Models and Simulation

Lecture notes, WISB134 Modellen & Simulatie

January 11, 2016

JASON FRANK

Mathematical Insitute, Utrecht University

Utrecht University

Contents

Acknowledgment

The course *Modellen en Simulatie* has a long history at Utrecht University. These notes incorporate material from the original lecture notes of Frits Beukers and the lecture slides of Gerard Sleijpen. I have translated that material, compressed it in some areas and expanded it in others, in preparation for my lectures in 2014, 2015, and 2016. Furthermore, some of the material on numerical methods was prepared in cooperation with Ben Leimkuhler.

Chapter 1

Introduction to models and modelling

The word *dynamic* means "marked by usually continuous and productive activity or change"[1]. In this course we think about systems that change in time: *dynamical systems*. One of our main goals is to see how mathematics plays a universal role in other disciplines. Examples of dynamical systems include the motion of celestial bodies under the influence of gravity, the changing population levels of plants and animals while 'eating or being eaten', the variation of prices under economic influences, the changing concentrations levels of chemical species during a reaction, and many, many more.

We encounter three classes of mathematical models for dynamical systems: *iterated maps*, *differential equations*, and *numerical methods*. The classes are interrelated, as we will see later. In this chapter we introduce each type of model in the context of population growth.

1.1 An iterated map

Let us describe a simple model for the population of a single species in an environment with limited resources, such as a limited amount of food. The population will be determined at discrete times t_0, t_1, t_2, etc. Let $P_n \geq 0$ denote the population size at time t_n. If there were an unlimited food supply, we assume the population would grow by the uninhibited growth factor $\beta = P_{n+1}/P_n > 1$. The quantity β is assumed constant, a *parameter* of the model. Due to limitations, rather, the maximum population that can be sustained by the environment (the so-called *carrying capacity* of the environment) is $P_c > 0$, a second parameter. A model for the population growth that accounts for limited resources in the

[1] Merriam-Webster online dictionary.

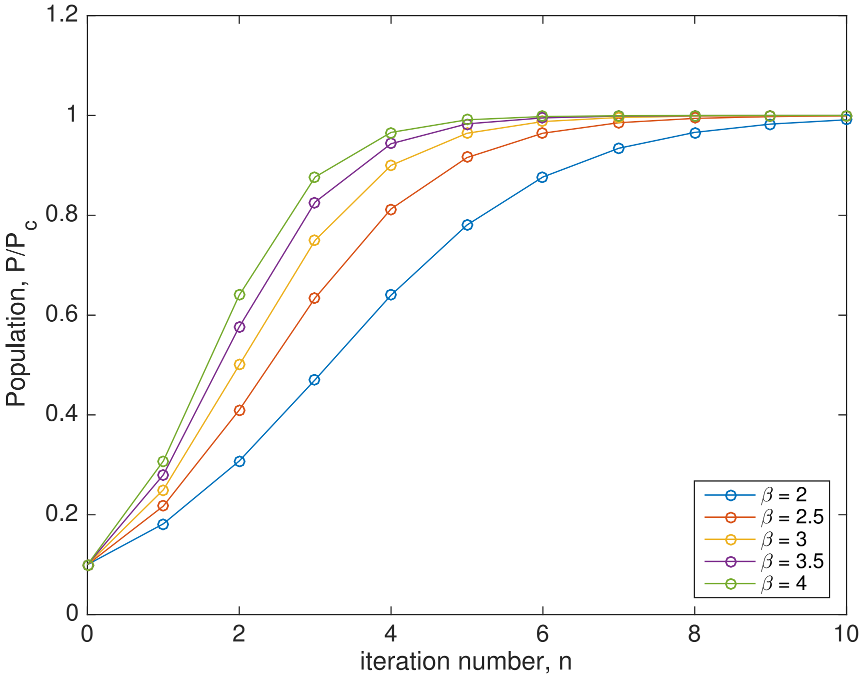

Figure 1.1: Iterates of the model (1.1) with $P_0/P_c = 0.1$ and several values of β. The iterates are connected by straight lines for visualization purposes. Independent of β, the populations eventually tend to the carrying capacity P_c.

environment is

$$P_{n+1} = \frac{\beta P_c P_n}{P_c + (\beta - 1)P_n}. \tag{1.1}$$

Given the initial population P_0 at t_0, the equation (1.1) tells us how to compute the population P_1 at time t_1, and from this we can iteratively compute P_2, P_3, etc. For example, Equation (1.1) provides us a first example of a model in the form of an *iterated map* or *recursion*. In Figure 1.1 some iterates of the model are shown for several values of β. For large β the population grows faster initially, but all populations level off at the carrying capacity.

1.2 A differential equation

The Verhulst model for population growth in an environment with limited resources is

$$\frac{dP}{dt} = rP\left(1 - \frac{P}{P_c}\right), \tag{1.2}$$

where $P(t)$ is the population size, $r > 0$ is the uninhibited growth rate, and $P_c > 0$ is the carrying capacity of the environment.

Suppose the population at time t_0 is $P(t_0) = P_0$, and we wish to find the population $P(t_1) = P_1$ at some later time $t_1 > t_0$. Dividing both sides by $P(1-P/P_c)$, and integrating with respect to t from time t_0 to t_1 we find

$$\begin{aligned}\int_{t_0}^{t_1} \frac{1}{P(1-P/P_c)} \frac{dP}{dt}\, dt &= \int_{t_0}^{t_1} r\, dt, \\ \int_{P(t_0)}^{P(t_1)} \frac{1}{P(1-P/P_c)}\, dP &= r(t_1 - t_0). \qquad (1.3)\end{aligned}$$

Using a partial fraction decomposition, the integral on the left can be evaluated as follows

$$\begin{aligned}\int_{P_0}^{P_1} \left(\frac{1}{P} + \frac{1/P_c}{1-P/P_c}\right) dP &= [\ln P - \ln(1-P/P_c)]_{P_0}^{P_1} \\ &= \ln \frac{P_1}{1-P_1/P_c} - \ln \frac{P_0}{1-P_0/P_c}.\end{aligned}$$

Using this result and applying the exponential function to both sides of (1.3) gives

$$\left(\frac{P_1}{1-P_1/P_c}\right)\left(\frac{P_0}{1-P_0/P_c}\right)^{-1} = e^{r(t_1-t_0)}.$$

Solving for P_1 yields

$$P_1 = \frac{e^{r(t_1-t_0)} P_c P_0}{P_c + (e^{r(t_1-t_0)} - 1)P_0}.$$

Note that, since t_1 is arbitrary, the above relation provides us with a formula for finding $P(t)$ for all $t > t_0$:

$$P(t) = \frac{e^{r(t-t_0)} P_c P_0}{P_c + (e^{r(t-t_0)} - 1)P_0}. \qquad (1.4)$$

Solutions $P(t)$ are shown in Figure 1.2, for the case $r = 1/4$ and different initial conditions P_0.

The Verhulst model (1.2) is our first example of a *differential equation*. Let us reflect for a moment on what we have done here. In contrast to the iterated map (1.1), for which the model itself provides a recipe for computing the solution, the model (1.2) does not directly specify how the population $P(t)$ varies as a function of time, but only specifies the relation between $P(t)$ and its derivative dP/dt. Consequently, we need to *solve* the differential equation to determine $P(t)$. Second, note that the solution (1.4) at time t is expressed as a function of the solution at another time t_0. If we do not know the solution at some other time, the differential equation defines a *class of solutions* parameterized by the unknown constant P_0 (that is, each P_0 defines a different solution).

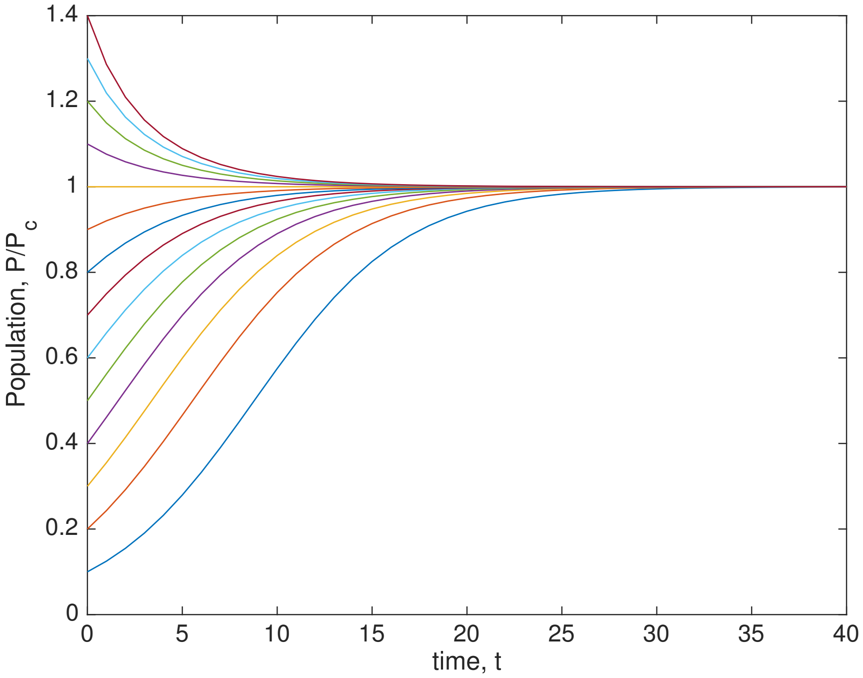

Figure 1.2: Solutions of the Verhulst model (1.2) with $r = 1/4$. Independent of the initial condition $P_0/P_c \in \{\frac{1}{10}, \frac{2}{10}, \ldots, \frac{14}{10}\}$, the populations eventually tend to the carrying capacity P_c.

1.3 A numerical method

Suppose we did not know how not evaluate the integral in (1.3). We might try to learn something about the solution using the following approach. Starting from the definition of the derivative

$$\frac{dP(t)}{dt} = \lim_{\Delta t \to 0} \frac{P(t+\Delta t) - P(t)}{\Delta t}, \tag{1.5}$$

we approximate this infinitessimal limit by a *finite difference*:

$$\frac{dP(t)}{dt} \approx \frac{P(t+\Delta t) - P(t)}{\Delta t}, \tag{1.6}$$

for some fixed, positive value of Δt. (That is, we refrain from taking the limit.) Substituting this approximation into (1.2) yields

$$\frac{P(t+\Delta t) - P(t)}{\Delta t} \approx r P(t) \left(1 - \frac{P(t)}{P_c}\right).$$

Denoting the approximation to $P(t_n)$ by P_n, where $t_{n+1} = t_n + \Delta t$, the above formula defines, for each $\Delta t > 0$, an iterated map or recursion

$$P_{n+1} = P_n + \Delta t \, r \, P_n \left(1 - \frac{P_n}{P_c}\right). \tag{1.7}$$

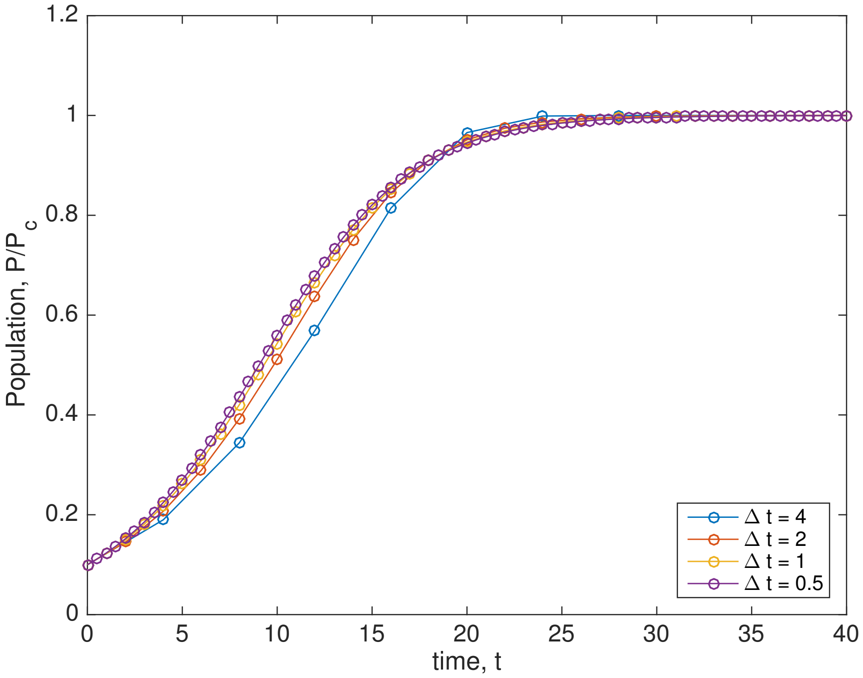

Figure 1.3: Iterates of the model (1.7) with $P_0/P_c = 0.1$, $r = 0.25$, and different stepsizes Δt. The iterates are connected by straight lines for visualization purposes. As Δt is decreased the iterates approach the exact solution shown in Figure 1.2.

This family of iterated maps, depending on the *time step* parameter Δt, is our first example of a *numerical method* or (more specifically) a *numerical integrator*. Whereas a differential equation such as (1.2) may or may not be easily integrated analytically, the numerical integrator is suitable for implementation on a computer.

See Figure 1.3 for numerical solutions computed with a range of stepsizes Δt. We know that as we make Δt smaller, the finite difference (1.6) becomes a better and better approximation of the derivative (1.5), and consequently we may expect that the approximation computed via (1.7) better and better approximates the solution (1.4). This intuition is correct, as we will see later, but it comes at a price: to compute the solution on the same interval with smaller time steps, the number of computations increases. The computer must do more work to improve the accuracy.

1.4 Relationships between the models

Note the similar form of the solution (1.4) of the differential equation model and the map (1.1). Given $r > 0$ and a time step Δt, if we define $\hat{\beta} = \exp(r\Delta t)$, then the solution of the differential equation at time $P(\Delta t)$ is given from (1.4)

$$P(\Delta t) = \frac{\hat{\beta} P_c P(0)}{P_c + (\hat{\beta} - 1) P(0)} = F(P(0)).$$

The function F defines a map taking the initial condition to the solution at time Δt. This function is the same as the map of the iteration (1.1). To compute the solution $P(2\Delta t)$, we may either choose a new $\hat{\beta} = \exp(r2\Delta t)$, or simply apply F again to the new 'initial condition' $P(\Delta t)$, i.e.

$$P(2\Delta t) = F(P(\Delta t)).$$

In other words, the map (1.1) just happens to be[2] the *time Δt-solution map* of the differential equation (1.2). In fact, for any fixed Δt, we can think of the solution of a differential equation as defining a map that takes initial conditions to solutions at time Δt. In this way, associated with a differential equation is a whole class of maps, one for each $\Delta t > 0$. The converse is not true however. There exist maps that are not the solution of any differential equation. This will become clear later in the course.

What about the numerical method (1.7)? Clearly, for each Δt the numerical method is just a map. We can say that a numerical method defines a one-parameter class of iterated maps. There is something special about these maps, however. As we let the parameter Δt tend to zero, we will demand that the associated map converge to the solution map of the differential equation.

In this course we study models in the form of iterated maps, differential equations, and numerical integrators. All three define dynamical systems under appropriate conditions.

[2] Well, it didn't "just happen"—it was constructed this way to illustrate a point.

Chapter 2

Definitions and elementary concepts

2.1 Modelling time dependent processes

We assume that the *state* of our system at any time t can be fully described by a set of d *variables* $\{y^{(1)}(t), y^{(2)}(t), \ldots, y^{(d)}(t)\}$ that vary with time. Generally, the $y^{(i)}(t)$ will be real numbers. When they denote populations or concentrations, they will be nonnegative. The language of linear algebra is convenient. We denote

$$y(t) = \begin{pmatrix} y^{(1)}(t) \\ \vdots \\ y^{(d)}(t) \end{pmatrix}, \qquad y \in \mathcal{D} \subset \mathbf{R}^d, \quad t \geq 0, \tag{2.1}$$

where $\mathcal{D}$ is sometimes referred to as the *state space* or *phase space*. In these notes we use superscripted parentheses to denote the index of a vector element. For the most part we try to avoid working with individual elements, however.

Example. The state of a chemical reaction may be specified by the concentrations of the reactants. Let $y^{(i)}(t)$ be the concentration (say, in grams per centiliter) of reactant i at time t. Then the state is give by (2.1), and $\mathcal{D} = \mathbf{R}_+^d$, where $\mathbf{R}_+$ is the set of nonnegative real numbers, since we do not permit negative concentrations.

Example. The state of the solar system may be modelled by the positions and velocities of all planets relative to the sun. Later we will explain why both positions and velocities are needed. Let $X_i(t) = (x^{(i)}(t), y^{(i)}(t), z^{(i)}(t))^T$ be the position vector of the ith planet, and $V_i(t) = (u^{(i)}(t), v^{(i)}(t), w^{(i)}(t))^T$

be its velocity vector. Then we may take

$$y(t) = \begin{pmatrix} X_1(t) \\ V_1(t) \\ \vdots \\ X_m(t) \\ V_m(t) \end{pmatrix},$$

where $m = 8$ or $m = 9$ depending on whether one considers Pluto a planet[1]. Here, $\mathcal{D} = \mathbf{R}^{3m}$.

For *discrete time models*, the state is only defined at distinct times t_0, t_1, t_2, ..., assumed uniformly spaced with step size Δt, i.e. $t_{n+1} = t_n + \Delta t$. In this case, we use superscript notation $y_n = y(t_n)$ to indicate time. Obviously, the effectiveness of such models depends on the existence of an appropriate Δt, which typically needs to be 'small enough' but not 'too small'.

The second ingredient to a dynamical system is a rule describing how the system state changes in time. In the next three sections we expound on the three 'rules' illustrated in the previous chapter: iterated maps, differential equations, and numerical integrators.

2.2 Iterated maps

The rule describing the change of the system state in discrete time may be given as a set of d algebraic functions of d variables:

$$x_{n+1}^{(1)} = F^{(1)}\left(x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(d)}\right), \tag{2.2}$$

$$x_{n+1}^{(2)} = F^{(2)}\left(x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(d)}\right), \tag{2.3}$$

$$\vdots \tag{2.4}$$

$$x_{n+1}^{(d)} = F^{(d)}\left(x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(d)}\right). \tag{2.5}$$

Here, each of the functions $F^{(i)}$ is a scalar valued function of d variables: $F^{(i)} : \mathbf{R}^d \to \mathbf{R}$. Analogous to (2.1) we define the discrete state vector $x_n = (x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(d)})^T$. Similarly we can define a vector function

$$F(x_n) = \begin{pmatrix} F^{(1)}(x_n) \\ \vdots \\ F^{(d)}(x_n) \end{pmatrix}, \qquad F : \mathbf{R}^d \to \mathbf{R}^d,$$

and write (2.2)–(2.5) in the compact (vector) notation

$$x_{n+1} = F(x_n), \quad n = 0, 1, 2, \ldots$$

[1] Pluto was discovered in 1930 and named the ninth planet. In 2005 the International Astronomical Union officially defined a 'planet'. Pluto fell short of the critera and, sadly, was demoted.

Given a mapping $F(x) : \mathbf{R}^d \to \mathbf{R}^d$, a subset $\mathcal{D} \subset \mathbf{R}^d$ is said to be *invariant* under the mapping F if $F(x) \in \mathcal{D}$ whenever $x \in \mathcal{D}$. We denote this by $F : \mathcal{D} \to \mathcal{D}$. In this situation, we define an *iteration* on $\mathcal{D}$ by

$$x_{n+1} = F(x_n), \qquad n = 0, 1, 2, \dots, \qquad x_0 \in \mathcal{D}. \tag{2.6}$$

A sequence of vectors satisfying (2.6) is referred to as an *orbit*:

$$\{x_0, x_1, x_2, \dots \,|\, x_{n+1} = F(x_n)\}.$$

An iterated map is also referred to as a *discrete map*, *recursion*, *recurrence relation*, *iterated function*, etc.

A recursion may also depend on more than one previous time level. For instance a *two-term recurrence* is defined via a rule such as

$$x_{n+1} = G(x_n, x_{n-1}), \qquad G : \mathbf{R}^d \times \mathbf{R}^d \to \mathbf{R}^d.$$

Example. One famous recursion model is the Fibonacci recursion

$$x_{n+1} = x_n + x_{n-1}. \tag{2.7}$$

Fibonacci used the model to describe the population dynamics of rabbits. The number x_n represents the number of rabbit *pairs* in month n. The model accounts for maturity of the rabbits, by assuming that newborn pairs must first mature by one month before they can produce offspring. Subsequently, all mature pairs produce precisely one pair of offspring. The model also assumes that all rabbit pairs remain alive. The number of rabbits in month $n+1$ is therefore equal to the number of rabbits in month n, plus the number of new offspring, which is equal to the number of rabbits in month $n-1$.

A general $(k+1)$-term recurrence assumes the form

$$x_{n+1} = G(x_n, x_{n-1}, \dots, x_{n-k}). \tag{2.8}$$

A k-term recurrence can always be written as a one-step iteration by introducing extra variables. To do so, define the vector

$$y_n = \begin{pmatrix} x_n \\ x_{n-1} \\ \vdots \\ x_{n-k} \end{pmatrix}$$

and the extended map

$$F(y_n) = \begin{pmatrix} G(x_n, x_{n-1}, \dots, x_{n-k}) \\ x_n \\ \vdots \\ x_{n-k+1} \end{pmatrix}.$$

Then the *companion form*

$$y_{n+1} = F(y_n)$$

is equivalent to (2.8) plus a number of trivial identities.

Example. We can rewrite the Fibonacci sequence (2.7) as a one-step model by introducing a second variable $z_n = x_{n-1}$. Then the sequence can be written

$$x_{n+1} = x_n + z_n, \qquad z_{n+1} = x_n.$$

In matrix-vector form, this becomes

$$\begin{pmatrix} x_{n+1} \\ z_{n+1} \end{pmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{pmatrix} x_n \\ z_n \end{pmatrix}.$$

The matrix above is the simplest example of a Leslie matrix, to be discussed in Chapter 4.

When $d = 1$ we speak of a scalar iteration. We will treat recursions on the real line in Chapter 3, and discuss the generalization to higher dimensions in Chapter 4.

2.3 Differential equations

In this section we will learn about models described by differential equations. The *motion* or *evolution* of the model in time is given by a function $y(t)$. Here, t varies over an interval $t \in [0, T]$, and $y(t)$ defines the state of the model at time t. For the Verhulst model (1.2), the state is specified by the population size $P(t)$.

Remark. Of note, $y(t)$ is a continuous (in fact, differentiable) function of t, which may not reflect the often discrete nature of the phenomenon being modeled. The number of individuals in a population is always a whole number, and cannot change continuously. The continuum approximation is most appropriate for a large population, or when the quantities involved reflect probabilities or averages.

Unlike models based on recursions, differential equation models do not directly define the motion $y(t)$, nor even provide a recipe for constructing it. Instead, the motion is implicitly defined by specifying the relationship between $y(t)$ and one or more of its derivatives. The general form of a differential equation is

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

The differential equation is seen as a *problem* whose *solution* yields the motion $y(t)$, where we mean by a solution a function $y(t)$ that identically satisfies the differential equation and any additional constraints, such as the initial condition.

Example. The model for exponential growth or decay, familiar from radioactive materials, interest on loans, and so forth, is

$$\frac{dy}{dt} = ay(t), \tag{2.9}$$

where a is a given constant. We can immediately verify by direct substitution that a solution to this differential equation is provided by the function

$$y(t) = C \exp(at) \tag{2.10}$$

with arbitrary constant C. Hence the differential equation is an implicit definition of a class of functions. By specifying the solution at some time, say $y(0) = y_0$, the constant C becomes fixed, since

$$y_0 = y(0) = Ce^{a \cdot 0} = C,$$

and therefore the solution through this point is $y(t) = y_0 \exp(at)$. The *initial value problem* is to find the solution $y(t)$, $t \in [0, T]$, of a differential equation given $y(0)$ at time $t = 0$.

Remark. We will often use shorthand notation $\dot{y}$ or y' to represent $\frac{dy}{dt}$.

In the above example, the function $f(t, y) = ay$ has no explicit dependence on t. In such a case, the differential equation is called *autonomous* and otherwise *non-autonomous*. An example of a non-autonomous differential equation is $\dot{y} = \sin(t)y$.

Most models require more than one piece of information to define the model state, and consequently $y(t)$ will be a vector in a d-dimensional subset:

$$y(t) = (y^{(1)}(t), y^{(2)}(t), \ldots, y^{(d)}(t))^T, \quad y(t) : [0, T] \to \mathcal{D} \subset \mathbf{R}^d.$$

For example, if the model describes the populations of aphids $A(t)$ and ladybugs $B(t)$ in a rose garden, then $y(t) = (A(t), B(t))^T$ is a vector with two components, the number of each species at time t.

Each of these variables $y^{(1)}(t)$, ..., $y^{(d)}(t)$ satisfies a differential equation that will typically depend on the other variables. The general form of this *system of differential equations* is

$$\begin{aligned}
\frac{dy^{(1)}}{dt} &= f^{(1)}\left(t, y^{(1)}(t), y^{(2)}(t), \ldots, y^{(d)}(t)\right), \\
\frac{dy^{(2)}}{dt} &= f^{(2)}\left(t, y^{(1)}(t), y^{(2)}(t), \ldots, y^{(d)}(t)\right), \\
&\vdots \\
\frac{dy^{(d)}}{dt} &= f^{(d)}\left(t, y^{(1)}(t), y^{(2)}(t), \ldots, y^{(d)}(t)\right).
\end{aligned}$$

Define the vector of time derivatives

$$\frac{dy}{dt} = \left(\frac{dy^{(1)}}{dt}, \frac{dy^{(2)}}{dt}, \ldots, \frac{dy^{(d)}}{dt} \right)^T$$

and the vector field

$$f(t, y) = (f^{(1)}(t, y), f^{(2)}(t, y), \ldots, f^{(d)}(t, y))^T.$$

Specifying an initial condition $y_0 = (y^{(1)}(0), y^{(2)}(0), \ldots, y^{(d)}(0))^T$, we pose the general form of *the initial value problem on $\mathcal{D} \subset \mathbf{R}^d$*:

$$\frac{dy}{dt} = f(t, y), \quad y(0) = y_0, \qquad t \in [0, T],\ y \in \mathcal{D},\ f : [0, T] \times \mathcal{D} \to \mathbf{R}^d. \tag{2.11}$$

The special case $d = 1$ is referred to as a *scalar differential equation*.

The *order* of a differential equation is the magnitude of the largest derivative of y present. For instance, the differential equation

$$\frac{d^2y}{dt^2} + a\frac{dy}{dt} + by = 0 \tag{2.12}$$

is a second order differential equation since the largest derivative of y is two. We can always rewrite a higher order differential equation as a first order system by introducing additional variables. In the above example, for instance, we define $v \equiv dy/dt$ and derive the system

$$\begin{aligned} \frac{dy}{dt} &= v, \\ \frac{dv}{dt} &= -av - by. \end{aligned}$$

Evidently, defining initial conditions for the first order form corresponds to specifying initial conditions on y and some of its derivatives in the higher order formulation. In this example, we would define $y(0) = y_0$ and $v(0) = \dot{y}(0) = v_0$.

2.3.1 Solution by separation of variables

There are a number of general strategies for solving scalar differential equations. For our purpose the most important strategy is the method of *separation of variables*, which has already been used to construct the solution of the Verhulst equation (1.4).

Suppose that the function f can be written as a quotient $f(t, y) = g(t)/h(y)$, where $g(t)$ and $h(y)$ have known anti-derivatives $G(t)$ and $H(y)$. Then we may separate the variables and integrate

$$\begin{aligned} \frac{dy}{dt} &= \frac{g(t)}{h(y)} \\ \int h(y)\, \frac{dy}{dt}\, dt &= \int g(t)\, dt \\ \int h(y)\, dy &= \int g(t)\, dt \\ H(y) &= G(t) + C, \end{aligned}$$

where C combines the integration constants and is to be determined from the initial condition. To express the solution $y(t)$ explicitly we also require $H(y)$ to be invertible.

