Mathematical modelling – an introduction

Leiv Øyehaug

October 12, 2021

Preface

In the course plan of ACIT4310 it is stated that

- *the course will provide the students with an understanding of what a mathematical model is and how we use models to gain insights into systems and processes in science and engineering*

and that

- *a student who has completed the course should be able to derive mathematical models from facts and first principles for a selection of dynamical systems, be aware of the usefulness and limitations of mathematical modelling as well as of pitfalls frequently encountered in modelling and simulation and to discuss properties of a system using the equations of the mathematical model*

The course literature [3, 6] focuses on methods and therefore falls short of fulfilling the above ambitions. Instead the present compendium will be used to

- explain why and how models are developed
- give some examples of mathematical models

Thus the text should prepare and motivate the students for later work with mathematical and computational methods.

Contents

Chapter 1

Introduction

> The miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve.
>
> E. P. Wigner

Scientists and engineers use mathematics as a tool when they want to describe the world around us. In engineering, mathematical models are used to improve existing and invent new technologies. In science, models are useful to understand a certain system and want to test the effects of changing the systems. In business and industry it is increasingly common that mathematicians, engineers and professionals from other disciplines work side-by-side to solve problems of societal, medical and environmental importance. The expert making the mathematical model translates laws of nature, rules of thumb, typical human or animal behaviour or other relations or interactions in nature or society into a mathematical model that can be analysed and simulated.

Mathematical models have been used for many centuries already to understand physical systems, which is why mathematics is often referred to as the *language of physics*. Increasing computational capabilities during the past decades have enabled analysis and simulation of highly complex models, for example models of oil and gas recovery from reservoirs below the North Sea, models used in designing fuel-efficient cars and aeroplanes, improved weather and climate models yielding accurate weather forecasts in the short term and predicting climate change in the very long term, and models ap-

plied to improve our understanding of biological systems leading to better treatment of disease and genetic disorders.

Advances in computer technologies also mean that much of our lives happens on the Internet, playing online computer games, interacting with friends on social media and spending our salaries shopping online. Many computer games rely on highly advanced mathematical models incorporated into so-called *physics engines* that e.g. ensure that the football in the FIFA computer game moves through the air and bounces in the ground in a realistic manner. Further, our online activity has led to an immense amount of data – *big data* – that are of interest to security agencies, actors in e-commerce and of course to giants of the Internet such as Google, Facebook, Amazon and others. The algorithms used to analyse these data require advanced mathematics and modelling. Models used in this setting are, however, quite different from the models that are the scope of this course.

Wikipedia [1] defines a mathematical model as "*a description of a system using mathematical concepts and language*». The process of developing a mathematical model is termed *mathematical modelling*. When formulating a mathematical model we translate our beliefs about how the world functions into mathematical language. This has advantages since

- mathematics is a very precise language that is appropriate for formulating our beliefs and assumptions about the system under study.
- mathematics has well-defined rules for manipulations
- all the results that mathematicians have proved over hundreds of years are at our disposal.
- the equations of the mathematical model can be solved using computers

Thus, translating a system into mathematical language is a natural and efficient way of improving our understanding of this system. This compendium will be concerned with *why* we need mathematical models and *how* we make them.

Chapter 2

Mathematical modelling – why and how?

All models are wrong, but some are useful

George Box

To varying degree, mathematics is a tool in all scientific and engineering disciplines and even surround us in our daily lives. This chapter provides some background concerning the purpose of making models, why compromises need to be made in modelling and give an overview of some types models, but first a case study of how two simple *ordinary differential equation* (ODE) models can be applied in archaelogy and forensic science, is given.

2.1 The purpose of modelling

What is the purpose of making mathematical models? The answer depends on both the state of knowledge about a system and how well the modelling is done. Examples of objectives are:

1. *Develop scientific understanding.* As in experimental research, it is often the case in modelling that one or more hypotheses are to be tested. Based on a comparison between the outcome of analysis and simulation and empirical data, the hypothesis is either rejected or confirmed.

2. *Predict outcomes of experiments.* Providing that the mathematical model has been validated and is known to behave very much like the

real system, experiments that are performed *in silico* instead of in real-life will reduce costs, alleviate the need for animal experiments and for performing experiments that might cause harm to the environment, such as nuclear testing.

3. *Estimate parameters in a system.* Models contain a number of parameters whose values are often not known. Imagine we have a mathematical model of a chemical system with unknown values for some reaction rates in the model. The values of these reaction rates that produce model results that most closely resemble empirical data will be our estimates for the reaction rates.

4. *Test the effect of changes in a system.* Just as in the experiment prediction case, a mathematical model can be used to test the effect of performing various changes to the system. In cell biology, for example, one might test the effect of injecting a certain substance into the cell using a mathematical model.

5. *Aid decision making.* Modelling and simulations are widely used in military planning to describe various possible scenarios. Generals may base their decisions on the results of such simulations.

Clearly many of these purposes are related, scientific understanding can be developed at the same time as estimating parameters, and effects of changes are often tested when generating new scientific insights.

2.2 Compromises in mathematical modelling

There is a large element of compromise in mathematical modelling. The majority of real world-systems are far too complicated to model in their entirety. Hence in all modelling processes one needs to identify the most important parts of the system, include these in the model and exclude the rest. Of course, *which* are the most important parts will depend on the questions one wants to answer. One modeller might be interested in studying one part of a system which another modeller would gladly exclude from her model since her interests lie elsewhere.

For example, when one models a stone falling from a low height, air resistance can be justifiably neglected. On the other hand, air resistance

is of course crucial when one describes the motion of a parachuter, both before and after opening of the parachute. Another example, quite frequently encountered in modelling, is when the time scales of the processes in the system are very different. Then one could choose to make a model valid on the small time scale and another model valid on larger time scales.

Another level of compromise concerns the complexity of mathematical expressions. A relatively simple model is more tractable to analysis than a complex model and, especially when it is difficult to quantify interactions between model actors, Occam's razor suggests that the simplest mathematical expression is the best since it requires the fewest assumptions. Thus a simple model should be favoured over a complex model as long as it captures key behaviours of the system under study.

2.3 The modelling cycle

Mathematical modelling is not a straightforward process. To arrive at a good model, several iterations of the *modelling cycle* is often needed. The modelling process can be summarised in the following five steps (see Fig. 1).

1. Describe the real-world model by identifying important actors and interactions between them.

2. Formulate the mathematical model. Make simplifying assumptions, choose variables, estimate magnitudes of inputs, justify decisions made.

3. Use mathematical analysis and/or numerical methods to calculate the mathematical solution of the model.

4. Interpret the solution. Consider mathematical results in terms of their real-world meanings.

5. Validate the model against real world observations.

Step 1 requires detailed knowledge of the system under study and a precise description of it. Step 2 requires skills in quantifying the processes of the system. This will in many cases involve formulating hypotheses regarding the speed of the process and regarding which actors of the model that interact. Step 3 requires skills in mathematics and numerics and abilty to implement the model on a computer. In steps 4 and 5, the interpretation and validation

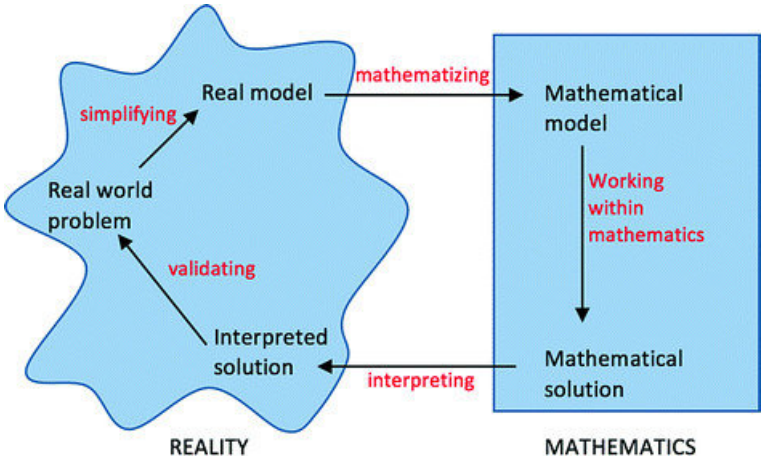


Figure 1: Modelling cycle of Maass [7].

steps, shortcomings of the mathematical model are identified. If the model does not behave in accordance with experimental data and fails to answer the original question(s), necessary modifications should be made and the cycle be repeated until an adequate solution has been found. Furthermore, although the stages are sequential, the cycle is not necessarily smooth, as the constant checking, testing and evaluating contained in each stage means that there is frequent movement within (and between) the stages potentially making the development of some models a very challenging exercise.

2.4 Classification of models

There are many types of mathematical models. In thw following some of the main classes of models are listed and contrasted to other model types.

Mechanistic vs empirical models

A model which uses a large amount of theoretical information is called a *mechanistic model*, because it accounts for the mechanisms through which changes occur. In *empirical models*, no account is taken of this mechanism. Instead, it is merely noted that they *do* occur, and the model trys to account quantitatively for changes associated with different conditions.

Deterministic vs stochastic models

Deterministic models ignore random variation, and so always predict the same outcome from a given starting point. On the other hand, the model may include terms that are more statistical in nature and so may predict the distribution of possible outcomes. Such models are said to be *stochastic*.

Discrete vs continuous models

A *discrete model* treats objects as discrete, such as cells in a biological model or persons in a social network model. A *continuous model* represents objects in a continuous manner such as a velocity field or the temperature in a solid.

Dynamic vs static models

A *dynamic model* accounts for the time-dependence of the system under study, whereas a *static model* describes the state of the system in a steady state (or equilibrium).

Black box vs white box models

The models considered in this course are *white box models*. This means that we know the details of the model and how the model produces predictions, the model is transparent. By contrast, with *black box models* we can observe input and output of the model, but what it is that produces the output is not visible to us. Many machine learning models are black-box models.

Mathematical vs computational vs simulation vs computer model

Much of the modelling literature refers to *computational, simulation and computer* models (rather than mathematical models). How do these relate to mathematical models? Most often the words "computational", "simulation" and "computer" in this setting refer to the way the model calculations are done - i.e. by computer simulation. The actual model of the system is not changed by the way solutions are obtained.

The system model

One further type of model, the system model, is worthy of mention. The system model is built from a series of sub-models, each of which describes some interacting components and each of which may be of different types. Thus, for example, a model may consist of one or more sub-models which are discrete and one or more which are continuous.

Chapter 3

Examples of models

> A good theory should be as simple as possible, but not simpler
>
> Albert Einstein

Using examples of ODE models, this section is concerned with how models are developed and evaluated and how the model's behaviour can be deduced from inspection of the model equations. Constructing the equations of the ODE model can be done in many ways. Models in physics are often first principles-based models, i.e. derived using laws of nature such as mass conservation, energy conservation, Newton's laws etc. By contrast, models in population biology rely on *assumptions* regarding the rate of population growth and decrease. Modelling in physiology represent a mixture of these modelling approaches: first principles are typically used to describe electrical currents in cell membranes and assumptions are required to model how ion channel permeabilities are regulated by membrane voltage.

3.1 Estimating time of death

Archaeologists frequently use carbon dating to estimate the age of an artefact or a the remains of a human skull. On a vastly different time-scale, a forensic scientist can apply Newton's law of cooling to estimate the time since the crime was committed in a murder investigation. In both cases, mathematical models are used to estimate the time of death.

Radioactivity

Given a sample of a particular radioisotope, the number of decay events dN expected to occur in a small interval of time dt is proportional to the number of atoms present N, that is

$$\frac{dN}{dt} = -\lambda N,$$

where dN/dt is the derivative of N with respect to the time t, the positive parameter λ is specific to the particular radioisotope and N is the size of the nuclei population. This quantity is a natural number, but for any physical sample N is so large that it can be treated as a continuous variable. The negative sign indicates that N decreases as time increases, as the decay events follow one after another. The solution to this first-order differential equation is the function

$$N(t) = N_0\, e^{-\lambda t},$$

where N_0 is the value of N at time $t = 0$.

Given a sample of a particular radionuclide, the half-life is the time taken for half the radionuclide's atoms to decay. i.e. the solution of the equation $N(t_{1/2}) = N_0/2$;

$$t_{1/2} = \frac{\ln 2}{\lambda}.$$

For example, a sample of the carbon isotope ^{14}C has a half-life of 5,730 years and thus a decay rate $\lambda = \ln 2/5730$ measured in years^{-1}. Living organisms contain a small and roughly constant concentration of ^{14}C due to intake of nutrients. When the organism dies, the amount of ^{14}C will slowly decrease because of radioactive decay. Measurements of ^{14}C in old wood, bones and skulls and comparison with the level of ^{14}C in living organisms provide estimates of when these organisms died.

Fig. 2, left, shows the curve that can be used to estimate the age of an artefact. For example, the skull and bones of the oldest human remains found in Norway (called Sol, found in Søgne in 1994), was estimated to contain roughly a third of the amount of ^{14}C in living organisms. Thus the estimated time of Sol's death was about 9000 years ago. On the other hand, the age of 100,000 year old samples of human skulls and bones can not be estimated accurately using carbon dating since the levels of ^{14}C would be too low in these samples.

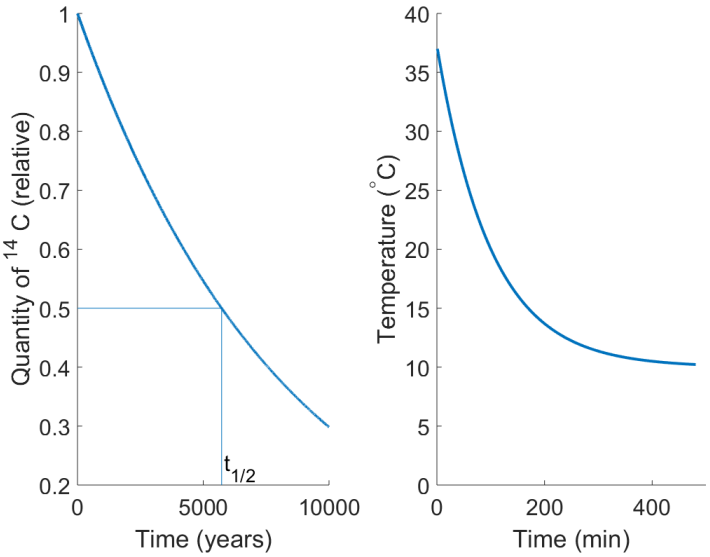


Figure 2: Examples of how the quantity ^{14}C decays with time (left) and how the temperature in a dead body may decrease with time when the temperature of the surroundings is 10 °C.

