Contents Volume 2

Contents Volume 1

Contents Volume 3

Volume 2

Integrals and Geometry in $\mathbb{R}^n$

$$u(x_N) - u(x_0) = \int_{x_0}^{x_N} u'(x)\,dx$$
$$\approx \sum_{j=1}^{N} u'(x_{j-1})(x_j - x_{j-1})$$

$$a \cdot b = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

27
The Integral

> The two questions, the first that of finding the description of the curve from its elements, the second that of finding the figure from the given differences, both reduce to the same thing. From this it can be taken that the whole of the theory of the inverse method of the tangents is reducible to quadratures. (Leibniz 1673)

> Utile erit scribit $\int$ pro omnia. (Leibniz, October 29 1675)

27.1 Primitive Functions and Integrals

In this chapter, we begin the study of the subject of *differential equations*, which is one of the common ties binding together all areas of science and engineering, and it would be hard to overstate its importance. We have been preparing for this chapter for a long time, starting from the beginning with Chapter *A very short course in Calculus*, through all of the chapters on functions, sequences, limits, real numbers, derivatives and basic differential equation models. So we hope the gentle reader is both excited and ready to embark on this new exploration.

We begin our study with the simplest kind of differential equation, which is of fundamental importance:

> Given the **function** $f : I \to \mathbb{R}$ defined on the interval $I = [a, b]$, find a **function** $u(x)$ on I, such that the derivative $u'(x)$ of $u(x)$ is equal to $f(x)$ for $x \in I$.

We can formulate this problem more concisely as: given $f : I \to \mathbb{R}$ find $u : I \to \mathbb{R}$ such that

$$u'(x) = f(x) \tag{27.1}$$

for all $x \in I$. We call the solution $u(x)$ of the differential equation $u'(x) = f(x)$ for $x \in I$, a *primitive function* of $f(x)$, or an *integral* of $f(x)$. Sometimes the term *antiderivative* is also used.

To understand what we mean by "solving" (27.1), we consider two simple examples. If $f(x)$=1 for $x \in \mathbb{R}$, then $u(x) = x$ is a solution of $u'(x) = f(x)$ for $x \in \mathbb{R}$, since $Dx = 1$ for all $x \in \mathbb{R}$. Likewise if $f(x) = x$, then $u(x) = x^2/2$ is a solution of $u'(x) = f(x)$ for $x \in \mathbb{R}$, since $Dx^2/2 = x$ for $x \in \mathbb{R}$. Thus the function x is a primitive function of the constant function 1, and $x^2/2$ is a primitive function of the function x.

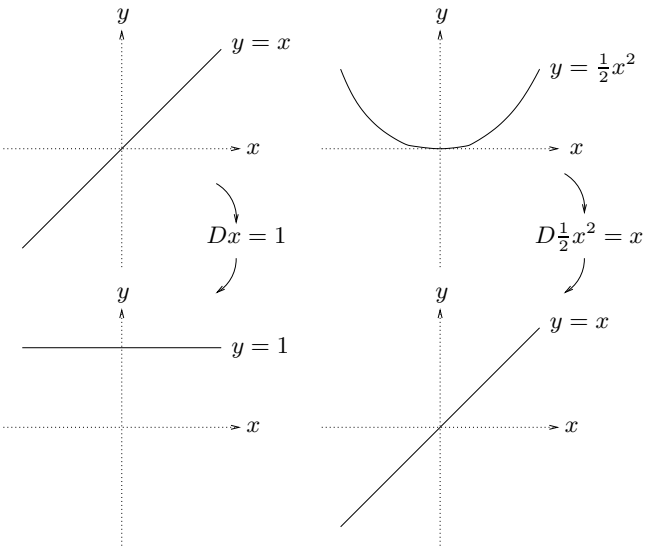

Fig. 27.1. $Dx = 1$ and $D(x^2/2) = x$

We emphasize that the solution of (27.1) is a **function** defined on an interval. We can interpret the problem in physical terms if we suppose that $u(x)$ represents some accumulated quantity like a sum of money in a bank, or an amount of rain, or the height of a tree, while x represents some changing quantity like time. Then solving (27.1) amounts to computing the total accumulated quantity $u(x)$ from knowledge of the rate of growth $u'(x) = f(x)$ at each instant x. This interpretation suggests that finding the total accumulated quantity $u(x)$ amounts to adding little pieces of momentary increments or changes of the quantity $u(x)$. Thus we expect that finding the integral $u(x)$ of a function $f(x)$ satisfying $u'(x) = f(x)$ will amount to some kind of *summation*.

A familiar example of this problem occurs when $f(x)$ is a velocity and x represents time so that the solution $u(x)$ of $u'(x) = f(x)$, represents the distance traveled by a body moving with instantaneous velocity $u'(x) = f(x)$. As the examples above show, we can solve this problem in simple cases, for example when the velocity $f(x)$ is equal to a constant v for all x and therefore the distance traveled during a time x is $u(x) = vx$. If we travel with constant velocity 4 miles/hour for two hours, then the distance traveled is 8 miles. We reach these 8 miles by accumulating distance foot-by-foot, which would be very apparent if we are walking!

An important observation is that the differential equation (27.1) alone is not sufficient information to determine the solution $u(x)$. Consider the interpretation when f represents velocity and u distance traveled by a body. If we want to know the position of the body, we need to know only the distance traveled but also the starting position. In general, a solution $u(x)$ to (27.1) is determined only up to a constant, because the derivative of a constant is zero. If $u'(x) = f(x)$, then also $(u(x) + c)' = f(x)$ for any constant c. For example, both $u(x) = x^2$ and $u(x) = x^2 + 1$ satisfy $u'(x) = 2x$. Graphically, we can see that there are many "parallel" functions that have the same slope at every point. The constant may be specified by specifying the value of the function $u(x)$ at some point. For example, the solution of $u'(x) = x$ is $u(x) = x^2 + c$ with c a constant, and specifying $u(0) = 1$ gives that $c = 1$.

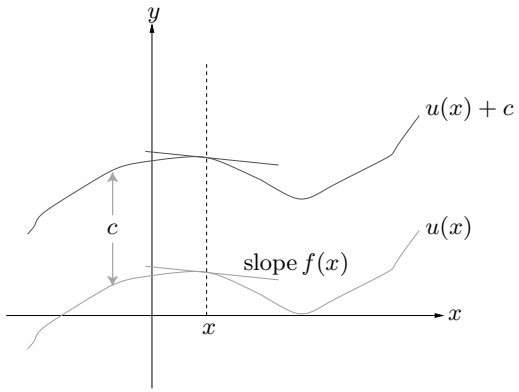

Fig. 27.2. Two functions that have the same slope at every point

More generally, we now formulate our basic problem as follows: Given $f : [a, b] \to \mathbb{R}$ and u_a, find $u : [a, b] \to \mathbb{R}$ such that

$$\begin{cases} u'(x) = f(x) & \text{for } a < x \leq b, \\ u(a) = u_a, \end{cases} \tag{27.2}$$

where u_a is a given *initial value*. The problem (27.2) is the simplest example of an *initial value problem* involving a differential equation and an initial value. The terminology naturally couples to situations in which x represents time and $u(a) = u_a$ amounts to specifying $u(x)$ at the initial time $x = a$. Note that we often keep the initial value terminology even if x represents a quantity different from time, and in case x represents a space coordinate we may alternatively refer to (27.2) as a *boundary value problem* with now $u(a) = u_a$ representing a given *boundary value.*

We shall now prove that the initial value problem (27.2) has a unique solution $u(x)$ if the given function $f(x)$ is Lipschitz continuous on $[a, b]$. This is the *Fundamental Theorem of Calculus*, which stated in words says that a Lipschitz continuous function has a (unique) primitive function. Leibniz referred to the Fundamental Theorem as the "inverse method of tangents" because he thought of the problem as trying to find a curve $y = u(x)$ given the slope $u'(x)$ of its tangent at every point x.