Example. We solve the initial value problem

$$y' = y^2, \qquad y(0) = y_0 \tag{2.13}$$

In this case $g(t) = 1$, $h(y) = y^{-2}$, $G(t) = t$ and $H(y) = -y^{-1}$, yielding

$$-\frac{1}{y} = t + C, \quad \text{or} \quad y(t) = \frac{1}{-C-t}.$$

Using the initial condition $y(0) = y_0$, we determine that $C = -y_0^{-1}$, and the general solution is

$$y(t) = \frac{1}{y_0^{-1} - t}. \tag{2.14}$$

Remark. Every autonomous scalar differential equation $y' = f(y)$ is separable with $g(t) = 1$ and $h(y) = 1/f(y)$. Explicit solution of the differential equation is then possible if an invertible antiderivative $H(y)$ exists. The solutions of many scalar differential equations have been studied and their solutions categorized and approximated in the theory of special functions.

2.3.2 Existence and uniqueness of solutions

Whether or not a solution can be explicitly constructed, it is often necessary to at least show that a solution exists and that it is unique. For differential equations, it is well understood that for a large class of initial value problems unique solutions exist. The following definition and theorem hold for scalar as well as multidimensional systems of differential equations:

Definition 1 *A vector field $f(y) : \mathcal{D} \to \mathbf{R}^d$ is said to be Lipschitz continuous on $\mathcal{D}$ if there exists a constant $L > 0$ such that*

$$\|f(x) - f(y)\| \le L\|x - y\|, \quad \forall x, y \in \mathcal{D}. \tag{2.15}$$

Theorem 2.3.1 *Let $\mathcal{D} \subset \mathbf{R}^d$ be an open set and $\bar{\mathcal{D}}$ its closure. If $f(y)$ is Lipschitz continuous on $\bar{\mathcal{D}}$ and $y_0 \in \mathcal{D}$, then the initial value problem (2.11) has a unique solution $y(t)$ satisfying $y(0) = y_0$. The solution exists on an interval in t containing the origin, and extending until such time as $y(t)$ reaches the boundary of $\mathcal{D}$.*

Remark. If f is continuously differentiable on a closed, bounded set $\bar{\mathcal{D}}$ then it is Lipschitz continuous on $\mathcal{D}$. If f is Lipschitz continuous, then it is continuous. In the scalar case, we can choose $L = \max_{\bar{\mathcal{D}}} |f'(y)|$. (And for $d > 1$, we can choose L to be the maximum normalized directional derivative of f on $\mathcal{D}$).

Example. For $y_0 < 0$, the differential equation (2.13) has Lipschitz constant $L = 2\bar{y}$ on any interval $\mathcal{D} = (-\bar{y}, \bar{y})$. There is a unique solution through any point in such an interval. If $y_0 < 0$ the solution exists for all $t > 0$. For $y_0 > 0$, the solution exists on $t \in (0, y_0^{-1})$, since $y(t)$ grows unbounded at $t \to y_0^{-1}$.

For the initial value problem

$$y' = 2\sqrt{y}, \quad y(0) = 0,$$

Lipschitz continuity fails on any $\mathcal{D}$ containing the origin (since $f'(y)$ is unbounded there), and no unique solution exists. It can be directly verified that two distinct solutions to this initial value problem are $y(t) \equiv 0$ and $y(t) = t^2$.

2.4 Numerical methods

Very few of the differential equations encountered in practice can be solved explicitly. Exact solutions are known for nonlinear differential equations in few dimensions, or when there is some mathematical structure that constrains the trajectories to lie at the intersections of certain surfaces. However, even for differential equations in $\mathbf{R}^3$, the trajectories can become so complex and chaotic, that no explicit solution is imaginable. The complexity increases with the dimension d. In this case, we may either try to prove qualitative statements about the solution, e.g. to determine the ultimate fate of all solution trajectories with initial condition in some open set; or we may employ numerical methods to approximately solve the differential equation using a computer for some specific initial condition. In this section we introduce numerical methods for differential equations.

The most fundamental property of the numerical approximation is that it must be computable in a finite number of operations (unless one is willing to wait an eternity for the solution). The process of replacing the continuous solution by a finite one is called *discretization*.

To discretize, we replace the interval $\mathcal{D} = [t_0, t_0 + T]$ by a finite number N of discrete times $t_n = t_0 + n\Delta t$, $n = 0, 1, \ldots, N$, where $\Delta t = T/N$ is the step size. Similarly, we replace the continuous solution $y(t)$ on D with the numerical solution $y_n \approx y(t_n)$, $n = 0, 1, \ldots, N$. The y_n can be thought of as snapshots of the system state at the discrete times, and the sequence $\{y_0, y_1, \ldots, y_N\}$ as a movie. Finally, we need a procedure to generate the y_n for $n > 0$ (y_0 is the given initial condition). In this section, we will consider methods that approximate y_{n+1} as a recursion given y_n, Δt, and the explicit form of $f(t, y)$. A time-stepping procedure is referred to as a *numerical integrator* or *scheme*.

There are several approaches to the construction of integrators. The oldest and simplest method is *Euler's method*, which we encoutered in (1.7). It is based on the rectangle rule for approximation of an integral. First consider the special case of (2.11) with $f(t, y) = f(t)$ (f is independent of y). The rectangle rule is just $y_{n+1} = y_n + \Delta t f(t_n)$. Generalizing this to y-dependent problems is straightforward, and results in Euler's method

$$y_{n+1} = y_n + \Delta t f(t_n, y_n). \tag{2.16}$$

Euler's method can be interpreted in several other ways: (1) as the direct extrapolation of the local slope of $y(t)$ through (t_n, y_n), (2) as the single term truncation of a Taylor expansion, or (3) as a finite difference, as we explained in Chapter 1. These different interpretations can be used to generalize this method and construct others.

To program Euler's method, we write a loop and compute the successive iterates as follows:

Algorithm 2.4.1 (Euler's Method) Given: *initial time t_0, initial value y_0, stepsize Δt, a vector field $f(t, y)$, and a number of time steps N,*

Output: *$y_1, y_2, \ldots, y_N$ approximating the solution of $\frac{dy}{dt} = f(y, t)$ at equally spaced points $t_1 := t_0 + \Delta t, t_2 = t_0 + 2\Delta t, \ldots$.*

```
for $n = 0, 1, \ldots, N-1$
$t_{n+1} := t_0 + n\Delta t$;
$y_{n+1} := y_n + \Delta t f(t_n, y_n)$;
end
```

□

Figure 1.3 illustrates the numerical solutions for a range of step sizes Δt. We can make a couple of observations. First, we are happy to see that all of the solutions tend to the stable equilibrium at $P = P_c = 1$. We will come back to this later. However, we may also notice that there is a quite a difference between the solutions leading up to the stable equilibrium, despite the fact that all of the numerical trajectories start from the same initial condition. We discuss this next.

2.4.1 Convergence

A numerical method is an approximation of the exact time-Δt solution map (cf. Section 1.4) of a differential equation. As such, there is always an element of error in the numerically generated solution. If the method is to be useful, we must be able to control this error, at least on short enough time intervals.

A tool of singular importance in numerical analysis is Taylor series, for which the relevant form here is

$$y(t + \Delta t) = y(t) + \Delta t\, y'(t) + \frac{\Delta t^2}{2!} y''(t) + \frac{\Delta t^3}{3!} y'''(t) + \cdots \tag{2.17}$$

for a perturbation $\Delta t > 0$ around t. For a *scalar function* ($d = 1$), assuming $y(t)$ is p-times continuosly differentiable, Taylor's theorem says there is a point $t^* \in [t, t + \Delta t]$ such that

$$y(t + \Delta t) = \sum_{i=0}^{p-1} \frac{\Delta t^i}{i!} \frac{d^i y}{dt^i}(t) + \frac{\Delta t^p}{p!} \frac{d^p y}{dt^p}(t^*).$$

For a vector function ($d > 1$), such a statement holds for each component, but in general the mean value will be attained at a different t^* for each component. Nonetheless the norm of the last (remainder) term is bounded on $[t, t + \Delta t]$, and we have

$$\| \frac{\Delta t^p}{p!} \frac{d^p y}{dt^p}(t) \| \leq C \Delta t^p,$$

for a positive constant C, and we write

$$y(t + \Delta t) = \sum_{i=0}^{p-1} \frac{\Delta t^i}{i!} \frac{d^i y}{dt^i}(t) + \mathcal{O}(\Delta t^p), \tag{2.18}$$

where the notation $\mathcal{O}(\Delta t^p)$ means the final terms converge to zero no slower than Δt^p, i.e.

$$\lim_{\Delta t \to 0} \left(\frac{1}{\Delta t^p} \sum_{i=p}^{\infty} \frac{\Delta t^i}{i!} \frac{d^i y}{dt^i}(t) \right) \leq \kappa,$$

for some contant $\kappa > 0$ independent of Δt.

We define the *global error* after n time steps to be the difference between the discrete approximation and the exact solution

$$e_n := y_n - y(t_n). \tag{2.19}$$

For an approximation we want the error to be small in norm at each step of simulation, that is, we would like to satisfy

$$\max_{n=0,1,\dots N} \|e_n\| \leq \delta,$$

for some choosable tolerance δ. For a given vector field f, initial value y_0, and time interval T, we have only one free parameter, the timestep $\Delta t = T/N$, which we can vary to make sure the norm of the error meets our tolerance. If the method is going to be useful, we must be able to vary Δt to meet any tolerance we choose.

Given a Lipschitz vector field f, a method is said to be *convergent* if, for every T,

$$\lim_{\substack{\Delta t \to 0, \\ \Delta t = T/N}} \max_{n=0,1,\dots,N} \|e_n\| = 0.$$

(Note that this definition considers only discrete values of Δt which are integral fractions of the time interval. Equivalently we could take the limit as $N \to \infty$ with $\Delta t = T/N$.)

We demonstrate convergence for Euler's method. Consider the initial value problem for an autonomous differential equation

$$\frac{dy}{dt} = f(y), \qquad y(t) \in \mathcal{D}, \quad f : \mathcal{D} \to \mathbf{R}^d,$$

and assume that f is Lipschitz with constant L, and all solutions $y(t)$ have bounded second derivatives on $\mathcal{D}$, i.e. $\|\frac{d^2y}{dt^2}\| \le C$.

For an initial value problem $y(0) = y_0$ and $t \in [0, T]$, we evaluate the global error (2.19) after $n+1$ steps of Euler's method, and expanding the exact solution $y(t)$ using Taylor's theorem:

$$\begin{aligned} e_{n+1} &= y_{n+1} - y(t_{n+1}) \\ &= y_n + \Delta t f(y_n) - \left[y(t_n) + \Delta t \frac{dy}{dt}(t_n) + \frac{\Delta t^2}{2} \frac{d^2 y}{dt^2}(t^*) \right] \\ &= y_n + \Delta t f(y_n) - \left[y(t_n) + \Delta t f(y(t_n)) + \frac{\Delta t^2}{2} \frac{d^2 y}{dt^2}(t^*) \right], \end{aligned}$$

for some $t^* \in [t_n, t_{n+1}]$. Applying the triangle inequality and the Lipschitz condition,

$$\|e_{n+1}\| \le \|y_n - y(t_n)\| + \Delta t \|f(y_n) - f(y(t_n))\| + \frac{\Delta t^2}{2} C = (1 + \Delta t L)\|e_n\| + \frac{\Delta t^2}{2} C. \quad (2.20)$$

To proceed we need the following lemma:

Lemma 2.4.1 *Let the sequence of nonnegative numbers x_n satisfy inequality recursion*

$$x_{n+1} \le a x_n + b, \qquad x_0 \ge 0,$$

where $a \ge 0$ and $b \ge 0$. Then

$$x_n \le a^n x_0 + \frac{a^n - 1}{a - 1} b.$$

The proof, by induction, is straightforward.

For the global error (2.20) the appropriate constants are $a = 1 + \Delta t L \le \exp(\Delta t L)$ and $b = C\Delta t^2/2$. Assuming the initial condition is exact, $e_0 = 0$ and we find

$$\|e_n\| \le \left(\frac{e^{n\Delta t L} - 1}{\Delta t L} \right) \frac{\Delta t^2}{2} C,$$

and in particular we find

$$\max_n \|e_n\| \le \Delta t \frac{C}{2L}(e^{TL} - 1).$$

The global error decreases at a rate proportional to Δt as $\Delta t \to 0$. Consequently, Euler's method converges as $\Delta t \to 0$.

2.4.2 Trapezoidal rule and other methods

Before concluding this section, we introduce a some additional numerical methods. From elementary calculus, you may remember that a better approximation than the rectangle rule is the trapezoidal rule. Again consider the equation $\frac{dy}{dt} = f(t)$. The trapezoidal rule approximation is $y_{n+1} = y_n + \frac{\Delta t}{2}\left(f(t_{n+1}) + f(t_n)\right)$. Generalizing to the ODE (2.11), the *trapezoidal rule* is

$$y_{n+1} = y_n + \frac{\Delta t}{2}\left(f(t_n, y_n) + f(t_{n+1}, y_{n+1})\right). \tag{2.21}$$

Notice that this method is of a character very different from Euler's method. It is not possible to evaluate the last term in the equation without knowing y_{n+1}, and it is not possible to compute y_{n+1} without evaluating this term. The trapezoidal rule defines y_{n+1} *implicitly* as function of y_n. In other words, we must solve the typically nonlinear system of algebraic equations

$$0 = r(y) := y - y_n + \frac{\Delta t}{2}\left(f(t_n, y_n) + f(t_{n+1}, y)\right) \tag{2.22}$$

for y to determine y_{n+1}. Such methods are termed *implicit methods*, and generally demand much more work from the computer per time step than an *explicit method* such as Euler's method. Methods for solving systems of algebraic equations are, for example Newton's method (which we will encounter in Chapter 4), or Picard iteration. Obviously an implicit method must be significantly advantageous in some other sense to justify its increased computational cost. We will say more about this later. Also, it is uncertain if (2.22) will possess any real solutions, and if so, how many. For instance, applying (2.21) to the differential equation (2.13) yields

$$y_{n+1} = y_n + \frac{\Delta t}{2}(y_n^2 + y_{n+1}^2) \qquad \Longleftrightarrow \qquad y_{n+1} - \frac{\Delta t}{2}y_{n+1}^2 = y_n + \frac{\Delta t}{2}y_n^2.$$

This is a quadratic equation in the unknown y_{n+1} and has two solution branches in general. Which one should we choose? For $\Delta t = 0$, there is the unique solution $y_{n+1} = y_n$. One of the two solution branches converges to this solution in the limit $\Delta t \to 0$, whereas the other branch diverges. We should choose the branch that remains bounded in the limit.

Two additional numerical methods that we will study are the *implicit Euler method*

$$y_{n+1} = y_n + \Delta t\, f(t_{n+1}, y_{n+1}), \tag{2.23}$$

which is similar to Euler's method, but implicit like Trapezoidal Rule; and the "extrapolated midpoint rule":

$$y_{n+1} = y_n + \Delta t\, f\left(t_{n+1/2}, y_n + \frac{\Delta t}{2} f(t_n, y_n)\right), \tag{2.24}$$

where $t_{n+1/2} = t_n + \Delta t/2$. Note that this method is explicit, as is Euler's method. However the extrapolated midpoint rule converges more rapidly.

Chapter 3

Models on the real line

3.1 Scalar iterations

In this chapter we focus on scalar models. We introduce the concepts of equilibrium and stability in this context. To motivate our discussion we first introduce a technique for visualizing scalar iterations.

3.1.1 Graphical analysis of iterated maps

One means of developing intuition about scalar maps is to construct the associated "cobweb diagram". This is illustrated in Figure 3.1. On the horizontal axis, we plot x_n and on the vertical axis x_{n+1}. The function $F(x)$ is shown in blue. Suppose $x_0 = 0.1$. We follow the dotted line upwards to the graph of $F(x_0)$, then to the left along the dotted line to the vertical axis to find x_1. To compute the following iterate, we first need to find the image of x_1 on the x_n axis. To facilitate this, the identity map $I(x) = x$ is also shown in the figure: the red line. Now we follow the dashed line from x_1 horizontally to the identity map, then down along the dotted line to find x_1 on the horizontal axis. From here we can repeat the whole process again to find x_2, etc. However, we can also take a short cut. From the figure it is clear that it is unnecessary to follow the dashed lines to the corresponding axes. Instead, we can iterate by alternately stepping *vertically to the graph of F*, and subsequently *horizontally to the identity map*. This *cobweb diagram* is illustrated with the yellow line.

Example. The *logistic map* is a scalar iteration related to the Verhulst model (1.2)

$$x_{n+1} = rx_n(1 - x_n), \qquad \Longleftrightarrow \qquad F(x) = rx(1 - x). \tag{3.1}$$

We choose $\mathcal{D} = [0, 1]$. The mapping $F(x)$ is a parabola, centered at $x = 1/2$ and concave downward. With $r > 0$, $F(x)$ is nonegative for all $x \in \mathcal{D}$, and attains its maximum at $x = 1/2$ with $F(1/2) = r/4$. Consequently, $\mathcal{D}$ is invariant under F for $r \in [0, 4]$.

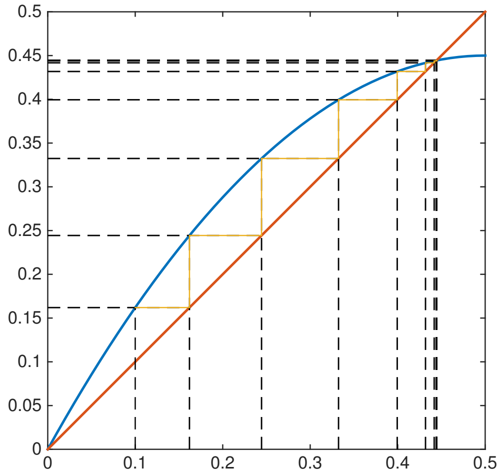

Figure 3.1: Graphical analysis method illustrated for $r = 1.8$.

Figure 3.2 illustrates cobweb diagrams for the logistic map for different values of the growth rate parameter r. The corresponding orbits are shown as a sequence $\{x_0, x_1, \dots\}$ as function of iteration number n on the right. In all three cases, the orbits appear to converge upon a fixed value. In the following sections we provide an explanation for this observed behavior.

3.1.2 Fixed points and stability

The points in a cobweb diagram at the intersection of the graphs of $F(x)$ and $I(x)$ (the blue and red lines, respectively) are special. These correspond to iterates $x_{n+1} = F(x_n) = I(x_n) = x_n$. The iterates are identical, and the orbit satisfies

$$x_n = x_{n-1} = \cdots = x_1 = x_0 = \alpha, \qquad \Longleftrightarrow \qquad \alpha = F(\alpha)$$

Such a point α is termed an *equilibrium* or *fixed point* of the recursion (2.6).

What are the fixed points of the logistic map (3.1)? These are the solutions of

$$x = F(x) = rx(1-x) \qquad \Longleftrightarrow \qquad x\,[r(1-x)-1] = 0.$$

That is,

$$\alpha \in \{0, 1 - r^{-1}\}.$$

Why are equilibria important? Looking at the cobweb diagrams in Figure 3.2, we notice that in all cases the orbits seem to eventually approach one of the two equilibria. For the case $r = 1$, the orbit approaches equilibrium $\alpha_1 = 0$, while for the cases $r = 1.8$ and $r = 2.8$ the orbit approaches $\alpha_2 = 1 - r^{-1}$. For $r = 1.8$ the iterates x_n tend to α_2 monotonically, whereas for $r = 2.8$ they oscillate about α_2 with decreasing amplitude.

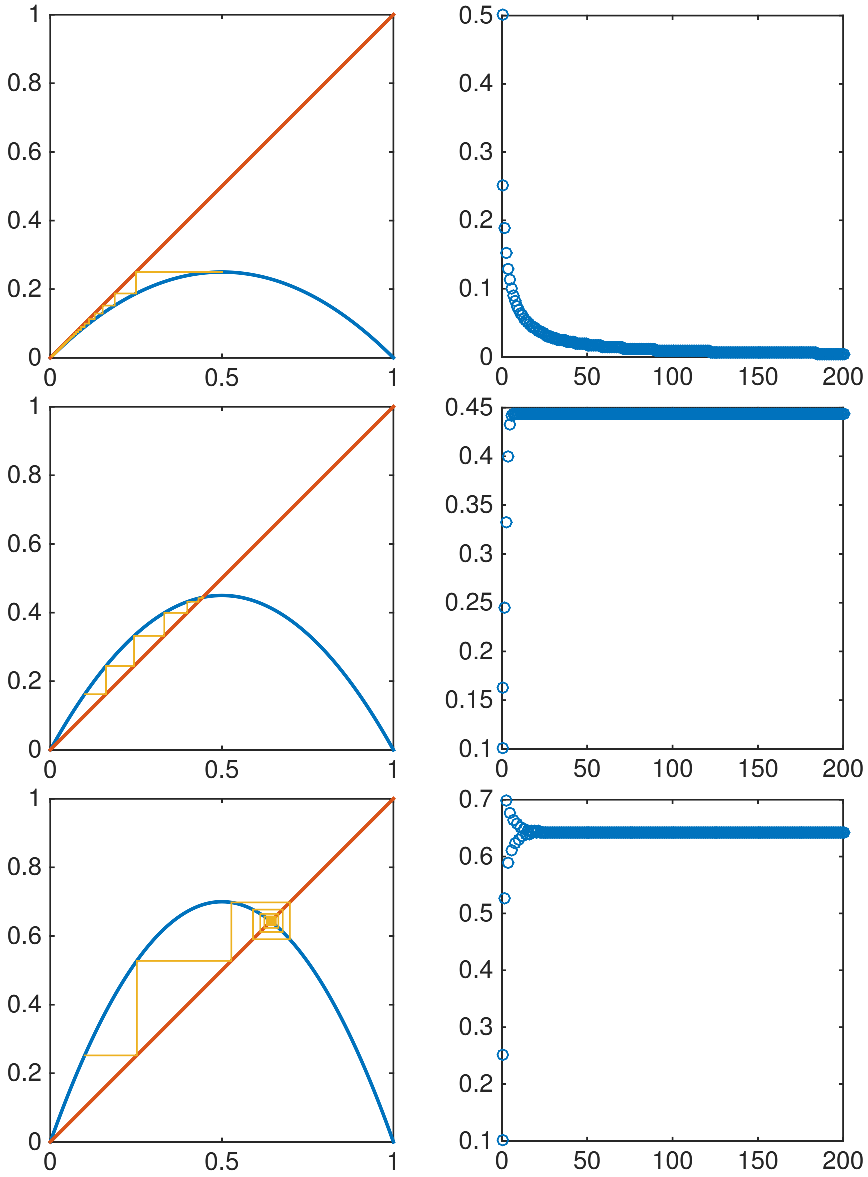

Figure 3.2: Cobweb diagrams (left) and orbits (right) for the logistic map (3.1) for $r = 1$ (top), $r = 1.8$ (middle) and $r = 2.8$ (bottom).

An equilibrium with the property that nearby iterates remain nearby is called a *stable* equilibrium. To be precise,

Definition 2 *An equilibrium $\alpha = F(\alpha)$ of the recursion (2.6) is Lyapunov stable if, for every $\varepsilon > 0$ there exists $\delta > 0$ such that $\|x_n - \alpha\| \leq \varepsilon$ for all $n = 1, 2, \ldots,$ whenever $\|x_0 - \alpha\| \leq \delta$. An equilibrium that is not stable is unstable.*

Can we use the definition to prove that the equilibrium $\alpha_1 = 0$ is stable for the logistic map

(3.1) with $r = 1$? We have already seen that the interval $\mathcal{D} = [0, 1]$ is invariant under the logistic map for all $r \in [0, 4]$. This implies that if $x_0 \in \mathcal{D}$, then $x_n \geq 0$, for all n. For $r = 1$, the map (3.1) becomes $x_{n+1} = x_n(1 - x_n) = x_n - x_n^2$, and since $x_n^2 \geq 0$, it follows that $x_{n+1} \leq x_n$, with strict inequality holding if $x_n \notin \{0, 1\}$. Hence $\{x_n\}$ is a monotonically decreasing sequence in $\mathcal{D}$ if $0 < x_0 < 1$. Consequently, for every $\varepsilon > 0$, we can choose $\delta = \min\{\varepsilon, 1\}$, and α_1 is stable.

Next, consider the trivial recursion $x_{n+1} = x_n$. For this iteration, every point is an equilibrium. Furthermore, every point is stable with $\delta = \varepsilon$. Yet the iterates do not 'tend to' some other value in the limit of large n, as we observed for the logistic map. We distinguish between stable equilibria and *asymptotically stable* equilibria.

Definition 3 *An equilibrium $\alpha = F(\alpha)$ of the recursion (2.6) is asymptotically stable if it is Lyapunov stable and, in addition,* $\lim_{n\to\infty} x_n = \alpha$.

In Figure 3.3 we provide two examples to illustrate some aspects of the stability definitions. The first recursion (upper plots of Fig. 3.3) is:

$$F(x) = \begin{cases} 1.2x, & 0 \leq x \leq \frac{1}{2}, \\ 0, & x > \frac{1}{2}. \end{cases}$$

In this case, for any small positive x_0, the iterates increase monotonically until $x_n > 1/2$, at which point $x_{n+1} = x_{n+2} = \cdots = 0$. That means that $\lim_{n\to\infty} x_n = 0$, for all x_0. However, it is not possible to find a $\delta > 0$ for, say, $\varepsilon = 1/4$ such that the iterates remain within a distance ε of 0 for all n. The equilibrium is unstable.

The second recursion (lower plots of Fig. 3.3) is

$$F(x) = \begin{cases} -\frac{1}{6}x, & x \leq 0, \\ -5x, & x > 0. \end{cases}$$

In this case the equilibrium is asymptotically stable, but the iterates make large excursions away from the equilibrium whenever $x_0 > 0$. Therefore, δ must be chosen much smaller than ε to meet the definition of stability.

The following theorem provides a useful criterion for identifying stable equilibria.

Theorem 3.1.1 *Let $\mathcal{D} = [a, b] \subset \mathbf{R}$ and $F : \mathcal{D} \to \mathcal{D}$ be continuously differentiable. A fixed point $\alpha = F(\alpha)$ is asymptotically stable if $|F'(\alpha)| < 1$ and unstable if $|F'(\alpha)| > 1$.*

Proof Let x_n be represented locally in a neighborhood of α as $x_n = \alpha + \eta_n$. Inserting this into the recursion (2.6) and applying Taylor's theorem yields

$$\alpha + \eta_{n+1} = F(\alpha + \eta_n) = F(\alpha) + F'(\xi)\eta_n,$$

for some ξ between α and $\alpha + \eta_n$. Since $\alpha = F(\alpha)$ this reduces to

$$\eta_{n+1} = F'(\xi)\eta_n.$$

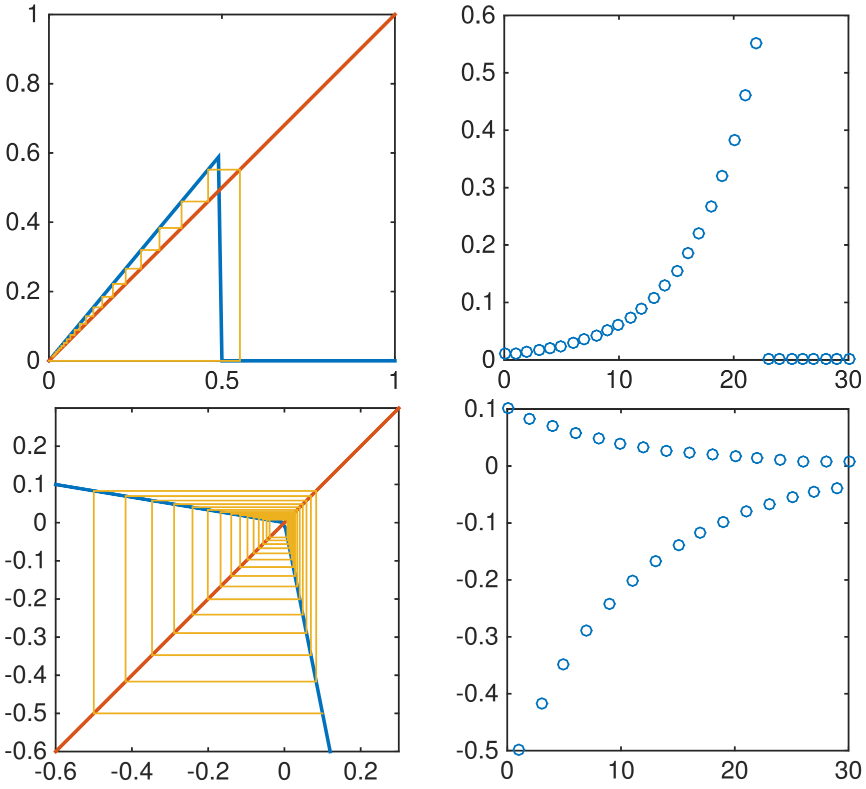

Figure 3.3:

Choose ρ such that $|F'(\alpha)| < \rho < 1$. Then there exist $\delta^+ > 0$ and $\delta^- > 0$ with $\delta = \min\{\delta^+, \delta^-\}$ such that $|F'(\xi)| \leq \rho$ for all $|\xi - \alpha| < \delta$ (see Figure 3.4). If $|x_n - \alpha| < \delta$, then

$$|x_{n+1} - \alpha| = |\eta_{n+1}| = |F'(\xi)||\eta_n| < \rho|\eta_n| = \rho|x_n - \alpha| < |x_n - \alpha|.$$

Iterating this argument shows that $|x_n - \alpha| < \rho^n |x_0 - \alpha|$ if $|x_0 - \alpha| < \delta$.