Newton's law of cooling

Due to its simplicity, Newton's law of cooling is a mathematical model often encountered in calculus courses. This law states that the rate of change of the temperature of a body is proportional to the difference in temperature between the body and its surroundings, i.e. the temperature of the body follows the ODE

$$\frac{dT}{dt} = -k\,(T - T_s)\,,$$

where T and T_s denote the temperature of the body and of the surroundings, respectively. The parameter k is the *heating coefficient* which depends on the material properties of the body. The solution of the initial value problem consisting of the ODE and the initial data $T(0) = T_0$ is

$$T(t) = T_s + (T_0 - T_s)\,e^{-kt}.$$

Assuming that the original (i.e. at $t = 0$) temperature of the coffee in a cup was 100 °C, knowing the value of k (equal to 0.1 min^{-1}, say) and measuring the coffee temperature (40 °C, say) in a room with temperature 20 °C allow

us to estimate how long it is since coffee was poured into the cup:

$$40 = 20 + (100 - 20)\, e^{-0.1\, t} \Leftrightarrow e^{-0.1\, t} = -\frac{20}{80} \Leftrightarrow t = \frac{\ln(20/80)}{-0.1} \approx 13.9,$$

measured in minutes.

In a similar manner forensic scientists can estimate the time of death in a murder investigation. When a person dies, the bodily functions that are responsible for keeping the body temperature at 37 °C, cease and the body starts to cool down roughly according to Newton's law of cooling. When the victim's body is found, the temperature is measured and, as in the coffee example, the time since death can be estimated. Figure 2, right, displays a possible time evolution of the temperature in a dead body when the temperature of the surroundings is 10 °C (for more information, see [9]).

3.2 Models of epidemics

The way the Covid-19 pandemic has impacted and changed the world cannot be understated. Public health institutions globally have invested huge resources into gathering and analysing epidemic data in order to be able to provide the best possible knowledge base for decision makers and advice for ordinary people in their everyday lives. Early on in the pandemic, around the world interdisciplinary teams of researchers gathered in order to record and predict disease spread using methods from statistics and mathematical modelling.

In an early phase of an epidemic, if the epidemic is allowed to spread unhindered, the increase in the number of infected individuals is proportional to the number of already infected individuals with proportionality k (known as the *relative growth rate*), i.e. if P is the number of infected individuals, the differential equation

$$\frac{dP}{dt} = kP$$

governs the dynamics of P. The general solution is $C \exp[kt]$ (where C is an arbitrary constant), that is, the infected part of the population will grow exponentially. Although this model is the simplest possible model for disease spread it is a reasonable model for the initial phase epidemics. It even produces interesting results when governments implement restrictions on the population. The strictness of restrictions can be reflected by the value of the

proportionality k. Thus, when there is strict regulation, the value of k could be much smaller than in absence of restrictions. Thus, *with* restrictions the growth would still be exponential, but much smaller.

The model above predicts that P goes to infinity with time (since $C\ \exp[kt]$ is a function that is not bounded) which is clearly not possible. One way of dealing with this is to make the reasonable assumption that the relative growth rate k will go to zero when the whole population is infected; therefore k is often replaced by $k(N-p)$, i.e. we consider the ODE

$$\frac{dP}{dt} = k\,(N-P)\,P.$$

Any solution of this ODE which has a nonzero initial value will go to N, the total population, as $t \to \infty$. This means that, according to this model, with time every individual will become infected. One problem with this model is that it does not keep track of the share of individuals that recover from disease, it merely separates the population into the "not yet infected" and "infected". To account for "not yet infected", "infected" and those who have recovered from disease and aquired immunity, we need what is called an *SIR model*.

SIR models

The SIR model is the most used mathematical model for predicting how infectious diseases spread in a population. It represents a step up in complexity compared to the above models, but it is still considered a relatively simple model. The SIR model is a so called *compartmental model* which separates the population into three compartments; one containing the *susceptible* individuals of the population (individuals that are healthy and have not yet been infected) described by the variable S, one containing the *infected* individuals (individuals that are infected and have not yet recovered from disease) described by the variable I, and one containing the *recovered* individuals (those that have recovered from disease and have developed immunity against it) described by the variable R (see Figure 3, top). The rate at which individuals are infected, that is, the rate at which individuals are moved from the "S" to the "I" compartment, is proportional to the product of the number of susceptible and infected individuals, with proportionality β, called the *infection rate*. The rate at which infected individuals are moved from the "I"

to the "R" compartment, i.e. the rate at which infected individuals recover from the disease, is proportional to the number of infected individuals, with proportionality γ, called the *recovery rate*. Thus the ODEs that govern the dynamics of the variables $s = S/N$, $i = I/N$ and $r = R/N$ (*shares of the population* that are susceptible, infected and recovered, respectively), are

$$\begin{aligned} \frac{ds}{dt} &= -\beta si, \\ \frac{di}{dt} &= \beta si - \gamma i \\ \frac{dr}{dt} &= \gamma i. \end{aligned}$$

Every individual belongs to one of the three compartments, i.e. $1 = s + i + r$ (which can be ascertained by summing the derivatives above and using the initial condition). The typical behaviour of these variables is that S is monotonely decreasing, I is increasing at first, then decreasing as S is becoming smaller, and that R is monotonely increasing (see Figure 3, bottom). The model equations show that the rate of infection, βsi, depends both on how many are susceptible and how many are already infected, which makes sense when you think about it: person A becomes infected when person A and an infected person B meet.

The *reproduction number* R_0 is defined as the ratio between the rate of infection and the rate of recovery from disease, i.e.

$$R_0 = \frac{\text{infection rate}}{\text{recovery rate}} = \frac{\beta si}{\gamma i} = \frac{\beta s}{\gamma} \approx \frac{\beta}{\gamma},$$

where the approximation comes from $s \approx 1$, i.e. almost the whole population is asumed to be susceptible. Thus the relation $R_0 = \beta/\gamma$ is most valid in the early phase of the disease spreading. Through media we keep hearing that R_0 should be below 1. The reason why can be demonstrated by considering the ODE of the variable i, representing the number of infected individuals;

$$\frac{di}{dt} = \beta si - \gamma i = \gamma i \left(\frac{\beta}{\gamma} s - 1 \right) \approx \gamma i \left(\frac{\beta}{\gamma} - 1 \right) \approx \gamma i \left(R_0 - 1 \right),$$

where, again, the approximation comes from $s \approx 1$. For $R_0 < 1$ the derivative of i is negative, i.e. the infected part of the population will be reduced if this

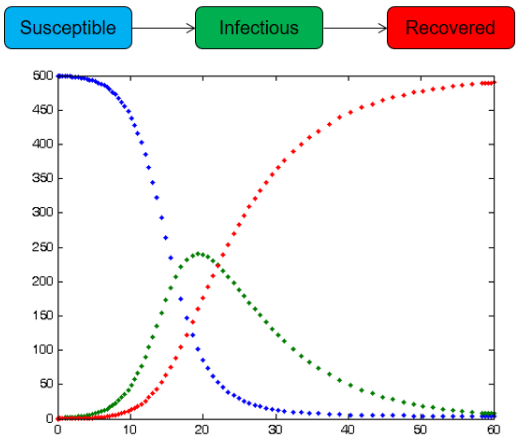


Figure 3: Top: The compartments of the SIR model. Bottom: Typical dynamics of the variables S (blue), I (green) and R (red).

is the case, as desired.

SIR models, originally developed in the 1920s by Kermack and McKendrick [5], have evolved into a large class of compartmental models that describe how individuals who are infected may move through various stages of disease. Such models have been applied to analyse and forecast the Covid-19 pandemic by modelling teams globally (see below). These models can also be applied to analyse and predict the effect of vaccination and of vaning immunity, both phenomena being highly relevant in the pandemic as of August 2021.

Modelling the Covid-19 pandemic

The Norwegian Public Health Institute (NIPH, in Norwegian "FHI") is one of many health agencies around the world which have developed a mathematical model of corona virus spread. Their model is an extended SEIR model (Figure 4, "E" stands for *exposed* and means infected, but not yet infectious) to achieve and maintain situational awareness about the coronavirus outbreak in Norway and to forecast future outbreaks and occurrences of the virus [2]. The NPHI model differs from conventional SEIR model in that the infected part of the population is divided into four groups: (i) exposed (i.e. infected but not yet infectious), (ii) presymptomatic infected, (iii) asymptomatic and (iv) symptomatic. Individuals of the groups (ii), (iii) and (iv) are assumed to be infectious to others. Further, the NPHI model

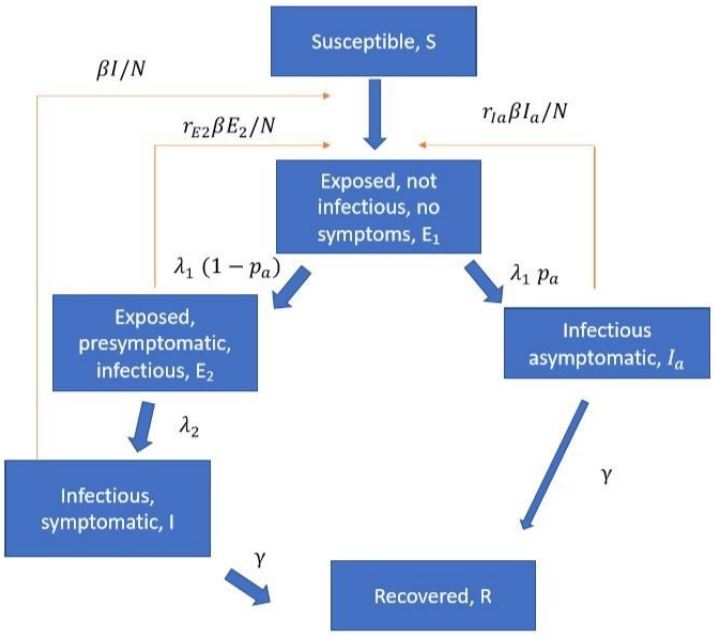


Figure 4: Compartments and transferral rates of the covid-19 model of the Norwegian Public Health Institute (FHI).

describes infections in all counties ("fylker") of Norway and account for mobility (the number of people moving back and forth) between counties using Telenor mobile data. When the researchers at NPHI put together results of model simulations with recordings of Covid-19 infections, they are able to estimate the values of parameters of the model including the parameters that constitute the reproduction ratio R_0. Throughout the pandemic in Norway, NIPH has published regularly the estimated value of R_0 and used this estimate (as well as other model results) in their advice to health authorities, government and the Norwegian population regarding level of restrictions and how one should behave.

3.3 Population dynamics

Thomas Robert Malthus already 200 years ago suggested that a population of size $p(t)$ with constant growth rate (i.e. following the differential equation $dp/dt = rp$, where r is constant and, as in the radioactivity example in the Introduction, the variable p is continuous despite the fact that the population size is a natural number) will experience exponential growth. However, the

growth will be limited by the scarcity of food, disease and harshness of the climate.

Logistic growth

In 1836 the mathematician Pierre François Verhulst came up with an alternative model which accounts for limited growth by replacing the constant growth rate in the exponential growth model by a population-dependent rate; $r \to r\,(1 - p/K)$, where $K > 0$ is the *carrying capacity* of the population. Thus the model becomes

$$p' = rp\,(1 - p/K)\,,$$

which is termed the *logistic differential equation*. Some observations can be made from the right hand side of this ODE:

- $p' > 0$ as long as $p < K$, i.e. the population is growing as long as $p < K$. That is, as long as the population has not reached the carrying capacity the population continues to grow.
- $p' < 0$ as long as $p > K$, i.e. the population is decreasing as long as $p > K$. That is, when the population for some reason is above the carrying capacity the population will decrease.
- p' reaches its maximum level when the function $f(p) = p\,(1 - p/K)$ reaches its maximum, i.e. at $p = K/2$.

From these observations it is possible to infer the shape of the population dynamics (how the population will change over time) given the start size of the population. If $p(0) > K$, the population will decrease and approach K. If $p(0) < K$, the population will increase and approach K. Finally, if $p(0)$ is also less than $K/2$, the growth will increase until $p = K/2$. After that the growth decreases giving an S-shaped solution. The solution can be found analytically and is

$$p(t) = \frac{p_0}{p_0 + (K - p_0)\,e^{-rt}} K,$$

where $p_0 = p(0)$, the initial population size. Fig. 5 shows two such solution curves, one with $p(0) > K$ and one with $p(0) < K/2$.

The derivative p' is zero when $p = 0$ (which is obvious since a population cannot grow from zero) and when $p = K$. Thus if the population equals

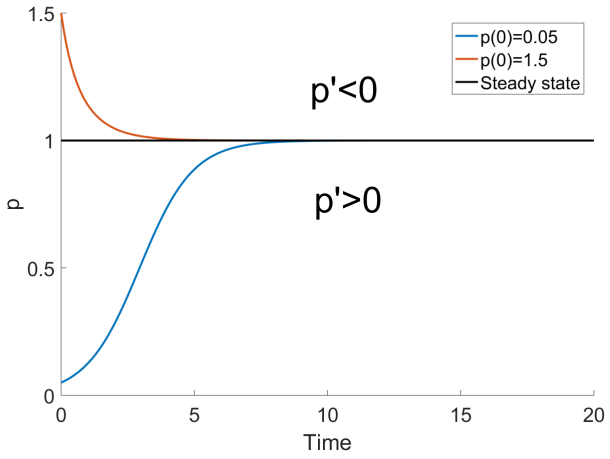


Figure 5: Solution curves of the differential equation () with $r = K = 1$ and $p(0) = 0.05$ (blue) and $p(0) = 1.5$ (red). The stable steady state $p = 1$ is provided for comparison.

the carrying capacity, it stays there. The points where $p' = 0$ are called *equilibrium points* (or just equilibria) or *steady states*. Moving slightly away from the steady state $p = 0$ in positive direction (since a negative value for p would not make sense) the derivative is positive, that is, p will not move back to $p = 0$ with time. Thus $p = 0$ is an *unstable* steady state. Meanwhile, $p = K$ is a *stable* steady state, since a small disturbance (often called a perturbation) in the positive direction causes the derivative to be negative, and p will move back towards $p = K$. Similarly, a small negative perturbation will cause the derivative to be positive and p will move back towards $p = K$.