We shall give a constructive proof of the Fundamental Theorem, which not only proves that $u : I \to \mathbb{R}$ exists, but also gives a way to compute $u(x)$ for any given $x \in [a, b]$ to any desired accuracy by computing a sum involving values of $f(x)$. Thus the version of the Fundamental Theorem we prove contains two results: (i) the existence of a primitive function and (ii) a way to compute a primitive function. Of course, (i) is really a consequence of (ii) since if we know how to compute a primitive function, we also know that it exists. These results are analogous to defining $\sqrt{2}$ by constructing a Cauchy sequence of approximate solutions of the equation $x^2 = 2$ by the Bisection algorithm. In the proof of the Fundamental Theorem we shall also construct a Cauchy sequence of approximate solutions of the differential equation (27.2) and show that the limit of the sequence is an exact solution of (27.2).

We shall express the solution $u(x)$ of (27.2) given by the Fundamental Theorem in terms of the data $f(x)$ and u_a as follows:

$$u(x) = \int_a^x f(y)\,dy + u_a \quad \text{for } a \le x \le b, \tag{27.3}$$

where we refer to

$$\int_a^x f(y)\,dy$$

as the *integral* of f over the interval $[a, x]$, a and x as the *lower and upper limits of integration* respectively, $f(y)$ as the *integrand* and y the *integration variable*. This notation was introduced on October 29 1675 by Leibniz, who thought of the integral sign $\int$ as representing "summation" and dy as the "increment" in the variable y. The notation of Leibniz is part of the big success of Calculus in science and education, and (like a good cover of a record) it gives a direct visual expression of the mathematical content of the integral in very suggestive form that indicates both the construction of

the integral and how to operate with integrals. Leibniz choice of notation plays an important role in making Calculus into a "machine" which "works by itself".

We recapitulate: There are two basic problems in Calculus. The first problem is to determine the derivative $u'(x)$ of a given function $u(x)$. We have met this problem above and we know a set of rules that we can use to attack this problem. The other problem is to find a function $u(x)$ given its derivative $u'(x)$. In the first problem we assume knowledge of $u(x)$ and we want to find $u'(x)$. In the second problem we assume knowledge of $u'(x)$ and we want to find $u(x)$.

As an interesting aside, the proof of the Fundamental Theorem also shows that the integral of a function over an interval may be interpreted as the area underneath the graph of the function over the interval. This couples the problem of finding a primitive function, or computing an integral, to that of computing an area, that is to *quadrature*. We expand on this geometric interpretation below.

Note that in (27.2), we require the differential equation $u'(x) = f(x)$ to be satisfied for x in the half-open interval $(a, b]$ excluding the left end-point $x = a$, where the differential equation is replaced by the specification $u(a) = u_a$. The proper motivation for this will become clear as we develop the proof of the Fundamental Theorem. Of course, the derivative $u'(b)$ at the right end-point $x = b$, is taken to be the left-hand derivative of u. By continuity, we will in fact have also $u'(a) = f(a)$, with $u'(a)$ the right-hand derivative.

27.2 Primitive Function of $f(x) = x^m$ for $m = 0, 1, 2, \ldots$

For some special functions $f(x)$, we can immediately find primitive functions $u(x)$ satisfying $u'(x) = f(x)$ for x in some interval. For example, if $f(x) = 1$, then $u(x) = x + c$, with c a constant, for $x \in \mathbb{R}$. Further, if $f(x) = x$, then $u(x) = x^2/2 + c$ for $x \in \mathbb{R}$. More generally, if $f(x) = x^m$, where $m = 0, 1, 2, 3, \ldots$[TS:h], then $u(x) = x^{m+1}/(m+1) + c$. Using the notation (27.3) for $x \in \mathbb{R}$ we write

$$\int_0^x 1\, dy = x, \quad \int_0^x y\, dy = \frac{x^2}{2}, \tag{27.4}$$

and more generally for $m = 0, 1, 2, \ldots$,

$$\int_0^x y^m\, dy = \frac{x^{m+1}}{m+1}, \tag{27.5}$$

because both right and left hand sides vanish for $x = 0$.

[TS:h] I have changed 2 dots to 3 dots.

Editor's or typesetter's annotations (will be removed before the final TEX run)

27.3 Primitive Function of $f(x) = x^m$ for $m = -2, -3, \ldots$

We recall that if $v(x) = x^{-n}$, where $n = 1, 2, 3, \ldots$, then $v'(x) = -nx^{-(n+1)}$, where now $x \neq 0$. Thus a primitive function of $f(x) = x^m$ for $m = -2, -3, \ldots$ is given by $u(x) = x^{m+1}/(m+1)$ for $x > 0$. We can state this fact as follows: For $m = -2, -3, \ldots$,

$$\int_1^x y^m \, dy = \frac{x^{m+1}}{m+1} - \frac{1}{m+1} \quad \text{for } x > 1, \tag{27.6}$$

where we start the integration arbitrarily at $x = 1$. The starting point really does not matter as long as we avoid 0. We have to avoid 0 because the function x^m with $m = -2, -3, \ldots$, tends to infinity as x tends to zero. To compensate for starting at $x = 1$, we subtract the corresponding value of $x^{m+1}/(m+1)$ at $x = 1$ from the right hand side. We can write analogous formulas for $0 < x < 1$ and $x < 0$.

Summing up, we see that the polynomials x^m with $m = 0, 1, 2, \ldots$, have the primitive functions $x^{m+1}/(m+1)$, which again are polynomials. Further, the rational functions x^m for $m = -2, -3, \ldots$, have the primitive functions $x^{m+1}/(m+1)$, which again are rational functions.

27.4 Primitive Function of $f(x) = x^r$ for $r \neq -1$

Given our success so far, it would be easy to get overconfident. But we encounter a serious difficulty even with these early examples. Extending the previous arguments to rational powers of x, since $Dx^s = sx^{s-1}$ for $s \neq 0$ and $x > 0$, we have for $r = s - 1 \neq -1$,

$$\int_1^x y^r \, dy = \frac{x^{r+1}}{r+1} - \frac{1}{r+1} \quad \text{for } x > 1. \tag{27.7}$$

This formula breaks down for $r = -1$ and therefore we do not know a primitive function of $f(x) = x^r$ with $r = -1$ and moreover we don't even know that one exists. In fact, it turns out that most of the time we cannot solve the differential equation (27.2) in the sense of writing out $u(x)$ in terms of known functions. Being able to integrate simple rational functions is special. The Fundamental Theorem of Calculus will give us a way past this difficulty by providing the means to approximate the unknown solution to any desired accuracy.

