, we have sequence $(t_n)_{n\in\mathbb{N}}$, $t_n \to \infty$ as $n \to \infty$ such that $\phi(t_n, \mathbf{y}_0) \to \mathbf{z} \in \overline{B(0,r)}$. We have to prove that $\mathbf{z} = 0$. To this end, first observe that $V(\phi(t, \mathbf{y}_0)) > V(\mathbf{z})$ for all t as V decreases along trajectories and $V(\phi(t_n, \mathbf{y}_0)) \to V(\mathbf{z})$ by continuity of V. If $\mathbf{z} \neq 0$ (and there are no other equilibria in $\overline{B(0,r)}$), we consider $\phi(t, \mathbf{z})$ which must satisfy $V(\phi(t, \mathbf{z}) < V(\mathbf{z})$. By continuity of the flow with respect to the initial condition and continuity of V, if $\mathbf{x}$ is close enough to $\mathbf{z}$, then for some $t > 0$ (not necessarily all) we will have also $V(\phi(t, \mathbf{x})) < V(\mathbf{z})$. We take $\mathbf{x} = \phi(t_n, \mathbf{y}_0)$ for t_n large enough. But then we obtain

$$V(\mathbf{z}) > V(\phi(t, \phi(t_n, \mathbf{y_0}))) = V(\phi(t + t_n, \mathbf{y_0})) > V(\mathbf{z})$$

which is a contradiction. Thus $\mathbf{z} = 0$. This shows asymptotic stability. Indeed, if there was a sequence $(t_n)_{n\in\mathbb{N}}$ converging to infinity, for which $\phi(t, \mathbf{y}_0)$ was not converging to zero (that is, staying outside some ball $B(0, r_0)$), then as above we could pick a subsequence of $\phi(t_n, \mathbf{y}_0)$ converging to some $\mathbf{z}$ which, by the above, must be 0. ∎

Example 5.49. Consider the system

$$\begin{aligned} x_1' &= -x_1 + x_1^2 - 2x_1x_2, \\ x_2' &= -2x_2 - 5x_1x_2 + x_2^2, \end{aligned} \tag{5.5.71}$$

The point $(0,0)$ clearly is a stationary point. Let us investigate its stability. We try the simplest Lyapunov function

$$V(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2).$$

We obtain

$$\begin{aligned} V'(x_1, x_2) &= x_1x_1' + x_2x_2' = x_1(-x_1 + x_1^2 - 2x_1x_2) + x_2(-2x_2 - 5x_1x_2 + x_2^2) \\ &= -x_1^2(1 - x_1 + 2x_2) - x_2^2(2 + 5x_1 - x_2). \end{aligned}$$

Hence V is a strict Lyapunov function provided $2+5x_1-x_2 > 0$ and $1-x_1+2x_2 > 0$ in some neighbourhood of $(0,0)$. We see that this set, say Ω, is a sector containing $(0,0)$. Hence, the origin is asymptotically stable.

Let us consider another example which, in conjunction with the previous one, provide background for a refinement of the theory developed so far.

Example 5.50. Consider a more realistic version of the nonlinear oscillator equation, which includes resistance of the medium proportional to the velocity.

$$u'' + u' + g(u) = 0.$$

This equation is called the Liénard equation. We adopt the same assumptions as before: g is a continuously differentiable function for $|u| < k$, with some constant $k > 0$, and $ug(u) > 0$ for $u \neq 0$. Thus, by continuity, $g(0) = 0$. Writing the equation as a system, we get

$$\begin{aligned} x_1' &= x_2, \\ x_2' &= -g(x_1) - x_2. \end{aligned} \tag{5.5.72}$$

Again, $(0,0)$ is an isolated equilibrium point. Since this equation differs from the previously considered one by a dissipative term, the total energy of the system is a good initial guess for a Lyapunov function. Hence, we take

$$V(x_1, x_2) = \frac{1}{2}x_2^2 + \int_0^{x_1} g(\sigma)d\sigma.$$

This function is defined and positive on the region

$$\Omega = \{(x_1, x_2);\ |x_1| < k, x_2 \in \mathbb{R}\},$$

with the derivative along trajectories given by

$$V'(x_1, x_2) = -x_2^2 + x_2x_2' + g(x_1)x_1' = -x_2^2 - g(x_1)x_2 + g(x_1)x_2 = -x_2^2.$$

Thus, again V is a Lyapunov function for (5.5.72) and the equilibrium at $(0,0)$ is stable. However, it fails to be a strict Lyapunov function as there is no neighbourhood of $(0,0)$ on which V' is strictly positive. Now, if we look closer at this example, we see that we should be able to prove something more. Namely, if we can ensure that a trajectory stays away from $L = \{(x_1, x_2);\ x_2 = 0, x_1 \neq 0\}$ then, following the proof of Theorem 5.47, we obtain that it must converge to $(0,0)$. On the other hand, at any point of L the vector field is transversal to L so the trajectory cannot stay on L as then the field would have to be tangent to L. Thus, it is to be expected that the trajectory must eventually reach $(0,0)$. We shall provide a rigorous and more general result of this type below.