The same reasoning with $<$ replaced by $>$ proves the converse. □

The proof of the theorem shows why strict inequality is necessary. The case $|F'(\alpha)| = 1$ must be analyzed individually.

Let us apply Theorem 3.1.1 to the logistic map (3.1). The derivative is

$$F'(x) = r(1 - 2x)$$

For the equilbrium $\alpha_1 = 0$, we find $F'(0) = r$, so the origin is stable if $r < 1$. At the equilibrium $\alpha_2 = 1 - r^{-1}$ we find $F'(1 - r^{-1}) = 2 - r$. This equilibrium is stable if

$$|2 - r| < 1 \quad \Longleftrightarrow \quad 1 < r < 3$$

Hence, we see that indeed α_2 is stable in the cases $r = 1.8$ and $r = 2.8$ as shown in Figure 3.2. Note also that for $1 < r < 2$, the slope $F'(\alpha_2)$ is positive and the convergence

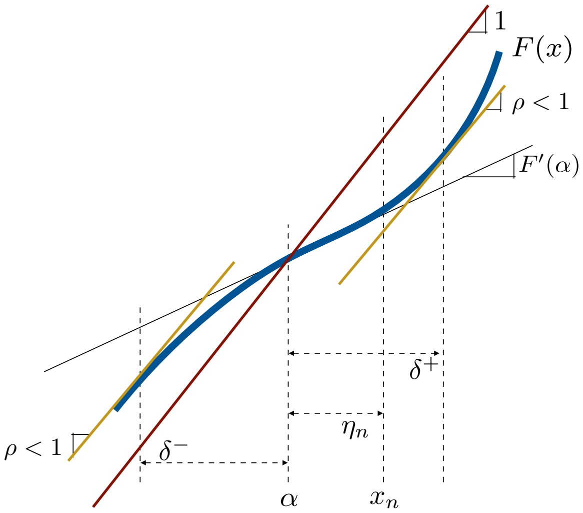

Figure 3.4: Illustration of the proof of Theorem 3.1.1.

monotone, whereas for $2 < r < 3$, the slope $F'(\alpha_2)$ is negative and the convergence oscillatory.

What about the cases $r = 1$ and $r = 3$? For these values $|F'(\alpha_2)| = 1$ and Theorem 3.1.1 does not apply. We have already seen that the origin is stable for $r = 1$. The case $r = 3$ is left to the reader to prove.

3.1.3 Periodic solutions and bifurcations

What is the desitny of the orbits of the logistic map (3.1) in case $r > 3$? We have seen that for all $0 \leq r \leq 4$, the function F maps $\mathcal{D}$ into itself, so the orbit $\{x_n\}$ remains in the set $\mathcal{D}$ for all n. Yet F possesses no stable equilibrium. Figure 3.5 illustrates the orbit for $r = 3.2$. After a short transient period, the orbit appears to alternate between two values. Such an orbit is called a *period-2* orbit.

A period-2 orbit consists of a pair of points $\beta_1 \in \mathcal{D}$ and $\beta_2 \in \mathcal{D}$ satisfying

$$F(\beta_1) = \beta_2, \quad F(\beta_2) = \beta_1, \quad \beta_1 \neq \beta_2.$$

Note that in this case, β_1 and β_2 are both fixed points of the composite map $F \circ F(x) = F(F(x))$, but not of $F(x)$ itself.

For the logistic map (3.1), the composite map is

$$F \circ F(x) = rF(x)(1 - F(x)) = r^2 x(1 - x)(1 - rx(1 - x)).$$

This function is shown in the leftmost plot Figure 3.6. We see that $F \circ F$ has four equilibria. The origin and the equilibrium near $x = 0.7$ are unstable, since the slope of $F \circ F$ is clearly

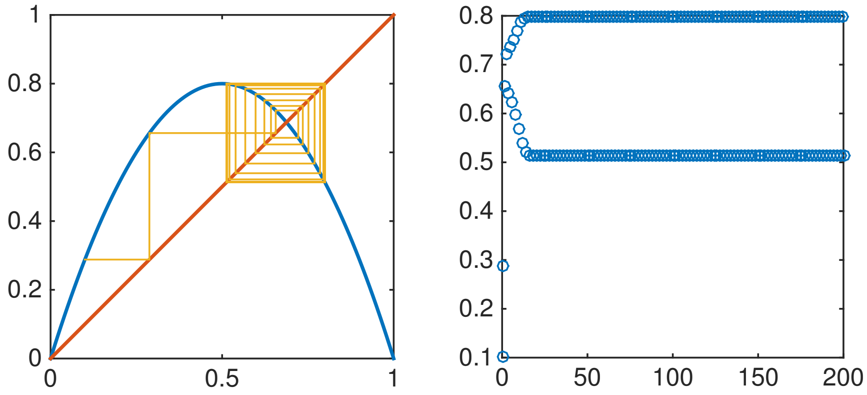

Figure 3.5:

greater than that of $I(x)$ there. The remaining two are stable and correspond to the observed 2-periodic orbit. In the middle plot of Figure 3.6 we see the logistic map $F(x)$ superimposed over the plot of $F \circ F$. The unstable equilibrium of $F \circ F$ is the equilibrium $\alpha_2 = 1 - r^{-1}$ of $F(x)$ which became unstable when we increased r above $r > 3$. In the rightmost plot of Figure 3.6 we show the superimposed functions F and $F \circ F$ for $r = 2.8$. The 2-periodic orbit arises when the slope of F increases above unity.

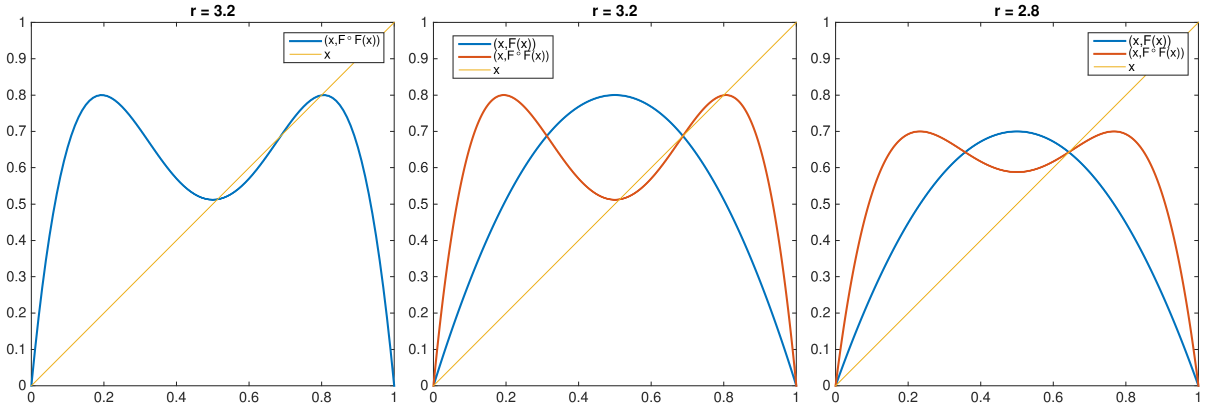

Figure 3.6:

As r is increased further to $r = 3.5$, the slope of $F \circ F$ at β_1 and β_2 again exceeds unity and the 2-periodic orbit becomes unstable. At this point a 4-periodic orbit appears, as shown in Figure 3.7.

The function $F^4(x) = F \circ F \circ F \circ F(x)$ possesses 8 equilibria: the unstable fixed points α_1 and α_2 of F, the unstable 2-periodic orbit β_1 and β_2, and the new 4-periodic orbit:

$$F(\gamma_1) = \gamma_2, \quad F(\gamma_2) = \gamma_3, \quad F(\gamma_3) = \gamma_4, \quad F(\gamma_4) = \gamma_1.$$

All eight equilibria of F^4 can be seen in the left plot of Figure 3.8. In the right plot, the functions F and $F \circ F$ are superimposed, illustrating the shared equilibria.

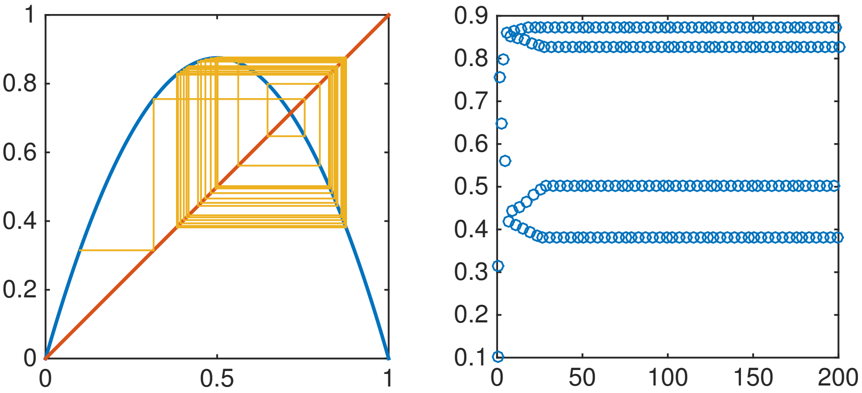

Figure 3.7:

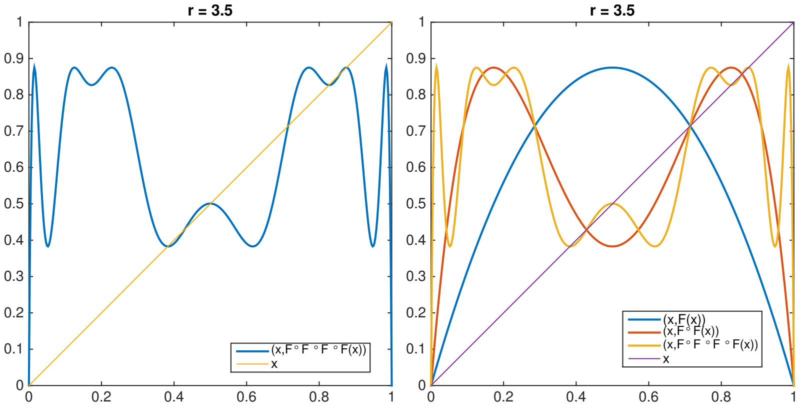

Figure 3.8:

Continuing to increase r in this way we find the 4-periodic orbit giving rise to an 8-periodic orbit, etc. An important theorem is that of Sharkovskii. Consider an ordering of the natural numbers as follows

$$\begin{array}{rrrrr} 3, & 5, & 7, & 9, & \cdots \\ 2\cdot 3, & 2\cdot 5, & 2\cdot 7, & 2\cdot 9, & \cdots \\ 2^2\cdot 3, & 2^2\cdot 5, & 2^2\cdot 7, & 2^2\cdot 9, & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots \\ \cdots & 2^3, & 2^2, & 2^1, & 1, \end{array}$$

where the first row is the odd numbers, the second row is two times the odd numbers, the third row is 2^2 times the odd numbers, etc., and the last row is all powers of 2 in decreasing magnitude. Then Sharkovskii's theorem states that any recursion that has a p-periodic orbit, also has a q-periodic orbit for every q that appears later in this ordering than p. In particular, a recursion that possesses a 3-periodic orbit, possesses a q-periodic orbit for every natural number q. Our examples above illustrate the last row of the ordering, i.e. the powers of 2. In the case $r = 3.5$ we saw that accompanying the 4-periodic orbit were unstable 2-periodic and 1-periodic orbits.

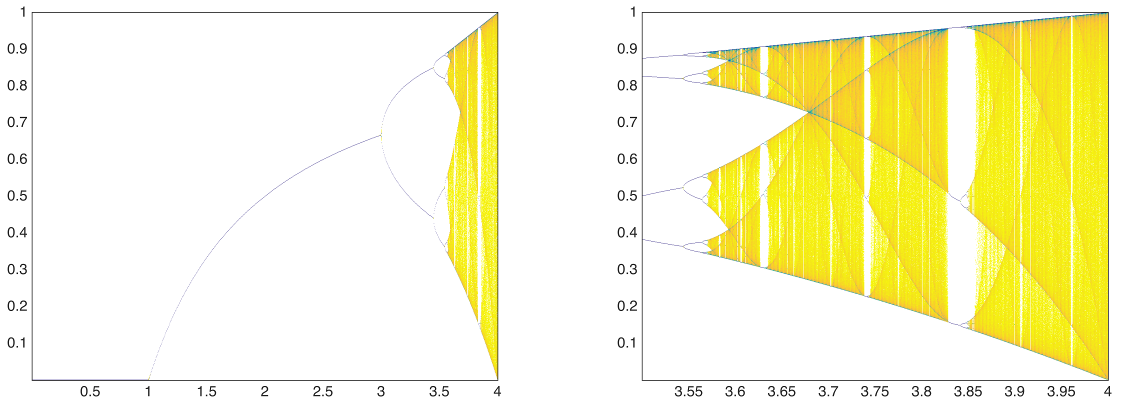

Figure 3.9: Bifurcation diagram for the Logistic map (3.1). Left: the entire parameter range $r \in [0, 4]$. Right: enlargement of the range $r \in [3.5, 4]$.

3.1.4 Chaotic maps

A map $F : \mathcal{D} \to \mathcal{D}$ is said to be *chaotic* or exhibit *chaos* if

1. For every x_0 there exist small perturbations that lead to large deviations of the orbits.
2. Every neighborhood of every $x \in \mathcal{D}$ there exists an x_0 such that the orbit $\{x_n\}$ is periodic.
3. There exists an orbit $\{x_n\}$ whose iterates come arbitrarily close to every $x \in \mathcal{D}$.

To demonstrate this definition, let us examine the map

$$x_{n+1} = 10\, x_n \bmod 1.$$

Given the decimal representation of a real number on $\mathcal{D} = [0, 1]$, the map just shifts the decimal one place to the right and removes the whole number part. For instance, given a decimal representation

$$\begin{aligned} x_0 &= 0.275783976844090\,87 \\ x_1 &= \quad 0.75783976844090873 \\ x_2 &= \qquad 0.57839768440908731 \\ x_3 &= \qquad\quad 0.78397684409087314 \end{aligned}$$

We check that this map satisfies all of the above criteria and is therefore a chaotic map.

First consider the sequence that arises by perturbing x_0 in the 13th digit:

$$\begin{aligned}
x_0 &= 0.27578397684409087 \\
\tilde{x}_0 &= 0.27578397684423296 \\
x_1 &= \quad 0.75783976844090873 \\
\tilde{x}_1 &= \quad 0.75783976844232963 \\
x_2 &= \qquad 0.57839768440908731 \\
\tilde{x}_2 &= \qquad 0.57839768442329638 \\
&\vdots \\
x_{12} &= 0.09087314399975 \\
\tilde{x}_{12} &= 0.23296388576409
\end{aligned}$$

We see that a perturbation of 10^{-13} leads to a deviation in the first digit after only 12 iterations. To illustrate criterion 2, we can construct a periodic orbit arbitrarily close to any initial condition by introducing a repeating decimal. For the initial condition x_0 above, for instance, a periodic orbit that differs from x_0 only in the 11th digit is:

$$\tilde{x}_0 = 0.2757839768\ 2757839768\ 2757839768\ldots$$

To illustrate criterion 3, we construct an orbit that passes arbitrarily close to every point in the unit interval $\mathcal{D} = [0, 1]$. To do so, we introduce an ordering on all rational numbers in $\mathcal{D}$ having a finite decimal representation as follows:

$$\begin{array}{lllll}
0.0 & 0.1 & 0.2 & \ldots & 0.9 \\
0.00 & 0.01 & 0.01 & \ldots & 0.99 \\
0.000 & 0.001 & 0.002 & \ldots & 0.999 \\
\ldots & & & &
\end{array}$$

Subsequently we construct an initial condition by concatenating the digits following the decimal to obtain

$$x_0 = 0.0123456789000102030\cdots 9899000001002 \cdots 998999 \cdots$$

The iterates of our map eventually approximate any real number to arbitrarily many digits. Consequently, the map satisfies all three conditions and is chaotic.

The logistic map is also chaotic for certain values of r.

3.2 Equilibria and stability for differential equations

Analogous to fixed points of maps, the simplest solutions of differential equations are *equilibria*. The study of equilibria and the behavior of solutions near equilibria are of fundamental importance to our understanding of model dynamics.

Definition 4 *A point $y^* \in \mathcal{D}$ satisfying $f(y^*) = 0$ is an equilibrium of the autonomous differential equation $y' = f(y)$. In particular, the initial value problem with $y(0) = y^*$ has solution $y(t) \equiv y^*$.*

Example. The differential equation (2.13) has a unique equilibrium at the origin. The Verhulst model (1.2) has two equilibria according to the two solutions of

$$f(P) = rP(1 - P/P_c) = 0 \qquad \Rightarrow \qquad P^* = 0 \text{ or } P^* = P_c.$$

These correspond to the biological states of extinction and full saturation of the environment, respectively.

Just as with maps, we shall be particularly interested in determining cases in which solutions that start in the neighborhood of an equilibrium stay there, and in which solutions that start in a neighborhood of an equilibrium leave that neighborhood.

Definition 5 *The equilibrium y^* is Lyapunov stable equilibrium if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that*

$$\|y(t) - y^*\| < \varepsilon, \forall t > 0, \quad \textit{whenever} \quad \|y(0) - y^*\| < \delta.$$

Otherwise, y^* is an unstable equilibrium. The equilibrium y^* is asymptotically stable *if, in addition, there exists a $\delta^* > 0$ such that $\lim_{t\to\infty} \|y(t) - y^*\| = 0$ whenever $\|y(0) - y^*\| < \delta^*$.*

Example. For the linear differential equation (2.9), the origin is stable if $a < 0$, since we can choose $\delta = \varepsilon$. If $a > 0$, the origin is unstable:

$$y' = ay, \qquad y^* = 0 \quad \text{is} \quad \begin{cases} \text{stable}, & a < 0 \\ \text{unstable}, & a > 0. \end{cases} \tag{3.2}$$

For the Verhulst model (1.2) with $r > 0$, we have $P'(t) < 0$ for $P < 0$ or $P > P_c$ and $P'(t) > 0$ for $0 < P < P_c$. Consequently, the equilibrium at $P = 0$ is unstable and the equilibrium at $P = P_c$ is stable.

Let us write the solution in the neighborhood of an equilibrium as $y(t) = y^* + \eta(t)$, where $\eta(t)$ represents the *perturbation from equilibrium.* Substituting this solution into the differential equation and expanding f in a Taylor series around y^* yields

$$\frac{d}{dt}(y^* + \eta(t)) = f(y^* + \eta(t)) = f(y^*) + f'(y^*)\eta(t) + \frac{1}{2}f''(y^*)\eta(t)^2 + \cdots .$$

We assume that η is small and disregard terms of order η^2 and higher.[1] Then, since $\frac{d}{dt}y^* = f(y^*) = 0$, the above expression leads to the linear differential equation governing the perturbation

$$\frac{d\eta}{dt} = f'(y^*)\eta.$$

[1] In the stable case $\eta(t) \to 0$ and this approximation is justified. In the unstable case, the linearization fails.

Since $f'(y^*)$ is constant, this differential equation is of the form (2.9), and the stability of the origin is determined by (3.2). If $f'(y^*) < 0$, then the perturbation $\eta(t)$ will decay to zero, and y^* is stable. If $f'(y^*) > 0$, then the perturbation $\eta(t)$ will grow, and the assumption that η is small will no longer hold. If $f'(y^*) = 0$, then we can say nothing about the stability based on these considerations. The reasoning is summarized in the following.

Theorem 3.2.1 *An equilibrium y^* of a scalar differential equation $y' = f(y)$ is asymptotically stable if $f'(y^*) < 0$ and unstable if $f'(y^*) > 0$.*

A similar statement holds for systems of differential equations. However, we must determine how to properly generalize the idea of the "sign of the derivative" to higher dimensions. More on this in Chapter 4.

The essence of our definition of stability above is that an equilibrium is stable when solutions emanating from a neighborhood of the equilibrium remain in a neighborhood of the equilibrium, and that this statement continues to hold as the neighborhoods are taken smaller and smaller.

For the problem (2.9), the solution is $y(t) = y_0 \exp(at)$. If $a < 0$, then

$$|y(t)| = e^{at}|y(0)| < |y(0)|,$$

and more generally

$$|y(t)| < |y(s)|, \quad \forall t > s.$$

Consequently, the interval $[-\delta, \delta]$ is invariant for any $\delta > 0$. Conversely, if $a > 0$, then $|y(t)| = \exp(at)|y(0)| > |y(0)|$, $|y(t)| > |y(s)|$ for all $t > s$. No finite interval is invariant.

3.3 Equilibria and stability for Euler's method

In this section we give a first impression of the stability of Euler's method. This subject will be treated in more generality in Chapter 4, and there we will also discuss other numerical methods.

Suppose that $f(y_n) = 0$. Then Euler's method gives

$$y_{n+1} = y_n + \Delta t\, f(y_n) = y_n,$$

i.e., such a y_n is a fixed point of Euler's method. Conversely, suppose y^* is a fixed point of Euler's method. Then,

$$y^* = y^* + \Delta t\, f(y^*) \qquad \Rightarrow \qquad f(y^*) = 0,$$

that is, y^* is an equilibrium of the vector field f. In conclusion, y^* is a fixed point of Euler's equation if and only if y^* is an equilibrium of the differential equation being solved.

Applying Theorem 3.1.1 we see find that y^* is asymptotically stable if

$$|1 + \Delta t f'(y^*)| < 1 \qquad \Rightarrow \qquad -2 < \Delta t f'(y^*) < 0.$$

The equilibrium y^* is asymptotically stable for the differential equation under Theorem 3.2.1 if $f'(y^*) < 0$. Under this condition, y^* is an asymptotically stable equilibrium of Euler's method if the stepsize satisfies the condition

$$0 < \Delta t < \frac{2}{-f'(y^*)}.$$

Be careful! This condition is *only* applicable for scalar, differential equations. The general case will be handled in Chapter 4. Nevertheless, an important observation is that in the case of Euler's method, stability depends on both the stability of the underlying differential equation *and* the step size Δt.

Chapter 4

Models in higher dimensions

4.1 Time series and phase space

A solution to an autonomous system of differential equations in d dimensions

$$\frac{dy}{dt} = f(y), \qquad y(t) \in \mathcal{D} \subset \mathbf{R}^d \tag{4.1}$$

is a vector-valued function of time $y(t)$. For each time t, we can think of the components of y, i.e. $y^{(1)}(t)$, $y^{(2)}(t)$, $\dots$, $y^{(d)}(t)$ as the coordinates of a point in d-dimensional space:

$$(y^{(1)}(t), y^{(2)}(t), \dots, y^{(d)}(t)) \in \mathbf{R}^d$$

As the solution evolves in time, the point moves in $\mathcal{D} \subset \mathbf{R}^d$, tracing out a curve. Such a solution curve is referred to as a *trajectory* or *orbit*. Due to the existence and uniqueness theorem, there is precisely one trajectory passing through each point in $\mathcal{D}$, and the curves may not cross each other, as the intersection of two distinct trajectories at a point would imply an initial condition with two solutions. In this context we refer to $\mathbf{R}^d$ as the *phase space*. For the specific case $d = 2$, we refer to $\mathbf{R}^2$ as the *phase plane*. We can get a good idea of how solutions to (4.1) behave by plotting several solutions in the phase plane. Such a plot is called a *phase portrait*. A function $f(y) : \mathbf{R}^d \to \mathbf{R}^d$ assigns to each point y in the phase plane a vector $f(y) \in \mathbf{R}^d$, a *vector field*.

Example. A pendulum of length ℓ makes an angle $\theta(t)$ with the downward vertical. The gravitational acceleration is g. The equation for pendulum motion is a second order differential equation

$$\frac{d^2\theta}{dt^2} = -\frac{g}{\ell}\sin\theta.$$

With the angular velocity $v = \frac{d\theta}{dt}$ the system is written in first order form

$$\frac{d\theta}{dt} = v, \qquad \frac{dv}{dt} = -\frac{g}{\ell}\sin\theta.$$

Some solutions (with $g/\ell = 1$) as functions of time t and in the phase plane are plotted in Figure 4.1.

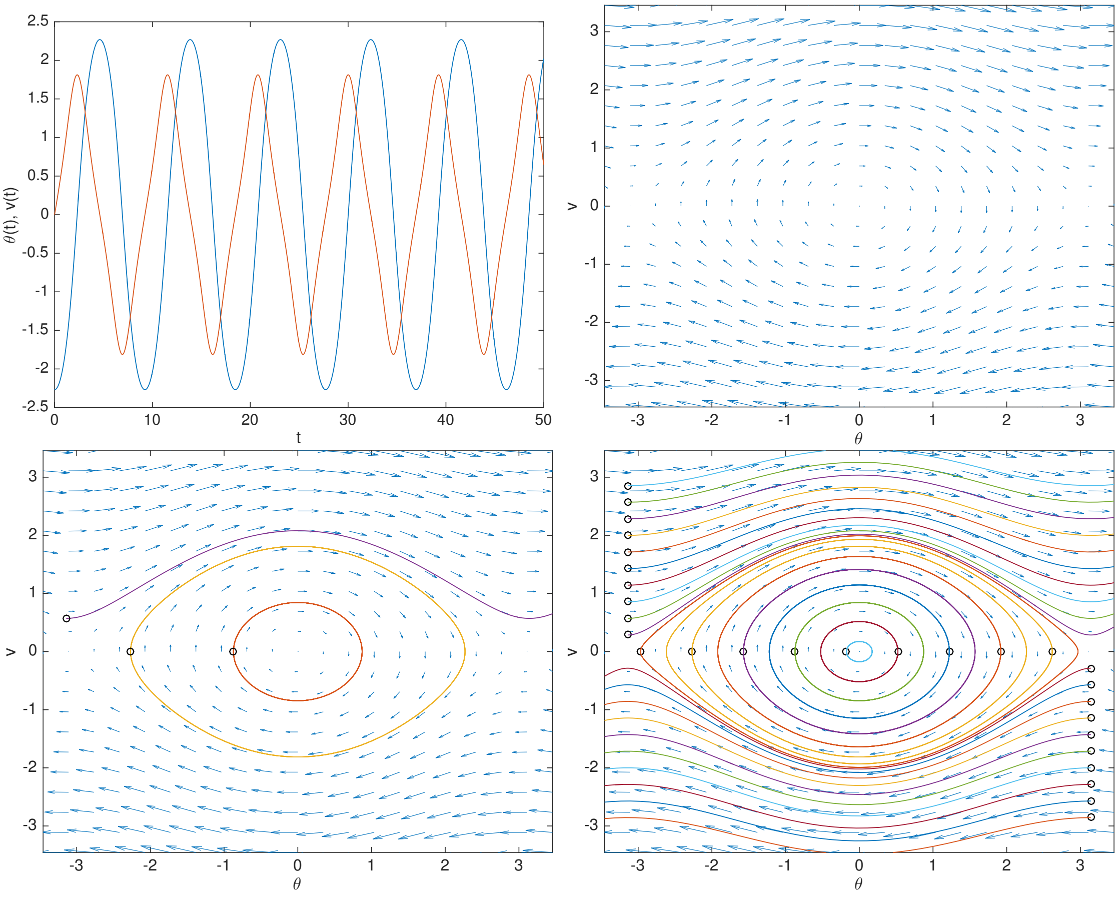

Figure 4.1: Solutions of the pendulum: (top left) time series solution showing $\theta(t)$ and $v(t)$ as functions of t; (top right) the vector field $f(\theta, v) = (v, -\sin\theta)^T$; (bottom left) some phase trajectories (θ, v); (bottom right) phase portrait. Initial conditions are identified with a circle.