27.5 A Quick Overview of the Progress So Far

Any function obtained by linear combinations, products, quotients and compositions of functions of the form x^r with rational power $r \neq 0$ and $x > 0$, can be differentiated analytically. If $u(x)$ is such a function, we thus obtain an analytical formula for $u'(x)$. If we now choose $f(x) = u'(x)$, then of course $u(x)$ satisfies the differential equation $u'(x) = f(x)$, so that we can write recalling Leibniz notation:

$$u(x) = \int_0^x f(y)\,dy + u(0) \quad \text{for } x \geq 0,$$

which apparently states that the function $u(x)$ is a primitive function of its derivative $f(x) = u'(x)$ (assuming that $u(x)$ is defined for all $x \geq 0$ so that no denominator vanishes for $x \geq 0$).

We give an example: Since $D(1+x^3)^{\frac{1}{3}} = (1+x^3)^{-\frac{2}{3}}x^2$ for $x \in \mathbb{R}$, we can write

$$(1+x^3)^{\frac{1}{3}} = \int_0^x \frac{y^2}{(1+y^3)^{\frac{2}{3}}}\,dy + 1 \quad \text{for } x \in \mathbb{R}.$$

In other words, we know primitive functions $u(x)$ satisfying the differential equation $u'(x) = f(x)$ for $x \in I$, for any function $f(x)$, which itself is a derivative of some function $v(x)$ so that $f(x) = v'(x)$ for $x \in I$. The relation between $u(x)$ and $v(x)$ is then

$$u(x) = v(x) + c \quad \text{for } x \in I,$$

for some constant c.

On the other hand, if $f(x)$ is an arbitrary function of another from, then we may not be able to produce an analytical formula for the corresponding primitive function $u(x)$ very easily or not at all. The Fundamental Theorem now tells us how to compute a primitive function of an arbitrary Lipschitz continuous function $f(x)$. We shall see that in particular, the function $f(x) = x^{-1}$ has a primitive function for $x > 0$ which is the famous *logarithm function* $\log(x)$. The Fundamental Theorem therefore gives in particular a constructive procedure for computing $\log(x)$ for $x > 0$.

27.6 A "Very Quick Proof" of the Fundamental Theorem

We shall now enter into the proof of the Fundamental Theorem. It is a good idea at this point to review the Chapter *A very short course in Calculus*. We shall give a sequence of successively more complete versions of the proof of the Fundamental Theorem with increasing precision and generality in each step.

The problem we are setting out to solve has the following form: given a function $f(x)$, find a function $u(x)$ such that $u'(x) = f(x)$ for all x in an interval. In this problem, we start with $f(x)$ and seek a function $u(x)$ such that $u'(x) = f(x)$. However in the early "quick" versions of the proofs, it will appear that we have turned the problem around by starting with a given function $u(x)$, differentiating u to get $f(x) = u'(x)$, and then recovering $u(x)$ as a primitive function of $f(x) = u'(x)$. This naturally appears to be quite meaningless circular reasoning, and some Calculus books completely fall into this trap. But we are doing this to make some points clear. In the final proof, we will in fact start with $f(x)$ and construct a function $u(x)$ that satisfies $u'(x) = f(x)$ as desired!

Let now $u(x)$ be differentiable on $[a, b]$, let $x \in [a, b]$, and let $a = y_0 < y_1 < \ldots < y_m = x$ be a *subdivision* of $[a, x]$ into subintervals $[a, y_1)$, $[y_1, y_2), \ldots, [y_{m-1}, x)$. By repeatedly subtracting and adding $u(y_j)$, we obtain the following identity which we refer to as a *telescoping sum* with the terms cancelling two by two:

$$\begin{aligned} u(x) - u(a) &= u(y_m) - u(y_0) \\ &= u(y_m) - u(y_{m-1}) + u(y_{m-1}) - u(y_{m-2}) + u(y_{m-2}) \\ &\qquad - \cdots + u(y_2) - u(y_1) + u(y_1) - u(y_0). \end{aligned} \tag{27.8}$$

We can write this identity in the form

$$u(x) - u(a) = \sum_{i=1}^{m} \frac{u(y_i) - u(y_{i-1})}{y_i - y_{i-1}}(y_i - y_{i-1}), \tag{27.9}$$

or as

$$u(x) - u(a) = \sum_{i=1}^{m} f(y_{i-1})(y_i - y_{i-1}), \tag{27.10}$$

if we set

$$f(y_{i-1}) = \frac{u(y_i) - u(y_{i-1})}{y_i - y_{i-1}} \quad \text{for } i = 1, \ldots, m. \tag{27.11}$$

Recalling the interpretation of the derivative as the ratio of the change in a function to a change in its input, we obtain our first version of the Fundamental Theorem as the following analog of (27.10) and (27.11):

$$u(x) - u(a) = \int_a^x f(y)\,dy \quad \text{where} \quad f(y) = u'(y) \quad \text{for } a < y < x.$$

In the integral notation, the sum $\sum$ corresponds to the integral sign $\int$, the increments $y_i - y_{i-1}$ correspond to dy, the y_{i-1} to the integration variable y, and the difference quotient $\frac{u(y_i)-u(y_{i-1})}{y_i-y_{i-1}}$ corresponds to the derivative $u'(y_{i-1})$.

This is the way that Leibniz was first led to the Fundamental Theorem at the age of 20 (without having studied any Calculus at all) as presented in his *Art of Combinations* from 1666.

Note that (27.8) expresses the idea that "the whole is equal to the sum of the parts" with "the whole" being equal to $u(x)-u(a)$ and the "parts" being the differences $(u(y_m)-u(y_{m-1}))$, $(u(y_{m-1})-u(y_{m-2}))$,...,$(u(y_2)-u(y_1))$ and $(u(y_1)-u(y_0))$. Compare to the discussion in Chapter *A very short Calculus course* including Leibniz' teen-age dream.