Predator-prey interactions

The size of the lynx population in the Canadian wilderness depends crucially on the size of the snowshoe hare population, the most important prey for the lynx. On the other hand the number of hares are significantly reduced when lynx are abundant. This in turn affects the lynx population negatively. Population data show that hare and lynx populations are strongly dependent

on each other (Fig. 6, left). The ODE system

$$\begin{aligned} x' &= ax - bxy, \\ y' &= cxy - dy, \end{aligned}$$

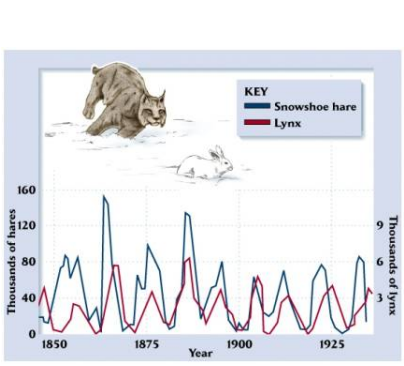
where all the parameters a, b, c and d are positive, is called the *Lotka-Volterra equations*, and may act as a model for the populations of snowshoe hare (x) and lynx (y). Having a closer look at the equations we note that

- in absence of lynx ($y = 0$) the hare population would grow exponentally which, in light of the above, is unrealistic
- in absence of hares ($x = 0$) the lynx population would experience exponential decay which may be realistic (unless lynx can find other prey)
- increasing the value of y sufficiently causes x' to become negative, i.e. when the lynx population is large enough, there will be a decrease in hares
- increasing the value of x sufficiently causes y' to become positive, i.e. when the hare population is large enough, there will be an increase in lynx

In the steady state of a two-variable ODE system, both derivatives will have to be zero. In the Lotka-Volterra equations, there are two steady states;

$$(x^*, y^*) = (0, 0) \quad \text{and} \quad (x^*, y^*) = \left(\frac{d}{c}, \frac{1}{b}\right).$$

The theory behind determining stability of steady states in ODE systems will be the topic later in the course (Sect 2 in [6]), but it can be shown that the former steady state is unstable and that the latter can be both stable and unstable, depending on parameter values. In fact, if the latter steady state is unstable, there will be oscillating solutions to the system. In the case of oscillatory solutions, the period of the oscillations is the same for both variables, but the magnitude of the predator variable y will lag behind the magnitude of the prey variable x. This is in accordance with the expectations since when prey reaches its maximum level, the predator



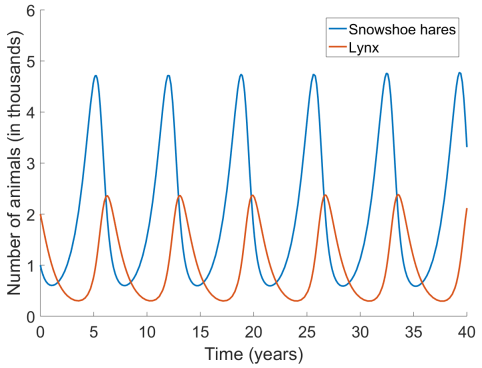


Figure 6: Prey-predator interactions. Left figure: Population data for snowshoe hare and lynx (from http://www.occc.edu). Right: Dynamics of snowshoe hare population size (x - blue) and lynx population size (y - red) obtained from numerical simulations of the Lotka-Volterra equations with parameter values $a = b = d = 1$ and $c = 1/2$ and initial values $x(0) = 1$ and $y(0) = 2$.

population will continue to grow for some time such that the peak in predator population will occur later than the peak in the prey population.

3.4 Second order linear ODE models

In general, a second order linear ODE with constant coefficients is of the form

$$ay'' + by' + cy = f(t).$$

The general solution is of the form $y = y_H + y_P$, where y_H is the general solution to the homogeneous counterpart

$$ay'' + by' + cy = 0,$$

and y_P is a particular solution of the inhomogeneous differential equation. Assuming there is a solution of the form $y_H = e^{\lambda t}$,

$$y_H = Ce^{\lambda_+ t} + De^{\lambda_- t},$$

where $\lambda_\pm$ are the roots of the polynomial $p(\lambda) = a\lambda^2 + b\lambda + c$;

$$\lambda_\pm = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

and C and D are arbitrary constants. In the case that $\lambda_\pm$ are complex, $\lambda_\pm = \mu \pm i\nu$,

$$y_H = e^{\mu t}\left(C\cos(\nu t) + D\sin(\nu t)\right),$$

that is, the general solution is oscillating with increasing or decreasing amplitude (depending on the sign of μ).

The second order ODE can be rewritten as a first order ODE system by introducing the new variable z that satisfies the ODE $y' = z$. Thus $z' = y''$ such that

$$az' + bz + cy = f(t).$$

Letting $v = [y \quad z]^T$, the ODE system

$$v' = \begin{bmatrix} 0 & 1 \\ -\frac{c}{a} & -\frac{b}{a} \end{bmatrix} v + \begin{bmatrix} 0 \\ \frac{1}{a}f(t) \end{bmatrix}$$

is obtained. Writing the second order ODE as a first order system has the advantage that generic numerical schemes can be applied to estimate the solution of the ODE and that the behaviour of its solution can more easily be analysed.

The second order ODE (??) is an extremely useful example of a mathematical model since it can be applied to various systems; here we consider the derivation of this ODE based on Newton's second law of motion a

Newton's second law of motion

Newton's second law of motion states that the sum of all forces acting on a body of mass m is equal to the mass times its acceleration a; $\Sigma F = ma$. Some very useful identities can be derived from this law. Using that the acceleration is the second derivative of position; $a = s''$, that the velocity is the first derivative; $v = s'$, and assuming uniform acceleration, $ma = mv' = ms''$ such that the equations of motion are derived:

$$\begin{aligned} a = v' &\Rightarrow v = at + v_0, \\ a = s'' &\Rightarrow s = \frac{1}{2}at^2 + v_0 t + s_0, \\ \frac{1}{2}vt = \frac{1}{2}at^2 + \frac{1}{2}v_0 t &\Rightarrow s = s_0 + \frac{1}{2}(v + v_0)\,t. \end{aligned}$$

The mass-spring-damper system sketched in Fig. 7 is a classical example of how a second order linear ODE model is derived. Several forces act on the mass; an external force F, the spring acts on it with a force proportional to the displacement x; kx (Hooke's law), and there is a damper which acts on the mass with a force proportional to the velocity $v = x'$; bx'. Thus Newton's second law asserts that, when positive direction is to the right,

$$mx'' = F(t) - kx - bx',$$

i.e.

$$mx'' + bx' + kx = F(t).$$

Thus we have a second order inhomogeneous linear ODE with constant coefficients. The homogeneous counterpart has general solution

$$x_H = Ce^{\lambda_+ t} + De^{\lambda_- t},$$

where the λ's are solutions of the equation $m\lambda^2 + b\lambda + k = 0$, i.e.

$$\lambda_\pm = \frac{-b \pm \sqrt{b^2 - 4mk}}{2m}.$$

In the case of zero damping ($b = 0$) the λ's are pure imaginary, i.e. x_H is a combination of sine and cosine functions, i.e. oscillations with frequency $\sqrt{k/m}$. In the case of large damping such that $b^2 - 4mk \geq 0$ both λ's are real and negative and x_H is a combination of two exponential functions that go to zero with time. Finally, in the intermediate case $b^2 - 4mk < 0$, the solutions consist of damped oscillations.

Consider next an oscillatory external force; $F(t) = F_0 \cos(\omega_0 t)$ and assume that the damping is zero: $b = 0$. Then, if $\omega_0 \neq \sqrt{k/m}$ the particular solution is a combination of sines and cosines with angular frequency ω_0. By contrast, if $\omega_0 = \sqrt{k/m}$ the particular solution is

$$x_P(t) = A\, t\, \sin(\omega_0 t),$$

that is, the oscillations will have linearly increasing amplitude, i.e. we have resonance in the system (see Fig. 9). The figure also shows that when the external force frequency ω_0 approaches $\sqrt{k/m}$, the amplitude increases.

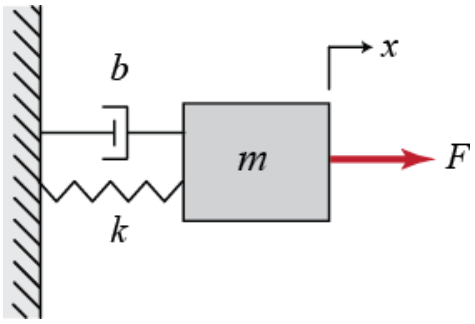


Figure 7: Sketch of a mass-spring-damper system. A mass is connected by a spring with constant k and a damper with damping b.

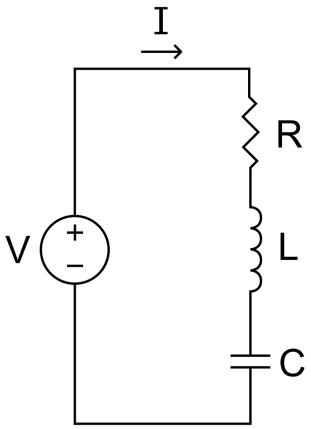


Figure 8: Sketch of an RLC circuit.

Kirchoffs law of electrical circuits

An RLC electrical circuit consists of a resistor with resistance R, a coil with inductance L and a capacitor with capacitance C (see fig, left). Denote the charge of the capacitor as Q and the current flowing in the circuit as I. Hence the voltage across the resistor, the coil and the capacitor, respectively, are RI, $LI'(t)$ and Q/C. Then by Kirchoff's law, stating that the voltage between any two points in the circuit is independent of the path used to connect these two points,

$$LI'(t) + RI + \frac{1}{C}Q = V(t),$$

where $V(t)$ is the voltage of the source. The current is the time derivative of the charge; $I = Q'$, resulting in the second order differential equation

$$LQ'' + RQ' + \frac{1}{C}Q = V(t).$$

This is an inhomogeneous second order linear differential equation which can in principle be solved analytically. The general solution is of the form $Q = Q_H + Q_P$, where Q_H and Q_P are the general solution to the homogeneous counterpart and a particular solution of the inhomogeneous differential equation, respectively. The structure of this ODE is the same as the mass-spring-damper system such that the following types of solution are found

$$Q_H = Ae^{\lambda_+ t} + Be^{\lambda_- t},$$

where the λ's are solutions of the equation $L\lambda^2 + R\lambda + 1/C = 0$, i.e.

$$\lambda_\pm = \frac{-R \pm \sqrt{R^2 - 4L/C}}{2L}.$$

In the case of zero resistance ($R = 0$) the λ's are pure imaginary, i.e. Q_H is a combination of sine and cosine functions with frequency $1/\sqrt{LC}$. In the case of large resistance such that $R^2 - 4L/C \geq 0$ both λ's are real and negative and Q_H is a combination of two exponential functions that go to zero with time. Finally, in the intermediate case $R^2 - 4L/C < 0$, the

Next, consider an oscillatory voltage source; $V(t) = V_0 \cos(\omega_0 t)$ and assume $R = 0$. Then, just as in the mass-spring system, if $\omega_0 \neq 1/\sqrt{LC}$ the particular solution is a combination of sines and cosines with angular

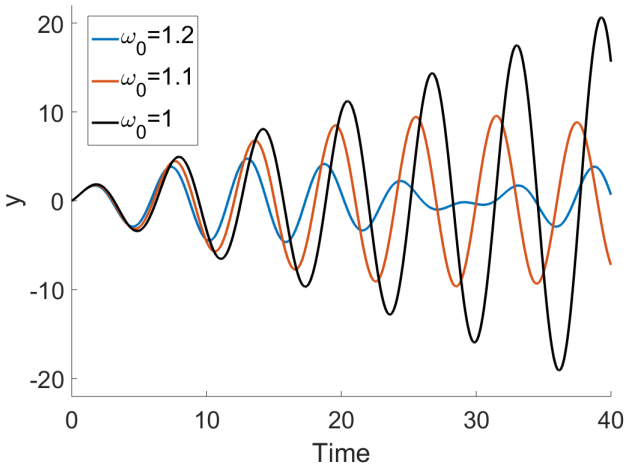


Figure 9: Solution of second order inhomogeneous linear ODE with parameters $a = 1$, $b = 0$ and $c = 1$ (corresponding to $m = 1$, $b = 0$ and $k = 1$ in the mass-spring ODE and $L = 1$, $R = 0$ and $C = 1$) with source term $F(t) = \cos(\omega_0 t)$ using several values for ω_0.

frequency ω_0. By contrast, if $\omega_0 = 1/\sqrt{LC}$ the particular solution is

$$x_P(t) = D\, t\, \sin(\omega_0 t)\,,$$

that is, oscillations with linearly increasing amplitude, which means we have resonance in the system (see Fig. 9). The figure also shows that when external force frequency ω_0 approaches $1/\sqrt{LC}$, the amplitude increases.

3.5 The Hodgkin-Huxley equations

Information is propagated throughout the brain by so-called action potentials in brain cells called neurons. These action potentials are generated by currents that travel through ion channels in the membrane of these cells. The physiologists Hodgkin and Huxley measured cross-membrane currents in the giant axon (a part of the neuron) of the squid and developed ODE models that described these currents. For this pioneering work they were awarded the Nobel prize in medicine and physiology in 1963.

The Hodgkin-Huxley equations read [4]

$$\begin{aligned} C_{\mathrm{M}}V' &= g_{\mathrm{Na}}m^3h\left(V-E_{\mathrm{Na}}\right)+g_{\mathrm{K}}n^4\left(V-E_{\mathrm{K}}\right)+g_{\mathrm{L}}\left(V-E_{\mathrm{L}}\right)+I_{\mathrm{App}}, \\ n' &= \alpha_n(V)\left(1-n\right)-\beta_n(V)n, \\ m' &= \alpha_m(V)\left(1-m\right)-\beta_m(V)m, \\ h' &= \alpha_h(V)\left(1-h\right)-\beta_h(V)h, \end{aligned}$$

where C_{M} is the membrane capacitance per unit area, V denotes the membrane potential, the terms $g_{\mathrm{Na}}m^3h\left(V-E_{\mathrm{Na}}\right)$ and $g_{\mathrm{K}}n^4\left(V-E_{\mathrm{K}}\right)$ are the sodium and potassium currents across the neuron membrane, $g_{\mathrm{L}}\left(V-E_{\mathrm{L}}\right)$ is a leak current and I_{App} denotes the current applied to the cell. The parameters g_{Na}, g_{K} and g_{L} are the sodium, potassium and leak conductances per unit area and E_{Na}, E_{K} and E_{L} are the so-called sodium, potassium and leak reversal potentials. The sodium and potassium channels are voltage-gated, that is, the quantity of ions that is allowed to pass through these channels is strongly voltage-dependent. Hodgkin and Huxley introduced the three so-called *gating variables* n, m and h to describe how the voltage-gated ion channels open and close to sodium and potassium. The magnitude of the variables m and h decides how open the sodium channel will be as the effective sodium conductance is $g_{\mathrm{Na}}m^3h$ and the magnitude of n decides how open the potassium channel will be as the effective potassium conductance is $g_{\mathrm{K}}n^4$.

The voltage-dependent rates in the ODEs governing the gating variables are

$$\begin{aligned} \alpha_n(V) &= 0.02\left(V-25\right)/\left(1-e^{-(V-25)/9}\right), \\ \alpha_m(V) &= 0.182\left(V+35\right)/\left(1-e^{-(V+35)/9}\right), \\ \alpha_h(V) &= 0.25e^{-(V+90)/12}, \end{aligned}$$

and

$$\begin{aligned} \beta_n(V) &= -0.002\left(V-25\right)/\left(1-e^{-(V-25)/9}\right), \\ \beta_m(V) &= -0.124\left(V+35\right)/\left(1-e^{(V+35)/9}\right), \\ \beta_h(V) &= 0.25e^{(V+62)/6}/e^{(V+90)/12}. \end{aligned}$$

The shape and magnitude of these rate functions were determined by Hodgkin and Huxley in order for their model to be able to mimic experimental data. One can imagine that they went through many iterations of the modelling cycle and a lot of tweaking of the above expressions in order to finally arrive at a model with which they were satisfied. Fig. 10 shows the dynamics of the membrane potential (top panel) and of gating variables (bottom). In the simulations the applied current I_{App} is sufficiently large that action potentials (the train of spikes in membrane potentials) are generated. When I_{App} is small, action potentials are not generated. The dynamics of the gating variables (Fig. 10, bottom panel) show that the most rapid changes in these variables occur at the same times as the spiking of the membrane potential. The details of the timing of the opening and closing of ion channels are not visible in the figure, but they are crucial for action potential generation.