4.2 Linear recursions in $\mathbf{R}^d$.

To understand recursions in $\mathbf{R}^d$ for $d > 1$, it is instructive to first study linear recursions. However, linear recursions are also important in their own right: we will meet an important class of linear recursions in the form of Markov chains in Chapter 6.

4.2.1 Eigenvector decomposition

A general linear recursion in $\mathbf{R}^d$, $d > 1$ takes the form

$$x_{n+1} = Ax_n,$$

where A is a $d \times d$ matrix. The solution of this recursion can easily be determined by induction. It is

$$x_n = A^n x_0.$$

We suppose that A has d independent eigenvectors $v_i \in \mathbf{C}^d$, $i = 1, \ldots, d$, each with corresponding eigenvalue $\lambda_i \in \mathbf{C}$, satisfying

$$Av_i = \lambda_i v_i, \quad i = 1, \ldots, d. \tag{4.2}$$

(Recall that the eigenvalues and eigenvectors of a generic matrix A are complex even when A is a real matrix.) Since the v_i can be chosen independently, every vector in $\mathbf{R}^d$ and in particular x_0 can be written as a linear combination of the v_i:

$$x_0 = \gamma_1 v_1 + \gamma_2 v_2 + \cdots + \gamma_d v_d, \qquad \gamma_i \in \mathbf{C}.$$

Note that $A^n v_i = \lambda_i^n v_i$. Consequently, the solution to the recursion is

$$x_n = A^n x_0 = \gamma_1 \lambda_1^n v_1 + \gamma_2 \lambda_2^n v_2 + \cdots + \gamma_d \lambda_d^n v_d$$

We say that λ_1 is a *dominant eigenvalue* if $|\lambda_j| < |\lambda_1|$, for all $j > 1$. Suppose A has dominant eigenvalue λ_1. Then we can rewrite the solution above as

$$x_n = \gamma_1 \lambda_1^n \left(v_1 + \frac{\gamma_2}{\gamma_1} \frac{\lambda_2^n}{\lambda_1^n} v_2 + \cdots + \frac{\gamma_d}{\gamma_1} \frac{\lambda_d^n}{\lambda_1^n} v_d \right)$$

Since λ_1 is dominant, we find[1]

$$\lim_{n \to \infty} \frac{\lambda_j^n}{\lambda_1^n} = \lim_{n \to \infty} \left(\frac{\lambda_j}{\lambda_1} \right)^n = 0.$$

[1] Write $\lambda_j = r_j e^{i\theta_j}$. Then $\frac{\lambda_j}{\lambda_1} = \frac{r_j}{r_1} e^{i(\theta_j - \theta_1)}$, where $r_j < r_1$ and $|e^{i(\theta_j - \theta_1)}| = 1$. Consequently, the quotient has modulus less than unity and $\lim_{n \to \infty} (\frac{\lambda_j}{\lambda_1})^n = 0$.

Consequently,

$$x_n \approx \gamma_1 \lambda_1^n v_1,$$

for large n. Specifically, the vector x_n converges to the direction of the dominant eigenvector v_1. Furthermore, if $|\lambda_1| < 1$, then $\lim_{n\to\infty} |\lambda_1|^n = 0$, and the origin is a stable equilibrium. If $\lambda_1 = 1$, then the vector $\gamma_1 v_1$ is a stable equilibrium (i.e. $x_n \to 0$). If $|\lambda_1| > 1$, then the magnitude of x_n grows unbounded, but in the biological context we may speak of a stable population distribution since the ratio of individuals in any two population groups is constant, equal to the ratio of the corresponding elements of the eigenvector.

Note that in the biological context (at least) the model is only sensible if the iterates x_n are nonnegative vectors (that is, each element of the vector x_n must be nonnegative). In particular this means that $\gamma_1 \lambda_1^n v_1$ must be nonnegative, Since $\gamma_1 v_1$ is a constant vector, it follows that λ_1 must be real and positive, and that it must be possible to choose $\gamma_1 v_1$ to be nonnegative as well.

4.3 Linear differential equations in $\mathbf{R}^d$

Let us consider a linear initial vaue problem in $\mathbf{R}^d$, $d > 1$

$$\frac{dy}{dt} = Ay, \qquad y(0) = y_0, \tag{4.3}$$

where A is again a $d \times d$ real matrix. We again suppose that A possesses d independent eigenvectors $v_i \in \mathbf{C}^d$, $i = 1, \ldots, d$, each with associated eigenvalue $\lambda_i \in \mathbf{C}$ satisfying (4.2). Since the v_i form a basis, we can expand any vector $y(t)$ as

$$y(t) = \gamma_1(t) v_1 + \cdots + \gamma_d(t) v_d \tag{4.4}$$

uniquely in terms of the coefficients $\gamma_j(t)$, $j = 1, \ldots, d$. Substituting (4.4) into both sides of (4.3) we find

$$\begin{aligned}\frac{d\gamma_1}{dt} v_1 + \cdots + \frac{d\gamma_d}{dt} v_d &= \gamma_1(t) A v_1 + \cdots + \gamma_d(t) A v_d,\\ &= \gamma_1(t)\lambda_1 v_1 + \cdots + \gamma_d(t)\lambda_d v_d,\end{aligned}$$

Gathering terms yields

$$\left(\frac{d\gamma_1}{dt} - \lambda_1\gamma_1\right) v_1 + \cdots + \left(\frac{d\gamma_d}{dt} - \lambda_d\gamma_d\right) v_d = 0.$$

The vectors are independent so all of the coefficients must equal zero, and consequently,

$$\frac{d\gamma_i}{dt} = \lambda_i \gamma_i, \quad i = 1, \ldots, d. \tag{4.5}$$

In particular, from the relation (4.4) we can discern the quantities $\gamma_j(0)$, $j = 1, \ldots, d$, from y_0. The solutions of (4.5) are

$$\gamma_i(t) = e^{\lambda_i t}\gamma_i(0), \quad i = 1, \ldots, d.$$

Consequently the solution of the differential equation is

$$y(t) = \gamma_1(0)e^{\lambda_1 t}v_1 + \cdots + \gamma_d(0)e^{\lambda_d t}v_d.$$

Clearly the origin is an equilibrium. Since A is invertible, it is also the unique equilibrium. Suppose, without loss of generality, that the eigenvalues are ordered such that $\operatorname{Re}\lambda_1 > \operatorname{Re}\lambda_2 \geq \operatorname{Re}\lambda_3 \geq \cdots \geq \operatorname{Re}\lambda_d$. We rewrite the above relation as

$$y(t) = \gamma_1(0)e^{\lambda_1 t}\left(v_1 + \frac{\gamma_2}{\gamma_1}e^{(\lambda_2-\lambda_1)t}v_2 + \cdots + \frac{\gamma_d}{\gamma_1}e^{(\lambda_d-\lambda_1)t}v_d\right).$$

Clearly[2]

$$\lim_{t\to\infty} e^{(\lambda_j-\lambda_1)t} = 0, \quad j > 1.$$

Consequently,

$$y(t) \approx \gamma_1(0)e^{\lambda_1 t}v_1, \quad t \gg 0.$$

That is, the vector $y(t)$ tends toward the direction of v_1 for large t. Furthermore, if $\operatorname{Re}\lambda_1 < 0$, then $\lim_{t\to\infty} y(t) = 0$, and the origin is stable. If $\operatorname{Re}\lambda_1 = 0$, then the norm of $y(t)$ is constant, and the origin is Lyapunov stable but not asymptotically so. Finally, if $\operatorname{Re}\lambda_1 > 0$, then $y(t)$ grows in magnitude without bound, and the origin is unstable.

4.4 Nonlinear recursions in $\mathbf{R}^d$

In this section we extend the stability results for linear recursions to nonlinear recursions, and provide an important application.

4.4.1 Multivariate Taylor series and the Jacobian matrix

For the material that follows, we need the multivariate form of Taylor series. For simplicity, we consider first a scalar function $f(x,y) : \mathbf{R}^2 \to \mathbf{R}$ of two variables and a perturbation (ξ, η). Expanding first with respect to x yields

$$f(x+\xi, y+\eta) = f(x, y+\eta) + \frac{\partial f}{\partial x}(x, y+\eta)\xi + \frac{1}{2!}\frac{\partial^2 f}{\partial x^2}(x, y+\eta)\xi^2 + \cdots$$

[2] To see that limit holds for negative real part, let $z = -x + iy$ where x and y are positive real numbers. Then $e^{zt} = \lim_{t\to\infty} e^{-xt}e^{iyt} = e^{-xt}(\cos yt + i \sin yt)$ and consequently $\lim_{t\to\infty} e^{zt} = 0$.

Next each term on the right is expanded with respect to y:

$$\begin{aligned}f(x, y+\eta) &= f(x,y) + \frac{\partial f}{\partial y}(x,y)\,\eta + \frac{1}{2!}\frac{\partial^2 f}{\partial y^2}(x,y)\,\eta^2,\\ \frac{\partial f}{\partial x}(x,y+\eta)\,\xi &= \frac{\partial f}{\partial x}(x,y)\,\xi + \frac{\partial^2 f}{\partial x \partial y}(x,y)\,\xi\,\eta + \cdots,\\ \frac{\partial^2 f}{\partial x^2}(x,y+\eta)\,\xi^2 &= \frac{\partial^2 f}{\partial x^2}(x,y)\,\xi^2 + \cdots.\end{aligned}$$

Combining these and suppressing the coordinates (x, y) gives

$$f(x+\xi, y+\eta) = f + \frac{\partial f}{\partial x}\xi + \frac{\partial f}{\partial y}\eta + \frac{1}{2!}\frac{\partial^2 f}{\partial x^2}\xi^2 + \frac{\partial^2 f}{\partial x \partial y}\xi\,\eta + \frac{1}{2!}\frac{\partial^2 f}{\partial y^2}\eta^2 + \cdots$$

With a lot of patience, one can work out the general form of Taylor series for a vector function $f(y) : \mathbf{R}^d \to \mathbf{R}^d$:

$$f(y+\eta) = f(y) + Df(y)\,\eta + \cdots \tag{4.6}$$

Here, the *Jacobian matrix* $Df(y)$ associated with a vector function of a vector variable $f(y) : \mathbf{R}^d \to \mathbf{R}^d$, is the matrix of partial derivatives:

$$Df(y) = \begin{bmatrix} \frac{\partial f^{(1)}}{\partial y^{(1)}} & \frac{\partial f^{(1)}}{\partial y^{(2)}} & \cdots & \frac{\partial f^{(1)}}{\partial y^{(d)}} \\ \frac{\partial f^{(2)}}{\partial y^{(1)}} & \frac{\partial f^{(2)}}{\partial y^{(2)}} & \cdots & \frac{\partial f^{(2)}}{\partial y^{(d)}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f^{(d)}}{\partial y^{(1)}} & \frac{\partial f^{(d)}}{\partial y^{(2)}} & \cdots & \frac{\partial f^{(d)}}{\partial y^{(d)}} \end{bmatrix}. \tag{4.7}$$

The Jacobian is the appropriate generalization of the derivative $f'(y)$ required for stability of nonlinear maps and differential equations.

Example. The Lorenz attractor is conceptual model of fluid convection. It is commonly used to as a differential equation exhibiting chaos. The Lorenz model is

$$\frac{dx}{dt} = \sigma(y-x), \tag{4.8}$$

$$\frac{dy}{dt} = x(\rho - z) - y, \tag{4.9}$$

$$\frac{dz}{dt} = xy - \beta z, \tag{4.10}$$

where σ, ρ and β are positive constants. (E.N. Lorenz studied the case $\rho = 28$, $\sigma = 10$, and $\beta = 8/3$ for which all trajectories approach a so-called "strange attractor"). The Jacobian matrix for the Lorenz model is obtained by taking partial derivatives of the right hand sides of the above equations:

$$Df(x,y,z) = \begin{bmatrix} -\sigma & \sigma & 0 \\ \rho - z & -1 & -x \\ y & x & -\beta \end{bmatrix}. \tag{4.11}$$

We will make use of this matrix shortly.

4.4.2 Fixed points and stability

Just as in the scalar case, an equilibrium or fixed point of a recursion in $\mathbf{R}^d$ is a point $\alpha \in \mathcal{D}$ satisfying

$$\alpha = F(\alpha).$$

To establish the stability of a fixed point, we again introduce a local representation

$$x_n = \alpha + \eta_n,$$

substitute this into the recursion and expand in Taylor series

$$\alpha + \eta_{n+1} = F(\alpha + \eta_n) = F(\alpha) + DF(\alpha)\eta_n + \dots,$$

where $DF(x)$ denotes the Jacobian matrix (4.7) of partial derivatives of $F(x)$. Ignoring terms of quadratic and higher order in η_n, this reduces to

$$\eta_{n+1} = DF(\alpha)\eta_n.$$

Again the eigenvalues of DF can be used to establish stability.

Theorem 4.4.1 *Let $F(x)$ be continuously differentiable at a fixed point $\alpha = F(\alpha)$. Let λ_j, $j = 1, \dots, d$, be the eigenvalues of the Jacobian $DF(\alpha)$ of $F(x)$ at α. Then*

- *α is asymptotically stable if $|\lambda_j| < 1$ for all j,*
- *α is unstable if $|\lambda_j| > 1$ for some j.*

The theorem says nothing about the case $|\lambda_j| \le 1$, $j = 1, \cdots, d$, with $|\lambda_k| = 1$ for some $1 \le k \le d$. Other analysis must be used to determine stability in these cases.

4.4.3 An application: Newton's method

An important example of a nonlinear recursion is Newton's method (or the Newton-Raphson method) for finding a zero (root) of a nonlinear system of equations. Suppose that we wish to find a zero of $G(x) : \mathbf{R}^d \to \mathbf{R}^d$, where G is continuously differentiable. Further suppose that after n iterations of Newton's method we have an estimate $x_n = x^* - \delta_n$ of a root $G(x^*) = 0$ of G. We do not know δ_n explicitly, or we would know x^*. However we can approximate it using Taylor series,

$$0 = G(x^*) = G(x_n + \delta_n) + DG(x_n)\delta_n + \cdots,$$

where the ellipses denote terms of quadratic and higher order in δ_n. If we assume that δ_n is small and that the tensor of second derivatives is bounded in the neighborhood of x^*, then neglecting these terms is an acceptable approximation, and we can define a new iterate to be $x_{n+1} \approx x_n + \delta_n$, or

$$x_{n+1} = x_n - (DG(x_n))^{-1}G(x_n). \tag{4.12}$$

Newton's method is a very powerful means of finding a root, because it can be shown that the error $e_n = \|x^* - x_n\|$ converges quadratically in the limit of large n: $e_{n+1} = e_n^2$. On the other hand, the rapid convergence must be weighed against the computational cost of solving a system of linear algebraic equations with the Jacobian matrix in each iteration.

4.5 Nonlinear differential equations in $\mathbf{R}^d$

Let y^* be an equilibrium of the autonomous differential equation $y' = f(y)$, i.e. $f(y^*) = 0$. Let $\eta(t)$ denote a perturbation solution and $y(t) = y^* + \eta(t)$. Inserting this solution in the differential equation and expanding in Taylor series (4.6) yields

$$\frac{dy}{dt} = \frac{d}{dt}(y^* + \eta(t)) = f(y^* + \eta(t)) = f(y^*) + Df(y^*)\eta(t) + \cdots$$

Ignoring the higher order terms, and noting $\frac{d}{dt}y^* = f(y^*) = 0$, the equation for the perturbation reduces to

$$\frac{d\eta}{dt} = Df(y^*)\eta$$

Since the Jacobian $Df(y^*)$ is a constant matrix, we are back in the setting of Section 4.3. The following general linear stability theorem can be proved:

Theorem 4.5.1 *Let $f(y)$ be continuously differentiable at an equilibirum $f(y^*) = 0$ of the autonomous differential equation $y' = f(y)$. Let λ_j, $j = 1, \ldots, d$, be the eigenvalues of the Jacobian $Df(y^*)$ of $f(y)$ at y^*. Then*

- *y^* is asymptotically stable if $\operatorname{Re}\lambda_j < 0$ for all j,*
- *y^* is unstable if $\operatorname{Re}\lambda_j > 0$ for some j.*

In particular the theorem tells us nothing about the case of imaginary eigenvalues.

Example. The Lorenz model (4.8)–(4.10) has three equilibria:

$$x = y = z = 0, \qquad x = y = \pm\sqrt{\beta(\rho - 1)},\ z = \rho - 1.$$

For instance, at the equilibrium $(0, 0, 0)$, the Jacobian (4.11) becomes

$$Df(0,0,0) = \begin{bmatrix} -\sigma & \sigma & 0 \\ \rho & -1 & 0 \\ 0 & 0 & -\beta \end{bmatrix}.$$

This matrix has all negative eigenvalues if $\rho < 2$. Otherwise it has at least one eigenvalue with positive real part, for which case the origin is an unstable equilibrium.

4.5.1 Summary of stability criteria for maps and differential equations

It is prudent to pause for a moment and compare the stability criteria for equilibria of iterated maps and differential equations.

Given a fixed point x^* of an iterated map, to analyze its stability we compute the Jacobian matrix $DF(x^*)$ and determine its eigenvalues λ_i, $i = 1, \ldots, d$. From Theorem 4.4.1 the fixed point is stable if these all have modulus less than unity:

$$|\lambda_i| < 1, \quad i = 1, \ldots, d.$$

In contrast, if y^* is an equilibrium of a differential equation, to analyze its stability we again compute the Jacobian matrix $Df(y^*)$ and determine its eigenvalues μ_i, $i = 1, \ldots, d$. From Theorem 4.5.1 the equilibrium is stable if all eigenvalues lie strictly in the left half of the complex plane:

$$\mathrm{Re}\,\mu_i < 0.$$

The binding factor is the exponential function which maps the left half-plane into the unit circle, i.e. $|\exp(z)| < 1$ if $\mathrm{Re}\, z < 0$. Essentially this is because the exponential function provides the time-t solution map of a linear differential equation (recall Section 1.4), and our analysis is based on linearization. That is, given the linear differential equation $\frac{dy}{dt} = \mu y$, the time-Δt solution is $y(t + \Delta t) = \exp(\Delta t \mu) y(t)$. Viewing the solution as an iterated map, its "eigenvalue" is precisely $\lambda = \exp(\Delta t \mu)$. Hence, if μ is in the left half-plane, $|\lambda| < 1$.

This is often a source of some confusion among students new to the subject.

4.6 Application to numerical integrators

4.6.1 Fixed points of numerical methods

In the previous section we have devoted a lot of attention to the equilibria of differential equations and maps. Considering the significance of equilibria, we would like them to be preserved by the numerical method. Given an autonomous differential equation, denote by $\mathcal{F}$ the set of equilibria

$$\mathcal{F} = \{y \in \mathbf{R}^d : f(y) = 0\}.$$

For a numerical method with map $\Psi_{\Delta t}(y)$, the fixed points may depend on Δt as well as f. We denote the set of fixed points of $\Psi_{\Delta t}(y)$ by $\mathcal{F}_{\Delta t}$:

$$\mathcal{F}_{\Delta t} = \{y \in \mathbf{R}^d : \Psi_{\Delta t}(y) = y\}.$$

An important practical question is whether the sets $\mathcal{F}$ and $\mathcal{F}_{\Delta t}$ coincide.

Consider Euler's method (2.16):

$$\Psi_{\Delta t}(y) = y + \Delta t f(y).$$

Suppose $y \in \mathcal{F}$. Then $f(y) = 0$ and therefore $\Psi_{\Delta t}(y) = y$. Thus, $F \subseteq F_{\Delta t}$. Suppose, conversely, that $y \in \mathcal{F}_{\Delta t}$. Then, $\Psi_{\Delta t} y = y = y + \Delta t f(y)$, so $\Delta t f(y) = 0$. It follows that for $\Delta t > 0$, $F_{\Delta t} \subseteq F$ and therefore, for Euler's method, $F_{\Delta t} = F$.

As a second example, consider the "implicit midpoint rule"

$$\frac{y_{n+1} - y_n}{\Delta t} = f\left(\frac{y_{n+1} + y_n}{2}\right).$$

First, suppose y_n and y_{n+1} are such that $(y_{n+1} + y_n)/2$ is an equilibrium. The right hand side evaluates to zero and therefore (for positive Δt) $y_{n+1} = y_n$, i.e. y_n is a fixed point of the method. Thus, $\mathcal{F} \subseteq \mathcal{F}_{\Delta t}$. Conversely, if y_n is a fixed point of the implicit map $\Psi_{\Delta t}$, then $y_{n+1} \equiv y_n$ and the left side evaluates to zero. We are left with $f(y_{n+1}) = f(y_n) = 0$, so again $\mathcal{F}_{\Delta t} \subseteq \mathcal{F}$ and the sets are equivalent.

Next, consider take the "extrapolated" Euler method:

$$y_{n+1} = y_n + \Delta t f\left(y_n + \Delta t f(y_n)\right).$$

It is clear that if $f(y_n) = 0$, then $y_{n+1} = y_n$, so $\mathcal{F} \subseteq \mathcal{F}_{\Delta t}$ here. The converse is not necessarily true, however, as we illustrate next. We apply this method to the Verhulst model (1.2) with $r = 1$:

$$y' = y(1-y)$$

to obtain the map

$$\Psi_{\Delta t}(y) = y + \Delta t\left[y + \Delta t y(1-y)\right]\left[1 - y - \Delta t y(1-y)\right].$$

For $y \in \mathcal{F}_{\Delta t}$ we have

$$\left[y + \Delta t y(1-y)\right]\left[1 - y - \Delta t y(1-y)\right] = 0,$$

so one of the terms in square brackets must equal zero. In the first case, we have

$$y(1 + \Delta t(1-y)) = 0 \quad \Rightarrow \quad \{y = 0 \text{ or } y = 1 + \frac{1}{\Delta t}\},$$

In the second case,

$$1 - y - \Delta t y(1-y) = 0 \quad \Rightarrow \quad \{y = 1 \text{ or } y = \frac{1}{\Delta t}\}.$$

Hence, $F_{\Delta t} = \{0, 1, 1 + \Delta t^{-1}, \Delta t^{-1}\}$. The fixed points $1 + \Delta t^{-1}$ and Δt^{-1}, which are not equilibrium points, are termed *extraneous fixed points*. Note that the extraneous fixed points grow unbounded as the time step is refined.

While the set of fixed points of a numerical method $\mathcal{F}_{\Delta t} = \{y \in \mathbf{R}^d : y = \Psi_{\Delta t}(y)\}$ is not generally the same as the set of fixed points $\mathcal{F} = \{y \in \mathbf{R}^d : f(y) = 0\}$ of the flow map, for many classes of methods, the fixed points of the numerical method are a superset of those of the flow map: $\mathcal{F}_{\Delta t} \supseteq \mathcal{F}$.

Thus $\mathcal{F}_{\Delta t}$ includes all the fixed points of $\mathcal{F}$ but it may have some extraneous ones. Notice that the existence of extraneous fixed points depends not only on the method but on the differential equation we are solving.

You might think that a method which admits extraneous fixed points would be of limited interest, since it is then quite possible that a numerical trajectory converges to a point that is nowhere near the stable equilibria of the original dynamical system. Nonetheless, such methods are used frequently in dynamical systems studies.

The reason that we can use methods like this is that in general (as in the example above), as $\Delta t \to 0$, all the extraneous fixed points tend to infinity. This gives us a simple way to test for extraneous fixed points: we simply vary the timestep and solve the problem repeatedly. The fixed points which stay put regardless of stepsize are genuine.

4.6.2 Linear stability of numerical methods

For a linear system of ODEs

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

where A is a $d \times d$ matrix with d linearly independent eigenvectors, we have seen that the general solution can be written in the compact form

$$y(t) = \sum_{i=1}^{d} \gamma_i(0) e^{\lambda_i t} v_i,$$

where $\lambda_1, \lambda_2, \ldots, \lambda_d$ are the eigenvalues, $v_1, v_2, \ldots, v_d$ are the corresponding eigenvectors, and $\gamma_1(0), \gamma_2(0), \ldots, \gamma_d(0)$ are coefficients. We have seen that, if all the eigenvalues have negative real part, then the origin is asymptotically stable.

A related statement can be shown to hold for many of the numerical methods in common use. For example, consider Euler's method applied to the linear problem $\frac{dy}{dt} = Ay$:

$$y_{n+1} = y_n + \Delta t A y_n = (I + \Delta t A) y_n.$$

If we let y_n be expanded in the eigenbasis (say $v_1, v_2, \ldots, v_d$), we may write

$$y_n = \alpha_n^{(1)} v_1 + \alpha_n^{(2)} v_2 + \ldots + \alpha_n^{(d)} v_d.$$

If we now apply Euler's method, we find

$$\begin{aligned} y_{n+1} &= (I + \Delta t A)(\alpha_n^{(1)} v_1 + \alpha_n^{(2)} v_2 + \ldots + \alpha_n^{(d)} v_d), \\ &= \alpha_n^{(1)} (I + \Delta t A) v_1 + \alpha_n^{(2)} (I + \Delta t A) v_2 + \ldots + \alpha_n^{(d)} (I + \Delta t A) v_d \end{aligned}$$

Now, since v_i is an eigenvector of A, we have

$$(I + \Delta t A)v_i = v_i + \Delta t A v_i = v_i + \Delta t \lambda_i v_i = (1 + \Delta t \lambda_i)v_i$$

and this implies

$$y_{n+1} = \sum_{i=1}^{d} \alpha_n^{(i)}(1 + \Delta t \lambda_i)v_i.$$

We may also write

$$y_{n+1} = \sum_{i=1}^{d} \alpha_{n+1}^{(i)} v_i,$$

and, comparing these last two equations and using the uniqueness of the basis representation, we must have

$$\alpha_{n+1}^{(i)} = (1 + \Delta t \lambda_i)\alpha_n^{(i)}, \quad i = 1, 2, \ldots, d.$$

It follows from this that the origin is a stable point if $|1 + \Delta t \lambda_i| \leq 1$, $i = 1, 2, \ldots, d$, and is asymptotically stable if $|1 + \Delta t \lambda_i| < 1$, $i = 1, 2, \ldots, d$. The condition for stability can be stated equivalently as requiring that for every eigenvalue λ of A, $\Delta t \lambda$ must lie inside a disk of radius 1 centered at $z = -1$ in the complex plane. This region is sketched below. We call it the *region of absolute stability* of Euler's method.

Thus, given the set of eigenvalues of A, this condition implies a restriction on the maximum stepsize Δt that can be used if the origin is to remain stable for the numerical map. This is illustrated in Figure 4.2.

4.6.3 Stability functions

To develop a general theory, let us consider first the scalar complex case. When applied to $y' = \lambda y$, $\lambda \in \mathbf{C}$, Euler's method has the form

$$y_{n+1} = R(\Delta t \lambda) y_n,$$

where $R(\mu) = I + \mu$.