Theorem 5.51. *(La Salle invariance principle) Let $\mathbf{y} = 0$ be a stationary point of (5.2.58) and let V be a Lyapunov function on some neighbourhood $\Omega \ni 0$. If, for $\mathbf{x} \in \Omega$, $\Gamma_{\mathbf{x}}^+$ is bounded with limit points in Ω and M is the largest invariant set of*

$$E = \{\mathbf{x} \in \Omega;\ V'(\mathbf{x}) = 0\}, \tag{5.5.73}$$

then

$$\phi(t, \mathbf{x}) \to M, \quad t \to \infty. \tag{5.5.74}$$

Proof. By assumptions, for any $\mathbf{x} \in \Omega$ satisfying the assumption of the theorem, $\emptyset \neq \omega(\Gamma_{\mathbf{x}}) \subset \Omega$. Since V is a Lyapunov function, $V(\phi(t, \mathbf{x}))$ in a non-increasing function of t which is bounded below by zero. Hence, there is $c \geq 0$ such that

$$\lim_{t\to\infty} V(\phi(t, \mathbf{x})) = c.$$

Now let $\mathbf{y} \in \omega(\Gamma_{\mathbf{x}})$. Then, for some $(t_n)_{n\in\mathbb{N}}$ converging to ∞ as $n \to \infty$ we have $\phi(t_n, \mathbf{x}) \to \mathbf{y}$. On the other hand, by continuity of V, we have

$$\lim_{t\to\infty} V(\phi(t_n, \mathbf{x})) = c.$$

Consequently, V is constant on $\omega(\Gamma_{\mathbf{x}})$.

Now, by Lemma 5.25(3), $\omega(\Gamma_{\mathbf{x}})$ is invariant so that if $\mathbf{y} \in \omega(\Gamma_{\mathbf{x}})$, then $\phi(t, \mathbf{y}) \in \omega(\Gamma_{\mathbf{x}})$ for all t. Thus, $V(\phi(t, \mathbf{y})) = c$ for all t and $\mathbf{y} \in \omega(\Gamma_{\mathbf{x}})$. Thus,

$$V'(\mathbf{y}) = 0, \quad for\ \mathbf{y} \in \omega(\Gamma_{\mathbf{x}})$$

and so

$$\omega(\Gamma_{\mathbf{x}}) \subset M \subset E.$$

But $\phi(t, \mathbf{x}) \to \omega(\Gamma_{\mathbf{x}})$ as $t \to \infty$, so $\phi(t, \mathbf{z}) \to M$ as $t \to \infty$ for all $\mathbf{z} \in \Omega$ satisfying $\omega(\Gamma_{\mathbf{z}}) \subset \Omega$. ■

Remark 5.52. A class of sets in Ω with forward trajectories in Ω is given by

$$V_k = \{\mathbf{x};\ V(\mathbf{x}) < k\}.$$

Indeed, since V is non-increasing along trajectories, $V(\phi(t,\mathbf{x})) \le V(\mathbf{x}) < k$ provided $\mathbf{x} \in V_k$.

Remark 5.53. La Salle principle gives immediate proof of Theorem 5.47. Indeed, in the domain of applicability of this theorem, $M = E = \{0\}$.

Corollary 5.54. *Assume that there is a Lyapunov function for (5.2.58) defined on the whole $\mathbb{R}^n$ which satisfies additionally $V(\mathbf{y}) \to \infty$ as $\|\mathbf{y}\| \to \infty$. If 0 is the only invariant set of $E = \{\mathbf{x} \in \mathbb{R}^n;\ V'(\mathbf{x}) = 0\}$, then 0 is globally asymptotically stable.*

Proof. From the properties of the Lyapunov function, we have

$$V(\phi(t,\mathbf{y})) \le V(\mathbf{y}), \quad \mathbf{y} \in \mathbb{R}^n,$$

independently of t. From the assumption on the behaviour of V at infinity, $\phi(t,\mathbf{y})$ must stay bounded and thus exist for all t. But then the limit points must belong to the set $\Omega = \mathbb{R}^n$ and the La Salle principle can be applied. ■