Hodgkin and Huxley were pioneers both as experimentalists and modellers. Their perhaps most impressive achievement was their ability to integrate experimental and theoretical work and their ingeniousness in translating hypotheses into mathematical equations. Their work is one of the most beautiful examples of mathematical modelling.

3.6 Climate models and tipping points

Global warming without doubt represents one of the major threats to human civilization. It has become increasingly clear that the earth's temperature will continue to rise in the next one hundred years, causing more hurricanes, flooding, drought and other natural disasters. Predictions for the earth's future climate are obtained by simulating highly complex climate models. Due to this complexity these models are beyond the scope of this text. However there exist numerous mathematical models that describe subsystems of the global climate system that can be studied using methods taught in this course. The *dieback model of forest vegatation* [8] is an example of such a model. It describes the dynamics of forest cover in regions which historically have been totally covered in forest, but where deforestation causes by rising temperatures threatens the ecosystem.

The dieback model is an ordinary differential equation of the form $v' =$

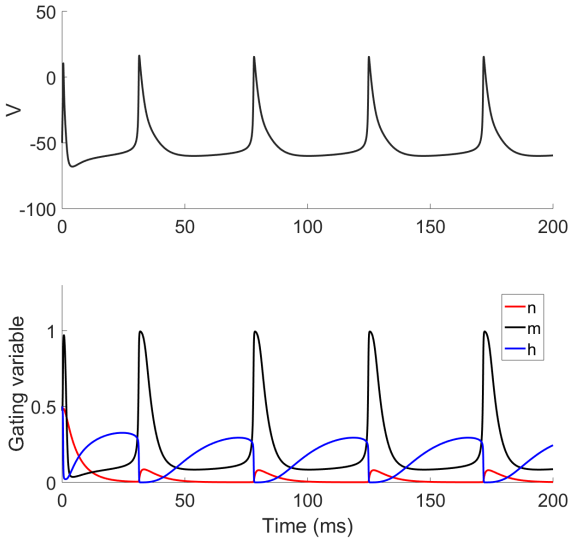


Figure 10: Solution curves of the Hodgkin-Huxley ODEs with parameter values $C_{\mathrm{M}} = 1$ (with unit $\mu\mathrm{F/cm}^2$), $g_{\mathrm{Na}} = 40$, $g_{\mathrm{K}} = 35$ and $g_{\mathrm{L}} = 0.3$ (unit $\mathrm{mS/cm}^2$), $E_{\mathrm{Na}} = 55$, $E_{\mathrm{K}} = -77$ and $E_{\mathrm{L}} = -65$ (unit mV) and $I_{\mathrm{App}} = 1$ (unit A). Top panel shows the dynamics of action potentials and bottom panel shows the dynamics of gating variables.

$F(v)$, where v denotes the fraction of the area that is covered by forest,

$$F(v) = F_0 \left[1 - \left(T_f + \alpha\,(1-v) - T_{opt}\right)^2 / \beta^2\right] v\,(1-v) - \gamma v. \qquad (1)$$

Here, γv is the deforestation rate and the effect of global warming on deforestation enters via the parameter T_f, *the forcing temperature*. At very low and very high values of T_f, the forest fraction is zero, see Figures 11 A and D, respectively. These figures plot the graph of $F(v)$ defined in (1) for $T_f = 12$ and $T_f = 36$ and show that there is only one *steady state* (values of forest cover fraction when the derivative is zero – depicted as filled black circles) in these cases and that this steady state is $v = 0$, corresponding to complete deforestation. For a moderate temperature ($T_f = 24$) there is one steady state at $v \approx 0.85$ (Figure 11 B) and at $T_f = 34$ there are two steady states, at $v \approx 0.8$ and $v = 0$ (Figure 11 C, two filled circles). The fact that there are two steady states can also be observed in Figures 11 E and F: the upper tail and lower tail of the curve in the temperature interval $[32.5, 34.5]$ correspond to these two steady states, the intermediate tail corresponds to

the unstable steady state (white circle in Figure 11 C). The temperature at which one loses the forest cover is where the curve in Figure 11 E turns. This is the *tipping point* of this system. That means the point where a highly dramatic change happens. This is especially dramatic in this system since if the temperature at a later stage should drop again, due to there being two possible steady states, the system would stay in the barren state without forest (corresponding to the lower tail at $v = 0$) in Figure 11 E, F.

Clearly the forest region described by the dieback model is affected by global warming. But in fact the global system is also affected by deforestation due to massive amounts of CO_2 being released into the atmosphere during deforestation (e.g. during forest fires) and the forest absorbing decreasing quantities of CO_2 with increasing deforestation. Thus there is a *positive feedback* (which is terribly negative) on the forest ecosystem from the global system; increased temperature (warming) means increased deforestation which in turn means increased levels of CO_2 which in turn cause more warming which mean increased deforestation, and so on:

$$\cdots \rightarrow \text{warming} \rightarrow \text{deforestation} \rightarrow \text{increased CO}_2 \rightarrow \text{warming} \rightarrow \cdots$$

This is an example of a positive feedback loop, where typically a tipping point migh occur. A change somewhere in the loop causes a change somewhere else in the loop, which in turn reinforces the first change and so the changes increase in strength. The tipping point is the point beyond which the system changes its state completely. In the forest ecosystem, there is a tipping point where the forcing temperature exceeds a certain threshold such that the system changes from a state where most of the area is forest-covered to a state where the area is barren and without trees.

3.7 First order differential equation models

In general many natural processes can be modeled by systems of ODEs of the form

$$\frac{dx}{dt} = f(t, x; p),$$

where x is a vector that comprises the variables corresponding to actors in the system under study (examples include position and velocity, number of individuls in animal populations, concentration of chemical substances,

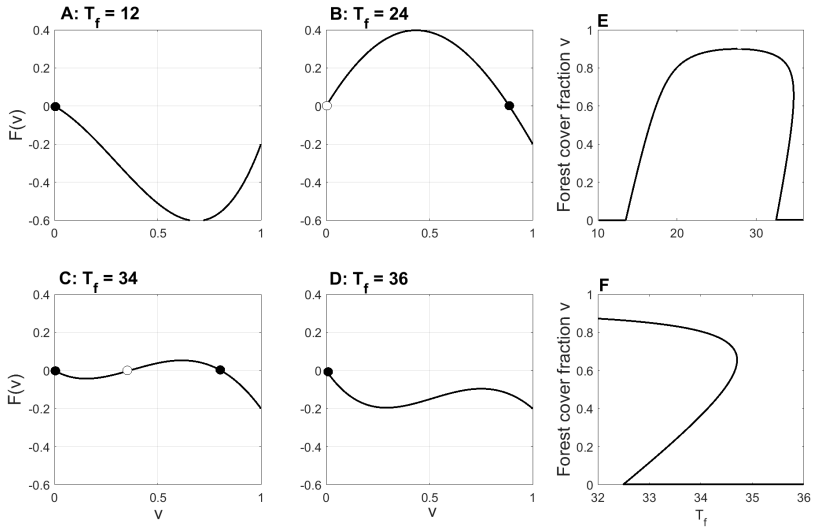


Figure 11: Graphs of the function $F(v)$ defined in (1) for $T_f = 12$ (A), $T_f = 24$ (B), $T_f = 34$ (C) and $T_f = 36$ (D), and the forest cover fraction as function of T_f (E) and zoom-in of forest cover fraction as function of T_f for $T_f \in [32, 36]$ (F). Stable and unstable steady states are indicated as filled and empty circles in (A) – (D). Parameter values used are $\alpha = 5$, $\beta = 10$, $F_0 = 2$, $\gamma = 0.2$ and $T_{opt} = 28$.

etc) and p is a vector that contains the parameters of the model. It is impossible in general to determine the analytical solution of such a system, therefore analysis and numerical methods used to compute approximations to the solution of these ODE systems are the main concerns of the course.

ODE systems cannot account for space-dependent physical processes such as heat flow, diffusion and convection. If the spatial variable x is introduced and assuming that

$$\frac{\partial u}{\partial t} = f(t, u; p) + D\frac{\partial^2 u}{\partial x^2},$$

where the term involving the second derivative is the *diffusion* term. Since this differential equation involves partial derivatives it is an example of a *partial differential equation* (PDE). PDEs are generally more difficult to to analyse than ODEs and certainly more time-consuming to estimate the solution of.

The rest of the course involves analysis of ODEs and PDEs and numerical methods for solving ODEs and PDEs.

Chapter 4

Numerical methods for first order ODEs

We consider the initial value theorem (IVP)

$$y' = f(t, y), \quad y(0) = y_0, \tag{2}$$

where the variable y may be a vector of any dimension. Although some IVPs can be solved analytically, for the vast majority of ODE systems we must resort to numerical methods to estimate the solution. This chapter deals with such methods. Without sacrifising generality, in the presentation only univariable ODEs are considered. The methods below are equally applicable to ODE systems where y is a vector of arbitrary dimension.

4.1 Obtaining solution estimates via linearisation

In elementary calculus courses students frequently encounter numerical methods for ODEs via vector fields such as the one depicted in Figure 12. The vector field is defined by the magnitude of the derivative $y'(t)$ in a given point (t^*, y^*) and visualised as short line segments with slope equal to $y'(t^*)$. Solution curves of the ODE in question thus can be sketched as curves that follow the vector field. The idea behind the *Euler method* for estimation of the solution of (2) can be illustrated using vector fields, since Euler's method is based on linearisation of the solution. The linearisation that passes through the initial point (t_0, y_0) can be used as an approximation to the exact so-

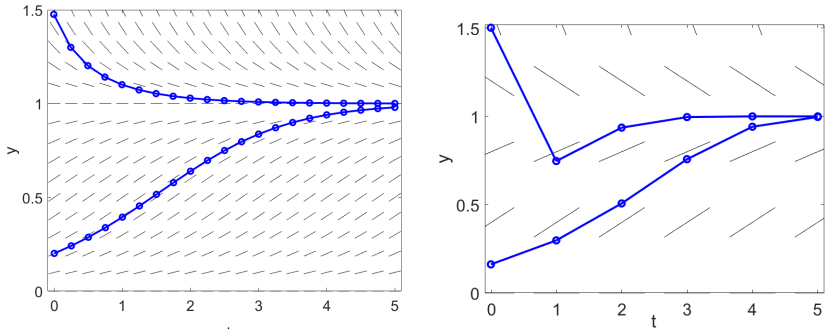


Figure 12: Examples of vector fields for the ODE $y' = y(1-y)$ for two different values of h. Left plot: $h = 0.2$, right plot: $h = 1$.

lution in the vicinity of the initial point. Specifically, the equation of this linearisation is $y_{lin} = y(0) + y'(t_0)\,(t - t_0)$ where from the ODE $y' = f(t, y)$, such that, for small values of t, we have the approximation

$$y(t) \approx y_0 + f(t_0, y_0)\,(t - t_0)\,.$$

Thus $y_1 = y_0 + f(t_0, y_0)\,(t_1 - t_0)$ is a relatively good approximation to the exact solution $y(t_1)$ of (2) at time $t = t_1$ (with t_1 being fairly close to t_0). Computing y_1 is the first step of the Euler algorithm. The next step, which consists of computing y_2 (the approximation to the exact solution $y(t)$ at time $t = t_2$) the linearisation of $y(t)$ around $t = t_1$ is used. In the same manner as for y_1, y_2 is found using $y_2 = y_1 + f(t_1, y_1)\,(t_2 - t_1)$.

4.2 The general approach

Integrating the IVP (2) from t_0 to t gives the *integral form* of the IVP

$$y(t) = y(t_0) + \int_{t_0}^{t} f(y(s))\,ds. \tag{3}$$

The idea of so-called Runge-Kutta methods is to estimate the solution one time step into the future using numerical integration to estimate the definite integral in (3), i.e. if y_n is the numerical approximation to the exact solution at $t = t_n$, the numerical approximation to the exact solution at $t = t_{n+1}$ is

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} f(y(s))\,ds.$$

However, the fact that the integrand is a function of the as yet unknown solution y on the interval $[t_n, t_{n+1}]$, necessitates making some clever choices in order to derive a simple to implement algorithm. To illustrate why this poses a challenge, consider using the trapezoidal rule

$$\int_a^{a+h} g(x)\,dx \approx h\frac{g(a)+g(a+h)}{2},$$

to approximate the integral. Then we would obtain

$$y_{n+1} = y_n + h\frac{f(t_n, y_n) + f(t_{n+1}, y_{n+1})}{2}. \tag{4}$$

This is in general a nonlinear equation where y_{n+1} is the unknown. It can be solved, but it is cumbersome since additional methods for solving algebraic equations (such as e.g. Newton's method) must be employed. This problem can be evaded using a simple method to estimate y_{n+1}, see below.

The accuracy of numerical methods for solving IVPs depends on the accuracy of the numerical integration technique used. In the following we denote by $\mathcal{I}_{[a,b]}(f)$ the numerical approximation of the definite integral $\int_a^b f(t)\,dt$ obtained by some numerical method (e.g. Euler's method, Heun's method etc). The integration method is said to be *accurate to order* $p+1$, provided that the absolute value of the discrepancy between exact and numerical approximation satisfies the inequality

$$\left|\int_a^b f(t)\,dt - \mathcal{I}_{[a,b]}(f)\right| \leq C_0 h^{p+1},$$

where C_0 is a constant that generally depends on the value of the p'th derivative of f on $[a,b]$. Thus we may write the numerical approximation of the integral as

$$\mathcal{I}_{[a,b]}(f) = \int_a^b f(t)\,dt + Ch^{p+1},$$

where C is related to C_0 above, but may be negative. This means that, for the first step,

$$y_1 = y_0 + \int_{t_0}^{t_1} f(t)\,dt + C_1 h^{p+1},$$

where C_1 is related to the integrand on the first interval $[t_0, t_1]$. Continuing,

for y_2,

$$\begin{aligned} y_2 &= y_1 + \int_{t_1}^{t_2} f(t)\,dt + C_2 h^{p+1} \\ &= y_0 + \int_{t_0}^{t_2} f(t)\,dt + (C_1 + C_2)\,h^{p+1}. \end{aligned}$$

Assuming now that the IVP is solved on some interval $[a, b]$ and separating this interval into n subintervals $[t_k, t_{k+1}]$, $k = 0, ..., n-1$, where $t_k = a + kh$, and $h = (b-a)/n$;

$$\begin{aligned} y_n &= y_{n-1} + \int_{t_{n-1}}^{t_n} f(t)\,dt + C_n h^{p+1} \\ &= y_0 + \int_{t_0}^{t_n} f(t)\,dt + (C_1 + C_2 + \cdots + C_n)\,h^{p+1} \\ &= y_0 + \int_a^b f(t)\,dt + \bar{C} n h^{p+1} = y_0 + \int_a^b f(t)\,dt + \bar{C}\,(b-a)\,h^p, \end{aligned}$$

where $\bar{C}$ is the mean of $C_1, ..., C_n$ and we used that $h = (b-a)/n$ gives $n = (b-a)/h$, such that the total error is $\bar{C}\,(b-a)\,h^p$. Thus if the integration method is order $p+1$ accurate, the numerical method for solving the IVP is order p accurate.