More generally, a very large class of one-step methods, when applied to the scalar complex equation

$$y' = \lambda y, \qquad \lambda \in \mathbf{C}$$

take the form

$$y_{n+1} = R(\Delta t \lambda) y_n,$$

where $R(\mu)$ is a rational function of μ, i.e. the ratio of two polynomials

$$R(\mu) = P(\mu)/Q(\mu),$$

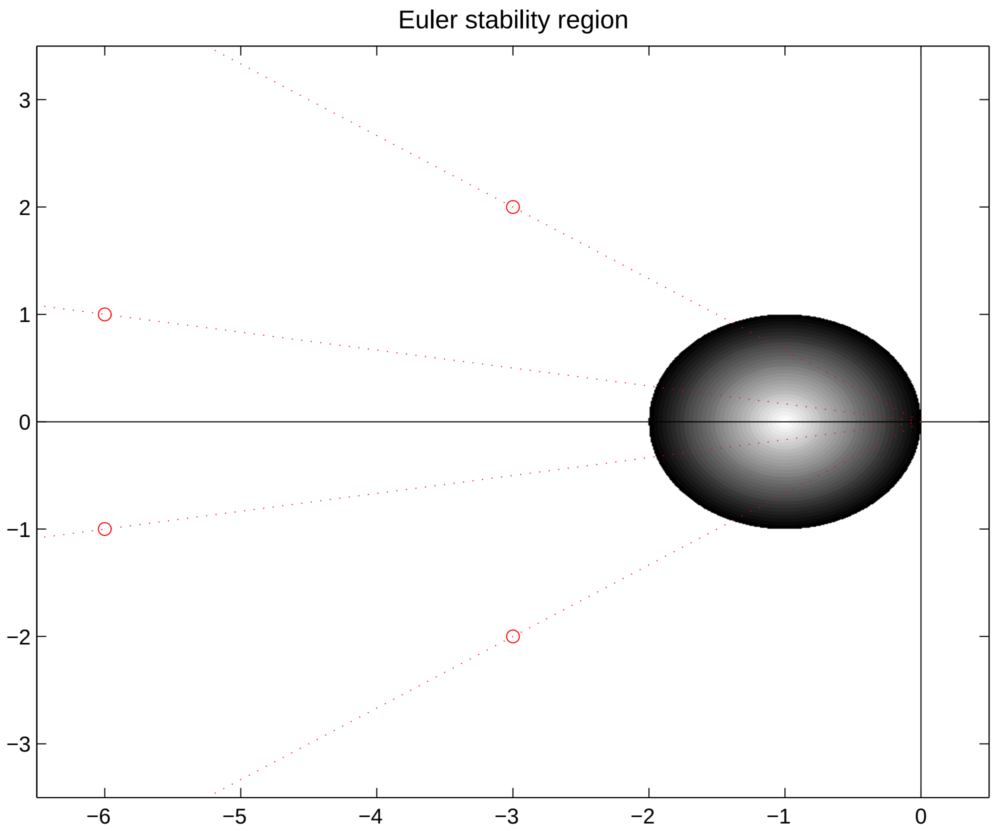

Figure 4.2: The spectrum of A is scaled by Δt. Stability of the origin is recovered if $\Delta t\lambda$ is in the region of absolute stability $|1+z|<1$ in the complex plane.

where P and Q are two polynomials. R is called the *stability function* of the one-step method.

For example, consider the trapezoidal rule (2.21):

$$y_{n+1} = y_n + \frac{\Delta t}{2}(\lambda y_n + \lambda y_{n+1}).$$

Solving for y_{n+1} gives

$$y_{n+1} = \frac{1-\frac{\Delta t}{2}\lambda}{1+\frac{\Delta t}{2}\lambda} y_n,$$

from which follows that $R(\mu) = \frac{1+\mu/2}{1-\mu/2}$.

Theorem 4.6.1 *Given the differential equation $y' = Ay$, where the $d \times d$ matrix A has d independent eigenvectors and the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_d$, consider applying a one step method. The method has a stable (asymptotically stable) fixed point at the origin when applied to*

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

if and only if the same method has a stable (asymptotically stable) equilibrium at the origin when applied to each of the scalar differential equations

$$\frac{dz}{dt} = \lambda_i z, \quad \textit{for all } i.$$

Since we know that a one-step method applied to $\frac{dz}{dt} = \lambda z$ can be written

$$z_{n+1} = R(\Delta t \lambda) z_n$$

we have the following corollary:

Corollary 4.6.1 *Consider a linear differential equation $\frac{dy}{dt} = Ay$ with diagonalizable matrix A. Let a one-step method be given with stability function R. The origin is stable for the numerical method applied to $\frac{dy}{dt} = Ay$ (at stepsize Δt) if and only if*

$$|R(\mu)| \le 1$$

for all $\mu = \Delta t\lambda$, λ an eigenvalue of A.

4.6.4 Stability regions

Evidently the key issue for understanding the long-term dynamics of one-step methods near fixed points concerns the region where $\hat{R}(\mu) = |R(\mu)| \le 1$. This is what we call the *stability region* of the numerical method. Let us examine a few stability functions and regions:

Euler's Method:

$$\hat{R}(\mu) = |1 + \mu|$$

The stability region is the set of points such that $\hat{R}(\mu) \le 1$. The condition

$$|1 + \mu| \le 1$$

means μ lies inside of a disc of radius 1, centred at the point -1.

Trapezoidal rule: the stability region is the region where:

$$\hat{R}(\mu) = \left|\frac{1 + \mu/2}{1 - \mu/2}\right| \le 1.$$

This occurs when

$$|1 + \mu/2| \le |1 - \mu/2|,$$

or, when $\mu/2$ is closer to -1 than to 1, which is just the entire left complex half-plane.

Another popular method is the *implicit Euler* method,

$$y_{n+1} = y_n + h\lambda y_{n+1}$$

$$\hat{R}(\mu) = |1-\mu|^{-1}.$$

which means the stability region is the *exterior* of the disk of radius 1 centered at 1 in the complex plane. These are some simple examples. All three of these are graphed in Figure 4.3.

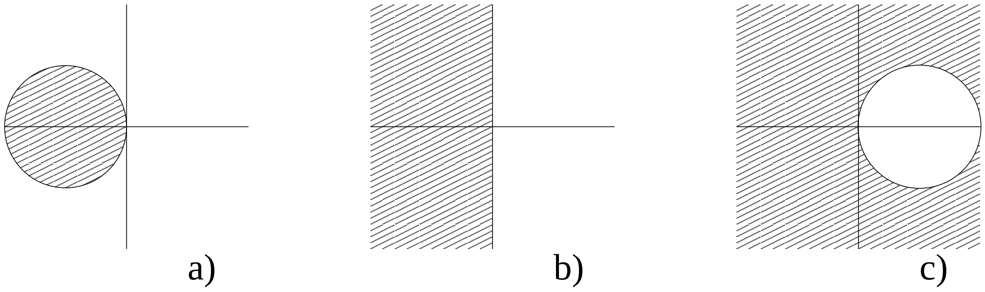

Figure 4.3: Stability Regions: (a) Euler's method, (b) trapezoidal rule, (c) implicit Euler

What do these diagrams tell us? Consider first a scalar differential equation $\frac{dx}{dt} = \lambda x$, with possibly complex λ. We know that for the differential equation, the origin is stable for λ lying in the left half plane. For a numerical method, the origin is stable if $\Delta t\, \lambda$ lies in the stability region. This is not usually going to happen independent of Δt, even if λ lies in the left half plane. For the methods seen above, this is true of trapezoidal rule and implicit Euler. For Euler's method the origin is only stable for the numerical method provided Δt is suitably restricted. Note that, in the case of Euler's method, if λ lies strictly in the left half plane, there exists Δt a sufficiently small to have $\hat{R}(\Delta t\lambda) < 1$, since the rescaling of λ by a smaller value of Δt moves the point $\Delta t\lambda$ towards the origin.

On the other hand, observe that this will be impossible to achieve if λ lies on the imaginary axis. Thus Euler's method is a very poor choice for integrating a problem like $dz/dt = i\omega z$, where ω is real.

A very valuable feature if we are interested in preserving the asymptotic stability of the origin under discretization is if the stability region includes the entire left half plane. In this case we say that the method is *unconditionally stable*. An unconditionally stable numerical method has the property that the origin is stable regardless of the stepsize.

Chapter 5

Linear dynamics in the plane

5.1 Linear differential equations on $\mathbf{R}^2$

In this section we provide a rather complete analysis of the solutions of the system of differential equations

$$\frac{dy^{(1)}}{dt} = a_{11}y^{(1)} + a_{12}y^{(2)} \tag{5.1}$$

$$\frac{dy^{(2)}}{dt} = a_{21}y^{(1)} + a_{22}y^{(2)}. \tag{5.2}$$

In matrix form, we write this system as

$$\frac{dy}{dt} = Ay, \qquad y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \end{pmatrix}, \qquad A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}. \tag{5.3}$$

The analysis of the solutions of such systems is greatly simplified once we have found the eigenvalues and eigenvectors of the matrix A. Recall that an eigenvector v and corresponding eigenvalue λ satisfy

$$Av = \lambda v.$$

The existence of such an eigenpair (λ, v) requires $(A - \lambda I)v = 0$. This in turn requires that the matrix $A - \lambda I$ be singular, with v in its null space. Hence, such λ necessarily satisfy

$$0 = \det(A - \lambda I) = \det \begin{bmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{bmatrix} = \lambda^2 - (a_{11} + a_{12})\lambda + (a_{11}a_{22} - a_{12}a_{21}).$$

The terms in parentheses are the trace[1] $s = \operatorname{tr} A = a_{11} + a_{12}$ and determinant $d = \det A = a_{11}a_{22} - a_{12}a_{21}$ of A. The characteristic polynomial is

$$\lambda^2 - s\lambda + d = 0, \qquad \lambda_{1,2} = \frac{s}{2} \pm \sqrt{\frac{s^2}{4} - d}, \tag{5.4}$$

[1] The trace of a matrix is the sum of the elements on the main diagonal.

where $\lambda_{1,2}$ denote the two roots of the quadratic equation.

Since A is a real matrix, the coefficients of characteristic polynomial (5.4) are real, which means that its roots are either real $\lambda_{1,2} \in \mathbf{R}$, or a complex conjugate pair $\lambda_1 = \bar{\lambda}_2 \in \mathbf{C}$. We treat all cases below:

5.1.1 Real, distinct eigenvalues

Suppose λ_1 and λ_2 are real and distinct. In this case the eigenvectors v_1 and v_2 are also real and linearly independent. Any vector in $\mathbf{R}^2$ can be written uniquely as a linear combination of v_1 and v_2. In particular,

$$y(t) = \alpha_1(t)v_1 + \alpha_2(t)v_2.$$

Substituting this vector function into the differential equation yields

$$\frac{dy}{dt} = \frac{d\alpha_1}{dt}v_1 + \frac{d\alpha_2}{dt}v_2 = \alpha_1(t)Av_1 + \alpha_2(t)Av_2 = \alpha_1(t)\lambda_1 v_1 + \alpha_2(t)\lambda v_2.$$

Since the projection onto v_1 and v_2 is unique, we can equate coefficients to give

$$\frac{d\alpha_1}{dt} = \lambda_1\alpha_1, \qquad \frac{d\alpha_2}{dt} = \lambda_2\alpha_2$$

The solutions of these scalar differential equations are

$$\alpha_1(t) = e^{\lambda_1 t}\alpha_1(0), \qquad \alpha_2(t) = e^{\lambda_2 t}\alpha_2(0),$$

which leads to the general solution

$$y(t) = e^{\lambda_1 t}\alpha_1(0)v_1 + e^{\lambda_2 t}\alpha_2(0)v_2. \tag{5.5}$$

Without loss of generality we assume that $\lambda_1 > \lambda_2$, and write

$$y(t) = e^{\lambda_1 t}\left(\alpha_1(0)v_1 + e^{(\lambda_2 - \lambda_1)t}\alpha_2(t)v_2\right).$$

The exponent in the second term in parentheses is negative by assumption, and in the limit $t \to \infty$, it decays to zero faster than the first term. Therefore we expect that for large t, the solution will point approximately in the direction of the first eigenvector v_1:

$$y(t) \approx e^{\lambda_1 t}\alpha_1(0)v_1, \quad \text{for } t \gg 0.$$

Since A is nonsingular, the origin is the unique equilibrium for this system. The stability of the origin depends on the eigenvalues. We distinguish three cases:

- *Stable node*: $\lambda_2 < \lambda_1 < 0$. Both terms in (5.5) decay monotonically to zero as $t \to \infty$. The origin is stable.

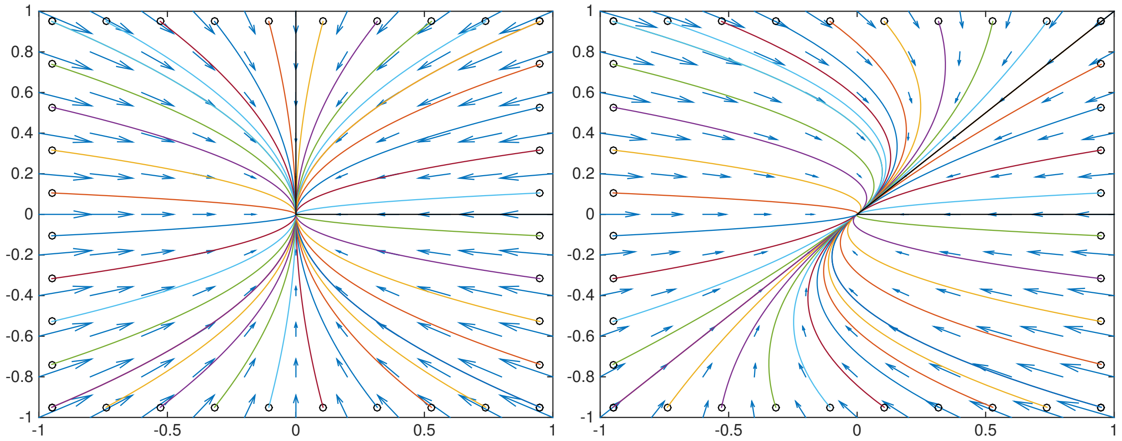

Figure 5.1: Stable nodes: canonical coordinates (left), transformed coordinates (right).

- *Unstable node*: $0 < \lambda_2 < \lambda_1$. Both terms in (5.5) grow unbounded monotonically as $t \to \infty$. The origin is unstable.
- *Saddle* $\lambda_2 < 0 < \lambda_1$. In this case, the first term v_1 in (5.5) grows unbounded while v_2 decays to zero for large $t \to \infty$ whereas the opposite occurs for $t \to -\infty$.

Example. Figure 5.1 shows phase plots (initial conditions denoted with a circle) for the cases

$$A = \begin{bmatrix} -1.5 & \\ & -0.8 \end{bmatrix}, \qquad \tilde{A} = \begin{bmatrix} -1.5 & 0.7 \\ 0 & -0.8 \end{bmatrix}. \tag{5.6}$$

For A the eigenpairs are $(\lambda_1 = -0.8, v_1 = (0,1)^T)$ and $(\lambda - 2 = -1.5, v_2 = (1,0)^T)$; for $\tilde{A}$ these are $(\lambda_1 = -0.8, v_1 = (1,1)^T)$ and $(\lambda - 2 = -1.5, v_2 = (1,0)^T)$. The eigenvectors are plotted as solid black lines in Figure 5.1. Note that for large t, the phase curves are tangent to v_1 at the origin.

We establish that the solution of an autonomous differential equation with the vector field negated, is a solution to the original differential in reversed time. Suppose $y(t)$ solves the initial value problem

$$y'(t) = f(y(t)), \quad y(0) = y_0, \quad t \in [0,T].$$

Define the function $\bar{y}(t) = y(T-t)$. Taking the derivative of this function with respect to t, applying the chain rule, and making use of the differential equation gives

$$\frac{d}{dt}\bar{y}(t) = -y'(T-t) = -f(y(T-t)) = -f(\bar{y}(t)),$$

from which is follows that $\bar{y}(t)$ solves the negated initial value problem

$$\bar{y}'(t) = -f(\bar{y}(t)), \quad \bar{y}(0) = y(T), \quad t \in [0,T]$$

with endpoint solution $\bar{y}(T) = y_0$. It follows that the phase plots for the differential equations $y' = f(y)$ and $y' = -f(y)$ are identical, but with the directions of traversal

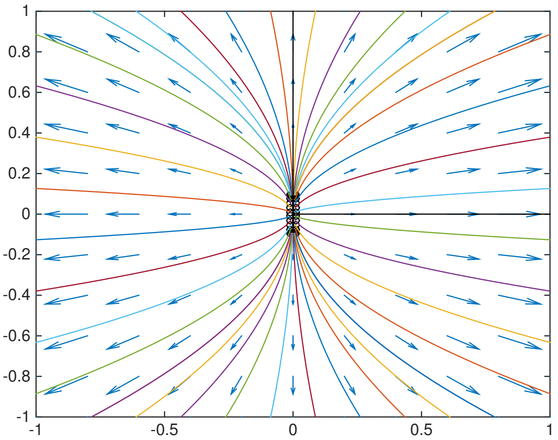

Figure 5.2: Unstable node.

reversed. In particular, a stable equilibrium of $y' = f(y)$ is an unstable equilibrium of $y' = -f(y)$ and vice versa. Also, trajectories emanating from an unstable equilibrium converge upon it in backwards time as $t \to -\infty$. The unstable node corresponding to the matrix

$$A = \begin{bmatrix} 1.5 & \\ & 0.8 \end{bmatrix}$$

is shown in Figure 5.2. In the rest of this section we only show one orientation of the flow (the stable one, when it exists).

Figure 5.3 illustrates saddle point phase plots corresponding to the matrices

$$A = \begin{bmatrix} -1.5 & \\ & 0.8 \end{bmatrix}, \qquad \tilde{A} = \begin{bmatrix} -1.5 & 2.3 \\ 0 & 0.8 \end{bmatrix},$$

with the same eigenvectors as (5.6).

Real distinct eigenvalues occur for $d < s^2/4$, since then the discriminant in (5.4) is positive. In the trace-determinant parameter space (Figure 5.4), the real distinct eigenvalues occur below the parabola $d = s^2/4$. The case $d < 0$ corresponds to a saddle point, since $d = \lambda_1\lambda_2$. For $0 < d < s^2/4$, the equilibrium is a node. The eigenvalues are like-signed, and since $s = \lambda_1 + \lambda_2$, the node is stable if $s < 0$ and unstable if $s > 0$. The borderline cases will be treated later.

If $d > s^2/4$, then $\lambda_{1,2}$ are complex conjugates, as are the eigenvectors $v_{1,2}$

$$\lambda_{1,2} = \mu \pm i\theta, \qquad v_{1,2} = r + is$$

Again the solution is given by (5.5). Note that since eigenvectors remain eigenvectors upon scaling, we can "absorb" the initial conditions $\alpha_{1,2}(0)$ into $v_{1,2}$ and further ignore

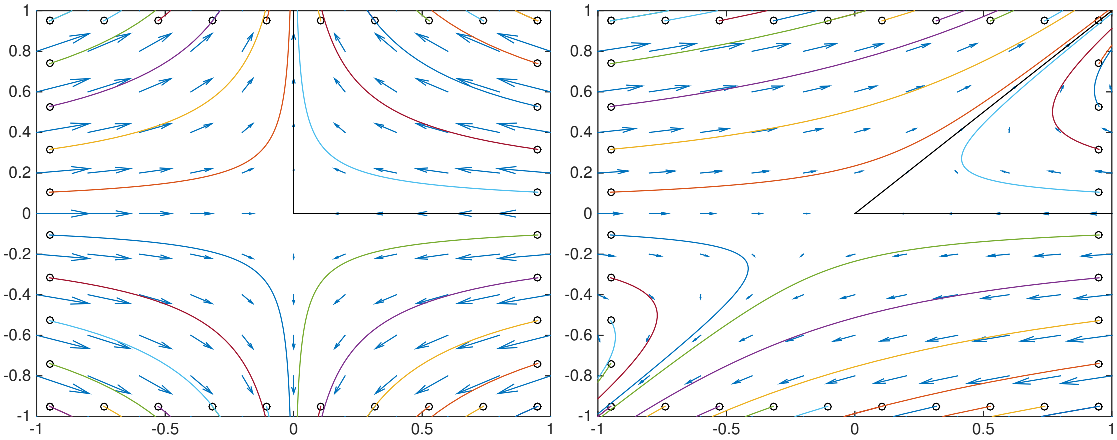

Figure 5.3: Saddle points: canonical coordinates (left), transformed coordinates (right).

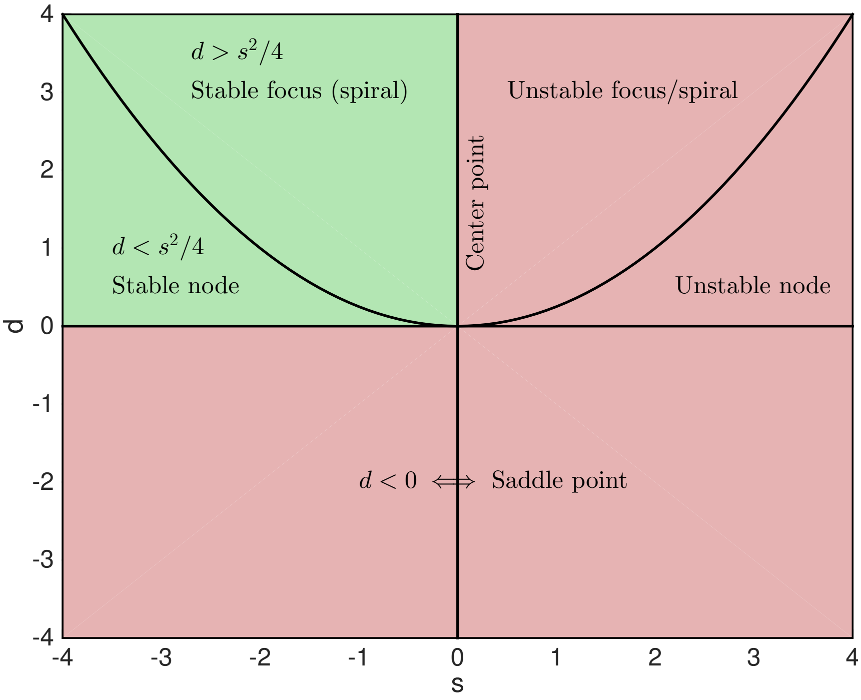

Figure 5.4: The trace(s)-determinant(d) parameter space.

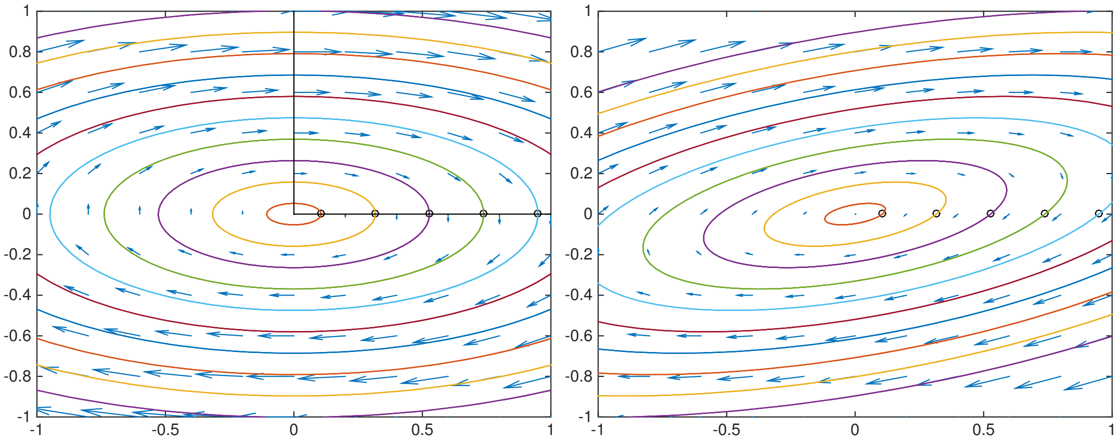

Figure 5.5: Center point: canonical coordinates (left), transformed coordinates (right).

these terms. Then (5.5) can be written as

$$
\begin{aligned}
y(t) &= e^{(\mu+i\theta)t}(r+is) + e^{(\mu-i\theta)}(r-is), \\
&= e^{\mu t}\left[(\cos\theta t + i\sin\theta t)(r+is) + (\cos\theta t - i\sin\theta t)(r-is)\right], \\
&= 2e^{\mu t}\left[(\cos\theta t)r - (\sin\theta t)s\right].
\end{aligned}
$$

Consider first the case $\mu = 0$ (purely imaginary eigenvalues). In this case the solution at times $t = \frac{k\pi}{2\theta}$, $k = 0, 1, 2, \dots$ are

$$
y(0) = r, \quad y(\frac{\pi}{2\theta}) = -s, \quad y(\frac{\pi}{\theta}) = -r, \quad y(\frac{3\pi}{2\theta}) = s, \quad y(\frac{2\pi}{\theta}) = r, \quad \dots
$$

Hence the solution is periodic. The equilibrium, called a *center point*, is shown in Figure 5.5. Clearly, the equilibrium is Lyapunov stable according to our definition of stability.

Next, suppose $\mu < 0$. Then the solution is as in the periodic case, except that the radius of the orbit is exponentially decreasing. The solution spirals inward towards the equilibrium as shown in Figure 5.6. In this case the equilibrium is called a *stable focus*. If $\mu > 0$, the trajectories spiral out away from the equilibrium, an *unstable focus*.

5.1.2 Degenerate cases

Finally, we consider a number of degenerate cases. The first is that of singular A:

$$
A = \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix}.
$$

Since A is singular, there is a nontrivial solution to $Ay = 0$, namely $y = (0,1)^T$. Any initial condition on the y_2 axis is an equilibrium. The corresponding differential equations

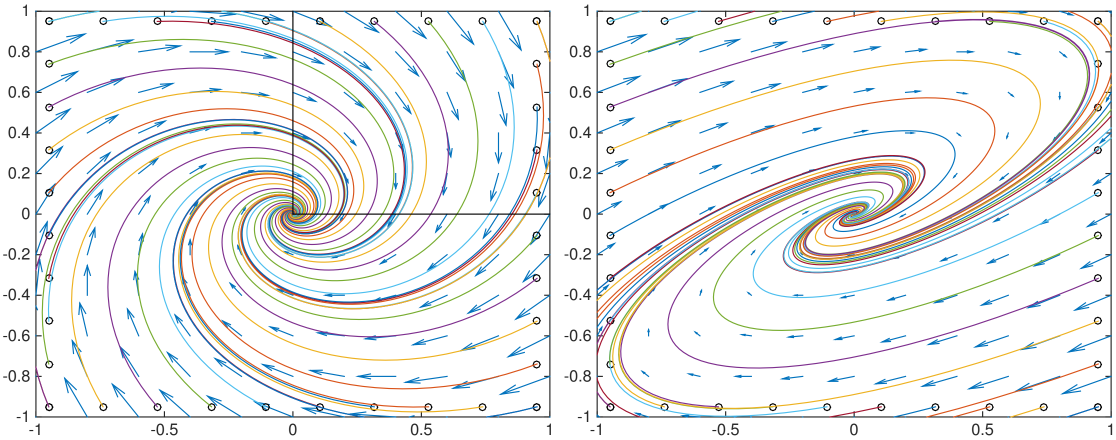

Figure 5.6: Stable focus: canonical coordinates (left), transformed coordinates (right).

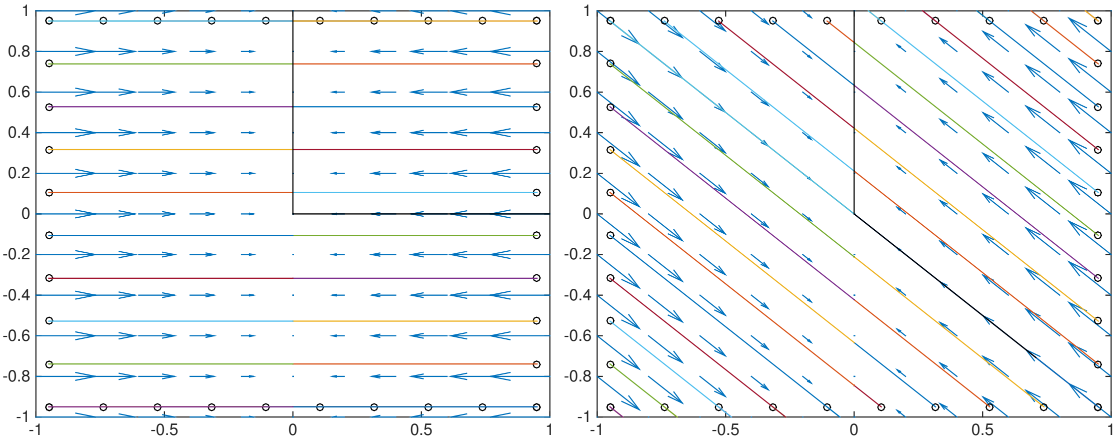

Figure 5.7: Singular flow (shear): canonical coordinates (left), transformed coordinates (right).

are

$$\frac{dy_1}{dt} = \lambda y_1, \qquad \frac{dy_2}{dt} = 0$$

If $\lambda < 0$, the trajectories tend to $(0, y_2)$ as $t \to 0$. See Figure 5.7.