27.7 A "Quick Proof" of the Fundamental Theorem

We now present a more precise version of the above "proof". To exercise flexibility in the notation, which is a useful ability, we change notation slightly. Let $u(x)$ be uniformly differentiable on $[a,b]$, let $\bar{x} \in [a,b]$, and let $a = x_0 < x_1 < \ldots < x_m = \bar{x}$ be a partition of $[a,\bar{x}]$. We thus change from y to x and from x to $\bar{x}$. With this notation x serves the role of a variable and $\bar{x}$ is a particular value of x. We recall the identity (27.9) in its new dress:

$$u(\bar{x}) - u(a) = \sum_{i=1}^{m} \frac{u(x_i) - u(x_{i-1})}{x_i - x_{i-1}}(x_i - x_{i-1}). \tag{27.12}$$

By the uniform differentiability of u:

$$u(x_i) - u(x_{i-1}) = u'(x_{i-1})(x_i - x_{i-1}) + E_u(x_i, x_{i-1}),$$

where

$$|E_u(x_i, x_{i-1})| \leq K_u(x_i - x_{i-1})^2, \tag{27.13}$$

with K_u a constant, we can write the identity as follows:

$$u(\bar{x}) - u(a) = \sum_{i=1}^{m} u'(x_{i-1})(x_i - x_{i-1}) + \sum_{i=1}^{m} E_u(x_i, x_{i-1}). \tag{27.14}$$

Setting h equal to the largest increment $x_i - x_{i-1}$, so that $x_i - x_{i-1} \leq h$ for all i, we find

$$\sum_{i=1}^{m} |E_u(x_i, x_{i-1})| \leq \sum_{i=1}^{m} K_u(x_i - x_{i-1})h = K_u(\bar{x} - a)h.$$

The formula (27.14) can thus be written

$$u(\bar{x}) - u(a) = \sum_{i=1}^{m} u'(x_{i-1})(x_i - x_{i-1}) + E_h, \tag{27.15}$$

where

$$|E_h| \le K_u(\bar{x} - a)h. \tag{27.16}$$

The Fundamental Theorem is the following analog of this formula:

$$u(\bar{x}) - u(a) = \int_a^{\bar{x}} u'(x)\,dx, \tag{27.17}$$

with the sum $\sum$ corresponding to the integral sign $\int$, the increments $x_i - x_{i-1}$ corresponding to dx, and x_i corresponding to the integration variable x. We see by (27.16) that the additional term E_h in (27.15) tends to zero as the maximal increment h tends to zero. We thus expect (27.17) to be a limit form of (27.15) as h tends to zero.

27.8 A Proof of the Fundamental Theorem of Calculus

We now give a full proof of the Fundamental theorem. We assume for simplicity that $[a, b] = [0, 1]$ and the initial value $u(0) = 0$. We comment on the general problem at the end of the proof. So the problem we consider is: Given a Lipschitz continuous function $f : [0, 1] \to \mathbb{R}$, find a solution $u(x)$ of the initial value problem,

$$\begin{cases} u'(x) = f(x) & \text{for } 0 < x \le 1, \\ u(0) = 0. \end{cases} \tag{27.18}$$

We shall now construct an approximation to the solution $u(x)$ and give a meaning to the solution formula

$$u(\bar{x}) = \int_0^{\bar{x}} f(x)\,dx \quad \text{for } 0 \le \bar{x} \le 1.$$

To this end, let n be a natural number and let $0 = x_0 < x_1 < \ldots < x_N = 1$ be the subdivision of the interval $[0, 1]$ with *nodes* $x_i^n = ih_n$, $i = 0, \ldots, N$, where $h_n = 2^{-n}$ and $N = 2^n$. We thus divide the given interval $[0, 1]$ into subintervals $I_i^n = (x_{i-1}^n, x_i^n]$ of equal lengths $h_n = 2^{-n}$, see Fig. 27.3.

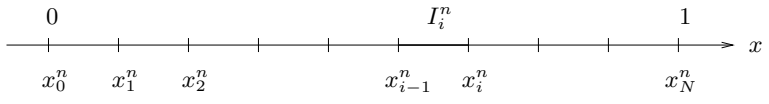

Fig. 27.3. Subintervals I_i^n of lengths $h_n = 2^{-n}$

The approximation to $u(x)$ is a continuous piecewise linear function $U^n(x)$ defined by the formula

$$U^n(x_j^n) = \sum_{i=1}^{j} f(x_{i-1}^n)h_n \qquad \text{for } j = 1, \ldots, N, \tag{27.19}$$

where $U^n(0) = 0$. This formula gives the values of $U^n(x)$ at the nodes $x = x_j^n$ and we extend $U^n(x)$ linearly between the nodes to get the rest of the values, see Fig. 27.4.

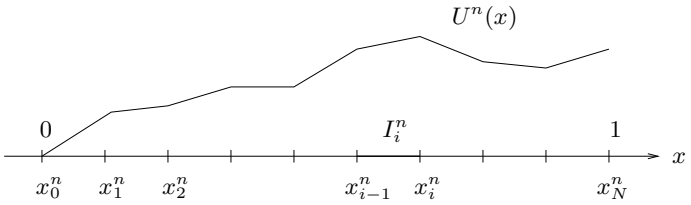

Fig. 27.4. Piecewise linear function $U^n(x)$

We see that $U^n(x_j^n)$ is a sum of contributions $f(x_{i-1}^n)h_n$ for all intervals I_i^n with $i \leq j$. By construction,

$$U^n(x_i^n) = U^n(x_{i-1}^n) + f(x_{i-1}^n)h_n \quad \text{for } i = 1, \ldots, N, \tag{27.20}$$

so given the function $f(x)$, we can compute the function $U^n(x)$ by using the formula (27.20) successively with $i = 1, 2, \ldots, N$, where we first compute $U^n(x_1^n)$ using the value $U^n(x_0^n) = U^n(0) = 0$, then $U^n(x_2^n)$ using the value $U^n(x_1^n)$ and so on. We may alternatively use the resulting formula (27.19) involving summation, which of course just amounts to computing the sum by successively adding the terms of the sum.

The function $U^n(x)$ defined by (27.19) is thus a continuous piecewise linear function, which is computable from the nodal values $f(x_i^n)$, and we shall now motivate why $U^n(x)$ should have a good chance of being an approximation of a function $u(x)$ satisfying (27.18). If $u(x)$ is uniformly differentiable on $[0, 1]$, then

$$u(x_i^n) = u(x_{i-1}^n) + u'(x_{i-1}^n)h_n + E_u(x_i^n, x_{i-1}^n) \quad \text{for } i = 1, \ldots, N, \tag{27.21}$$

where $|E_u(x_i^n, x_{i-1}^n)| \leq K_u(x_i^n - x_{i-1}^n)^2 = K_u h_n^2$, and consequently

$$u(x_j^n) = \sum_{i=1}^{j} u'(x_{i-1}^n)h_n + E_h \quad \text{for } j = 1, \ldots, N, \tag{27.22}$$

where $|E_h| \leq K_u h_n$, since $\sum_{i=1}^{j} h_n = jh_n \leq 1$. Assuming that $u'(x) = f(x)$ for $0 < x \leq 1$, the connection between (27.20) and (27.21) and (27.19) and (27.22) becomes clear considering that the terms $E_u(x_i^n, x_{i-1}^n)$ and E_h are small. We thus expect $U^n(x_j^n)$ to be an approximation of $u(x_j^n)$ at the nodes x_j^n, and therefore $U^n(x)$ should be an increasingly accurate approximation of $u(x)$ as n increases and $h_n = 2^{-n}$ decreases.