Example 5.55. Consider the so-called van der Pol equation

$$z'' - az'(z^2 - 1) + z = 0, \tag{5.5.75}$$

where $a > 0$ is a constant. In this case it is easier to work with the so-called Lièнard coordinates which are applicable to any equation of the form

$$x'' + f(x)x' + g(x) = 0.$$

Let us define $F(x) = \int^x f(\xi)d\xi$. Then

$$dF/dt = f(x)x'.$$

Hence, if we define $x_1 = x$ and $x_2 = x_1' + F(x_1)$, then $x_2' = x_1'' + f(x_1)x_1' = -g(x_1)$. The differential equation can then be written as

$$\begin{aligned} x_1' &= x_2 - F(x_1), \\ x_2' &= -g(x_1). \end{aligned}$$

In our case, we obtain

$$\begin{aligned} x_1' &= x_2 + a\left(\frac{1}{3}x_1^3 - x_1\right), \\ x_2' &= -x_1. \end{aligned}$$

Let us use the standard Lyapunov function

$$V(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2).$$

With this choice we get

$$V'(x_1, x_2) = x_1\left(x_2 + a\left(\frac{1}{3}x_1^3 - x_1\right)\right) - x_1x_2 = ax_1^2\left(\frac{1}{3}x_1^2 - 1\right).$$

Thus, $V' \le 0$ for $x_1^2 < 3$. The largest domain of the form

$$V_k = \{(x_1, x_2);\ V(x_1, x_2) < k\}$$

which lies entirely in the region $\{(x_1, x_2);\ V' \le 0\}$ is given by $V_{3/2}$. Furthermore, $V' = 0$ on $x_1 = 0$ and on this line $x_1' = x_2$. Hence, the trajectories will stay on this line only if the x_1-coordinate of the tangent is zero; that is, only of $x_1' = x_2 = 0$. Thus, the largest invariant subset of $V_{3/2} \cap \{(x_1, x_2);\ V' = 0\}$ is $(0,0)$. Thus, by the La Salle principle, $(0,0)$ is asymptotically stable and $V_{3/2}$ is a basin of attraction.

Example 5.56. Consider the equation

$$z'' + 2az' + z + z^3 = 0, \tag{5.5.76}$$

where a is a constant satisfying $0 < a < 1$. Equivalent system is given by

$$\begin{aligned} x_1' &= x_2, \\ x_2' &= -x_1 - 2ax_2 - x_1^3. \end{aligned} \tag{5.5.77}$$

*The origin $(0,0)$ is the only equilibrium. If $a = 0$, then (5.5.76) is the same as in the Example **??** and thus we know that*

$$V(x_1, x_2) = \frac{x_2^2}{2} + \frac{x_1^2}{2} + \frac{x_1^4}{4}$$

is the first integral (energy) and $V' = 0$ on the trajectories. The addition of $2az'$ makes the system dissipative and thus it is reasonable to take V as the trial Lyapunov function for (5.5.77). We get

$$V'(x_1, x_2) = -2ax_2^2 \leq 0$$

for any (x_1, x_2). Let us apply Corollary 5.54. It is clear that $V(x_1, x_2) \to \infty$ as $\|(x_1, x_2)\| \to \infty$. Furthermore, $V' = 0$ on $x_2 = 0$. We have to find the largest invariant subset of this set. To stay on $x_2 = 0$ we must have $x_2' = 0$. But, if $x_2 = 0$, then $x_1' = 0$ hence $x_1 =$ constant. This, from the second equation of (5.5.77) we obtain $x_1 = 0$. Consequently, $(0,0)$ is the largest invariant subset of $\{(x_1, x_2);\ V' = 0\}$ and thus $(0,0)$ is globally asymptotically stable.