The case $d = s^2/4$ corresponds to the parabola in Figure 5.4. In this case, there may be infinitely many eigenvectors or just one. The first case corresponds to a matrix A with canonical form

$$A = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

In this case, since A is a multiple of the identity, every vector is an eigenvector of A. The solutions are $y(t) = \exp(\lambda t) y(0)$. The stable case $\lambda < 0$ is shown in Figure 5.8.

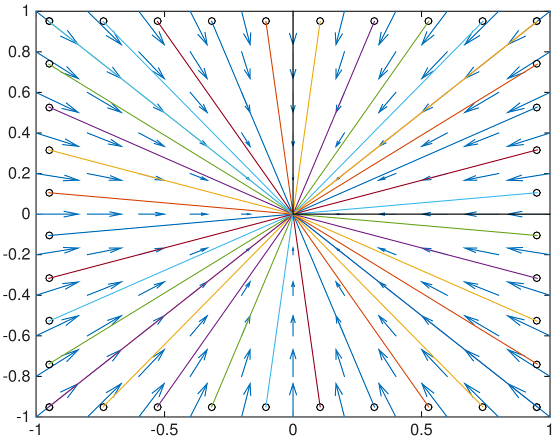

Figure 5.8: Node corresponding to $\lambda_1 = \lambda_2$ with independent eigenvectors.

Alternatively, A may have Jordan form

$$A = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}.$$

In this case, there is a single eigenvector $v = (1, 0)^T$. The corresponding differential equations are

$$\frac{dy^{(1)}}{dt} = \lambda y^{(1)} + y^{(2)}, \qquad \frac{dy^{(2)}}{dt} = \lambda y^{(2)}.$$

The second of these has solution $y^{(2)}(t) = \exp(\lambda t) y^{(2)}(0)$. Substituting this solution into the first equation yields

$$\frac{dy^{(1)}}{dt} = \lambda y^{(1)} + \exp(\lambda t) y^{(2)}(0).$$

To solve this differential equation, we multiply through by $\exp(-\lambda t)$ and integrate

$$\begin{aligned} e^{-\lambda t}\frac{dy^{(1)}}{dt} - \lambda e^{-\lambda t} y^{(1)} &= y^{(2)}(0), \\ \frac{d}{dt}\left[e^{-\lambda t} y^{(1)}\right] &= y^{(2)}(0), \\ \int_0^T \frac{d}{dt}\left[e^{-\lambda t} y^{(1)}\right]\, dt &= \int_0^T y^{(2)}(0)\, dt, \\ e^{-\lambda T} y^{(1)}(T) - y^{(1)}(0) &= y^{(2)}(0) T, \\ y^{(1)}(T) &= e^{\lambda T}(y^{(1)}(0) + T y^{(2)}(0)). \end{aligned}$$

Consequently the general solution is

$$\begin{aligned} y^{(1)}(t) &= e^{\lambda t}(y^{(1)}(0) + t y^{(2)}(0)), \\ y^{(2)}(t) &= e^{\lambda t} y^{(2)}(0). \end{aligned}$$

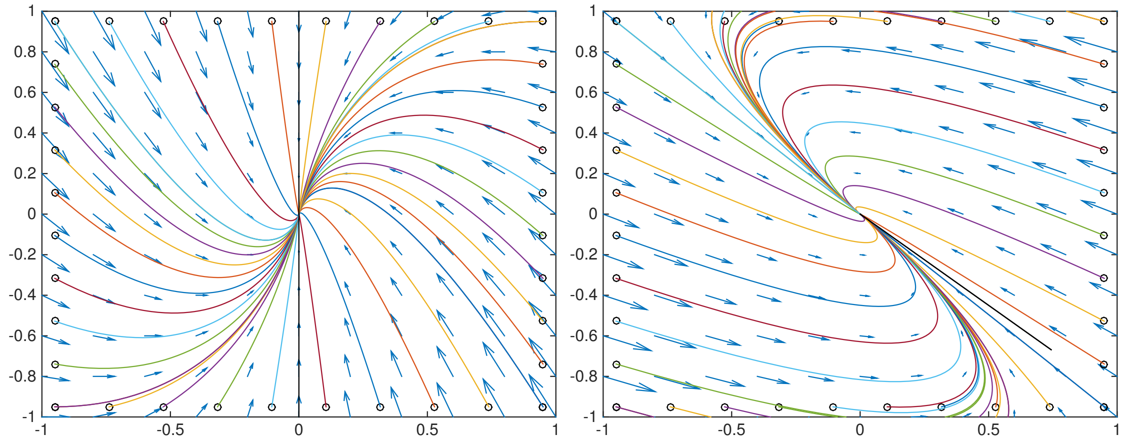

Figure 5.9: Node corresponding to a non-simple eigenvalue: canonical coordinates (left), transformed coordinates (right).

In vector form,

$$y(t) = e^{\lambda t}(y(0) + ty^{(2)}(0)v)$$

As $t \to \infty$, the second term dominates and the trajectories approach the approximate solution $y(t) \approx t\exp(\lambda t)y^{(2)}(0)v$. See Figure 5.9.

To summarize, the equilibrium is stable in all cases for which both eigenvalues satisfy $\operatorname{Re}\lambda < 0$. It is unstable if any of the eigenvalues has $\operatorname{Re}\lambda > 0$.

Chapter 6

Markov chains

6.1 Graph theory of matrices and the Perron-Frobenius Theorem

Following the reasoning of the previous section, we see that it is useful to have a means of identifying when a matrix A possesses a dominant eigenvalue. The theorem at the end of this section provides criteria for this, but we first need to understand some basic ideas from graph theory.

Let $A = (a_{ij})$ denote a matrix with elements a_{ij}. By $A \geq 0$ we mean $a_{ij} \in \mathbf{R}$ and $a_{ij} \geq 0$ for all $i, j = 1, \ldots, d$. Similarly we use the notation $v \geq 0$ to denote a vector v whose elements are real and nonnegative.

A directed graph consists of a set of vertices and a set of edges that express connections between the vertices. For a $d \times d$ matrix we define d vertices. For every nonzero element a_{ij} of A, we define a directed edge from j to i.

For an *incidence matrix*, $a_{ij} = 1$ if there is an edge from node j to node i, and $a_{ij} = 0$ otherwise. The nonzero element of the matrix A^p counts the number of paths of length p between pairs of nodes. For instance, denote by A^2_{ij} the element (i, j) of the matrix A^2. If $A^2_{ij} = 3$ then there are three paths of length 2 connecting node j to node i.

Example. The incidence matrix of a Leslie matrix

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \tag{6.1}$$

has the associated graph shown in Figure 6.1. Due to the nonzero elements in the first row, there is a directed edge from each vertex to vertex 1, including a self-directed edge (*loop*) from 1 to itself. The nonzero elements on the first sub-diagonal imply edges from vertex j to vertex $j+1$, for $j = 1, \ldots, 4$.

The matrices A^2 and A^3 are

$$A^2 = \begin{bmatrix} 2 & 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}, \qquad A^3 = \begin{bmatrix} 4 & 4 & 4 & 3 & 2 \\ 2 & 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

For example, the two paths of length 2 to from node 2 to node 1 (i.e. $A^2_{12} = 2$) are $(2 \to 1 \to 1)$ and $(2 \to 3 \to 1)$. The 4 paths of length 3 from node 3 to node 1 (i.e. $A^3_{13} = 4$) are: $(3 \to 1 \to 1 \to 1)$, $(3 \to 1 \to 2 \to 1)$, $(3 \to 4 \to 1 \to 1)$, and $(3 \to 4 \to 5 \to 1)$.

For a *weighted graph*, each connection receives a weight a_{ij}. In this case, the edge weights of a path are multiplied together, and the various paths added up.

Example. Consider the Leslie matrix with same nonzero structure as the previous example:

$$A = \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 & \phi_4 & \phi_5 \\ \sigma_1 & 0 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_3 & 0 & 0 \\ 0 & 0 & 0 & \sigma_4 & 0 \end{bmatrix}, \tag{6.2}$$

where the ϕ_i and σ_i are positive constants. In this case, the (1,2) element of A^2 contains the sum of the two weighted paths $(2 \to 1 \to 1)$ and $(2 \to 3 \to 1)$:

$$A^2_{12} = (2 \to 1) \times (1 \to 1) + (2 \to 3) \times (3 \to 1) = \phi_2\phi_1 + \sigma_2\phi_3,$$

which is the same result as obtained through matrix multiplication.

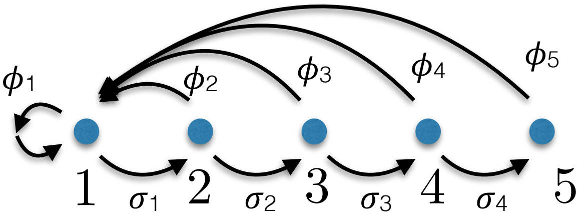

Figure 6.1: Graph associated with the matrix (6.2).

A *walk* on a directed graph is a sequence of consecutive vertices connected by directed edges, following the orientation of the edges. For example, in the graph in Figure 6.1, one possible walk is $(3 \to 4 \to 1 \to 1 \to 2 \to 1)$. In particular, vertices and edges may appear multiple times in a walk. The *length* of a walk is the number of "steps" taken, that is the number of vertices visited excluding the starting vertex (including multiplicity). In the above example, the walk has length 5 (i.e. $\{4, 1, 1, 2, 1\}$).

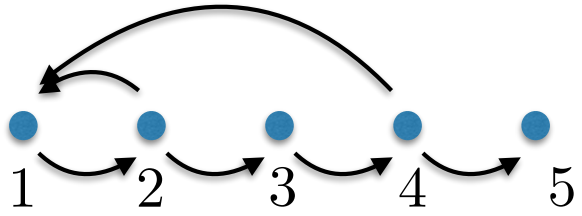

Figure 6.2: This graph is reducible, since there are no walks from vertex 5 to any other vertex, and it is of period 2, since all cycles have length 2 or 4.

The graph is *irreducible* if there exists a walk from every vertex to every other vertex. Otherwise it is *reducible*. The graph in Figure 6.1 is irreducible. By comparison, the graph in Figure 6.2 is reducible, because there is no walk from vertex 5 to any other vertex.

A *cycle* is a walk that starts and ends at the same vertex, but otherwise has no repeated vertices. The *periodicity* of a graph is the greatest common divisor of the lengths of all cycles. A graph with period one is called *aperiodic*. The graph in Figure 6.1 is aperiodic. The graph in Figure 6.2 has period 2, since all cycles have even length. If a graph contains a loop (edge from a vertex to itself), then this is a cycle of length one, and the graph is aperiodic. Equivalently, if a matrix has a nonzero element on its main diagonal, then its graph is aperiodic.

The following theorem allows us to establish when a linear recursion has a dominant eigenvalue. Its proof is beyond the scope of these notes.

Theorem 6.1.1 (Perron & Frobenius) *Suppose $A \geq 0$. Then:*

1. *A has a real eigenvalue $\lambda_1 \geq 0$ such that $|\lambda_j| \leq \lambda_1$, $j = 2, \ldots, d$.*
2. *There exists a constant $\gamma_1 \in \mathbf{C}$ such that associated eigenvector $\gamma_1 v_1 \geq 0$.*
3. *If A is irreducible and aperiodic, then λ_1 is the dominant eigenvalue: $\|\lambda_j\| < \lambda_1$, $j = 2, \ldots, d$, holds with strict inequality.*

The theorem also shows when the linear recursion possesses a nonnegative eigenvector, which may be important in applications where negative or complex values have no meaning (for example, when the iterates x_n denote population numbers or probabilities).

6.2 Probability vectors and transition matrices

A very important class of linear recursions are Markov chains. A Markov chain is a linear probabilistic model of a system that attains only a finite number d of states. The probability of observing the system in state j at discrete time n is denoted $p_n^{(j)}$. Since probabilities are never negative, and never greater than 1, we require $0 \leq p_n^{(j)} \leq 1$. Also, since the system must be found in one of the d states, we require

$$\sum_{j=1}^{d} p_n^{(j)} = 1$$

We prefer to work with vectors, and introduce the vector $p_n = (p_n^{(1)}, p_n^{(2)}, \ldots, p_n^{(d)})^T$ to denote the probabilities of all states at time n. We will also make use of the vectors $\mathbb{1} = (1, 1, \ldots, 1)^T \in \mathbf{R}^d$ and e_j to denote the jth canonical unit vector in $\mathbf{R}^d$, i.e. the vector whose elements are all zero except the jth element, which is 1 (equivalently, the jth column of the identity matrix). Using this notation, the requirements on p_n can be expressed as

$$0 \leq e_j^T p_n \leq 1, \ \forall j, \qquad \mathbb{1}^T p_n = 1. \tag{6.3}$$

A vector satisfying these properties is called a *probability vector*.

A Markov chain specifies that the probabilities of observing the system in its various states changes as a function of time. It is assumed that the system is likely to change from state j to state i with probability P_{ij}, and that this probability is independent of time. Given a probability vector p_0 over the possible states at time zero, the probabilities evolve according to the recursion

$$p_{n+1} = P p_n, \qquad P = (P_{ij}) \tag{6.4}$$

where P is the *transition matrix* with elements P_{ij}. To gain more understanding of this, suppose our system is known to be in state $j = 1$ at time $n = 0$. We denote this as $p_0 = e_1$. Then, according to (6.4), the probabilities of observing the system in each system state at time $n = 1$ are given by

$$p_1 = P p_0 = P e_1,$$

which is just the first column of the matrix P. In words, the first column of P, with elements P_{i1} specifies the likelihood of finding the system in each state i, given that the system was in state $j = 1$ at the previous time. That is, P_{i1} denotes the likelihood of a *transition* from state 1 to state i. Similarly, if $p_0 = e_j$, then $p_1 = P e_j$ is the jth column of P, which specifies the likelihood of observing the system in state i given that was in state j at the previous time. Hence P_{ij} specifies the likelihood of a transition from state j to state i. It also follows from this reasoning that the columns of P must be probability vectors:

$$0 \leq P_{ij} \leq 1, \qquad \sum_i P_{ij} = 1, \forall j,$$

where column j specifies the likely states of the system at time $n + 1$ given that it was in state j at time n. We express this also as

$$P \geq 0, \qquad \mathbb{1}^T P = \mathbb{1}^T. \tag{6.5}$$

As an example, we consider a Markov chain model for weather prediction. We assume the weather is observed in one of three mutually exclusive states: (1) sunshine, (2) overcast, and (3) rain. At time n, the probability of observing the weather in state j is given by $p_n^{(j)}$. The transition matrix is

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{6} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}. \tag{6.6}$$

The first column means that: following a sunny day, with equal likelihood, either another sunny day or a cloudy day will be observed. The second column is interpreted as: following a cloudy day, there is a 1/6 chance of a sunny day, a 1/3 chance of another cloudy day, and a 1/2 chance of a rainy day. The third columns states that after rain, with equal probability, it either rains again or ("after rain comes sunshine") sunshine follows. Figure 6.3 illustrates the graph associated with P.

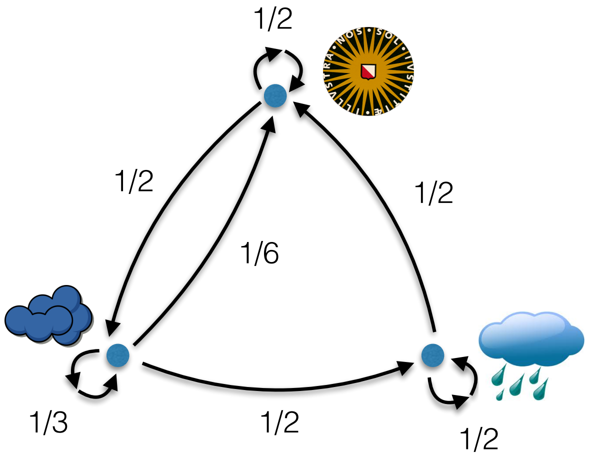

Figure 6.3: Graph of the transition matrix (6.6) of the weather model.

The following theorem ensures that the Markov chain is well defined, and gives us a condition for the existence of a stable steady state (i.e. a background "climate" for our weather model).

Theorem 6.2.1 *Let $P \in \mathbf{R}^{d\times d}$ be a transition matrix satisfying (6.5) with eigenvalues λ_j, $j = 1, \dots, d$. Let p_0 be a probability vector, and p_n be the solution of the Markov chain (6.4). Then,*

1. *p_n is a probability vector for all n.*
2. *$\lambda_1 = 1$ is an eigenvalue of P, and $|\lambda_j| \le \lambda_1$, for all j.*
3. *if P is in addition irreducible and aperiodic, then $\lambda_1 = 1$ is the dominant eigenvalue.*

Proof The elements of p_1 are

$$p_1^{(j)} = \sum_{k=1}^{d} P_{jk} p_0^{(k)}.$$

Since each term in the the sum is a product of nonnegative elements, the sum is nonnegative. Furthermore, using (6.5) and (6.3),

$$\mathbb{1}^T p_1 = \mathbb{1}^T P p_0 = \mathbb{1}^T p_0 = 1.$$

Consequently, p_1 is again a probability vector. That p_n is also a probability vector for all n follows by induction. This establishes conclusion 1 of the theorem.

Note that taking the transpose of the second relation in (6.5) gives

$$P^T \mathbb{1} = \mathbb{1},$$

from which follows immediately that $\lambda_1 = 1$ is an eigenvalue of P^T. But since, for an arbitrary square matrix A,

$$\det(A - \lambda I) = \det(A^T - \lambda I),$$

the characteristic polynomials of any square matrix A and its transpose A^T are identical, and have identical roots. Hence A and A^T have the same eigenvalues. In particular $\lambda_1 = 1$ is an eigenvalue of P.

Next, we define the absolute value of a vector or matrix to hold element-wise:

$$|v| = (|v_1|, \ldots, |v_d|)^T$$

Now, taking absolute values the following inequality holds, where λ is an eigenvalue of P with associated eigenvector v:

$$|\lambda||v| = |\lambda v| = |Pv| \leq P|v|. \tag{6.7}$$

The last relation follows from the triangle inequality, i.e. for the ith row,

$$|Pv|_i = |\sum_{j=1}^{d} P_{ij} v_j| = |P_{i1} v_1 + \cdots + P_{id} v_d| \leq |P_{i1} v_1| + \cdots + |P_{id} v_d|.$$

Furthermore, since the P_{ij} are nonnegative, we can extract them from the absolute values. Now, multiplying both sides of inequality (6.7) by $\mathbb{1}^T$ gives

$$|\lambda| \mathbb{1}^T |v| \leq \mathbb{1}^T P |v| = \mathbb{1}^T |v|.$$

From which follows $|\lambda| \leq 1$ for all eigenvalues of P. This establishes conclusion 2 of the theorem. (Alternatively, we can apply Theorem 6.1.1. Since $P \geq 0$, there exists a nonnegative, real eigenvalue $\lambda_1 \geq |\lambda_j|$ with associated nonnegative, real eigenvector v_1. For this pair, we have

$$\lambda_1 \mathbb{1}^T v_1 = \mathbb{1}^T (\lambda_1 v_1) = \mathbb{1}^T P v_1 = \mathbb{1}^T v_1$$

from which follows $\lambda_1 = 1$.)

Conclusion 3 is an immediate consequence of conclusions 1 and 2 and Theorem 6.1.1. □

How can we use this model? Suppose we wish to know what the likelihood of rain is five days after a sunny day. To determine this, we choose $p_0 = e_1$, which specifies that the sun is shining with probability one on day 0. The likelihoods of the various weather states five days later is given by p_5,

$$p_5 = P^5 p_0 = P^5 e_1.$$

The likelihood of rain is the third component of p_5, i.e. $p_5^{(3)} = e_3^T p^5$:

$$p_5^{(3)} = e_3^T P^5 e_1.$$

Equivalently, the likelihood is given by element $(3,1)$ of the matrix P^5.

6.3 Defective transition matrices

We can also address a question like: *what is the likelihood that it rains for the first time 5 days after sunshine?* To determine this, we want to consider the joint probability of all walks of length 5 originating at vertex 1 and ending precisely at vertex 3, that do not visit vertex three prior to the 5th step. We can achieve this by removing all edges leaving vertex 3.[1] The resulting graph is shown in Figure 6.4. The associated matrix is

$$\tilde{P} = \begin{bmatrix} \frac{1}{2} & \frac{1}{6} & 0 \\ \frac{1}{2} & \frac{1}{3} & 0 \\ 0 & \frac{1}{2} & 0 \end{bmatrix}.$$

Such a matrix satisfying

$$\tilde{P} \geq 0, \qquad \mathbb{1}^T \tilde{P} \leq \mathbb{1}^T,$$

is termed a *defective transition matrix*.

The chance of a first rain precisely five days after sunshine is given by

$$e_3^T p_5 = e_3^T \tilde{P}^5 e_1,$$

that is, the $(3,1)$ element of $\tilde{P}^5$. Other statistics are, for instance: *on average, how many days after sunshine is the first rain?* This statistic is determined as follows:

$$\sum_{n=1}^{\infty} n \times (\text{chance of first rain on day } n) = \sum_{n=1}^{\infty} n e_3^T \tilde{P}^n e_1.$$

For computing such statistics, the following identities, which can be shown to hold if all eigenvalues of a matrix A satisfy $|\lambda| < 1$, are useful:

$$\sum_{n=1}^{\infty} A^n = A(I - A)^{-1}, \qquad \sum_{n=1}^{\infty} n A^n = A(I - A)^{-2}.$$

[1] In terms of paths on graphs the transition matrix entry P_{ij} gives the likelihood when standing at node j of following the path to node i.

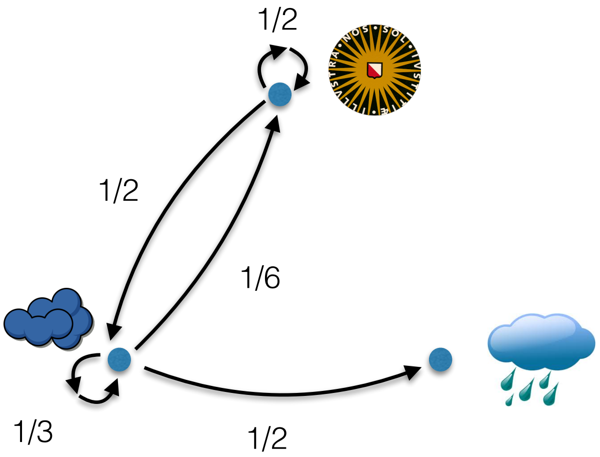

Figure 6.4:

To use the identities, we need to establish that the eigenvalues of the defective transition matrix $\tilde{P}$ are strictly less than unity in modulus, the proof of which uses arguments similar to those of the proof of Theorem 6.2.1.

Chapter 7

Resonance

We consider a spring-mass-damper system with spring constant $\kappa > 0$, damping (friction) constant $c \geq 0$ and mass $m > 0$. Let $x(t)$ denote the elongation of the spring from its rest state. *Hooke's law* states that the force exerted by the spring on the mass is proportional to the elongation and opposite in sign:

$$\text{force} = -\kappa x(t).$$

Newton's Second Law states that the rate of change of momentum is equal to the sum of the applied forces. For a constant mass, this is

$$m\ddot{x} = \sum(\text{forces})$$

A damper is a device that applies a force proportional to velocity and opposite in sign. The motion of the spring mass damper system is

$$m\ddot{x} = -c\dot{x} - \kappa x, \qquad x(0) = x_0, \quad \dot{x}(0) = y_0. \tag{7.1}$$

We want to investigate the case in which an external force is applied to the system. For ease of notation later, we introduce new variables

$$\rho = \frac{c}{2m}, \qquad \nu^2 = \frac{\kappa}{m},$$

and rewrite (7.1) with applied forcing $f(t)$ in the form

$$\ddot{x} + 2\rho\dot{x} + \nu^2 x = f(t). \tag{7.2}$$

Given a linear differential equation with external forcing, such as (7.1), we associate the *homogeneous problem* defined by $f(t) \equiv 0$, whose solution $x_h(t)$ satisfies

$$\ddot{x}_h + 2\rho\dot{x}_h + \nu^2 x_h = 0.$$

Suppose $x_p(t)$ is a *particular solution* to (7.2), and let $x(t) = x_p(t) + x_h(t)$. Then $x(t)$ is also a solution to (7.2) as follows from substitution

$$\begin{aligned}(\ddot{x}_p + \ddot{x}_h) + 2\rho(\dot{x}_p + \dot{x}_h) + \nu^2(x_p + x_h) &= f(t),\\ (\ddot{x}_h + 2\rho\dot{x}_h + \nu^2 x_h) + \ddot{x}_h + 2\rho\dot{x}_p + \nu^2 x_p &= f(t),\end{aligned}$$

since the term in parentheses in the last line is zero, and the remaining terms are precisely the differential equation for the particular solution. Given a particular solution, the homogeneous solution is chosen to ensure the initial conditions are satisfied.

Hence, the solution to the inhomogeneous problem is a sum of a particular solution and a homogeneous solution. We consider each of these separately.

7.1 The homogeneous solution

As we did with the second order differential equation (2.12) we introduce the velocity $y_h(t) = \dot{x}_h(t)$ and write the homogeneous problem as a first order system

$$\begin{pmatrix}\dot{x}_h\\ \dot{y}_h\end{pmatrix} = \begin{bmatrix}0 & 1\\ -\nu^2 & -2\rho\end{bmatrix}\begin{pmatrix}x_h\\ y_h\end{pmatrix}.$$

This is a problem in the form (5.3) with matrix A given in square brackets above. Let λ_1 and λ_2 and be the eigenvalues corresponding to the eigenvectors v_1 and v_2 of A. Then, assuming $\lambda_1 \neq \lambda_2$, the solution can be written

$$\begin{pmatrix}x_h(t)\\ y_h(t)\end{pmatrix} = e^{\lambda_1 t}v_1 + e^{\lambda_2 t}v_2.$$

In particular, the solution for $x_h(t)$ is

$$x_h(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t},$$

for some constants C_1 and C_2, where λ_1 and λ_2 satisfy

$$0 = \det\begin{bmatrix}-\lambda & 1\\ -\nu^2 & -2\rho - \lambda\end{bmatrix} = \lambda^2 + 2\rho\lambda + \nu^2 = p(\lambda),$$

where $p(\lambda)$ is the characteristic polynomial.

Remark. As an aside, note that we can define a differential operator by applying p to the derivative $\frac{d}{dt}$ as follows:

$$p\left[\frac{d}{dt}\right]x_h(t) = \left[\left(\frac{d}{dt}\right)^2 + 2\rho\frac{d}{dt} + \nu^2\right]x_h(t) = \frac{d^2x_h}{dt^2} + 2\rho\frac{dx_h}{dt} + \nu^2 x_h = 0.$$

The same holds for higher order linear differential equations, and allows us to immediately construct the characteristic polynomial without going to the trouble of introducing auxiliary variables and matrix formulations:

$$\begin{aligned}&\frac{d^kx}{dt^k}+a_{k-1}\frac{d^{k-1}x}{dt^{k-1}}+\cdots+a_1\frac{dx}{dt}+a_0=0\\ \Rightarrow\quad & p(\lambda)=\lambda^k+a_{k-1}\lambda^{k-1}+\cdots+a_1\lambda+a_0=0\\ \Rightarrow\quad & x(t)=C_1e^{\lambda_1t}+C_2e^{\lambda_2t}+\cdots+C_ke^{\lambda_kt},\end{aligned}$$

assuming the eigenvalues are distinct.