4.3 Euler's method

Approximating the integral $\int_{t_n}^{t_{n+1}} f(y)\,dt$ using the Riemann left sum, the iteration

$$y_{n+1} = y_n + hf(t_n, y_n)$$

is obtained. This is the same as the iteration that follows when linearising around (t_n, y_n), as depicted above. This method is *explicit* (since y_{n+1} is explicitly given as function of the previous step y_n) and the method is sometimes referred to as Euler's explicit method, or the Euler forward method. Alternatively, using the Riemann right sum, the iteration

$$y_{n+1} = y_n + hf(t_{n+1}, y_{n+1})$$

is obtained. Here y_{n+1} is *implicitly* given as function of y_n and this version is therefore called Euler's implicit method, or the Euler backward method.

As is evident from the two iterations above, an explicit iteration is preferred since an implicit iteration in general requires solving a set of nonlinear equations.

To illustrate the difference between explicit and implicit methods, we consider the IVP

$$y' = -10y, \quad y(0) = y_0,$$

whose analytical solution $y = y_0 \exp[-10t]$ converges to zero with increasing t. The explicit Euler method for this IVP reads

$$y_{n+1} = y_n - h \cdot 10 y_n = (1 - 10h)\, y_n.$$

This difference equation has solution $y_n^{\text{expl}} = y_0 (1 - 10h)^n$. By contrast, the implicit Euler method reads

$$y_{n+1} = y_n - h \cdot 10 y_{n+1} \Leftrightarrow y_{n+1} = \frac{y_n}{1 + 10h},$$

which has solution $y_n^{\text{impl}} = y_0 / (1 + 10h)^n$. The latter solution goes to zero when $n \to \infty$ regardless of time step size h, whereas the former solution will not converge to zero if $1 - 10h < -1$. This is the case when $10h > 2$, i.e. when $h > 1/5$. Thus, for this specific ODE the requirement to compute a sensible numerical solution is $h < 1/5$, i.e. the time step must be sufficiently small in the case of the explicit Euler method.

4.4 Heun's method

If one instead applies the trapezoidal rule for numerically computing the integral in (3), the result is equation (4) above, which represents a nonlinear equation in y_{n+1}. To avoid having to solve this equation, in *Heun's method* (also sometimes known as Runge-Kutta's second order method), y_{n+1} is first estimated using Euler's method to achieve a temporary estimate $\tilde{y}_{n+1}$; $\tilde{y}_{n+1} = y_n + hf(t_n, y_n)$. This estimate replaces the right endpoint in the formulation of the trapezoidal rule. In summary, Heun's method consists of the following two statements:

$$\begin{aligned} \tilde{y}_{n+1} &= y_n + hf(t_n, y_n), \\ y_{n+1} &= y_n + h\frac{f(t_n, y_n) + f(t_{n+1}, \tilde{y}_{n+1})}{2}. \end{aligned}$$

Due to using the trapezoidal rule, which has error of order h^3, the error at each time step is of order h^3. Thus the global error is of order h^2.

4.5 Runge-Kutta's 4th order method

The last method we look at, Runge-Kutta's 4th order method, uses Simpson's rule to estimate the integral on the right hand side in (3). When using Simpson's rule, the function to be integrated is approximated to a parabolic function on each double interval (Figure 13, bottom). It is therefore by far the most accurate of the methods considered: Simpson's rule is order 5 accurate, i.e. the error is $\leq Ch^5$, where C is a constant that depends on the magnitude of the fifth derivative on the interval in question (note that for this accuracy to be achieved the integrand must be at least five times differentiable). The error in Runge Kutta's method is therefore fourth order. However, just as in Heun's method, to render the method explixit, we make use of the same trick as before. On the interval $[t_n, t_{n+1}]$, the approximation directly applying Simpson's rule is

$$\int_{t_n}^{t_{n+1}} f(t,y)\,dt \approx \frac{h}{3}\left[f(t_n,y_n) + 4f(t_{n+1/2},y_{n+1/2}) + f(t_{n+1},y_{n+1})\right],$$

such that one possibility is to define

$$y_{n+1} \approx y_n + \frac{h}{6}\left[f(t_n,y_n) + 4f(t_{n+1/2},y_{n+1/2}) + f(t_{n+1},y_{n+1})\right], \qquad (5)$$

where $t_{n+1/2} = t_n + h/2$. This expression cannot be directly used due to the half-point $y_{n+1/2}$ and is cumbersome due to y_{n+1} being given implicitly. The way forward is thus, similar to in Heun's method, to estimate $y_{n+1/2}$ og y_{n+1} using the known value y_n and then inserting these estimates into the right side of (5).

In the following, we consider the expression within the brackets in (5), a weighted average of f across $[t_n, t_{n+1}]$ where most weight is placed on the middle point. Using Euler's method to estimate $y_{n+1/2}$ (as in Heun's method above), we have $y_{n+1/2} \approx \tilde{y}_{n+1/2} = y_n + (h/2)f(t_n, y_n)$. Once this estimate is obtained, an improved estimate of $y_{n+1/2}$ is obtained by inserting the first estimate; $y_{n+1/2} \approx \tilde{\tilde{y}}_{n+1/2} = y_n + (h/2)f(t_n + h/2, \tilde{y}_{n+1/2})$. The mean of these estimates is used to estimate the midpoint-value $f(t_{n+1/2}, y_{n+1/2})$.

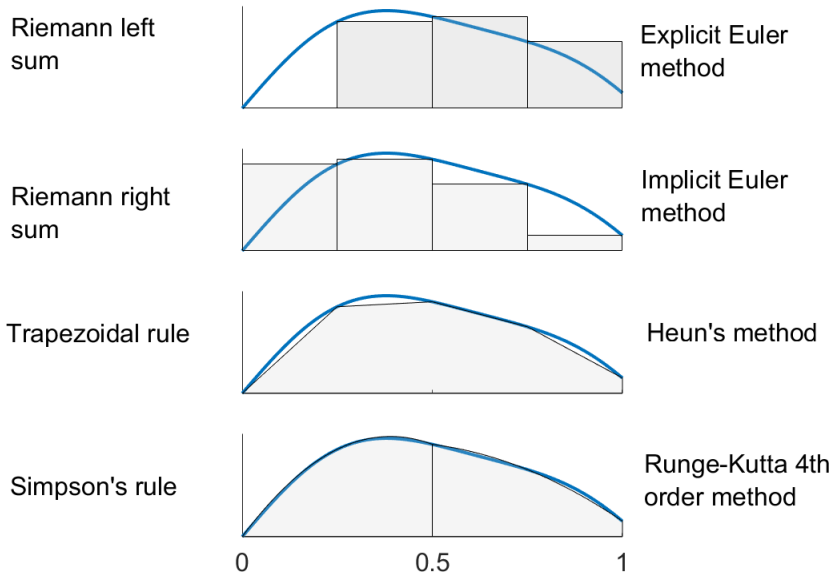


Figure 13: Summary of numerical methods for solving the initial value problem $y' = f(t, y)$, $y(t_0) = y_0$, and the numerical integration methods used to derive them.

Finally, y_{n+1} is estimated using the Euler method on $[t_n, t_{n+1}]$ with the last estimate replacing $f(t_n, y_n)$. Thus, the algoritm for Runge-Kutta's 4th order algorithm is

$$y_{n+1} = y_n + \frac{h}{6}\left[k_1 + 2k_2 + 2k_3 + k_4\right],$$

where

$$\begin{aligned} k_1 &= f(t_n, y_n), \\ k_2 &= f(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_1), \\ k_3 &= f(t_n + \frac{h}{2}, y_n + \frac{h}{2}k_2), \\ k_4 &= f(t_{n+1}, y_n + hk_3). \end{aligned}$$

4.6 Example with comparisons between different methods

We consider the IVP (the logistic ODE)

$$y' = y\,(1-y)\,, \quad y(0) = 1/2 \tag{6}$$

which has analytical solution $y(t) = 1/\left(1+\exp\left[-t\right]\right)$. The solution of the IVP is estimated using three of the methods above (Euler, Heun and Runge-Kutta's fourth order method) for the values $h = 2^{-k}$, $k = 1, 2, ..., 5$ for the time step. Figure 14 A compares the numerical solutions for $h = 0.5$ and the analytical solution for the Euler method, the Heun method and Runge-Kutta 4th order method. The RK estimate is virtually impossible to distuinguish from the analytical solution, whereas the Euler and Heun estimated are easily distuinguished. Figure 14 B displays the logarithm of the error $E(T) = |y(T) - y_\Delta(T)|$ against the logarithm of h. The reason for plotting these quantities is that we expect, due to the above reasoning concerning method accuracies, that this error goes like $E = Ch^p$. Thus $\ln E = \ln C + p \ln h$, i.e. the logarithm of the error is expected to be a linear function of $\ln h$ with slope equal to p.

The results show, as expexted, that the error is roughly first order in the Euler method, second order in the Heun method and fourth order in the RK4 method (Figure 14 B).

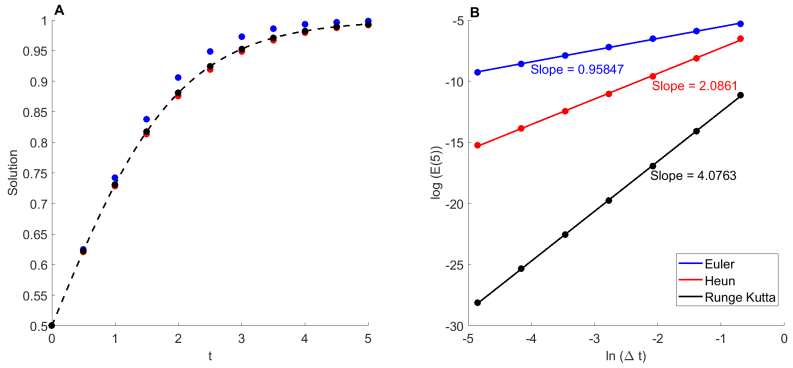


Figure 14: A: Analytical solution to (6) (dashed curve) compared to the numerical solution obtained by Euler's explicit method (filled blue circles), by Heun's method (filled red) and by Runge-Kutta's 4th order method (filled black) using time step $h = 0.5$. B: Logarithm plot of the error $E(t) = |y(t) - y_\Delta(t)|$ evaluated at $t = 5$ against time step h (filled circles, same colour coding as in A) and best fitted straight line. Slopes of straight lines are indicated.

Chapter 5

Boundary value problems: analysis and numerics

Consider the following second order ODEs with constraints;

$$y'' = 12x^2, \quad y(0) = 0, \quad y'(0) = 1.$$

$$v'' = 12x^2, \quad v(0) = 0, \quad v(1) = 0,$$

The former of these problems is an *initial value problem* (IVP – constraints are given initially) and the latter is a *boundary value problem* (BVP – constraints are given at the boundaries of the domain on which the ODE is to be solved). Both have the same solution and can be solved by hand to give

$$y(x) = v(x) = x^4 - x.$$

5.1 A shooting method

One might think that if the BVP had been an ODE IVP then one could apply a numerical ODE solver like Euler's or Runge-Kutta's method to estimate its solution. In fact, the related ODE IVP

$$w'' = 12x^2, \quad w(0) = 0, \quad w'(0) = A,$$

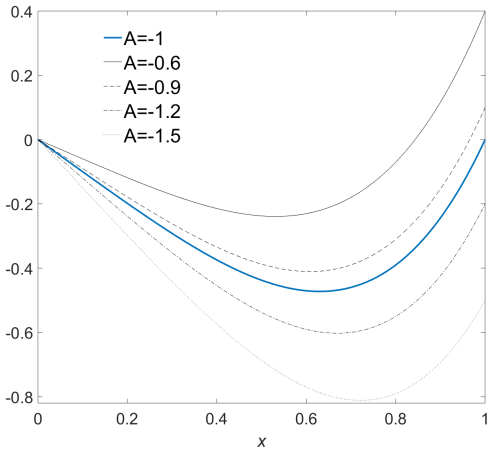


Figure 15: The exact solution of the BVP (full blue curve) and solutions for the IVP with various values for A (thin black curves).

can be solved by a numerical ODE solver (or analytically, by hand), for any value of A. The question is: will the solution of () match the solution of () for some value of A? Solving this ODE IVP involves aiming at the correct boundary value at the right end of the interval (at $x = 1$) by guessing an initial slope $w'(0) = A$. This is the idea of *shooting methods*. Using this kind of method, we will hit the target if we end up with $w(1) = 0$. In Figure 15, curves corresponding to various values of A; $A = -0.6, -0.9, -1.2, -1.5$, are plotted together with the exact solution of the BVP (which matches with the case $A = -1$).