Example 5.57. **Stability by linearization.** W shall give a proof of Theorem 5.13 1 using the Lyapunov function method. Consider again

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}). \tag{5.5.78}$$

If $\mathbf{f}$ has continuous partial derivatives of the first order in some neighbourhood of $\mathbf{y}^0$, then

$$\mathbf{f}(\mathbf{x} + \mathbf{y}^0) = \mathbf{f}(\mathbf{y}^0) + \mathcal{A}\mathbf{x} + \mathbf{g}(\mathbf{x}) \tag{5.5.79}$$

where

$$\mathcal{A} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{y^0}) & \dots & \frac{\partial f_1}{\partial x_n}(\mathbf{y^0}) \\ \vdots & & \vdots \\ \frac{\partial f_1}{\partial x_n}(\mathbf{y^0}) & \dots & \frac{\partial f_n}{\partial x_n}(\mathbf{y^0}) \end{pmatrix},$$

and $\mathbf{g}(\mathbf{x})/\|\mathbf{x}\|$ is continuous in some neighbourhood of $\mathbf{y}^0$ and vanishes at $\mathbf{x} = \mathbf{y}^0$. This follows from the Taylor expansion theorem. Note that if $\mathbf{y}^0$ is an equilibrium of (5.5.78), then $f(\mathbf{y}^0) = 0$ and we can write

$$\mathbf{x}' = \mathcal{A}\mathbf{x} + o(\|\mathbf{x}\|). \tag{5.5.80}$$

where $\mathbf{x} = \mathbf{y} - \mathbf{y}^0$.

Theorem 5.58. *Suppose that $\mathbf{f}$ is differentiable function in some neighbourhood of the equilibrium point $\mathbf{y}^0$. Then, the equilibrium point $\mathbf{y}^0$ is asymptotically stable if all the eigenvalues of the matrix $\mathcal{A}$ have negative real parts, that is, if the equilibrium solution $\mathbf{x}(t) = \mathbf{0}$ of the linearized system is asymptotically stable.*

Proof. We shall give the proof for the case when $\mathcal{A}$ has distinct eigenvalues. The general case also can be proved using Lyapunov function method but it is much more involved.

Let $\{\lambda_1, \dots, \lambda_n\}$ be distinct eigenvalues of $\mathcal{A}$ with $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ being the corresponding eigenvalues (since the eigenvalues are distinct, they must be simple so that to each there corresponds exactly one eigenvector). Now, denoting by $< \cdot, \cdot >$ the dot product in $\mathbb{C}^n$

$$< \mathbf{x}, \mathbf{y} >= \sum_{i=1}^{n} x_i y_i,$$

we have

$$< \mathbf{y}, \mathcal{A}\mathbf{x} >=< \mathcal{A}^T\mathbf{y}, \mathbf{x} >$$

where A^T is the transpose of $A = \{a_{ij}\}_{1\le i,j\le n}$, $A^T\{a_{ji}\}_{1\le i,j\le n}$ (thanks to the fact that entries of A are real). A^T has the same eigenvalues as A; denote by $\{\mathbf{f}_1, \dots, \mathbf{f}_n\}$ the corresponding eigenvectors. It is easy to see that $< \mathbf{f}_j, \mathbf{e}_i >= 0$ provided $j \neq i$. Indeed,

$$\lambda_i < \mathbf{f}_i, \mathbf{e}_j >=< A^T\mathbf{f}_i, \mathbf{e}_j >=< \mathbf{f}_i, A\mathbf{e}_j >= \lambda_j < \mathbf{f}_i, \mathbf{e}_j >$$

so that $(\lambda_i - \lambda_j) < \mathbf{f}_i, \mathbf{e}_j >= 0$ and the statement follows if $\lambda_i \neq \lambda_j$. It is then possible to normalize the eigenvectors so that

$$< \mathbf{f}_i, \mathbf{e}_j >= \delta_{ij},$$

(that is, 1 for $i = 1$ and 0 for $i \neq j$). We can expand any $\mathbf{x}$ as

$$\mathbf{x} = \sum_{i=1}^{n} < \mathbf{f}_i, \mathbf{x} > \mathbf{e}_i.$$

Then

$$\mathbf{x}' = \sum_{i=1}^{n} \frac{d}{dt} < \mathbf{f}_i, \mathbf{x} > \mathbf{e}_i,$$

and

$$A\mathbf{x} = \sum_{i=1}^{n} \lambda_i < \mathbf{f}_i, \mathbf{x} > \mathbf{e}_i$$

so that

$$\frac{d}{dt} < \mathbf{f}_i, \mathbf{x} >= \lambda_i < \mathbf{f}_i, \mathbf{x} > +o(\|\mathbf{x}\|)$$

(as multiplying $o(\|\mathbf{x}\|)$ by $\mathbf{e}$ does not change the asymptotic behaviour of the '$o(\cdot)$' symbol).