For the homogeneous problem the roots of the characteristic polynomial are

$$p(\lambda)=0\quad\Rightarrow\quad\lambda_{1,2}=-\rho\pm\sqrt{\rho^2-\nu^2}.$$

We consider now what happens as we increase the damping parameter ρ from zero.

For $\rho=0$, the eigenvalues are purely imaginary

$$\lambda_{1,2}=\pm i\nu,$$

and the homogeneous solution is

$$x_h(t)=C_1e^{i\nu t}+C_2e^{-i\nu t}=A\cos\nu t+B\sin\nu t$$

where the constant pairs C_1 and C_2 or A and B may be determined from the initial conditions. It is useful to note the following identity: Let $\gamma\in\mathbf{C}$ alternatively be written $\gamma=\gamma_1+i\gamma_2=|\gamma|\,e^{-i\delta}$ in real/imaginary and polar coordinates respectively. Then the following are equivalent

$$\mathrm{Re}\left(\gamma e^{i\nu t}\right)=\gamma_1\cos\nu t-\gamma_2\sin\nu t=|\gamma|\cos(\nu t-\delta).$$

In particular, an alternative representation of the homogeneous solution is

$$x_h(t)=C\cos(\nu t-\delta),$$

where constants C and δ can again be determined from the initial conditions. We call ν the *frequency*, δ the *phase*, and C the *amplitude*. This representation emphasizes the fact that $x_h(t)$ is a shifted and scaled cosine wave.

As the damping parameter ρ is increased from zero we have, initially, $\rho<\nu$. In this case, the eigenvalues are complex conjugates $\lambda_1=\bar{\lambda}_2$:

$$\lambda_{1,2}=-\rho\pm i\sqrt{\nu^2-\rho^2}.$$

In this case, the homogeneous solution is

$$x_h(t)=e^{-\rho t}(C_1e^{i\sqrt{\nu^2-\rho^2}t}+C_2e^{-i\sqrt{\nu^2-\rho^2}t})=C\cos(\sqrt{\nu^2-\rho^2}t-\delta)$$

for some constants C and δ to be determined from the initial conditions. The solution is oscillatory with decaying amplitude $C\exp(-\rho t)$.

Note however that $|\lambda_1| = |\lambda_2| = \nu$, i.e. the motion of the eigenvalues (as we increase ρ) is constrained to a circle. As we continue to increase ρ, the eigenvalues move symmetrically along the circle until they 'collide' for $\rho = \nu$ at $\lambda_1 = \lambda_2 = -\nu$. If we increase ρ beyond this point, we have $\rho > \nu$, and two real eigenvalues emerge:

$$\lambda_{1,2} = -\rho \pm \sqrt{\rho^2 - \nu^2}.$$

The solution is

$$x_h(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}$$

Both eigenvalues are negative, hence the solution decays to $x_h \to 0$. The eigenvalues are centered at $-\rho$, and the greatest of these, $\lambda_1 = -\rho + \sqrt{\rho^2 - \nu^2}$, dominates the convergence eventually:

$$x_h(t) \approx C_1 e^{(-\rho + \sqrt{\rho^2 - \nu^2})t}, \quad \text{as} \quad t \to \infty.$$

Since it also holds for this eigenvalue that $\lambda_1 > -\nu$, we see that the most rapid damping rate occurs at the collision point $\rho = \nu$.

These considerations lead us to define the cases:

- Undamped: $\rho = 0$,
- Underdamped (oscillatory): $0 < \rho < \nu$,
- Critically damped: $\rho = \nu$,
- Overdamped: $\rho > \nu$.

7.2 The particular solution

We now return to the inhomogeneous problem (7.2). We will be interested in a case with some amount of damping, $\rho > 0$. From the results of the previous section, we see that the homogeneous solution is eventually damped out $x_h(t) \to 0$. This implies (i) eventually only the particular solution survives $x(t) \approx x_p(t)$ as $t \to \infty$, and (ii) since x_h satisfies the initial conditions, these eventually become irrelevant.

Let us make a concrete choice of the forcing function, $f(t) = \cos \mu t$, with forcing frequency μ. It is somewhat more convenient to extend the differential equation (7.2) to complex valued $x(t)$ and $f(t) = \exp(i\mu t)$. If we do this, then the real part of the solution $\operatorname{Re} x(t)$ gives the correct motion (check this!).

Since the particular solution $x_p(t)$ eventually dominates, we attempt a solution of the form

$$x_p(t) = \text{Re}\,\left(He^{i\mu t}\right),$$

where the constant H is a complex number, which may depend on μ. Substituting into (7.2) yields

$$(-\mu^2)He^{i\mu t} + 2\rho(i\mu)He^{i\mu t} + \nu^2 He^{i\mu t} = e^{i\mu t}.$$

Solving for H leaves

$$H = \frac{1}{p(i\mu)},$$

where p is the characteristic polynomial. In signal theory, an external driving term $e^{i\mu t}$ is referred to as the *input signal*, and the particular solution $x_p(t)$ is referred to as the *output signal* or *response* of the system. The function H is in general complex, and can be written as $H = |H|\exp(-i\delta)$. Then, the particular solution may be written

$$x_p(t) = \text{Re}\left[He^{i\mu t}\right] = |H|\cos(\mu t - \delta)$$

Consequently, the response of the system to an input signal $f(t) = \cos\mu t$ is a signal $x_p(t)$ with phase shift δ and amplitude scaling or *gain*

$$|H| = \frac{1}{|p(i\mu)|} = \frac{1}{\sqrt{(\nu^2-\mu^2)^2 + 4\rho^2\mu^2}}.$$

The characteristic polynomial is present in the denominator of the gain $|H|$. Since the characteristic polynomial takes value 0 at the eigenvalues $\lambda_{1,2}$ of the homogeneous part, the gain may be very large if $i\mu \approx \lambda_{1,2}$. In this case we speak of *resonance*. Since $i\mu$ is purely imaginary, resonance may only occur if $\lambda_{1,2}$ are near the imaginary axis. We saw in the previous section that this is the case when the damping is small $\rho \ll \nu$. Therefore resonance may be observed when $\rho \ll \nu$ and $\mu \approx \nu$.

We plot the response function for the case $\rho = 0.1$ and $\nu = 1$ in Figure 7.1. We see that the gain is a factor 5 for $\mu = 1$.

From Fourier theory, it is known that a large class of functions may be approximated as a sum of sinusoidal functions

$$f(t) \approx \sum_{j=1}^{J} \alpha_j \cos(\mu_j t)$$

If the input signal is such a composite function, then resonance may occur if any of the frequencies $\mu_j \approx \nu$, $\rho \ll \nu$.

7.3 Modelling spring-mass-damper systems

Springs-mass-damper systems are used to model all kinds of linear interactions in mechanical systems. Classical mechanics is founded on the laws of Newton. Newton's second law states that the change in momentum (mass times velocity) of an object is equal to the

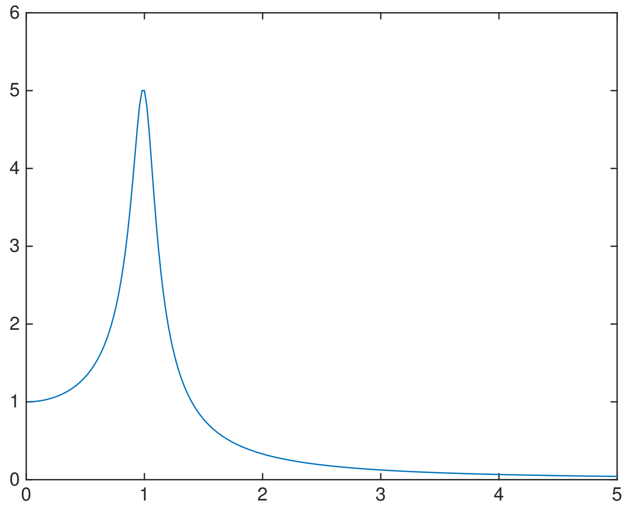

Figure 7.1: Response function for spring-mass-damper system.

net force applied to it. If a spring is inserted between two masses, and these are pressed together or pulled apart, the spring exerts an equal and opposite force on each mass that opposes this motion. A damper opposes all motion with a force that is proportional to the velocity.

Suppose we have a system of two masses (m_1 and m_2) connected by three springs (spring constants k_1, k_2, and k_3) connecting them to each other and to the walls on either side (see Figure 7.2), and the friction encountered by the masses as they slide over the floor is modelled as damping (damping constants c_1 and c_2).

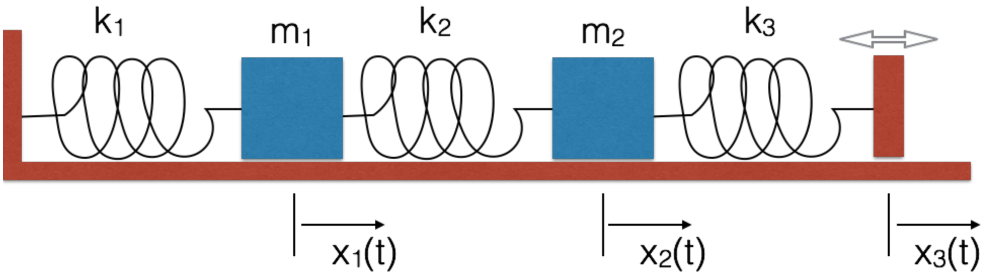

Figure 7.2: Spring-mass-damper system with moving right wall.

Let the displacement of the masses be denoted x_1 and x_2 with positive orientation rightward, and suppose the wall on the right side can move be moved with a given (external) motion $x_3(t)$.

Then the motion of the masses is given by

$$\begin{aligned} m_1\ddot{x}_1 &= -c_1\dot{x}_1 + \text{forces}, \\ m_2\ddot{x}_2 &= -c_2\dot{x}_2 + \text{forces}, \end{aligned}$$

where the forces due to the springs are yet to be determined.

Consider first the force exerted on mass m_1 by spring k_1. As the mass moves rightward, the spring elongates from rest by an amount x_1. The spring pulls back with a force k_1x_1 in the negative (leftward) direction. The force on the mass from this spring is $-k_1x_1$.

What is the force of spring k_2? Suppose $x_1 = x_2$, that is, the masses are displaced by the same amount. Then the middle spring is not elongated and exerts no force. On the other hand, if the mass on the left is displaced to the right ($x_1 > 0$), and the mass on the right is displaced to the left ($x_2 < 0$), then the middle spring is compressed by an amount $x_1 - x_2$. It exerts a force of magnitude $k_2(x_1 - x_2)$ on each mass, but in *opposite directions*, again opposing this motions of each. This means the force on the left mass is $-k_2(x_1 - x_2)$ and the force on the right mass is $k_2(x_1 - x_2) = -k_2(x_2 - x_1)$.

Applying the same reasoning to the third spring, we arrive at the equations of motion

$$\begin{aligned} m_1\ddot{x}_1 &= -c_1\dot{x}_1 - k_1x_1 - k_2(x_1 - x_2), \\ m_2\ddot{x}_2 &= -c_2\dot{x}_2 - k_2(x_2 - x_1) - k_3(x_2 - x_3). \end{aligned}$$

Notice that in the equation of motion of each mass, the force of exerted by each spring is *always opposite to the displacement of that mass.* (In other words, in the equation for x_1, we see $-k_1x_1 - k_2x_1$ on the right side, and in the equation for x_2 we see $-k_2x_2 - k_3x_2$ on the right side.)

Let us rewrite this system as a first order system in matrix notation. Introduce the velocities $y_1 = \dot{x}_1$, $y_2 = \dot{x}_2$. Then we can check that

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{y}_1 \\ \dot{y}_2 \end{pmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ K_{11} & K_{12} & -\frac{c_1}{m_1} & 0 \\ K_{21} & K_{22} & 0 & -\frac{c_2}{m_2} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} f(t),$$

where we have set $f(t) = k_3x_3(t)$ to be a generic forcing function, and assigned constants

$$K_{11} = -\frac{k_1 + k_2}{m_1}, \quad K_{12} = \frac{k_2}{m_1}, \quad K_{21} = \frac{k_2}{m_2}, \quad K_{22} = -\frac{k_2 + k_3}{m_2}.$$

Further denoting the matrix in square brackets by A, and letting $z = (x_1, x_2, y_1, y_2)^T$, $b = (0, 0, 0, 1)^T$, the system becomes

$$\dot{z} = Az + bf(t). \tag{7.3}$$

We can perform a similar resonance analysis for this coupled system as for the system in the previous section. Choose the input signal $f(t) = e^{i\mu t}$, keeping in mind that we then are interested in the real part of $z(t)$. We can again see that the solution is only determined up to a homogeneous solution $z_h(t)$ that solves

$$\dot{z}_h = Az_h$$

and satisfies the boundary conditions. The eigenvalues λ_j, $j = 1, \ldots, 4$, of A are the roots of the (quartic) characteristic polynomial defined by the condition

$$\det(\lambda I - A) = 0.$$

Assuming these satisfy $\operatorname{Re}\lambda_j < 0$, for all j (which they do if $c_{1,2} > 0$), the homogeneous solution eventually damps out: $\lim_{t\to\infty} z_h(t) = 0$.

To determine the particular solution $z_p(t)$, we try

$$z_p(t) = He^{i\mu t},$$

where H is a constant vector that depends on the input frequency μ. Inserting this solution in (7.3) gives

$$i\mu He^{i\mu t} = AHe^{i\mu t} + be^{i\mu t}.$$

Solving for H gives

$$H = (i\mu I - A)^{-1}b.$$

If we are interested in the response of a particular variable, say x_2, to this forcing, note that we can extract this from z using $x_2(t) = e_2^T z(t)$, where $e_2 = (0, 1, 0, 0)^T$ is a canonical unit vector. Then the response function for x_2 is

$$e_2^T H = e_2^T (i\mu I - A)^{-1}b.$$

Again, if the damping is small, then the eigenvalues of A are near the imaginary axis. Then if $i\mu$ comes close to an eigenvalue, the matrix $(i\mu I - A) \approx (\lambda I - A)$ becomes nearly singular, and its inverse is large. This can lead to large values of H and resonance.

As an example, we take $m_1 = 1$, $m_2 = 2$, $k_1 = k_3 = 1$, $k_2 = 1.5$, $c_1 = c_2 = 0.1$. In this case, the eigenvalues of A are $\lambda_{1,2} = -0.0312 \pm 0.8019$ and $\lambda_{3,4} = -0.0438 \pm 1.7617$. The response functions for all four components of z are shown in Figure 7.3. Resonances occur near the eigenvalues (natural frequencies).

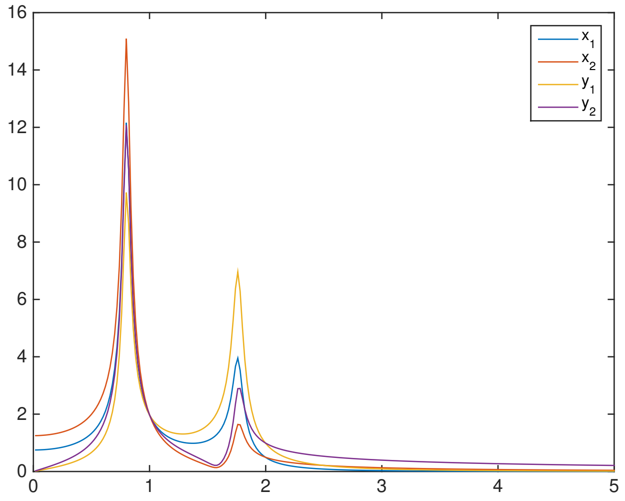

Figure 7.3: Response functions for coupled system.

Chapter 8

Higher order numerical methods

8.1 The Group of Flow Maps

Let us next consider the general autonomous initial value problem

$$\frac{dy}{dt} = f(y), \qquad y \in \mathcal{D},\ f : \mathcal{D} \to D, \quad t \in [0, T], \quad y(0) = y_0. \tag{8.1}$$

Given a fixed time Δt, we can consider any point of phase space, y_0, as a starting point of a trajectory $(y(t) = y(t; y_0))$ which is continued up to time τ, assuming the solution exists on the entire interval. We think of the action of solving the differential equation as defining a map from starting points of trajectories to their endpoints. Associated to a differential equation (8.1) we define the *flow map*

$$\Phi_\tau(y_0) = y(\tau)$$

where $y(t)$ is the solution of the initial value problem

$$\frac{dy}{dt} = f(y), \quad y(0) = y_0, \quad t \in [0, \tau].$$

When the system is nonlinear there is usually not a simple formula for the flow map. Nonetheless the concept is valuable.

Consider solving the ODE (8.1) with initial condition $y(0) = y_0$ for a time t_1. Denote solution at t_1 by y_1. Next, solve initial value problem (8.1) with initial condition $y(0) = y_1$ on the interval $[0, t_2]$. Denote the solution at time t_2 by y_2. It is clear that y_2 is also the same solution you would obtain if you would solve (8.1) with initial condition $y(0) = y_0$ over the interval $[0, t_1 + t_2]$. In terms of the flow map,

$$\Phi_{t_2} \circ \Phi_{t_1} = \Phi_{t_1+t_2} = \Phi_{t_1} \circ \Phi_{t_2}, \qquad t_1, t_2 > 0.$$

Thus the composition of the two maps is another map of the same family. If we can solve the differential equations for all positive or negative time, then we can write

$$\Phi_t \circ \Phi_{-t} = \Phi_{t+-t} = \Phi_0,$$

but Φ_0 is evaluated by solving the differential equations over a zero length interval, so is just the identity map. Hence we see that $\Phi_{-t} = \Phi_t^{-1}$. The set of all such maps $\{\Phi_t \,|\, t \in \mathbf{R}\}$ is what is termed a *one-parameter group* with the (commutative) group operation being composition of maps.

The flow map can also be applied to an open set $B \subset \mathbf{R}^d$, and then we view the flow map as a function taking sets on phase space to other sets on phase space, t time units later. In other words, each element of B is taken as the initial condition of (8.1), and is mapped to the solution at the end of the interval of time t.

$$\Phi_{\Delta t}(B) = \{z \in \mathbf{R}^d : z = \Phi_{\Delta t}(y); y \in B\}$$

Implicit in this usage is the assumption that the trajectories through every point of B exist for at least t units of time.

For example, in Figure 8.1 a rectangular set of initial conditions in $\mathbf{R}^2$ is mapped under the flow map $\Phi_{\Delta t}$ of a differential equation of the form

$$\frac{dx}{dt} = y, \qquad \frac{dy}{dt} = \frac{a}{x^7} - \frac{b}{x^{13}}, \tag{8.2}$$

for $\Delta t = 0,\ 0.35,\ 0.9$. The solutions are periodic.

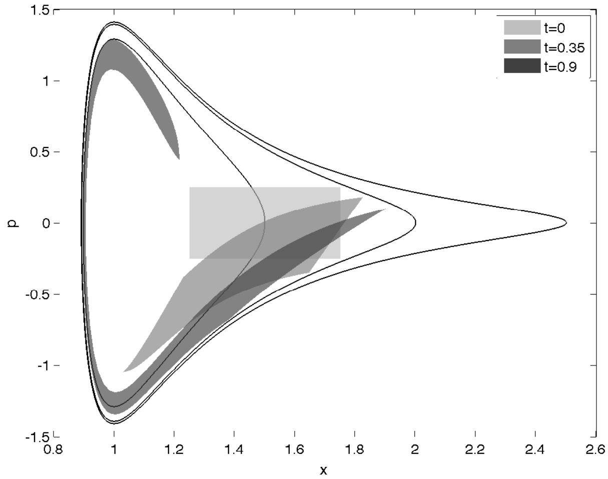

Figure 8.1: Evolution of a rectangular set B under the flow map of (8.2) at times $t = 0$, $t = 0.35$ and $t = 0.9$.

8.2 Numerical Flow Maps

When we discretize an ODE, we normally replace it by a recursion which describes in an approximate sense how the solution evolves from timestep to timestep, as with Euler's method

$$y_{n+1} = y_n + \Delta t f(y_n).$$

We could also think of regularly sampling the *exact solution* on a succession of time intervals of length Δt, according to the rule

$$y(t_{n+1}) = \Phi_{\Delta t}(y(t_n))$$

If $y(t; y_0)$ represents the solution of the differential equation (8.1) with initial condition $y(0) = y_0$ after t units of time, we can write

$$y(\Delta t; y_0) = \Phi_{\Delta t}(y_0).$$

Moreover,

$$y(2\Delta t; y_0) = \Phi_{\Delta t} \circ \Phi_{\Delta t}(y_0),$$

and so on. This means that we can view the iteration of $\Phi_{\Delta t}$ as producing snapshots of the solution at equally spaced points in time. Similarly, Euler's method can be seen as iteration of a map

$$\Psi_{\Delta t}(y) = y + \Delta t f(y)$$

which approximates the flow map $\Phi_{\Delta t}$. We refer to $\Psi_{\Delta t}$ as the *numerical flow map* for Euler's method.

Do the mappings $\Psi_{\Delta t}$ form a one-parameter group? The answer is no. Quite simply,

$$\Psi_{\Delta t_1} \circ \Psi_{\Delta t_2} \neq \Psi_{\Delta t_1 + \Delta t_2}.$$

Indeed, even

$$\Psi_{\Delta t} \circ \Psi_{\Delta t} \neq \Psi_{2\Delta t}.$$

This is a fundamental difference between the exact flow map and its numerical approximation.

There are a wide variety of numerical methods available for solving the ODE (2.11) or (8.1). In some cases these are based on using the values of solutions computed at two or more successive points in time (so-called *multistep methods*). For now we restrict ourselves to *generalized one-step methods* which can always be associated to the recurrence relation

$$y_{n+1} = \Psi_{\Delta t}(y_n). \tag{8.3}$$

The discrete approximation would satisfy

$$y_1 = \Psi_{\Delta t}(y_0), \qquad y_2 = \Psi_{\Delta t}(y_1) = \Psi_{\Delta t} \circ \Psi_{\Delta t}(y_0),$$

and so on. It is useful to have a notation for applying a map n times, recursively. We will use a superscript to indicate this:

$$\Psi^2_{\Delta t} = \Psi_{\Delta t} \circ \Psi_{\Delta t}, \qquad \Psi^3_{\Delta t} = \Psi_{\Delta t} \circ \Psi_{\Delta t} \circ \Psi_{\Delta t},$$

etc.

For explicit methods such as Euler's method, the numerical flow map is obvious. For implicit methods such as the Trapezoidal rule (2.21), it is unclear in advance whether such a map is well-defined. It follows from the implicit function theorem that for Δt small enough and appropriate f, there exists a solution to the trapezoidal rule, but it may not be unique. Typically, there is a branch of solutions that converge to y_n as $\Delta t \to 0$. If we take this branch, then the flow map (8.3) is well-defined.

8.3 Convergence of generalized one-step methods

In this section we consider convergence of the numerical flow map $\Psi_{\Delta t}$ to the exact flow map $\Phi_{\Delta t}$ in the approximation limit $\Delta t \to 0$, thus establishing conditions for the convergence of a large class of one-step methods.

Define the *local error* of a numerical method as the difference between the flow-map and its discrete approximation:

$$le(y, \Delta t) = \Psi_{\Delta t}(y) - \Phi_{\Delta t}(y).$$

The local error measures just how much error is introduced in a single timestep of size Δt. Let us assume that, on our (invariant) domain of interest $\mathcal{D}$, we can expand le in powers of Δt (typically using Taylor series), and that it satisfies

$$\|le(y, \Delta t)\| \leq C\Delta t^{p+1}, \tag{8.4}$$

where C is a constant that depends on $y(t)$ and its derivatives, and $p \geq 1$. A method that meets this criterion is said to be *consistent*.

We will further suppose that our numerical method satisfies a Δt-dependent Lipschitz condition on $\mathcal{D}$

$$\|\Psi_{\Delta t}(u) - \Psi_{\Delta t}(v)\| \leq (1 + \Delta t\hat{L})\|u - v\|, \qquad \forall u, v \in \mathcal{D}. \tag{8.5}$$

The constant $\hat{L}$ is not necessarily the same as the Lipschitz constant for the vector field.

The error can be viewed as the difference between n iterations of $\Psi_{\Delta t}$ and n iterations of $\Phi_{\Delta t}$, thus we define it to be

$$e_n = y_n - y(t_n),$$

so

$$e_{n+1} = y_{n+1} - y(t_{n+1}) = \Psi_{\Delta t}(y_n) - \Phi_{\Delta t}(y(t_n)).$$

To this expression we add and subtract $\Psi_{\Delta t}(y(t_n))$, which is the numerical solution started from a point on the solution trajectory, then take norms to obtain

$$\begin{aligned} \|e_{n+1}\| &= \|\Psi_{\Delta t}(y_n) - \Psi_{\Delta t}(y(t_n)) + \Psi_{\Delta t}(y(t_n)) - \Phi_{\Delta t}(y(t_n))\| \\ &\le \|\Psi_{\Delta t}(y_n) - \Psi_{\Delta t}(y(t_n))\| + \|\Psi_{\Delta t}(y(t_n)) - \Phi_{\Delta t}(y(t_n))\|. \end{aligned}$$

Now we use our two assumptions (8.4) and (8.5) to write

$$\begin{aligned} \|e_{n+1}\| &\le (1 + \Delta t\hat{L})\|y_n - y(t_n)\| + C\Delta t^{p+1} \\ &= (1 + \Delta t\hat{L})\|e_n\| + C\Delta t^{p+1}. \end{aligned}$$

Applying Lemma 2.4.1 then yields

$$\|e_n\| \le \frac{C\Delta t^{p+1}}{\kappa - 1}(\kappa^n - 1) + \kappa^n\|e_0\|$$

where $\kappa = 1 + \Delta t\hat{L}$. Finally, since $1 + \Delta t\hat{L} \le e^{\Delta t\hat{L}}$ and therefore $\kappa^n \le e^{T\hat{L}}$, if we assume the initial condition is exact $e_0 = 0$, we get the uniform bound

$$\|e_n\| \le \Delta t^p \frac{C}{\hat{L}}(e^{T\hat{L}} - 1), \tag{8.6}$$

which proves convergence at order p. We summarize this result in an important convergence theorem:

Theorem 8.3.1 (Convergence of One-Step Methods) *Given a differential equation (8.1) and a generalized one-step method $\Psi_{\Delta t}$ which satisfies conditions (8.4) and (8.5), the global error satisfies*

$$\max_{n=0,\ldots,N} \|e_n\| = \mathcal{O}(\Delta t^p).$$

This theorem is powerful. Without specifying anything about the construction of the method, it guarantees the convergence of any one-step method that is consistent and satisfies an Δt-dependent Lipschitz condition.

As an example, let us again prove convergence of Euler's method (2.16) for smooth vector fields f, making use of Theorem 8.3.1. Consider a compact domain $\mathcal{D} \subset \mathbf{R}^d$ and suppose f is smooth on $\mathcal{D}$ and has Lipschitz constant L on $\mathcal{D}$ (since f is smooth, we can take $L = \max_{\mathcal{D}} \|\frac{\partial f}{\partial y}\|$). Then, since

$$\|\Psi_{\Delta t}(y) - \Psi_{\Delta t}(z)\| = \|y + \Delta t f(y) - (z + \Delta t f(z))\| \le \|y - z\| + \Delta t L\|y - z\| = (1 + \Delta t L)\|y - z\|,$$

The numerical flow map is Lipschitz with $\hat{L} = L$.

The exact solution satisfies

$$\Phi_{\Delta t}(y) = y + \Delta t\frac{dy}{dt} + \frac{\Delta t^2}{2}\frac{d^2y}{dt^2} + \mathcal{O}(\Delta t^3) = y + \Delta t f(y) + \frac{\Delta t^2}{2}\frac{d^2y}{dt^2} + \mathcal{O}(\Delta t^3)$$

Therefore the local error is

$$le(y, \Delta t) = y + \Delta t f(y) - \left[y + \Delta t f(y) + \frac{\Delta t^2}{2}\frac{d^2y}{dt^2} + \mathcal{O}(\Delta t^3)\right] = \mathcal{O}(\Delta t^2),$$

and we can apply Theorem 8.3.1 with $C = \max_{\mathcal{D}} \|\frac{d^2y}{dt^2}\|$ to show that Euler's method is convergent with order $p = 1$.