To make this approximation idea precise, we first study the convergence of the functions $U^n(x)$ as n tends to infinity. To do this, we fix $\bar{x} \in [0, 1]$

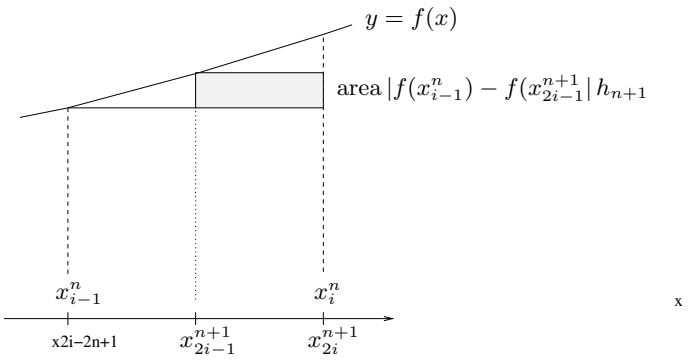

Fig. 27.5. The difference between $U^{n+1}(x)$ and $U^n(x)$

TS i

and consider the sequence of numbers $\{U^n(\bar{x})\}_{n=1}^{\infty}$. We want to prove that this is a Cauchy sequence and thus we want to estimate $|U^n(\bar{x}) - U^m(\bar{x})|$ for $m > n$.

We begin by estimating the difference $|U^n(\bar{x}) - U^{n+1}(\bar{x})|$ for two consecutive indices n and $m = n + 1$. Recall that we used this approach in the proof of the Contraction Mapping theorem. We have

$$U^n(x_i^n) = U^n(x_{i-1}^n) + f(x_{i-1}^n)h_n,$$

and since $x_{2i}^{n+1} = x_i^n$ and $x_{2i-2}^{n+1} = x_{i-1}^n$,

$$\begin{aligned} U^{n+1}(x_i^n) &= U^{n+1}(x_{2i}^{n+1}) = U^{n+1}(x_{2i-1}^{n+1}) + f(x_{2i-1}^{n+1})h_{n+1} \\ &= U^{n+1}(x_{i-1}^n) + f(x_{2i-2}^{n+1})h_{n+1} + f(x_{2i-1}^{n+1})h_{n+1}. \end{aligned}$$

Subtracting and setting $e_i^n = U^n(x_i^n) - U^{n+1}(x_i^n)$, we have

$$e_i^n = e_{i-1}^n + (f(x_{i-1}^n)h_n - f(x_{2i-2}^{n+1})h_{n+1} - f(x_{2i-1}^{n+1})h_{n+1}),$$

that is, since $h_{n+1} = \frac{1}{2}h_n$,

$$e_i^n - e_{i-1}^n = (f(x_{i-1}^n) - f(x_{2i-1}^{n+1}))h_{n+1}. \tag{27.23}$$

Assuming that $\bar{x} = x_j^n$ and using (27.23) and the facts that $e_0^n = 0$ and $|f(x_{i-1}^n) - f(x_{2i-1}^{n+1})| \le L_f h_{n+1}$, we get

$$\begin{aligned} |U^n(\bar{x}) - U^{n+1}(\bar{x})| = |e_j^n| &= |\sum_{i=1}^{j}(e_i^n - e_{i-1}^n)| \\ &\le \sum_{i=1}^{j}|e_i^n - e_{i-1}^n| = \sum_{i=1}^{j}|f(x_{i-1}^n) - f(x_{2i-1}^{n+1})|h_{n+1} \\ &\le \sum_{i=1}^{j} L_f h_{n+1}^2 = \frac{1}{4}L_f h_n \sum_{i=1}^{j} h_n = \frac{1}{4}L_f \bar{x} h_n, \end{aligned} \tag{27.24}$$

TS i

Please check the x in Fig. 27.5 on the right side.

Editor's or typesetter's annotations (will be removed before the final TEX run)

where we also used the fact that $\sum_{i=1}^{j} h_n = \bar{x}$. Iterating this estimate and using the formula for a geometric sum, we get

$$|U^n(\bar{x}) - U^m(\bar{x})| \leq \frac{1}{4} L_f \bar{x} \sum_{k=n}^{m-1} h_k = \frac{1}{4} L_f \bar{x} (2^{-n} + \ldots + 2^{-m+1})$$
$$= \frac{1}{4} L_f \bar{x} 2^{-n} \frac{1 - 2^{-m+n}}{1 - 2^{-1}} \leq \frac{1}{4} L_f \bar{x} 2^{-n} 2 = \frac{1}{2} L_f \bar{x} h_n,$$

that is

$$|U^n(\bar{x}) - U^m(\bar{x})| \leq \frac{1}{2} L_f \bar{x} h_n. \tag{27.25}$$

This estimate shows that $\{U^n(\bar{x})\}_{n=1}^{\infty}$ is a Cauchy sequence and thus converges to a real number. We decide, following Leibniz, to denote this real number by

$$\int_0^{\bar{x}} f(x)\, dx,$$

which thus is the limit of

$$U^n(\bar{x}) = \sum_{i=1}^{j} f(x_{i-1}^n) h_n$$

as n tends to infinity, where $\bar{x} = x_j^n$. In other words,

$$\int_0^{\bar{x}} f(x)\, dx = \lim_{n \to \infty} \sum_{i=1}^{j} f(x_{i-1}^n) h_n.$$

Letting m tend to infinity in (27.25), we can express this relation in quantitative form as follows:

$$\left| \int_0^{\bar{x}} f(x)\, dx - \sum_{i=1}^{j} f(x_{i-1}^n) h_n \right| \leq \frac{1}{2} L_f \bar{x} h_n.$$

At this point, we have defined the integral $\int_0^{\bar{x}} f(x)\, dx$ for a given Lipschitz continuous function $f(x)$ on $[0,1]$ and a given $\bar{x} \in [0,1]$, as the limit of the sequence $\{U^n(\bar{x})\}_{n=1}^{\infty}$ as n tends to infinity. We can thus define a function $u : [0,1] \to \mathbb{R}$ by the formula

$$u(\bar{x}) = \int_0^{\bar{x}} f(x)\, dx \quad \text{for } \bar{x} \in [0,1]. \tag{27.26}$$

We now proceed to check that the function $u(x)$ defined in this way indeed satisfies the differential equation $u'(x) = f(x)$. We proceed in two steps. First we show that the function $u(x)$ is Lipschitz continuous on $[0,1]$ and then we show that $u'(x) = f(x)$.

Before plunging into these proofs, we need to address a subtle point. Looking back at the construction of $u(x)$, we see that we have defined $u(\bar{x})$ for $\bar{x}$ of the form $\bar{x} = x_j^n$, where $j = 0, 1, \ldots, 2^n$, $n = 1, 2, \ldots,$. These are the rational numbers with finite decimal expansion in the base of 2, and they are *dense* in the sense that for any real number $x \in [0, 1]$ and any $\epsilon > 0$, there is a point of the form x_j^n so that $|x - x_j^n| \leq \epsilon$. Recalling the Chapter *Real numbers*, we understand that if we can show that $u(x)$ is Lipschitz continuous on the dense set of numbers of the form x_j^n, then we can extend $u(x)$ as a Lipschitz function to the set of real numbers $[0, 1]$.