This allows to define a Lyapunov function. Let $\alpha_1, \dots, \alpha_n$ be positive numbers and put

$$V(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \overline{< \mathbf{f}_i, \mathbf{x} >} < \mathbf{f}_i, \mathbf{x} > .$$

This is clearly differentiable function which is positive for $\mathbf{x} \neq 0$. Differentiating along trajectories, we get

$$\begin{aligned} V'(\mathbf{x}) &= \sum_{i=1}^{n} \alpha_i \left(\frac{d}{dt} \overline{< \mathbf{f}_i, \mathbf{x} >} < \mathbf{f}_i, \mathbf{x} > \right) \\ &= + \sum_{i=1}^{n} \alpha_i \left(\left(\frac{d}{dt} \overline{< \mathbf{f}_i, \mathbf{x} >} \right) < \mathbf{f}_i, \mathbf{x} > + \overline{< \mathbf{f}_i, \mathbf{x} >} \left(\frac{d}{dt} < \mathbf{f}_i, \mathbf{x} > \right) \right) \\ &= \sum_{i=1}^{n} \alpha_i \left(\bar{\lambda_i} + \lambda_i \right) \overline{< \mathbf{f}_i, \mathbf{x} >} < \mathbf{f}_i, \mathbf{x} > +o(\|\mathbf{x}\|^2). \end{aligned}$$

Since $\bar{\lambda_i} + \lambda_i = 2\Re\lambda_i < 0$, the first term is negative of second order and the other term is of higher order than 2,and thus for sufficiently small $\|\mathbf{x}\|$ the derivative $V'(\mathbf{x})$ is strictly negative in some neighbourhood Ω of 0. Hence, 0 is asymptotically stable. Constants α_i can be changed to fine-tune the estimate a basin of attraction.

The SIS model

If the disease does not induce immunity but, instead, after recovery the infected individuals become again susceptible, then the SIR model turns into the SIS model

$$\begin{aligned} S' &= -\beta SI + \alpha I, \\ I' &= \beta SI - \alpha I, \end{aligned} \tag{5.5.81}$$

where α is the rate of recovery. Here, again, if we add the equations, we will find that the total population $N = S + I$ is constant in time. Thus, we can write

$$S = N - I$$

and thus (5.5.81) reduces to

$$I' = \beta I(N - I) - \alpha I = (\beta N - \alpha) I \left(1 - \frac{I}{\frac{\beta N - \alpha}{\beta}}\right) = rI\left(1 - \frac{I}{K}\right). \tag{5.5.82}$$

This is the logistic equation that was analysed earlier. In particular, we have the following cases

a) $r = \beta N - \alpha < 0$, or $\frac{\beta N}{\alpha} < 1$, then the solution has only one nonnegative equilibrium, 0, that is attractive. It can be easily seen as then $K < 0$ and thus

$$I' \leq rI;$$

that is

$$I(t) \leq I(0)e^{rt}.$$

Hence $I(t) \to 0$ faster than e^{rt} and thus the disease will die out.

b) if $r > 0$, then the properties of the logistic equation shows that

$$\lim_{t\to\infty} I(t) = K = \frac{\beta N - \alpha}{\beta}.$$

Hence, the disease will permanently stay in the population.

Remark 5.59. In any epidemiological model the equilibrium $I = 0$, that always exists, is called the disease free equilibrium. A positive equilibrium, if it exists, is called an endemic equilibrium.

Remark 5.60. In both models there is a parameter $\mathcal{R}_0$ that determines the progression of the disease: if $\mathcal{R}_0 < 1$, the disease will die out and if $\mathcal{R}_0 > 1$ it will spread. In the SIR model we have

$$\mathcal{R}_0 = \frac{\beta S(0)}{\nu}$$

while in the SIS model

$$\mathcal{R}_0 = \frac{\beta N}{\alpha}.$$

Seemingly these two constants are unrelated. However, let us look at their biological meaning. The coefficient β gives the number of infections per unit time induced by one infective whereas $1/\nu$ (respectively $1/\alpha$ is the average time an infective remains infectious. Finally, if we assume that consider a population that at time $t = 0$ had no infective individuals, then the number of susceptibles at the beginning is $S(0)$ in the first case and $N = N(0)$ in the second. Thus, we have arrived at the common interpretation of $\mathcal{R}_0$

Definition 5.61. *The basic reproduction number $\mathcal{R}_0$ is the number of infections that one infectious individual will introduce in a population consisting only of susceptible individuals.*

For such a defined parameter $\mathcal{R}_0$, the condition $\mathcal{R}_0 \lesseqgtr 1$ determines the stability of the disease free equilibrium. It is easy to understand from the biological point of view: if one infective is producing less than one secondary infection, then the disease cannot spread, whereas if it is more than one than the disease will spread at the geometrical rate.