In the proof of the Theorem 8.3.1, the relation (8.6) indicates that the magnitude of the global error bound will be reduced in proportion to Δt^p. For example, when using Euler's method in practice, we typically observe that halving the stepsize reduces the error by a factor of two. We say for this reason that Euler's method is *1st order accurate.* The error incurred in each time step is $\mathcal{O}(\Delta t^{p+1})$, and in fact this bound holds for any fixed number of time steps. The loss of one order of Δt occurs because the number of time steps needed to cover a fixed interval of length T increases as $\Delta t \to 0$ at a rate proportional to $1/\Delta t$.

Note the proof suggests that—although in the limit $\Delta t \to 0$, T fixed, the error can be made as small as possible—in the limit $T \to \infty$, Δt fixed, the global error may grow at an exponential rate.

8.4 Local accuracy of higher-order methods

We will now investigate the accuracy of the methods introduced above in terms of their approximation of the numerical solution over one timestep.

To do so, we need to work with higher order derivatives of the function $f(y) : \mathbf{R}^d \to \mathbf{R}^d$. We will denote by f' the Jacobian matrix Df, which can be seen as an linear operator $f' : \mathbf{R}^d \to \mathbf{R}^d$.

The second derivative

$$f'' = (\frac{\partial^2 f^{(i)}}{\partial y^{(j)}\partial y^{(k)}})$$

is a bilinear operator. We denote its contraction with two vectors $a \in \mathbf{R}^d$ and $b \in \mathbf{R}^d$ by

$$f''(a, b) = \sum_{j,k} \frac{\partial^2 f^{(i)}}{\partial y^{(j)}\partial y^{(k)}} a^{(j)} b^{(k)}.$$

This contraction is symmetric by the equivalence of mixed partial derivatives, so the order of the arguments a and b is irrelevant. The third derivative is similarly trilinear and symmetric in all three arguments, denoted $f'''(\cdot, \cdot, \cdot)$.

In general, we cannot have an exact expression for $le(y;\Delta t)$, but we can approximate this by computing its Taylor series in powers of Δt. For the continuous dynamics (i.e. the exact solution of (8.1)) we have the Taylor series expansion

$$y(t+\Delta t) = y(t) + \Delta t y'(t) + \frac{1}{2}\Delta t^2 y''(t) + \frac{1}{6}\Delta t^3 y'''(t) + \mathcal{O}(\Delta t^4).$$

The derivatives can be related directly to the solution itself by using the differential equation

$$\begin{aligned} y'(t) &= f(y(t)) \\ y''(t) &= f'(y(t))y'(t) \\ y'''(t) &= f''(y(t))(y'(t),y'(t)) + f'(y(t))y''(t) \\ &\vdots \end{aligned}$$

Then we can recursively use the differential equation again to obtain

$$\begin{aligned} y' &= f, \\ y'' &= f'f, \\ y''' &= f''(f,f) + f'f'f, \\ &\vdots \end{aligned}$$

where we have dropped the arguments of the various expressions. In all cases, y and its derivatives are assumed to be evaluated at t and f and its derivatives at y.

Alternatively, we can write

$$\Phi_{\Delta t}(y) = y + \Delta t f + \frac{\Delta t^2}{2} f'f + \frac{\Delta t^3}{6}\left[f''(f,f) + f'f'f\right] + \mathcal{O}(\Delta t^4) \qquad (8.7)$$

The same procedure can be carried out for the Runge-Kutta method itself. For example, for Euler's method, we have

$$\Psi_{\Delta t}(y) = y + \Delta t f(y)$$

(that is it!). This means that the discrete and continuous series match to $\mathcal{O}(\Delta t^2)$ and the local error expansion can be written

$$le(y,\Delta t) = \frac{\Delta t^2}{2} f'f + \mathcal{O}(\Delta t^3).$$

again, we have dropped the argument y of f for notational compactness.

For the trapezoidal rule (2.21), and other implicit schemes, the derivatives need to be computed by implicit differentiation. For simplicity, with y fixed, write $z = z(\Delta t) = \Psi_{\Delta t}(y)$. Then $z(\Delta t)$ must satisfy the implicit relation

$$z = y + \frac{\Delta t}{2}\left(f(y) + f(z)\right)$$

Observe that $z(0) = y$. Next we differentiate the expression for z,

$$z' = \frac{dz}{d\Delta t} = \frac{1}{2}\left(f(y) + f(z)\right) + \frac{\Delta t}{2} f'(z)z', \tag{8.8}$$

(Note that z does not satisfy the differential equation, so we *cannot* replace z' here by $f(z)$!) For $\Delta t = 0$, the last term of (8.8) vanishes and we have

$$z'(0) = f(y).$$

Differentiate (8.8) once more with respect to Δt to obtain

$$z'' = \frac{1}{2} f'(z)z' + \frac{1}{2} f'(z)z' + \frac{\Delta t}{2}(f''(z)(z', z') + f'(z)z''), \tag{8.9}$$

for $\Delta t = 0$, the last term of (8.9) drops out and we have

$$z''(0) = f'(z(0))z'(0) = f'(y)f(y).$$

Using these expressions, we may write the first few terms of the Taylor series in Δt of $\Psi_{\Delta t}(y)$:

$$\Psi_{\Delta t}(y) = z(\Delta t) = z(0) + \Delta t z'(0) + \frac{\Delta t^2}{2} z''(0) + \ldots = y + \Delta t f + \frac{\Delta t^2}{2} f' f + \ldots \tag{8.10}$$

Comparing this with the expansion for the exact solution (8.7) we see that the first few terms are identical. Thus the local error vanishes to *at least* $\mathcal{O}(\Delta t^3)$. To get the next term in the local error expansion, we differentiate (8.9):

$$\begin{aligned} z''' &= f''(z)(z', z') + f'(z)z'' + (1/2)\left(f''(z)(z', z') + f'(z)z''\right) + \\ &\quad \frac{\Delta t}{2}\left(f'''(z)(z', z', z') + 2f''(z)(z', z'') + f''(z)(z', z'') + f'(z)z'''\right), \end{aligned}$$

which evaluates to

$$z'''(0) = (3/2)\left(f''(f, f) + f' f' f\right),$$

which means that the expansions (8.10) and (8.7) *do not* match in the term of 3rd order: for trapezoidal rule, we have

$$le(y, \Delta t) = -\frac{1}{12}\left(f''(f, f) + f' f' f\right)\Delta t^3 + \mathcal{O}(\Delta t^4).$$

Chapter 9

Multistep methods

In this section we introduce one of the two main classes of numerical integrators for dynamical systems: the linear multistep methods. The section contains an overview of the analytical issues relevant to these methods. The other great class of numerical integrators are Runge-Kutta methods. These generalize the one-step methods, of which we have already seen some examples.

Assuming the solution has been computed for some number of time steps, the main idea of multistep methods is to use the previous k step values to approximate the solution at the next step. A linear k-step method is defined as

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = \Delta t \sum_{j=0}^{k} \beta_j f(y_{n+j}). \tag{9.1}$$

For a linear k-step method, we require that $\alpha_k \neq 0$ and either $\alpha_0 \neq 0$ or $\beta_0 \neq 0$. Furthermore, the coefficients in (9.1) are not uniquely defined, since multiplication of both sides by a constant defines the same method. Usually the coefficients are normalized such that either $\alpha_k = 1$, or $\sum_j \beta_j = 1$.

When using (9.1) in a computer code, at time step $n+k-1$ it is assumed that the values y_{n+j}, $j = 0, \ldots, k-1$ are already computed, so y_{n+k} is the only unknown in the formula. If β_k is nonzero the method is implicit, otherwise it is explicit. Since the initial value problem (8.1) only specifies $y_0 = y(t_0)$, it is necessary to first generate data y_j, $j = 1, \ldots, k-1$ before the formula (9.1) can be applied. This is done, for example, by using forward Euler or another one-step method in the first $k-1$ steps.

9.1 Some multistep methods

Some examples of linear multistep methods are:

- The θ-method generalizes all linear one-step methods

$$y_{n+1} - y_n = \Delta t(1-\theta)f(y_n) + \Delta t\theta f(y_{n+1}). \tag{9.2}$$

  Here we have $\alpha_0 = -1$, $\alpha_1 = 1$, $\beta_0 = 1-\theta$ and $\beta_1 = \theta$. Important methods are forward Euler ($\theta = 0$), backward Euler ($\theta = 1$) and trapezoidal rule ($\theta = 1/2$). For any $\theta > 0$, this method is implicit.

- *Leapfrog* is an explicit two-step method ($k = 2$) given by $\alpha_0 = -1$, $\alpha_1 = 0$, $\alpha_2 = 1$ and $\beta_1 = 2$:

$$y_{n+2} - y_n = 2\Delta t f(y_{n+1}) \tag{9.3}$$

- The class of *Adams methods* have $\alpha_k = 1$, $\alpha_{k-1} = -1$ and $\alpha_j = 0$ for $j < k-1$. *Adams-Bashforth methods* are explicit, additionally satisfying $\beta_k = 0$. Examples of 1, 2 and 3-step methods are (using notation $f_n \equiv f(y_n)$):

$$y_{n+1} - y_n = \Delta t f_n \tag{9.4}$$

$$y_{n+2} - y_{n+1} = \Delta t\left(\frac{3}{2}f_{n+1} - \frac{1}{2}f_n\right) \tag{9.5}$$

$$y_{n+3} - y_{n+2} = \Delta t\left(\frac{23}{12}f_{n+2} - \frac{4}{3}f_{n+1} + \frac{5}{12}f_n\right) \tag{9.6}$$

  Adams-Moulton methods are implicit, with $\beta_k \neq 0$.

- The *Backward differentiation formulae (BDF)* are a class of linear multistep methods satisfying $\beta_j = 0$, $j < k$ and generalizing backward Euler. The two-step method (BDF-2) is

$$y_{n+2} - \frac{4}{3}y_{n+1} + \frac{1}{3}y_n = \Delta t\frac{2}{3}f(y_{n+2}). \tag{9.7}$$

9.2 Order of accuracy and convergence

Associated with the linear multistep method (9.1) are the polynomials

$$\rho(\zeta) = \sum_{j=0}^{k}\alpha_j\zeta^j, \qquad \sigma(\zeta) = \sum_{j=0}^{k}\beta_j\zeta^j. \tag{9.8}$$

These are important for understanding the dynamics of multistep methods, and will be used later.

The *residual* of a linear multistep method at time t_{n+k} may be defined in a number of ways. We obtain it by substituting the exact solution $y(t)$ of (8.1) at times $y(t_{n+j})$, $j = 0, \ldots, k$ into (9.1), i.e.

$$r_n := \sum_{j=0}^{k}\alpha_j y(t_{n+j}) - \Delta t\sum_{j=0}^{k}\beta_j y'(t_{n+j}). \tag{9.9}$$

(This is actually the residual accumulated in the $(n+k-1)$th step, but for notational convenience we will denote it r_n.) A linear multistep method has maximal *order of accuracy* p if $r_n = \mathcal{O}(\Delta t^{p+1})$ for all sufficiently smooth f.

Write the Taylor series expansions of $y(t_{n+j})$ and $y'(t_{n+j})$ as

$$y(t_{n+j}) = \sum_{i=0}^{\infty} \frac{(j\Delta t)^i}{i!} y^{(i)}(t_n), \qquad y'(t_{n+j}) = \sum_{i=0}^{\infty} \frac{(j\Delta t)^i}{i!} y^{(i+1)}(t_n),$$

where in this section $y^{(i)}(t)$ means the ith derivative of $y(t)$. Substituting these into (9.9) and manipulating,

$$\begin{aligned} r_n &= \sum_{j=0}^{k} \alpha_j \sum_{i=0}^{\infty} \frac{(j\Delta t)^i}{i!} y^{(i)}(t_n) - \Delta t \sum_{j=0}^{k} \beta_j \sum_{i=0}^{\infty} \frac{(j\Delta t)^i}{i!} y^{(i+1)}(t_n) \\ &= \sum_{j=0}^{k} \alpha_j y(t_n) + \sum_{i=1}^{\infty} \sum_{j=0}^{k} \alpha_j \frac{(j\Delta t)^i}{i!} y^{(i)}(t_n) - \sum_{i=1}^{\infty} \sum_{j=0}^{k} \beta_j \frac{(j\Delta t)^i}{j(i-1)!} y^{(i)}(t_n) \\ &= \sum_{j=0}^{k} \alpha_j y(t_n) + \sum_{i=1}^{\infty} \frac{1}{i!} \Delta t^i y^{(i)}(t_n) \left[\sum_{j=0}^{k} \alpha_j j^i - i \sum_{j=0}^{k} \beta_j j^{i-1} \right]. \end{aligned}$$

Equivalent conditions for a linear multistep method to have order of accuracy p are:

- The coefficients α_j and β_j satisfy (where $0^0 = 1$)

$$\sum_{j=0}^{k} \alpha_j = 0 \qquad \text{and} \qquad \sum_{j=0}^{k} \alpha_j j^i = i \sum_{j=0}^{k} \beta_j j^{i-1} \quad \text{for} \quad i = 1, \ldots, p. \tag{9.10}$$

- The polynomials $\rho(\zeta)$ and $\sigma(\zeta)$ satisfy

$$\rho(e^z) - z\sigma(e^z) = \mathcal{O}(z^{p+1}). \tag{9.11}$$

- The polynomials $\rho(\zeta)$ and $\sigma(\zeta)$ satisfy

$$\frac{\rho(z)}{\log z} - \sigma(z) = \mathcal{O}((z-1)^p). \tag{9.12}$$

The first of these follows from the considerations above. For proofs of the second and third forms, see Hairer, Nørsett and Wanner (1993).

Examples. The method (9.3) has order 2. The methods (9.5) and (9.6) have orders 2 and 3, respectively. The method (9.7) has order 2.

9.3 The root condition, a counter-example

Earlier we saw that for Euler's method, convergence follows from the fact that the residual is $\mathcal{O}(\Delta t^2)$, leading to a global error of $\mathcal{O}(\Delta t)$ on a fixed interval. For linear multistep methods, first order accuracy alone is insufficient to ensure convergence. We will not prove convergence for this class of methods, but will simply state the convergence theorem and show where it can go wrong.

The method (9.1) is said to satisfy the *root condition*, if all roots ζ of

$$\rho(\zeta) = 0,$$

lie on the unit disc ($|\zeta| \leq 1$), and any root of modulus one ($|\zeta| = 1$) has multiplicity one. Note that $\rho(\zeta)$ is the characteristic polynomial of the recursion defined by the left side of (9.1) and corresponds to the multistep method applied to $y' = 0$. The root condition is necessary to ensure that the origin is stable when integrating the trivial differential equation.

Furthermore, a linear multistep method is incomplete without a starting procedure to generate the first $k-1$ iterates $y_1, \ldots, y_{k-1}$.

Theorem 9.3.1 *Suppose a linear multistep method (9.1) is equipped with a starting procedure satisfying* $\lim_{\Delta t \to 0} y_j = y(t_0 + j\Delta t)$ *for* $j = 1, \ldots, k-1$. *Then the method converges to the exact solution of (8.1) on a fixed interval as* $\Delta t \to 0$ *if and only if it has order of accuracy* $p \geq 1$ *and satisfies the root condition.*

The proof of this theorem will not be handled in these notes. See, e.g., the monograph of Hairer, Nørsett, & Wanner (1993).

To illustrate the necessity of the root condition, consider the method

$$y_{n+3} + y_{n+2} - y_{n+1} - y_n = \Delta t \left(\frac{8}{3} f(y_{n+2}) + \frac{2}{3} f(y_{n+1}) + \frac{2}{3} f(y_n) \right). \tag{9.13}$$

Substituting $\rho(\zeta) = \zeta^3 + \zeta^2 - \zeta - 1$ and $\sigma(\zeta) = \frac{8}{3}\zeta^2 + \frac{2}{3}\zeta + \frac{2}{3}$ into (9.11) gives

$$\rho(e^z) - z\sigma(e^z) = \frac{1}{3} z^4 + \mathcal{O}(z^5),$$

so the method is third order accurate. Now applying the method to the easiest of all initial value problems

$$y' = 0, \quad y(0) = 1, \quad t \geq 0$$

yields the linear difference equation

$$y_{n+3} + y_{n+2} - y_{n+1} - y_n = 0. \tag{9.14}$$

If the roots ζ_1, ζ_2 and ζ_3 of the characteristic polynomial $\rho(\zeta) = 0$ were distinct, the exact solution of such a recursion would be

$$y_n = c_1\zeta_1^n + c_2\zeta_2^n + c_3\zeta_3^n.$$

If any root ζ_i had modulus greater than 1, then the recursion would satisfy $|y_n| \to \infty$ unless the corresponding constant c_i were identically 0. For the current case, $\rho(\zeta) = (\zeta - 1)(\zeta + 1)^2$, and there is a double root at -1. For a double root $\zeta_3 \equiv \zeta_2$, the solution of the recursion (9.14) becomes

$$y_n = c_1\zeta_1^n + c_2\zeta_2^n + c_3 n\zeta_2^n.$$

One still has $|y_n| \to \infty$ (but with linear growth), unless $c_3 = 0$.

The constants c_1, c_2 and c_3 are determined by the initial conditions necessary to start the multistep method. Suppose we take $y_0 = y_1 = y_2 = 1$, consistent with the exact solution. Then this yields

$$c_1 = 1, \quad c_2 = c_3 = 0,$$

and the solution is $y_n = 1$, for all n. The method is exact.

Suppose, however, that the initial conditions are perturbed slightly. We take $y_0 = 1 + \varepsilon$, $y_1 = y_2 = 1$. Then we find

$$c_1 = 1 + \frac{3}{4}\varepsilon, \quad c_2 = \frac{1}{4}\varepsilon, \quad c_3 = -\frac{1}{2}\varepsilon,$$

and the solution is unbounded. The numerical solution sequence is unstable to perturbations in the initial conditions. As a consequence, any errors incurred will destabilize the solution. The method works only for the trivial differential equation, $y' = 0$, and then only if the starting procedure is exact. It fails to converge for all other differential equations.

As an example we compute the solution of $y' = -y$, $y(0) = 1$, $t \in [0, 1.5]$ using (9.13) for $\Delta t = 1/10$, $1/100$ and $1/1000$. The instability gets *worse* as Δt decreases.

Theorem 9.3.2 *The maximum order of a k-step method satisfying the root condition is $p = k$ for explicit methods and, for implicit methods, $p = k + 1$ for odd k and $p = k + 2$ for even k.*

9.4 Stability

An important criterion for distinguishing between different methods is their ability to preserve the stability of a stable equilibrium. To test this, we check under what conditions

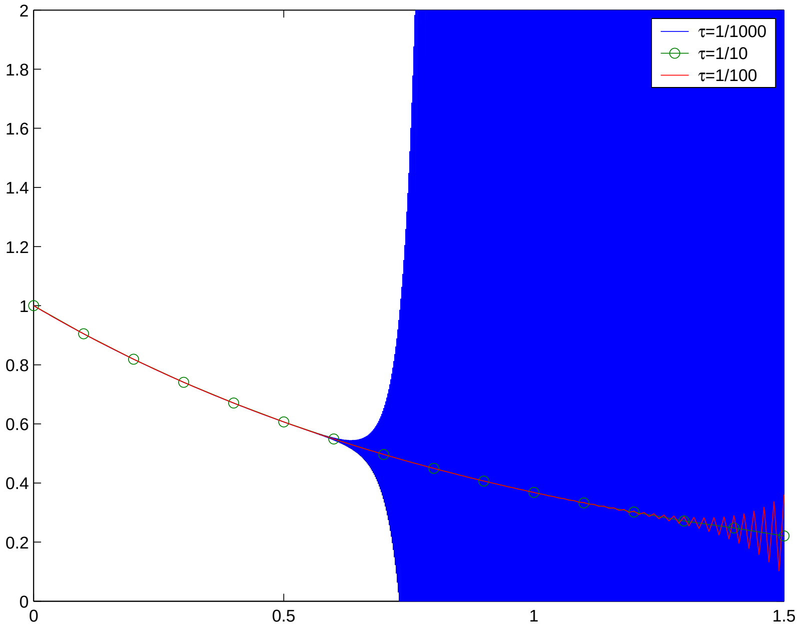

Figure 9.1: Solution of $y' = -y$, $y(0) = 1$ with (9.13) for various Δt.

the numerical solution converges to zero when we apply (9.1) to the scalar complex linear test problem $y' = \lambda y$, $\lambda \in \mathbf{C}$:

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = \Delta t \lambda \sum_{j=0}^{k} \beta_j y_{n+j}.$$

Letting $z = \Delta t \lambda$ we write

$$\sum_{j=0}^{k} (\alpha_j - z\beta_j) y_{n+j} = 0.$$

For any z this is a linear difference equation with characteristic polynomial

$$\sum_{j=0}^{k} (\alpha_j - z\beta_j) \zeta^j = 0 = \rho(\zeta) - z\sigma(\zeta).$$

The *stability region* $\mathcal{S}$ of a linear multistep method is the set of all points $z \in \mathbf{C}$ such that all roots ζ of the polynomial equation $\rho(\zeta) - z\sigma(\zeta) = 0$ lie on the unit disc $|\zeta| \leq 1$, and those with modulus one are simple.

On the boundary of the stability region $\mathcal{S}$, precisely one root has modulus one, say $\zeta = e^{i\theta}$. Therefore an explicit representation for the boundary of $\mathcal{S}$ is easily derived:

$$\partial\mathcal{S} = \left\{ z = \frac{\rho(e^{i\theta})}{\sigma(e^{i\theta})},\ \theta[-\pi, \pi] \right\}.$$

Figure 9.2 shows plots of the stability regions for the Adams-Bashforth methods of orders $p = 1$, 2 and 3.

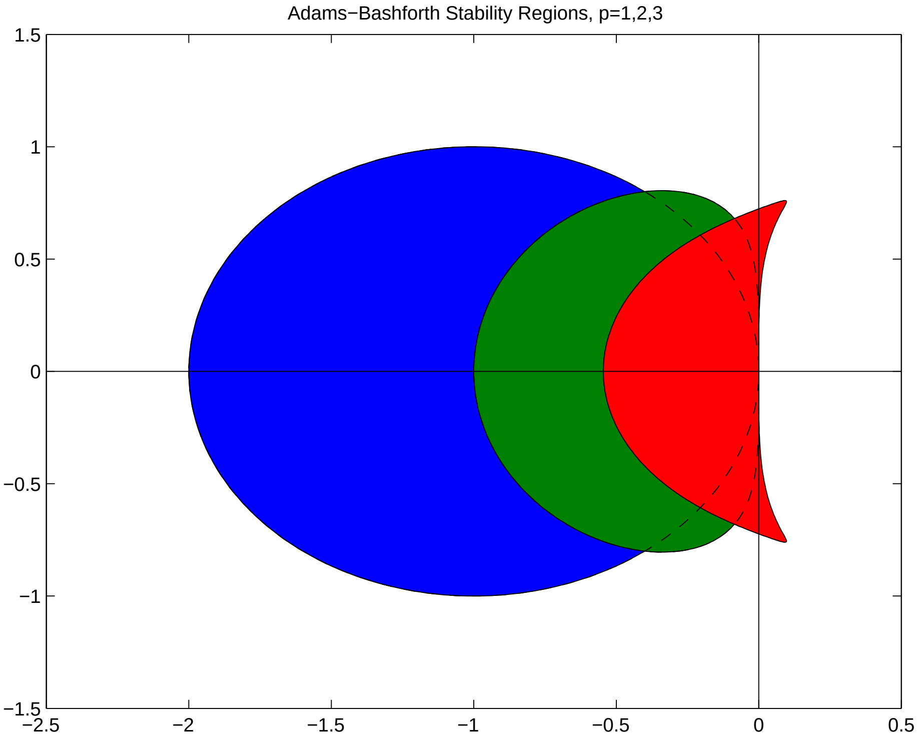

Figure 9.2: Stability regions of the Adams-Bashforth methods of orders $p = 1$ (blue), 2 (green) and 3 (red).

A linear multistep method is called *A-stable* (or *unconditionally stable*) if the stability domain $\mathcal{S}$ contains the entire left half-plane

$$\{z \in \mathbf{C} : \operatorname{Re} z \leq 0\} \subset \mathcal{S}.$$

Theorem 9.4.1 *An A-stable linear multistep method has order $p \leq 2$.*

This restriction on the maximum order of a linear multistep method was an important result in numerical analysis, proved by G. Dahlquist.

For stiff problems in which the stiff components have eigenvalues near the real axis, A-stability is too strong a requirement. Instead a weaker concept is introduced:

The linear multistep method (9.1) is A(α)-stable, for $\alpha \in (0, \pi/2)$, if the stability domain contains a wedge in the left half-plane:

$$\{z \in \mathbf{C} : |\arg(z) - \pi| < \alpha\} \subset \mathcal{S}.$$

The important point is that for λ lying within the wedge of stability, the method is unconditionally stable (norm-nonincreasing for any Δt).

Figure 9.3 shows plots of the stability regions for the backward differentiation formulae (BDF) of orders $p = 1, \ldots, 6$. The first and second order methods are A-stable. The rest are A(α)-stable with α (approximately):

Order, p	α
3	86.03°
4	73.35°
5	51.84°
6	17.84°

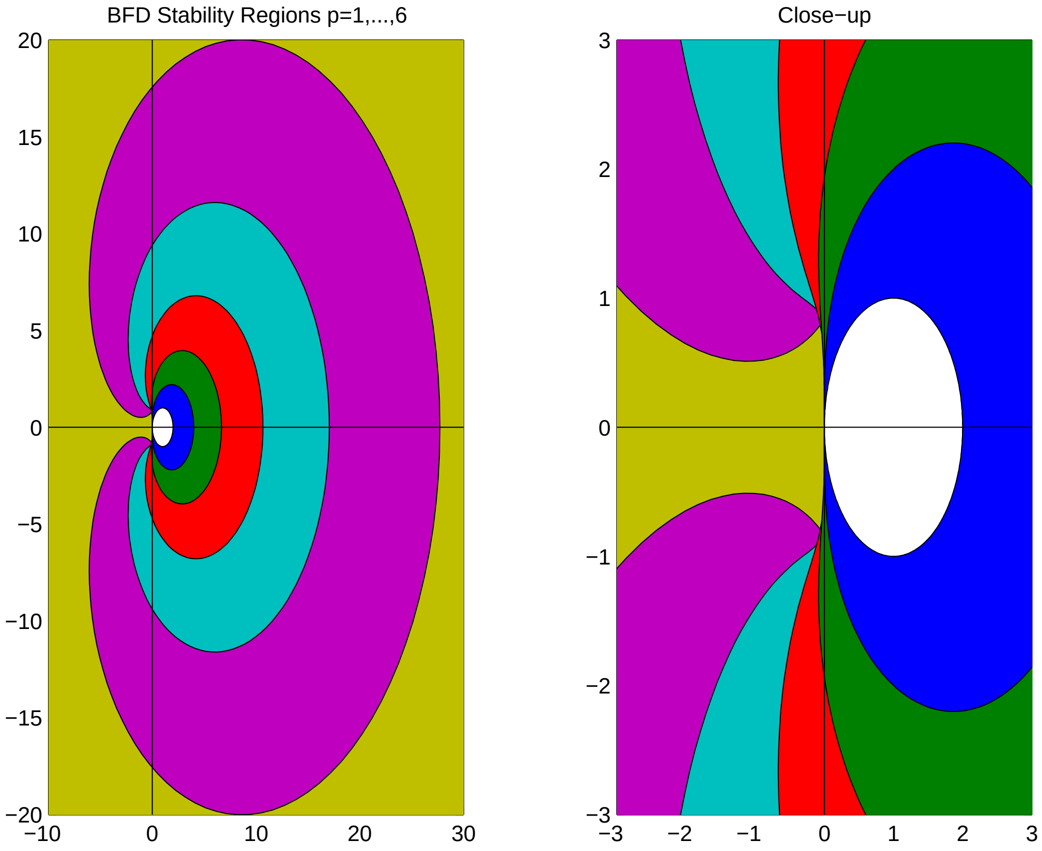

Figure 9.3: Stability regions of the BDF methods of orders $p = 1, \ldots, 6$. For BDF-1 $\mathcal{S}$ is everything outside the white region; for BDF-2 it is everything outside the white and blue regions; etc.