We thus assume that $\bar{x} = x_j^n$ and $\bar{y} = x_k^n$ with $j > k$, and we note that

$$U^n(\bar{x}) - U^n(\bar{y}) = \sum_{i=1}^{j} f(x_{i-1}^n)h_n - \sum_{i=1}^{k} f(x_{i-1}^n)h_n = \sum_{i=k+1}^{j} f(x_{i-1}^n)h_n$$

and using the triangle inequality

$$|U^n(\bar{x}) - U^n(\bar{y})| \leq \sum_{i=k+1}^{j} |f(x_{i-1}^n)|h_n \leq M_f \sum_{i=k+1}^{j} h_n = M_f|\bar{x} - \bar{y}|,$$

where M_f is a positive constant such that $|f(x)| \leq M_f$ for all $x \in [0, 1]$. Letting n tend to infinity, we see that

$$u(\bar{x}) - u(\bar{y}) = \int_0^{\bar{x}} f(x)\,dx - \int_0^{\bar{y}} f(x)\,dx = \int_{\bar{y}}^{\bar{x}} f(x)\,dx, \tag{27.27}$$

where of course,

$$\int_{\bar{y}}^{\bar{x}} f(x)\,dx = \lim_{n\to\infty} \sum_{i=k+1}^{j} f(x_{i-1}^n)h_n,$$

and also

$$|u(\bar{x}) - u(\bar{y})| \leq \left|\int_{\bar{y}}^{\bar{x}} f(x)\,dx\right| \leq \int_{\bar{y}}^{\bar{x}} |f(x)|\,dx \leq M_f|\bar{x} - \bar{y}|, \tag{27.28}$$

where the second inequality is the so-called *triangle inequality for integrals* to be proved in the next section. We thus have

$$|u(\bar{x}) - u(\bar{y})| \leq M_f|\bar{x} - \bar{y}|, \tag{27.29}$$

which proves the Lipschitz continuity of $u(x)$.

We now prove that the function $u(x)$ defined for $x \in [0, 1]$ by the formula

$$u(x) = \int_a^x f(y)\,dy,$$

where $f : [0,1] \to \mathbb{R}$ is Lipschitz continuous, satisfies the differential equation

$$u'(x) = f(x) \quad \text{for } x \in [0,1],$$

that is

$$\frac{d}{dx}\int_0^x f(y)\,dy = f(x). \tag{27.30}$$

To this end, we choose $x, \bar{x} \in [0,1]$ with $x \geq \bar{x}$ and use (27.27) and (27.28) to see that

$$u(x) - u(\bar{x}) = \int_0^x f(z)dz - \int_0^{\bar{x}} f(y)dy = \int_{\bar{x}}^x f(y)dy,$$

and

$$\begin{aligned} |u(x) - u(\bar{x}) - f(\bar{x})(x-\bar{x})| &= \Big| \int_{\bar{x}}^x f(y)\,dy - f(\bar{x})(x-\bar{x}) \Big| \\ &= \Big| \int_{\bar{x}}^x (f(y) - f(\bar{x}))\,dy \Big| \leq \int_{\bar{x}}^x |f(y) - f(\bar{x})|\,dy \\ &\leq \int_{\bar{x}}^x L_f|y-\bar{x}|\,dy = \frac{1}{2}L_f(x-\bar{x})^2, \end{aligned}$$

where we again used the triangle inequality for integrals. This proves that u is uniformly differentiable on $[0,1]$, and that $K_u \leq \frac{1}{2}L_f$.

Finally to prove uniqueness, we recall from (27.15) and (27.16) that a function $u : [0,1] \to \mathbb{R}$ with Lipschitz continuous derivative $u'(x)$ and $u(0) = 0$, can be represented as

$$u(\bar{x}) = \sum_{i=1}^m u'(x_{i-1})(x_i - x_{i-1}) + E_h,$$

where

$$|E_h| \leq K_u(\bar{x} - a)h.$$

Letting n tend to infinity, we find that

$$u(\bar{x}) = \int_0^{\bar{x}} u'(x)\,dx \quad \text{for } \bar{x} \in [0,1], \tag{27.31}$$

which expresses the fact that a uniformly differentiable function with Lipschitz continuous derivative is the integral of its derivative. Suppose now that $u(x)$ and $v(x)$ are two uniformly differentiable functions on $[0,1]$ satisfying $u'(x) = f(x)$, and $v'(x) = f(x)$ for $0 < x \leq 1$, and $u(0) = u_0$, $v(0) = u_0$, where $f : [0,1] \to \mathbb{R}$ is Lipschitz continuous. Then the difference $w(x) = u(x) - v(x)$ is a uniformly differentiable function on $[0,1]$ satisfying $w'(x) = 0$ for $a < x \leq b$ and $w(0) = 0$. But we just showed that

$$w(x) = \int_a^x w'(y)\,dy,$$

and thus $w(x) = 0$ for $x \in [0,1]$. This proves that $u(x) = v(x)$ for $x \in [0,1]$ and the uniqueness follows.

Recall that we proved the Fundamental Theorem for special circumstances, namely on the interval $[0,1]$ with initial value 0. We can directly generalize the construction above by replacing $[0,1]$ by an arbitrary bounded interval $[a,b]$, replacing h_n by $h_n = 2^{-n}(b-a)$, and assuming instead of $u(0) = 0$ that $u(a) = u_a$, where u_a is a given real number. We have now proved the formidable Fundamental Theorem of Calculus.

Theorem 27.1 (Fundamental Theorem of Calculus) *If $f : [a,b] \to \mathbb{R}$ is Lipschitz continuous, then there is a unique uniformly differentiable function $u : [a,b] \to \mathbb{R}$, which solves the initial value problem*

$$\begin{cases} u'(x) = f(x) \quad \text{for } x \in (a,b], \\ u(a) = u_a, \end{cases} \tag{27.32}$$

where $u_a \in \mathbb{R}$ is given. The function $u : [a,b] \to \mathbb{R}$ can be expressed as

$$u(\bar{x}) = u_a + \int_a^{\bar{x}} f(x)\,dx \quad \text{for } \bar{x} \in [a,b],$$

where

$$\int_0^{\bar{x}} f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{j} f(x_{i-1}^n)h_n,$$

with $\bar{x} = x_j^n$, $x_i^n = a + ih_n$, $h_n = 2^{-n}(b-a)$. More precisely, if the Lipschitz constant of $f : [a,b] \to \mathbb{R}$ is L_f, then for $n = 1, 2, \ldots,$

$$\Big| \int_a^{\bar{x}} f(x)\,dx - \sum_{i=1}^{j} f(x_{i-1}^n)h_n \Big| \leq \frac{1}{2}(\bar{x}-a)L_f h_n. \tag{27.33}$$

Furthermore if $|f(x)| \leq M_f$ for $x \in [a,b]$, then $u(x)$ is Lipschitz continuous with Lipschitz constant M_f and $K_u \leq \frac{1}{2}L_f$, where K_u is the constant of uniform differentiability of $u : [a,b] \to \mathbb{R}$.