SIS model with treatment

In many cases the return of an infective to the susceptible class is due to a treatment. In the simplest case we can assume that the constant α in (5.5.81) represents the efficacy of the treatment. A more realistic model takes into account that the treatment of a single patient takes some time and thus the rate of recovery should be rather modelled by the Holling type functional response. As before, let the number of treated infectives in time T by one nurse be given by

$$C = \nu\gamma I T_a,$$

where the constant γ is the rate at which the infectives are treated (number per unit time), ν is the efficacy of the treatment and T_a is the time available for administering the treatment. Since

$$T = T_a + \gamma I T_a T_t = T_a(1 + \gamma I T_t),$$

where T_t is the average time of treatment,

$$C = \frac{\nu\gamma}{1 + \gamma T_t I} I.$$

However, $\gamma = 1/T_t$, hence we obtain the SIS model with saturated treatment as

$$\begin{aligned} S' &= -\beta SI + \frac{\nu\gamma M}{1+I} I, \\ I' &= \beta SI - \frac{\nu\gamma M}{1+I} I, \end{aligned} \tag{5.5.83}$$

where M is the number of the available medical personnel. By defining $\alpha = \nu\gamma M$, we have

$$\begin{aligned} S' &= -\beta SI + \frac{\alpha}{1+I} I, \\ I' &= \beta SI - \frac{\alpha}{1+I} I. \end{aligned} \tag{5.5.84}$$

As in the previous subsection, $N(t) = S(t) + I(t) = N = S_0 + I_0$ is constant. We assume $N > 1$ (note that if N denoted the density, this assumption would not be obvious).

Hence, substituting $S(t) = N - I(t)$ we obtain the single equation

$$I'(t) = \beta I(N - I) - \frac{\alpha I}{1+I}. \tag{5.5.85}$$

Eqn (5.5.85) is in the form of the Allee model. It is a separable equation that, in principle, can be solved. This, however, one one hand would produce a messy and difficult to analyse formula and, on the other, would hide a general structure that can be utilised in cases when an explicit solution is not available.

We use general one dimensional 'phase-plane' analysis to find the properties of equilibria. Denote

$$F(I) = \beta I(N - I) - \frac{\alpha I}{1+I} = I\left(\beta(N - I) - \frac{\alpha}{1+I}\right) = If(I).$$

Clearly, $I = 0$ is an equilibrium so, in particular, any solution originating from $I(0) = I_0 > 0$ satisfies $I(t) > 0$. We see that

$$F'(I) = f(I) + If'(I) \tag{5.5.86}$$

and hence $F'(0) = f(0) = \beta N - \alpha$ we obtain that if $\beta N/\alpha > 1$, then $I = 0$ is a repelling equilibrium and if $\beta N/\alpha < 1$, it is an asymptotically stable equilibrium. In the expression

$$\mathcal{R}_0 = \frac{\beta N}{\alpha}$$

we recognize the basic reproduction number. Here it requires some explanation as the average duration of the disease is I dependent. However, the definition requires the basic reproduction number to be calculated

in a population consisting only of susceptible individuals; that is, whenever in calculations of $\mathcal{R}_0$ we have an I dependent term, we put $I = 0$.

To find stability for $N = \alpha/\beta$, we use the geometrical argument. In this case

$$F(I) = -\frac{\beta I^2}{1+I}(I + N - 1)$$

so $F(I) < 0$ for $I < 0$ (as $N > 1$). Hence, $I = 0$ is repelling.

Consider now endemic equilibria. These are the solutions to the quadratic equation

$$g(I) := (N - I)(1 + I) = \frac{\alpha}{\beta}. \tag{5.5.87}$$

The graph of g is the downward parabola with roots at $I = -1$ and $I = N$. The maximum of g is taken at

$$I_{\max} = \frac{N-1}{2}$$

and equals

$$g(I_{\max}) = \frac{(N+1)^2}{4}.$$

We observe that for $\mathcal{R}_0 \geq 1$ there is a unique positive solution to (5.5.87), see Fig. 5.8. If, however, $\mathcal{R}_0 < 1$, (5.5.87) may have two, one, or no solutions. The first case occurs if

$$N < \frac{\alpha}{\beta} < \frac{(N+1)^2}{4}, \tag{5.5.88}$$

see Fig. 5.9. Equivalently, in terms of $\mathcal{R}_0$,

$$\frac{4N}{(N+1)^2} < \mathcal{R}_0 < 1. \tag{5.5.89}$$

Then, if

$$\frac{\alpha}{\beta} = \frac{(N+1)^2}{4}, \tag{5.5.90}$$

then again we have one positive equilibrium and, finally, for

$$\frac{\alpha}{\beta} > \frac{(N+1)^2}{4} \tag{5.5.91}$$

there is no positive equilibrium, see Fig. 5.10.