5.2 A finite difference method

Given the more general boundary value ODE

$$-u'' = f(x), \quad u(0) = u(1) = 0. \tag{7}$$

There are ways of dealing with this BVP analytically, but in the following a numerical method is derived. Using Taylor's theorem and assuming that u is four times differentiable, the second derivative can be approximated by the ratio $[u(x + \Delta x) - 2u(x) + u(x - \Delta x)]/\Delta x^2$;

$$\frac{u(x + \Delta x) - 2u(x) + u(x - \Delta x)}{\Delta x^2} = u''(x) + \frac{1}{12}\Delta x^2 u^{(4)}(\xi),$$

whereξ is a number in the small interval $[x-\Delta x, x+\Delta x]$. This approximation is the basis for a numerical method for estimating the solution of (7). Now, separate $[0,1]$ into N subintervals of length $\Delta x = 1/N$ and define v_j as the approximation to $u(j\Delta x)$ such that $v_0 = v_N = 0$. Then,

$$-\frac{v_{j-1} - 2v_j + v_{j+1}}{\Delta x^2} = f_j, \quad j = 1, ..., N,$$

where $f_j = f(j\Delta x)$. That is, we have the following set of equations

$$\begin{aligned} 2v_1 - v_2 &= \Delta x^2 f_1, \\ &\vdots \\ -v_{j-1} + 2v_j - v_{j+1} &= \Delta x^2 f_j, \\ &\vdots \\ -v_{N-2} + 2v_{N-1} &= \Delta x^2 f_{N-1}, \end{aligned}$$

These $N+1$ equations can be written on the form

$$A\mathbf{v} = \Delta x^2 \mathbf{f},$$

where

$$A = \begin{bmatrix} -2 & 1 & 0 & \cdots & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 1 & 0 \\ 0 & \cdots & 0 & 1 & -2 & 1 \\ 0 & \cdots & \cdots & 0 & 1 & -2 \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{N-2} \\ v_{N-1} \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} 0 \\ f_1 \\ \vdots \\ f_{N-1} \\ 0 \end{bmatrix}.$$

Note that the matrix A is *tridiagonal*. This means that it has nonzero elements only on the diagonal and on the sub- and superdiagonals (the sets of elements directly below resp. above the diagonal). This is due to the way the second derivative in x_j is approximated using the solution in $x = x_j$ and in the two neighbouring points. The solution of the linear system of equations $A\mathbf{v} = \Delta x^2 \mathbf{f}$ is the numerical solution of the BVP and is plotted as filled blue circles together with the exact solution (full black curve) in Figure 16 for $N = 10$ and $\Delta x = 0.1$. The agreement between exact and numerical

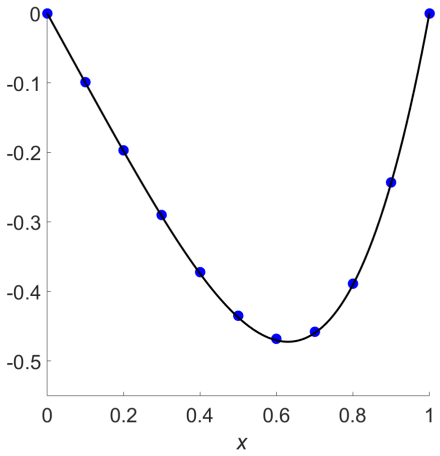


Figure 16: Comparison of exact (full black curve) and numerical (filled blue circles) solutions of the BVP.

solution is very good. In large implementations of BVPs (and a lot more so in PDEs, as we will see later), the sparsity of the matrix A is exploited to reduce computational time.

5.3 A "serious" example

The partial differential equation (PDE)

$$\rho c \frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(k(x)\frac{\partial u}{\partial x}\right) + s(x),$$

describes heat flow in a rod of length L, where ρ is material density, c denotes the specific heat capacity, $k(x)$ is the conductivity and $s(x)$ is a source term that corresponds to heat that is applied to the rod. In the equilibrum the time derivative is zero such that

$$\big(k(x)u'\big)' + s(x) = 0,$$

and we impose zero-flux condition at the left end and fixed temperature (equal to zero) at the right end of the rod;

$$u'(0) = 0, \quad u(L) = 0.$$

In order to determine the numerical solution, define q as $q(x) = k(x)u'$. Then,

$$q' + s(x) = 0.$$

Separate $[0, L]$ into N subintervals of length $\Delta x = L/N$, define k_j and s_j as $k(j\Delta x)$ and $s(j\Delta x)$, respectively, and v_j as the approximation to $u(j\Delta x)$ for $j = 0, 1, ..., N$. Approximating the derivative of q by $\left(q_{j+1/2} - q_{j-1/2}\right)/\Delta x$ since this gives a smaller numerical error, that is,

$$\frac{q_{j+1/2} - q_{j-1/2}}{\Delta x} + s_j = 0,$$

where $q_{j+1/2} = k_{j+1/2}\left(v_{j+1} - v_j\right)/\Delta x$ and $q_{j-1/2} = k_{j-1/2}\left(v_j - v_{j-1}\right)/\Delta x$. Then,

$$\frac{k_{j+1/2}v_{j+1} - d_j v_j + k_{j-1/2}v_{j-1}}{\Delta x^2} + s_j = 0,$$

where $d_j = k_{j-1/2} + k_{j+1/2}$, i.e.

$$k_{j-1/2}v_{j-1} - d_j v_j + k_{j+1/2}v_{j+1} = -\Delta x^2 s_j.$$

The boundary conditions $u'(0) = 0$ and $u(L) = 0$ translate to $\left(v_1 - v_0\right)/\Delta x = 0$, i.e. $-v_0 + v_1 = 0$ and $v_N = 0$. Thus we have the following set of equations;

$$\begin{aligned}
-v_0 + v_1 &= 0, \\
k_{1/2}v_0 - d_1 v_1 + k_{3/2}v_2 &= -\Delta x^2 s_1, \\
&\vdots \\
k_{j-1/2}v_{j-1} - d_j v_j + k_{j+1/2}v_{j+1} &= -\Delta x^2 s_j, \\
&\vdots \\
k_{N-3/2}v_{N-2} - d_{N-1}v_{N-1} + k_{N-1/2}v_N &= -\Delta x^2 s_{N-1}, \\
v_N &= 0.
\end{aligned}$$

.

These N equations can be written on the form of a linear system of equations;

$$A\mathbf{v} = -\Delta x^2 \mathbf{s},$$

where

$$A = \begin{bmatrix} -1 & 1 & 0 & \cdots & \cdots & 0 \\ k_{1/2} & -d_1 & k_{3/2} & 0 & \cdots & \vdots \\ 0 & k_{3/2} & -d_2 & \ddots & \ddots & \vdots \\ \vdots & 0 & \ddots & \ddots & k_{N-3/2} & 0 \\ \vdots & \cdots & \ddots & k_{N-3/2} & -d_{N-1} & k_{N-1/2} \\ 0 & \cdots & \cdots & 0 & 0 & 1 \end{bmatrix}.$$

and

$$\mathbf{v} = \begin{bmatrix} v_0 \\ v_1 \\ \vdots \\ v_{N-1} \\ v_N \end{bmatrix}, \quad \mathbf{s} = \begin{bmatrix} s_0 \\ s_1 \\ \vdots \\ s_{N-1} \\ s_N \end{bmatrix}.$$

The numerical solution of the BVP is the solution of the linear system. Note that, again, A is a tridiagonal matrix, this time with non-constant values on the diagonal and the sub- and super-diagonals.

For a specific example, we assume that $L = 3$, $k(x) = 0.5 \arctan(20(x-1)) + 1$ and $s(x) = e^{-(x-2)^2}$ and impose the boundary conditions

$$u'(0) = 0, \qquad u(3) = 0.$$

The functions $k(x)$ and $s(x)$ are plotted in the upper and middle panels in Figure 17. The numerical solution of the BVP, obtained by implementing the procedure outlined above, is plotted in the bottom panel of Figure 17, for $N = 100$ and $\Delta x = 3/100$.

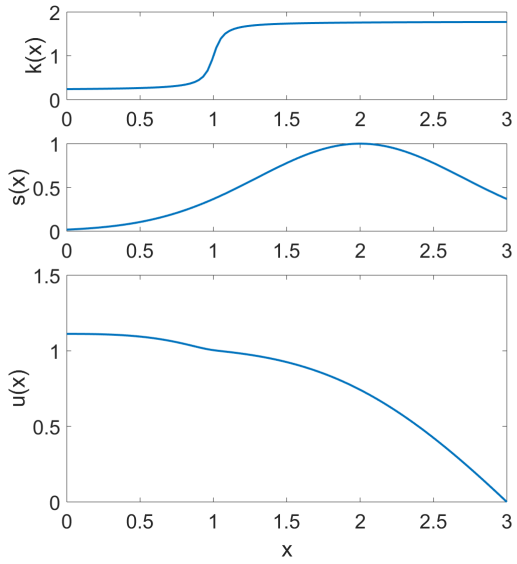


Figure 17: Graphs of the conductivity $k(x)$ and the external heat source $s(x)$ (top and middle panels, respectively) and of the numerical solution of the BVP $u(x)$ (bottom panel).

Chapter 6

Difference schemes for the diffusion equation

ODE systems cannot account for space-dependent physical processes such as heat flow, diffusion and convection. If we assume that the solution of our differential equation u represents the concentration of a chemical substance and that this substance is allowed to diffuse freely, the differential equation that describes this situation can be formulated as

$$\frac{\partial u}{\partial t} = f(t, u; p) + D\frac{\partial^2 u}{\partial x^2},$$

where the term involving the second derivative is the *diffusion* term and the function f represents a production rate. Since this differential equation involves partial derivatives it is an example of a *partial differential equation* (PDE). PDEs are generally more difficult to to analyse than ODEs and certainly more time-consuming to solve numerically.

We will consider the *1D diffusion equation* (also known as the *1D heat equation*) and use subscript notation for partial derivatives, that is

$$u_t = \frac{\partial u}{\partial t}, \quad u_{xx} = \frac{\partial^2 u}{\partial x^2}.$$

The domain on which the PDE is solved is defined by the inequalities $t \geq 0$ and $0 \leq x \leq 1$ and we impose initial and boundary values on the boundaries of this domain, depicted in Figure 18. The solution $u = u(x, t)$ is then the

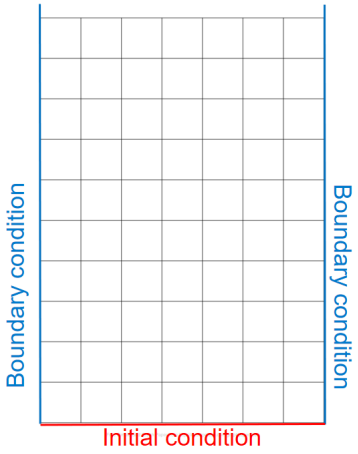


Figure 18: Rectangular, open-ended domain on which the diffusion equation is to be solved.

solution of the *initial boundary value problem* (IBVP)

$$\begin{aligned} u_t &= u_{xx}. \\ u(0,t) &= 0, \quad t > 0, \\ u(1,t) &= 0, \quad t > 0, \\ u(x,0) &= g(x), \quad 0 \le x \le 1. \end{aligned} \tag{8}$$

6.1 Approximations of partial derivatives

Using Taylor's theorem and assuming that u is four times differentiable, the second derivative can be approximated by a second order central difference;

$$\frac{u(x+\Delta x,t) - 2u(x,t) + u(x-\Delta x,t)}{\Delta x^2} = u_{xx}(x,t) + \frac{1}{12}\Delta x^2 u_{xxxx}(\xi,t),$$

where ξ is a number in the interval $[x-\Delta x, x+\Delta x]$. The one-sided (forward) approximation of the temporal derivative is, providing that u is two times differentiable with respect to t,

$$\frac{u(x,t+\Delta t) - u(x,t)}{\Delta t} = u_t(x,t) + \frac{1}{2}\Delta t\, u_{tt}(x,\tau_1),$$

where τ_1 is a number in the interval $[t, t+\Delta t]$. The central approximation of the temporal derivative is more accurate, providing that u is three times

differentiable with respect to t, since

$$\frac{u(x,t+\Delta t)-u(x,t-\Delta t)}{2\Delta t} = u_t(x,t) + \frac{1}{6}\Delta t^2 u_{ttt}(x,\tau_2),$$

where τ_2 is a number in the interval $[t-\Delta t, t+\Delta t]$.

We prove the last identity (the proofs of the other identities are almost identical). Using Taylor's theorem,

$$u(x,t+\Delta t) = u(x,t) + \Delta t u_t(x,t) + \frac{1}{2}\Delta t^2 u_{tt}(x,t) + \frac{1}{6}\Delta t^3 u_{ttt}(x,\tau'),$$

where τ' is somewhere in $[t, t+\Delta t]$. Similarly,

$$u(x,t-\Delta t) = u(x,t) - \Delta t u_t(x,t) + \frac{1}{2}\Delta t^2 u_{tt}(x,t) - \frac{1}{6}\Delta t^3 u_{ttt}(x,\tau''),$$

where τ'' is somewhere in $[t-\Delta t, t]$. Then,

$$u(x,t+\Delta t) - u(x,t-\Delta t) = 2\Delta t u_t(x,t) + \frac{2}{6}\Delta t^3 u_{ttt}(x,\tau_2),$$

where τ_2 is somewhere in $[t-\Delta t, t+\Delta t]$. Then,

$$\begin{aligned}\frac{u(x,t+\Delta t)-u(x,t-\Delta t)}{2\Delta t} &= \frac{2\Delta t u_t(x,t) + \frac{2}{6}\Delta t^3 u_{ttt}(x,\tau_2)}{2\Delta t} \\ &= u_t(x,t) + \frac{1}{6}\Delta t^2 u_{ttt}(x,\tau_2),\end{aligned}$$

which proves the identity.

6.2 An explicit scheme for the 1D heat equation

This subsection details the simplest scheme possible for numerically solving the IVBP (8). To derive the scheme, we separate $[0,1]$ into N subintervals of length $\Delta x = 1/N$ and define $v_{j,n}$ as the approximation to $u(j\Delta x, n\Delta t)$ such that, using the boundary conditions in (8), $v_{0,n} = v_{N,n} = 0$ for all values of n. Approximating u_t and u_{xx} as described in the previous subsection, we have

$$\frac{v_{j,n+1}-v_{j,n}}{\Delta t} = \frac{v_{j+1,n}-2v_{j,n}-v_{j-1,n}}{\Delta x^2}, \quad j = 1, ..., N-1,\ n = 1, 2, ...$$

Rewriting, we have the following set of equations;

$$v_{j,n+1} = v_{j,n} - \alpha\left(-v_{j+1,n} + 2v_{j,n} - v_{j-1,n}\right), \quad j = 1, ..., N-1,\ n = 1, 2, ...,$$

where $\alpha = \Delta t/\Delta x^2$. With $\mathbf{v}_n$ defined as the numerical solution vector at time $t_n = n\Delta t$; $\mathbf{v}_n = [v_{1,n}\, v_{2,n}, ...v_{N-1,n}]^T$, these equations can be written on matrix form as

$$\mathbf{v}_{n+1} = \mathbf{v}_n - \alpha B\mathbf{v}_n = (I - \alpha B)\,\mathbf{v}_n, \quad n = 1, 2, ..., \tag{9}$$

where I is the identity matrix of dimension $N-1$ and B is the *tridiagonal matrix* with 2 on the diagonal and -1 on the super- and subdiagonals;

$$B = \begin{bmatrix} 2 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & -1 & 0 \\ 0 & \cdots & 0 & -1 & 2 & -1 \\ 0 & \cdots & \cdots & 0 & -1 & 2 \end{bmatrix}, \quad \mathbf{v}_n = \begin{bmatrix} v_{1,n} \\ v_{2,n} \\ \vdots \\ v_{N-2,n} \\ v_{N-1,n} \end{bmatrix}.$$

The tridiagonal structure of B comes from the approximate expression $\left(v_{j+1,n} - 2v_{j,n} - v_{j-1,n}\right)/\Delta x^2$ for the second derivative u_{xx}. The scheme in (9) is the *explicit scheme*. It can be shown that the eigenvalues μ_j of B are

$$\mu_j = 2 + 2\cos\theta_j, \quad \theta_j = \frac{j\pi}{N}, \quad j = 1, ..., N-1.$$

Thus, since $0 < \theta_j < \pi$, the inequality$-1 < \cos\theta_j < 1$ holds such that $0 < \mu_j < 4$.