27.9 Comments on the Notation

We can change the names of the variables and rewrite (27.26) as

$$u(x) = \int_0^x f(y)\,dy. \tag{27.34}$$

We will often use the Fundamental Theorem in the form

$$\int_a^b u'(x)\,dx = u(b) - u(a), \tag{27.35}$$

which states that the integral $\int_a^b f(x)\,dx$ is equal to the difference $u(b) - u(a)$, where $u(x)$ is a primitive function of $f(x)$. We will sometimes use the notation $[u(x)]_{x=a}^{x=b} = u(b) - u(a)$ or shorter $[u(x)]_a^b = u(b) - u(a)$, and write

$$\int_a^b u'(x)\,dx = \big[u(x)\big]_{x=a}^{x=b} = \big[u(x)\big]_a^b.$$

Sometimes the notation

$$\int f(x)\,dx,$$

without limits of integration, is used to denote a primitive function of $f(x)$. With this notation we would have for example

$$\int dx = x + C, \quad \int x\,dx = \frac{x^2}{2} + C, \quad \int x^2\,dx = \frac{x^3}{3} + C,$$

where C is a constant. We will not use this notation in this book. Note that the formula $x = \int dx$ may be viewed to express that "the whole is equal to the sum of the parts".

27.10 Alternative Computational Methods

Note that we might as well compute $U^n(x_i^n)$ from knowledge of $U^n(x_{i-1}^n)$, using the formula

$$U^n(x_i^n) = U^n(x_{i-1}^n) + f(x_i^n)h_n, \tag{27.36}$$

obtained by replacing $f(x_{i-1}^n)$ by $f(x_i^n)$, or

$$U^n(x_i^n) = U^n(x_{i-1}^n) + \frac{1}{2}(f(x_{i-1}^n) + f(x_i^n))h_n \tag{27.37}$$

using the mean value $\frac{1}{2}(f(x_{i-1}^n) + f(x_i^n))$. These alternatives may bring certain advantages, and we will return to them in Chapter *Numerical quadrature*. The proof of the Fundamental Theorem is basically the same with these variants and by uniqueness all the alternative constructions give the same result.

27.11 The Cyclist's Speedometer

An example of a physical situation modeled by the initial value problem (27.2) is a cyclist biking along a straight line with $u(x)$ representing the position at time x, $u'(x)$ being the speed at time x and specifying the position $u(a) = u_a$ at the initial time $x = a$. Solving the differential equation (27.2)

amounts to determining the position $u(x)$ of the cyclist at time $a < x \leq b$, after specifying the position at the initial time $x = a$ and knowing the speed $f(x)$ at each time x. A standard bicycle speedometer may be viewed to solve this problem, viewing the speedometer as a device which measures the instantaneous speed $f(x)$, and then outputs the total traveled distance $u(x)$. Or is this a good example? Isn't it rather so that the speedometer measures the traveled distance and then reports the momentary (average) speed? To answer this question would seem to require a more precise study of how a speedometer actually works, and we urge the reader to investigate this problem.

27.12 Geometrical Interpretation of the Integral

In this section, we interpret the proof of the Fundamental Theorem as saying that the integral of a function is the area underneath the graph of the function. More precisely, the solution $u(\bar{x})$ given by (27.3) is equal to the area under the graph of the function $f(x)$ on the interval $[a, \bar{x}]$, see Fig. 27.6. For the purpose of this discussion, it is natural to assume that $f(x) \geq 0$.

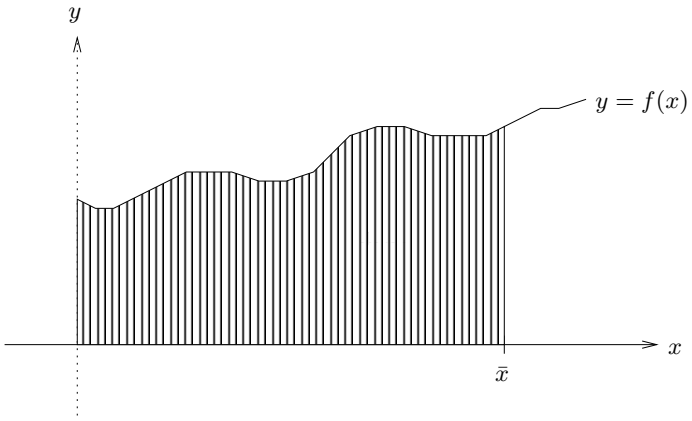

Fig. 27.6. Area under $y = f(x)$

Of course, we also have to explain what we mean by the area under the graph of the function $f(x)$ on the interval $[a, \bar{x}]$. To do this, we first interpret the approximation $U^n(\bar{x})$ of $u(\bar{x})$ as an area. We recall from the previous section that

$$U^n(x_j^n) = \sum_{i=1}^{j} f(x_{i-1}^n)h_n,$$

where $x_j^n = \bar{x}$. Now, we can view $f(x_{i-1}^n)h_n$ as the area of a rectangle with base h_n and height $f(x_{i-1}^n)$, see Fig. 27.7.

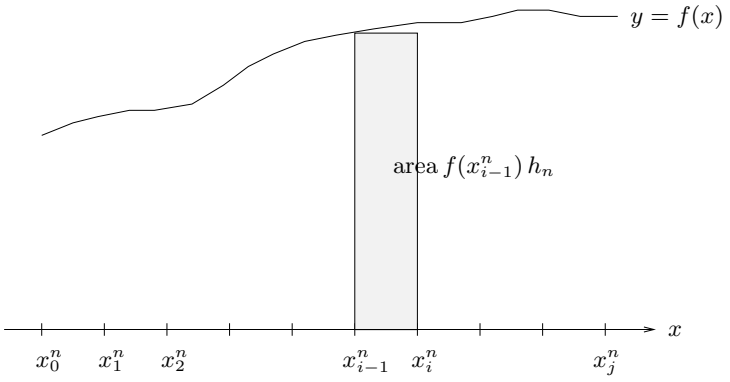

Fig. 27.7. Area $f(x_{i-1}^n)\, h_n$ of rectangle

We can thus interpret the sum

$$\sum_{i=1}^{j} f(x_{i-1}^n)h_n$$

as the area of a collection of rectangles which form a staircase approximation of $f(x)$, as displayed in Fig. 27.8. The sum is also referred to as a *Riemann sum.*

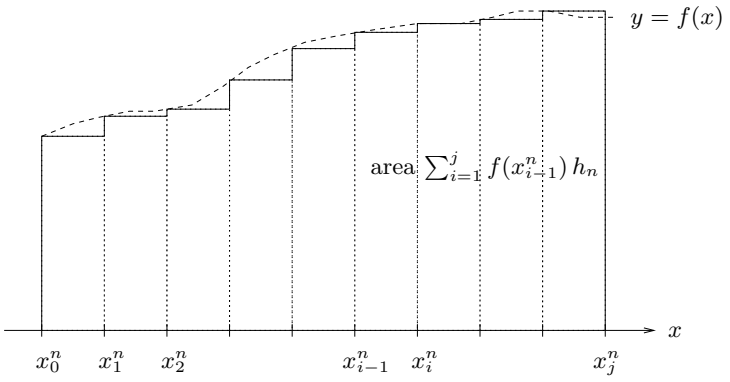

Fig. 27.8. Area $\sum_{i=1}^{j} f(x_{i-1}^n)\, h_n$ under a staircase approximation of $f(x)$