To find the stability of the equilibria, we write

$$F(I) = \frac{\beta I}{1+I}\left((N-I)(1+I) - \frac{\alpha}{\beta}\right) = \frac{\beta I}{1+I}\left(g(I) - \frac{\alpha}{\beta}\right).$$

Let us denote by I_2^* the equilibrium larger than $I_{\max}$, by I_1^* the one smaller than $I_{\max}$ and by I^* the equilibrium equal to $I_{\max}$.

- $\mathcal{R}_0 < 1$ and $\alpha/\beta > (N+1)^2/4$. There is only the disease free equilibrium that, as above, is globally asymptotically stable, see Fig. 5.14.
- $\mathcal{R}_0 < 1$ and $\alpha/\beta = (N+1)^2/4$. There is a disease free equilibrium and an endemic equilibrium I^*. The disease free equilibrium is asymptotically stable, as above, but not globally asymptotically stable. The endemic equilibrium is unstable (precisely, semi-stable – it repels solutions smaller than I^* and attracts solutions bigger than I^*, see Fig. 5.13.

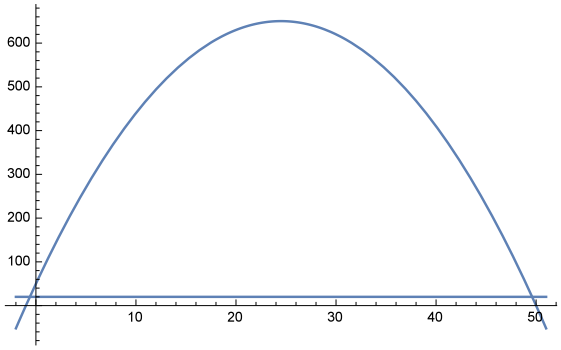


Fig. 5.8. The unique endemic equilibrium for $\mathcal{R}_0 > 1$ ($N = 50 > 1$ and $\alpha/\beta = 20$.)

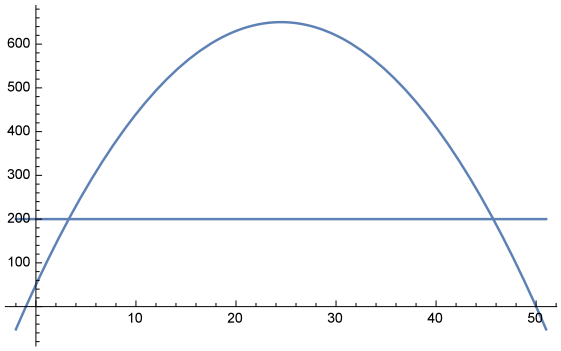


Fig. 5.9. Two endemic equilibria in the case (5.5.88).

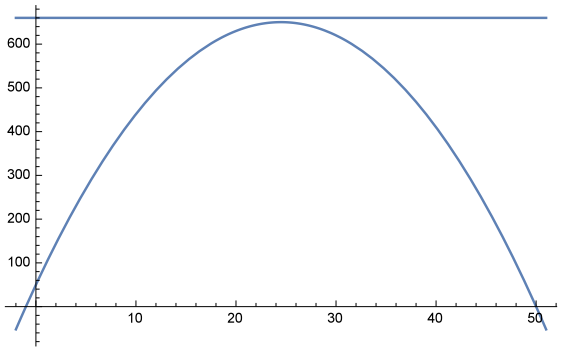


Fig. 5.10. No endemic equilibria in the case (5.5.91).

$\mathcal{R}_0 < 1$ and $\alpha/\beta < (N+1)^2/4$. There is a disease free equilibrium and two endemic equilibria I_1^*, I_2^*. The disease free equilibrium is asymptotically stable, as above, but not globally asymptotically stable. The endemic equilibrium I_1^* is unstable and I_2^* is asymptotically stable. Neither stable equilibrium is globally asymptotically stable: $I = 0$ attracts solutions in $[0, I_1^*)$ while I_2^* attracts solutions from (I_1^*, ∞). The intervals $[0, I_1^*)$ and (I_1^*, ∞) are called basins of attraction of respective equilibria, see Fig. 5.12.