From (9) we deduce that the numerical solution at time step $t = t_n$ for $0 \le x \le 1$ can be written explicitly taking the n'th power of the matrix $I - \alpha B$;

$$\mathbf{v}_n = (I - \alpha B)^n\,\mathbf{v}_0, \quad n = 1, 2, ...,$$

where the elements of $\mathbf{v}_0$ are calculated using the initial condition $u(x, 0) = g(x)$.

At this point the reader is reminded of the following result from linear algebra. A matrix C is said to be *diagonalisable* providing that it is *similar*

to a diagonal matrix D, i.e. $D = R^{-1}CR$ (such that $C = RDR^{-1}$), where the diagonal elements of D are the eigenvalues of A and the columns of the matrix R are the associated eigenvectors. This can be used to calculate the powers of the matrix C;

$$\begin{aligned} C^n &= \left(RDR^{-1}\right)^n = \underbrace{\left(RDR^{-1}\right)\left(RDR^{-1}\right)\cdots\left(RDR^{-1}\right)}_{n\,\text{factors}} \\ &= RD\left(R^{-1}R\right)D\left(R^{-1}R\right)DR^{-1}\cdots RDR^{-1} \\ &= R\underbrace{DD\cdots D}_{n\,\text{factors}}R^{-1} = RD^nR^{-1}. \end{aligned}$$

Thus, provided that the eigenvalue with largest absolute value (or length when complex eigenvalues are considered) is smaller than 1, C^n will go to the zero matrix when $n \to \infty$. On the other hand, if the largest absolute value is larger than 1, C^n will blow up when $n \to \infty$.

We know that the solution of the initial value boundary problem under study will converge to zero. This feature must be shared by the numerical solution, i.e. we expect that all elements of $\mathbf{v}_n$ go to zero as $n \to \infty$. We will investigate whether this is in fact the case. First, take μ to be the eigenvalue of B. Then the corresponding eigenvalue of $I - \alpha B$ is $\lambda = 1 - \alpha\mu$;

$$(I - \alpha B)\,\mathbf{w} = \mathbf{w} - \alpha B\mathbf{w} = \mathbf{w} - \alpha\mu\mathbf{w} = (1 - \alpha\mu)\,\mathbf{w}.$$

Using this and the identity above,

$$\mathbf{v}_n = (I - \alpha B)^n\,\mathbf{v}_0 = RD^nR^{-1}\mathbf{v}_0,$$

where the diagonal entries of the diagonal matrix D are the eigenvalues of $I - \alpha B$, i.e. these entries are of the form $\lambda = 1 - \alpha\mu$, where we know that μ is in the interval $\langle 0, 4\rangle$. To be sure that the solution does not blow up with increasing n, we therefore require

$$-1 < 1 - \alpha\mu < 1 \Leftrightarrow -2 < -\alpha\mu < 0 \Leftrightarrow 2 > \alpha\mu > 0.$$

The inequality $\alpha\mu > 0$ is always satisfied, so the inequality to be examined is $\alpha\mu < 2$, i.e. $\alpha < 2/\mu$. This must hold for all possible eigenvalues, specifically for the largest possible eigenvalue 4. Thus the requirement for *stability* of

the solution is

$$\alpha < \frac{2}{4} = \frac{1}{2}.$$

This means that we will expect the solution to blow up providing that $\alpha > 1/2$. If this is the case the scheme is *unstable*.

6.3 An implicit scheme

In the implicit scheme, instead of using the solutions at times t_n and t_{n+1} to approximate u_t (the forward difference), instead the solutions at times t_{n-1} and t_n are used (the backward difference), i.e.

$$u_t \approx \frac{v_{j,n} - v_{j,n-1}}{\Delta t} \quad j = 1, ..., N-1,\ n = 1, 2, ...$$

Thus, by the same reasoning as above,

$$v_{j,n-1} = v_{j,n} + \alpha \left(-v_{j+1,n} + 2v_{j,n} - v_{j-1,n}\right), \quad j = 1, ..., N-1,\ n = 1, 2, ...$$

On matrix form the equations can be written

$$\mathbf{v}_{n-1} = \mathbf{v}_n + \alpha B \mathbf{v}_n = (I + \alpha B)\, \mathbf{v}_n, \quad n = 1, 2, ..., \tag{10}$$

This is the *implicit scheme* since $\mathbf{v}_n$ is implicitly given. From (10) we deduce that the numerical solution at time step $t = t_n$ for $0 \leq x \leq 1$ can be written explicitly taking the n'th power of the inverse matrix $(I + \alpha B)^{-1}$;

$$\mathbf{v}_n = \left[(I + \alpha B)^{-1}\right]^n \mathbf{v}_0, \quad n = 1, 2, ...$$

We know that the solution of the initial value boundary problem under study will converge to zero. This feature must be shared by the numerical solution, i.e. we expect that all elements of $\mathbf{v}_n$ go to zero as $n \to \infty$. This is indeed the case, as proven in the following argument. The eigenvalues λ of $I + \alpha B$ are $\lambda = 1 + \alpha\mu$, i.e. the eigenvalues are larger than 1. Using that a matrix and its inverse share eigenvectors and have eigenvalues that are inverses of each other;

$$A\mathbf{w} = \lambda\mathbf{w} \Leftrightarrow I\mathbf{w} = A^{-1}A\mathbf{w} = A^{-1}\lambda\mathbf{w} \Rightarrow A^{-1}\lambda\mathbf{w} = \mathbf{w} \Rightarrow A^{-1}\mathbf{w} = \lambda^{-1}\mathbf{w},$$

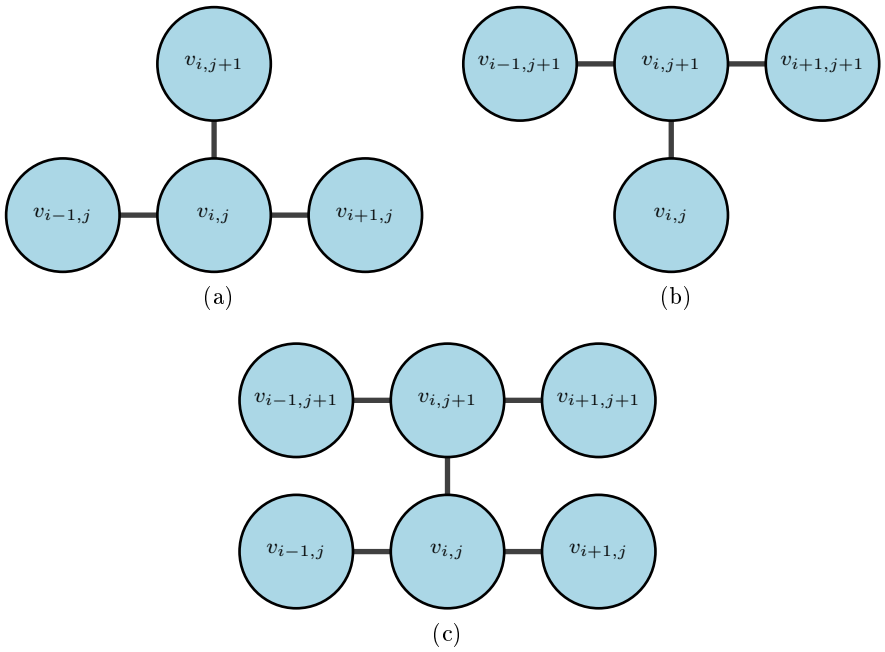


Figure 19: Computational molecules of the explicit scheme (a), of the implicit scheme (b) and of the Crank-Nicolson scheme (c).

the eigenvalues of $(I + \alpha B)^{-1}$ are $\lambda^{-1} = (1 + \alpha\mu)^{-1}$. Thus all eigenvalues of $(I + \alpha B)^{-1}$ are in the interval $\langle 0, 1\rangle$ such that the implicit scheme (10) is unconditionally stable.

6.4 The Crank-Nicolson scheme

In the *Crank-Nicolson scheme,* $u_t(x_{j,}, t_n)$ is approximated using the backward difference as in the implicit scheme and the second derivative $u_{xx}(x_{j,}, t_n)$ is approximated by a weighted average of the midpoint approximation at time steps t_{n-1} and t_n;

$$u_{xx} \approx \theta \frac{v_{j+1,n} - 2v_{j,n} + v_{j-1,n}}{\Delta x^2} + (1-\theta)\frac{v_{j+1,n-1} - 2v_{j,n-1} + v_{j-1,n-1}}{\Delta x^2}, \quad j = 1, ..., N-1,\ n = 1, 2, ...$$

where $0 \leq \theta \leq 1$. If the value $\theta = 1/2$ is chosen, the error of the approximation is second order;

$$\frac{u(x+\Delta x,t)-2u(x,t)+u(x-\Delta x,t)}{\Delta x^2} = u_{xx}(x,t) + \frac{1}{12}\Delta x^2 u_{xxxx}(\xi,t),$$

whereξ is a number in the interval $[x-\Delta x, x+\Delta x]$. The one-sided backward approximation of the temporal derivative is second order at the points $(x_j, t_{n-1/2})$, where $t_{n-1/2} = t_{n-1} + \Delta t/2 = t_n - \Delta t/2$:

$$\begin{aligned}\frac{v_{j,n}-v_{j,n-1}}{\Delta t} &\approx \frac{u(x_j,t_n)-u(x_j,t_n-\Delta t)}{\Delta t} \\ &= \frac{u(x_j,t_{n-1/2}+\Delta t/2)-u(x,t_{n-1/2}-\Delta t/2))}{\Delta t} \\ &= u_t(x_j,t_{n-1/2}) + \frac{1}{6}\Delta t^2 u_{ttt}(x,\tau)\end{aligned}$$

where τ is a number in the interval $[t, t+\Delta t]$.

Thus, by the same reasoning as in the implicit scheme,

$$v_{j,n-1} = v_{j,n} + \frac{1}{2}\alpha\left(-v_{j+1,n} + 2v_{j,n} - v_{j-1,n}\right) +$$
$$\frac{1}{2}\alpha\left(-v_{j+1,n-1} + 2v_{j,n-1} - v_{j-1,n-1}\right), \quad j = 1, ..., N-1,\ n = 1, 2, ...$$

which on matrix form can be written

$$\begin{aligned}\mathbf{v}_{n-1} &= \mathbf{v}_n + \frac{1}{2}\alpha B\mathbf{v}_n + \frac{1}{2}\alpha B\mathbf{v}_{n-1} \\ \Leftrightarrow \left(I - \frac{1}{2}\alpha B\right)\mathbf{v}_{n-1} &= \left(I + \frac{1}{2}\alpha B\right)\mathbf{v}_n \\ \Leftrightarrow \mathbf{v}_n &= \left(2I+\alpha B\right)^{-1}\left(2I-\alpha B\right)\mathbf{v}_{n-1}, \quad n = 1, 2, ...\end{aligned}$$

This is the *Crank-Nicolson scheme*. The numerical solution at time step $t = t_n$ can be written explicitly taking the n'th power of the inverse matrix $(I+\alpha B)^{-1}$;

$$\mathbf{v}_n = \left[\left(2I+\alpha B\right)^{-1}\left(2I-\alpha B\right)\right]^n \mathbf{v}_0, \quad n = 1, 2, ...$$

To show stability of the Crank-Nicolson scheme we show that the largest eigenvalue of $\left(2I+\alpha B\right)^{-1}\left(2I-\alpha B\right)$ is in the interval $\langle -1, 1\rangle$. First, we assume that two matrices C_1 and C_2 share eigenvectors, i.e. they have

eigenvalue-eigenvector pairs (λ_1, v) and (λ_2, v), respectively. Then

$$C_1 C_2 v = C_1 \lambda_2 v = \lambda_2 C_1 v = \lambda_1 \lambda_2 v,$$

that is $C_1 C_2$ has the same eigenvectors as C_1 and C_2 with corresponding eigenvalue $\lambda_1 \lambda_2$. We apply this results to the present case where we seek an expression for the eigenvalues of the product $(2I + \alpha B)^{-1} (2I - \alpha B)$. The matrices $(2I + \alpha B)^{-1}$ and $2I - \alpha B$ share eigenvectors and have eigenvalues $(2+\alpha\mu)^{-1}$ and $2-\alpha\mu$, respectively. Thus the product matrix $(2I + \alpha B)^{-1} (2I - \alpha B)$ has eigenvalues of the form

$$\lambda = \frac{2 - \alpha\mu}{2 + \alpha\mu} = \frac{2 + \alpha\mu - 2\alpha\mu}{2 + \alpha\mu} = 1 - 2\frac{\alpha\mu}{2 + \alpha\mu}.$$

Here the fraction $\alpha\mu/(2 + \alpha\mu)$ is in the interval $\langle 0, 1 \rangle$ such that $-1 < \lambda < 1$. This ensures stability for the Crank-Nicolson method.

6.5 A simple example

In this section the schemes above have been implemented in Matlab in order to estimate the solution of the IVBP

$$\begin{aligned} u_t &= u_{xx}. \\ u(0,t) &= 0, \quad t > 0, \\ u(1,t) &= 0, \quad t > 0, \\ u(x,0) &= \sin(\pi x), \quad 0 \le x \le 1. \end{aligned} \tag{11}$$

Using the method of separation of variables we find that the analytical solution is (this is easily confirmed by checking that the function satisfies the IBVP)

$$u(x,t) = \exp\left[-\pi^2 t\right] \sin(\pi x).$$

Numerical and analytical solutions are plotted in Fig. 20 for a selection of values for Δx and Δt and thus for $\alpha = \Delta t/\Delta x^2$. For $\alpha = 1$ and $\alpha = 0.8$, the explicit scheme is unstable, hence the associated numerical solutions are not depicted. For the smallest value of Δx, the numerical solutions obtained by the Crank-Nicolson scheme are quite accurate (compare the diamonds and the blue curve in the lower panel figures). Interestingly, for larger Δx, the solutions obtained by the explicit scheme seem to be more accurate than

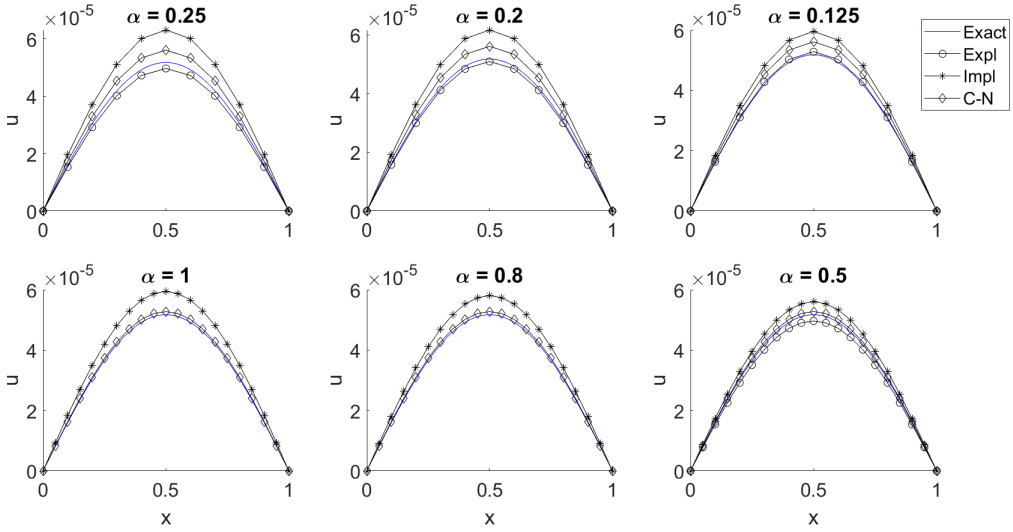


Figure 20: Analytical (blue curves) and numerical solutions (circles, diamonds and stars connected by black straight lines) of the IBVP (11) for a wide range of values for the spatial step Δx and time step Δt and thus for $\alpha = \Delta t/\Delta x^2$; for $\alpha = 0.25$: $\Delta x = 0.1$,$\Delta t = 0.0025$. $\alpha = 0.2$: $\Delta x = 0.1$,$\Delta t = 0.002$. $\alpha = 0.125$: $\Delta x = 0.1$,$\Delta t = 0.00125$. $\alpha = 1$: $\Delta x = 0.05$,$\Delta t = 0.0025$. $\alpha = 0.8$: $\Delta x = 0.05$,$\Delta t = 0.002$. $\alpha = 0.5$: $\Delta x = 0.05$,$\Delta t = 0.0125$.

the two other methods (compare circles and blue curves in the upper panel figures). Furthermore, for some reason the implicit scheme appears to be consistently less accurate than the two other schemes (stars in all figures).