$\mathcal{R}_0 \geq 1$. There is a disease free equilibrium and an endemic equilibrium I_2^*. The disease free equilibrium is unstable. The endemic equilibrium is asymptotically stable (and globally asymptotically stable in $(0, \infty)$), see Fig. 5.11.

Remark 5.62. It is a common (mis)perception that to control a disease if is sufficient to bring $\mathcal{R}_0$ below 1. We have seen that, indeed, the disease free equilibrium is asymptotically stable in this case but, nevertheless,

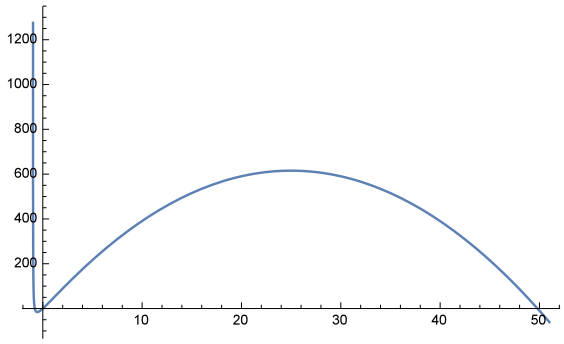


Fig. 5.11. The graph of $F(I)$ for $N = 50 > 1, \beta = 1, \alpha = 10, \mathcal{R}_0 = 5$

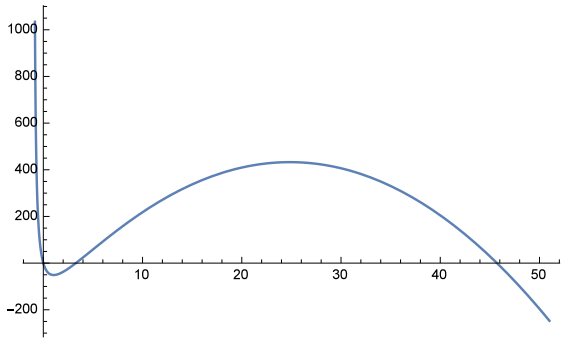


Fig. 5.12. The graph of $F(I)$ for $N = 50 > 1, \beta = 1, \alpha = 200, \mathcal{R}_0 = 0.25$

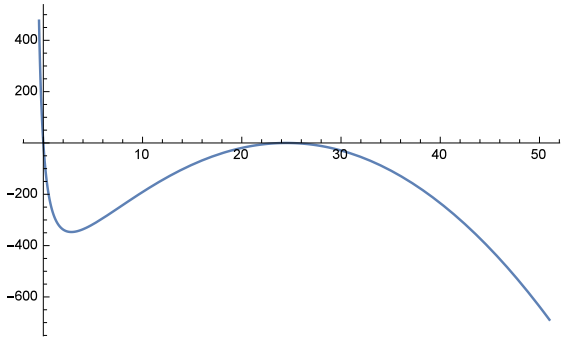


Fig. 5.13. The graph of $F(I)$ for $N = 50 > 1, \beta = 1, \alpha = 650.25 = (N+1)^2/4, \mathcal{R}_0 = 0.077$

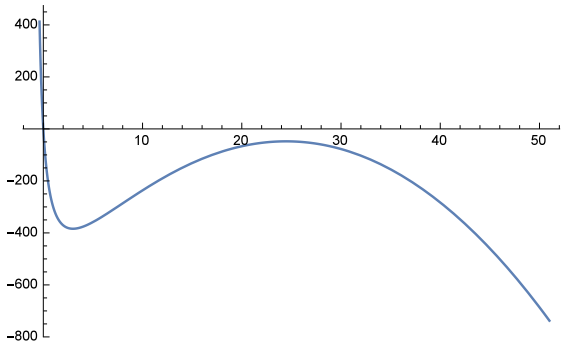


Fig. 5.14. The graph of $F(I)$ for $N = 50 > 1, \beta = 1, \alpha = 700, \mathcal{R}_0 = 0.0714$

the disease can persist – if the population of infectives is sufficiently large, then it will be attracted to the endemic equilibrium and the disease will not be eradicated. Only by bringing $\mathcal{R}_0$ down below $4N/(N+1)^2$ we will make the disease free equilibrium globally asymptotically stable and thus the disease will be eradicated.

Question: Assume that we have a disease that is spreading. We found the basic reproduction number

$$\mathcal{R}_0 = \frac{\beta N}{\alpha} = \frac{\beta N}{\nu\gamma M} = \frac{\beta N T_t}{\nu M} > 1.$$

What interventions can the health authorities undertake to stop the disease?