Chapter 7

Schemes and methods for other PDE problems

Two more examples of IBVPs are considered: the time-independent 2D Poisson equation and the time-dependent reaction-diffusion equation.

7.1 Scheme for the Poisson equation in 2D

The two-dimensional Poisson equation may be written on the form

$$\Delta u = -f,$$

where the 2D Laplacian Δ is defined by $\Delta u = u_{xx} + u_{yy}$. The Laplacian thus represents 2D diffusion. The physical interpretation of the Poisson equation is that it describes the steady state distribution of the concentration of a substance in steady state where all transients have died out such that the time derivative of u that appears in the diffusion PDE is zero. The PDE is defined for $(x, y) \in \Omega$ and the solution or its derivative is specified on the domain boundary $\partial\Omega$. Let us for simpicity assume that Ω is rectangular; $\Omega = [0, L_x] \times [0, L_y]$, and that the solution u is equal to some specified function on the boundary. Then, the boundary value problem to be solved

is

$$\begin{aligned} \Delta u &= -f(x,y), \quad 0 < x < L_x,\, 0 < y < L_y, \\ u(x,0) &= g_0(x), \quad 0 \le x \le L_x, \\ u(x,L_y) &= g_1(x), \quad 0 \le x \le L_x, \\ u(0,y) &= h_0(y), \quad 0 \le y \le L_y, \\ u(L_x,y) &= h_1(y), \quad 0 \le y \le L_y. \end{aligned} \tag{12}$$

In the following we will outline how this problem can be solved numerically using a difference method.

We let $v_{i,j}$ be an approximation to the exact solution $u(x_i, y_j)$, where $x_i = ih$ and $y_j = jh$ and assume that $L_x = L_y = L$ and $h = L/N$. As previously, we use the approximations

$$\begin{aligned} u_{xx}(x_i,y_j) &\approx \frac{v_{i+1,j} - 2v_{i,j} + v_{i-1,j}}{h^2} \\ u_{yy}(x_i,y_j) &\approx \frac{v_{i,j+1} - 2v_{i,j} + v_{i,j-1}}{h^2} \end{aligned}$$

Then

$$\frac{v_{i+1,j} - 2v_{i,j} + v_{i-1,j}}{h^2} + \frac{v_{i,j+1} - 2v_{i,j} + v_{i,j-1}}{h^2} = -f(x_i,y_j)$$

which is more easily expressed as

$$v_{i,j} = \frac{1}{4}\left(v_{i+1,j} + v_{i-1,j} + v_{i,j+1} + v_{i,j-1} + h^2 f(x_i,y_j)\right),$$

i.e. the value at node (i,j) is the average of the surrounding nodes added to the quantity $h^2 f(x_i,y_j)/4$. The system of equations will be written on matrix-vector form.

For pedagogical purposes we set $L = 3$ and $N = 3$ such that $h = 1$. Then there are only four nodes in the interior of the domain at which the solutions must be calculated, namely the nodes $(1,1)$, $(1,2)$,$(2,1)$,$(2,2)$. The corresponding set of equations reads

$$\begin{aligned} 4v_{1,1} - v_{1,2} - v_{2,1} + 0 \times v_{2,2} &= v_{0,1} + v_{1,0} + h^2 f(x_1,y_1) \\ -v_{1,1} + 4v_{1,2} + 0 \times v_{2,1} - v_{2,2} &= v_{0,2} + v_{1,3} + h^2 f(x_1,y_2) \\ -v_{1,1} + 0 \times v_{2,1} + 4v_{2,1} - v_{2,2} &= v_{3,1} + v_{2,0} + h^2 f(x_2,y_1) \\ 0 \times v_{1,1} - v_{1,2} - v_{2,1} + 4v_{2,2} &= v_{3,2} + v_{2,3} + h^2 f(x_2,y_2) \end{aligned}$$

On matrix-vector form, these equations are written

$$\begin{bmatrix} 4 & -1 & -1 & 0 \\ -1 & 4 & 0 & -1 \\ -1 & 0 & 4 & -1 \\ 0 & -1 & -1 & 4 \end{bmatrix} \begin{bmatrix} v_{1,1} \\ v_{1,2} \\ v_{2,1} \\ v_{2,2} \end{bmatrix} = \begin{bmatrix} v_{0,1} + v_{1,0} + h^2 f(x_1, y_1) \\ v_{0,2} + v_{1,3} + h^2 f(x_1, y_2) \\ v_{3,1} + v_{2,0} + h^2 f(x_2, y_1) \\ v_{3,2} + v_{2,3} + h^2 f(x_2, y_2) \end{bmatrix}$$

If we define A as the matrix on the left, $\mathbf{x}$ as the vector containing the unknowns and $\mathbf{b}$ the vector on the right side of the equality, the system to be solved is $A\mathbf{x} = \mathbf{b}$. Generally A is $(n-1)^2 \times (n-1)^2$, i.e. even for moderate values of n the linear system is computationally expensive to solve. The computational complexity suggests an alternative approach to solving the system. Let $A = D + L + U$, where D is a diagonal matrix, L is a lower triangular and U is an upper triangular matrix. With the matrix A defined above,

$$D = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}, L = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & -1 & 0 \end{bmatrix}, U = \begin{bmatrix} 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Note that our system can now be written $(D + L + U)\,\mathbf{x} = \mathbf{b}$ or, alternatively, $D\mathbf{x} = \mathbf{b} - (L + U)\,\mathbf{x}$. This formulation motivates an iterative approach, where we make an initial guess $\mathbf{x}^{(0)}$ of the solution and then solve the iteration

$$D\mathbf{x}^{(r+1)} = \mathbf{b} - (L + U)\,\mathbf{x}^{(r)}$$

for $r = 0, 1, 2, \ldots$. Due to D being diagonal, the $(r+1)$'th iteration can be explicitly calculated;

$$\mathbf{x}^{(r+1)} = D^{-1}\left(\mathbf{b} - (L + U)\,\mathbf{x}^{(r)}\right),$$

where, in the present case, D^{-1} is diagonal with diagonal entries $1/4$.

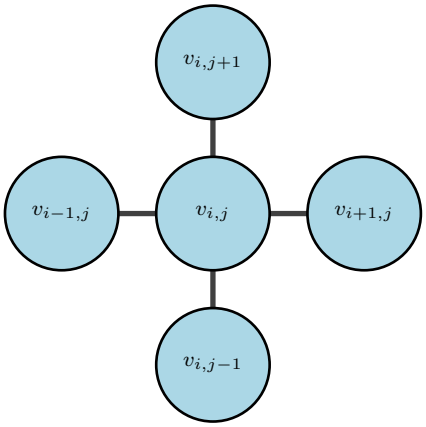


Figure 21: Computational molecule of the explicit scheme for the ....

7.2 Numerical scheme for nonlinear reaction-diffusion equations: The method of lines

When we study physical systems that involve chemical reactions, the quantity under study, typically the concentration of some chemical, is one of the reactants of a chemical reaction at the same time as it diffuses freely. The scalar reaction-diffusion equation to be considered here is therefore

$$u_t = Du_{xx} + f(u), \tag{13}$$

where the variable u represents the concentration of the chemical, $f(u)$ is a nonlinear function of u which may describe the dynamics of the chemical reaction and D is the diffusion coefficient of the chemical. We assume that (13) is defined on $\langle 0, 1\rangle$ and has initial condition $u(x, 0) = g(x)$ and boundary conditions $u(0, t)) = h_l$ and $u(1, t) = h_r$.

In the method of lines, we approximate the time-dependent solution $u(x_i, t)$ at the spatial point $x_i = i\Delta x$ by the function $v_i(t)$, where $[0, 1]$ has been partitioned into $n + 1$ subintervals of width $\Delta x = 1/(n + 1)$. Thus for the interior points,

$$v_i' = D\frac{v_{i-1} - 2v_i + v_{i+1}}{\Delta x^2} + f(v_i), \quad i = 2, 3, ..., n - 1.$$

Due to the initial and boundary conditions, we have $v_i(0) = g(x_i)$, $v_0(t) = h_l$

and $v_{n+1}(t) = h_r$. Thus, we obtain the large ODE system

$$\begin{aligned}
v_1' &= D\frac{h_l - 2v_1 + v_2}{\Delta x^2} + f(v_1), \\
&\vdots \\
v_i' &= D\frac{v_{i-1} - 2v_i + v_{i+1}}{\Delta x^2} + f(v_i), \quad i = 2, 3, ..., n-1, \\
&\vdots \\
v_n' &= D\frac{v_{n-1} - 2v_n + h_r}{\Delta x^2} + f(v_n)
\end{aligned}$$

On matrix-vector form, with $\mathbf{v}(t) = (v_1(t), v_2(t), ..., v_n(t))^T$, the IVP which we wish to solve is of the form

$$\mathbf{v}' = \frac{D}{\Delta x^2} A\mathbf{v} + \mathbf{F}(\mathbf{v}), \quad \mathbf{v}(0) = \mathbf{g},$$

where A is the tridiagonal matrix defined by

$$A = \begin{bmatrix}
-2 & 1 & 0 & & \\
1 & -2 & 1 & \ddots & \\
0 & 1 & \ddots & \ddots & 0 \\
 & \ddots & \ddots & \ddots & 1 \\
 & & 0 & 1 & -2
\end{bmatrix},$$

and

$$\mathbf{F}(\mathbf{v}) = \begin{bmatrix}
(D/\Delta x^2)h_l + f(v_1) \\
\vdots \\
f(v_i) \\
\vdots \\
(D/\Delta x^2)h_r + f(v_n)
\end{bmatrix}, \quad \mathbf{g} = \begin{bmatrix}
g(x_1) \\
\vdots \\
g(x_i) \\
\vdots \\
g(x_n)
\end{bmatrix}.$$

The IVP can in principle be solved using one of the numerical methods for ODE systems.

In the simplest possible case, with f zero and zero boundary conditions ($h_l = h_r = 0$), (13) reduces to the diffusion equation. Setting the initial function $g(x) = \sin(\pi x)$, we again have the IBVP (11) whose exact solution

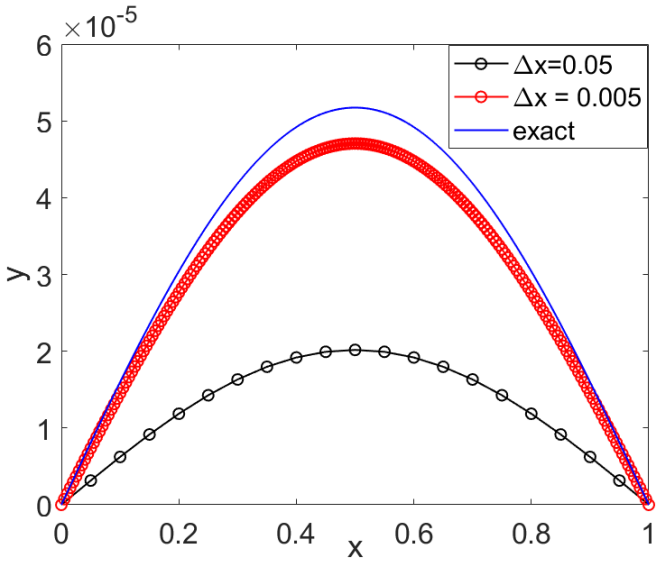


Figure 22: Analytical (blue curve) and numerical solutions (black and red circles connected by straight lines) of the IBVP (11) for $\Delta x = 0.05$ (black) and $\Delta x = 0.005$ (red) and time step $\Delta t = 0.001$ using the method of lines applying the Runge Kutta fourth order method ro numerically solve the IVP.

is $\exp\left[-\pi^2 t\right] \sin(\pi x)$. Then, the method of lines amounts to solving the IVP

$$\mathbf{v}' = \frac{D}{\Delta x^2} A\mathbf{v}, \quad \mathbf{v}(0) = \mathbf{g}.$$

We use Runge Kutta's fourth order method to estimate the numerical solution of this IVP. Figure 22 depicts the graph of the exact solution $u(x, 1)$ and the numerical solutions at $t = 1$ obtained for $\Delta x = 0.05$ and $\Delta x = 0.005$ using $\Delta t = 0.0001$. Despite the relatively small spatial and time steps, the method does not seem to be very accurate.

Bibliography

[1] Mathematical model. https://en.wikipedia.org/wiki/Mathematical_model, .

[2] Coronavirus modelling at the NIPH. https://www.fhi.no/en/id/infectious-diseases/coronavirus/coronavirus-modelling-at-the-niph-fhi/, .

[3] M. Hjorth-Jensen. Computational physics. lecture notes. https://github.com/CompPhysics/ComputationalPhysics/blob/master/doc/Lectures/lectures2015.pdf, 2015.

[4] A.L. Hodgkin and A.F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117(4):500–544, 1952.

[5] W.O. Kermack and A.G. McKendrick. A contribution to the mathematical theory of epidemics. *Proc Roy Soc Lond A*, 115:700–721, 1927.

[6] J.D. Logan. *Applied Mathematics. Fourth Edition.* Wiley, 2013.

[7] K. Maass. What are modelling competencies? *The international journal on mathematics education*, 38(2):113–142, 2006.

[8] P.D.L. Ritchie, J.J. Clarke, P.M. Cox, and C. Huntingford. Overshooting tipping point thresholds in a changing climate. *Nature*, 592:517–523, 2021.

[9] T. Woodson, C. Tyndall, and C. Stepheson. Estimating time of death. http://people.uncw.edu/lugo/MCP/DIFF_EQ/deproj/death/death.htm.