Intuitively, the area under the staircase approximation of $f(x)$ on $[a, \bar{x}]$, which is $U^n(\bar{x})$, will approach the area under the graph of $f(x)$ on $[a, \bar{x}]$ as n tends to infinity and $h_n = 2^{-n}(b-a)$ tends to zero. Since $\lim_{n\to\infty} U^n(\bar{x}) =$

$u(\bar{x})$, this leads us to *define* the area under $f(x)$ on the interval $[0, \bar{x}]$ as the limit $u(\bar{x})$.

Note the logic used here: The value $U^n(\bar{x})$ represents the area under a staircase approximation of $f(x)$ on $[a, \bar{x}]$. We know that $U^n(\bar{x})$ tends to $u(\bar{x})$ as n tends to infinity, and on intuitive grounds we feel that the limit of the area under the staircase should be equal to the area under the graph of $f(x)$ on $[a, \bar{x}]$. We then simply define the area under $f(x)$ on $[a, \bar{x}]$ to be $u(\bar{x})$. By definition we thus interpret the integral of $f(x)$ on $[0, \bar{x}]$ as the area under the graph of the function $f(x)$ on $[a, \bar{x}]$. Note that *this is an interpretation*. It is not a good idea to say the integral *is* an area. This is because the integral can represent many things, such as a distance, a quantity of money, a weight, or some thing else. If we interpret the integral as an area, then we also interpret a distance, a quantity of money, a weight, or some thing else, as an area. We understand that we cannot take this interpretation to be literally true, because a distance cannot *be equal* to an area, but it can be *interpreted* as an area. We hope the reader gets the (subtle) difference.

As an example, we compute the area A under the graph of the function $f(x) = x^2$ between $x = 0$ and $x = 1$ as follows

$$A = \int_0^1 x^2\,dx = \left[\frac{x^3}{3}\right]_{x=0}^{x=1} = \frac{1}{3}.$$

This is an example of the magic of Calculus, behind its enormous success. We were able to compute an area, which in principle is the sum of very many very small pieces, without actually having to do the tedious and laborious computation of the sum. We just found a primitive function $u(x)$ of x^2 and computed $A = u(3) - u(0)$ without any effort at all. Of course we understand the telescoping sum behind this illusion, but if you don't see this, you must be impressed, right? To get a perspective and close a circle, we recall the material in Leibniz' teen-age dream in Chapter *A very short course in Calculus*.

27.13 The Integral as a Limit of Riemann Sums

The Fundamental Theorem of Calculus states that the integral of $f(x)$ over the interval $[a, b]$ is equal to a limit of Riemann sums:

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{2^n} f(x_{i-1}^n)h_n,$$

where $x_i^n = a + ih_n$, $h_n = 2^{-n}(b-a)$, or more precisely, for $n = 1, 2, \ldots,$

$$\Big|\int_a^b f(x)\,dx - \sum_{i=1}^{2^n} f(x_{i-1}^n)h_n\Big| \le \frac{1}{2}(b-a)L_f h_n,$$ TS[j]

TS[j] In the hardcopy, this equation is labelled (27.38), please check it.

Editor's or typesetter's annotations (will be removed before the final TEX run)

where L_f is the Lipschitz constant of f. We can thus define the integral $\int_a^b f(x)\,dx$ as a limit of Riemann sums without invoking the underlying differential equation $u'(x) = f(x)$. This approach is useful in defining integrals of functions of several variables (so-called multiple integrals like double integrals and triple integrals), because in these generalizations there is no underlying differential equation.

In our formulation of the Fundamental Theorem of Calculus, we emphasized the coupling of the integral $\int_a^x f(y)\,dy$ to the related differential equation $u'(x) = f(x)$, but as we just said, we could put this coupling in the back-ground, and define the integral as a limit of Riemann sums without invoking the underlying differential equation. This connects with the idea that the integral of a function can be interpreted as the area under the graph of the function, and will find a natural extension to multiple integrals in Chapters *Double integrals* and *Multiple integrals*.

Defining the integral as a limit of Riemann sums poses a question of uniqueness: since there are different ways of constructing Riemann sums one may ask if all limits will be the same. We will return to this question in Chapter *Numerical quadrature* and (of course) give an affirmative answer.

27.14 An Analog Integrator

James Thompson, brother of Lord Kelvin, constructed in 1876 an analog mechanical integrator based on a rotating disc coupled to a cylinder through another orthogonal disc adjustable along the radius of the first disc, see Fig. 27.9.TS^k The idea was to get around the difficulties of realizing the Analytical Engine, the mechanical digital computer envisioned by Babbage in the 1830s. Lord Kelvin tried to use a system of such analog integrators to compute different problems of practical interest including that of tide prediction, but met serious problems to reach sufficient accuracy. Similar ideas ideas were taken up by Vannevar Bush at MIT Massachussetts Institute of Technology in the 1930s, who constructed a *Differential Analyzer* consisting of a collection of analog integrators, which was programmable to solve differential equations, and was used during the Second World War for computing trajectories of projectiles. A decade later the digital computer took over the scene, and the battle between arithmetic and geometry initiated between the Pythagorean and Euclidean schools more than 2000 years ago, had finally come an end.

TS^k In the hardcopy, is here a reference of Fig. 27.10, please check it.

Editor's or typesetter's annotations (will be removed before the final TEX run)

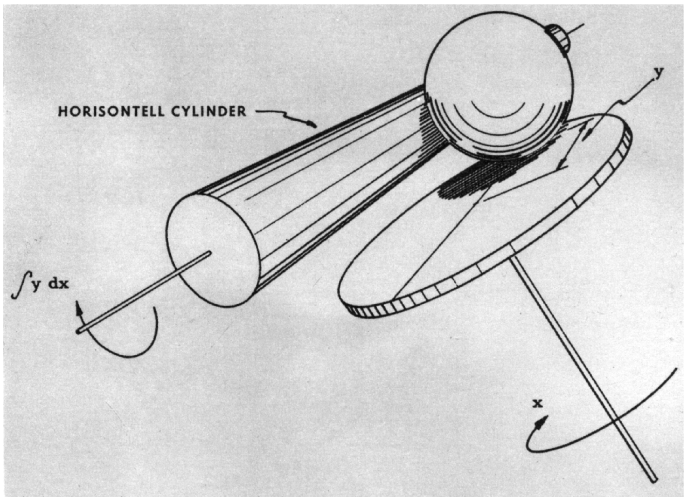

Fig. 27.9. The principle of an Analog Integrator

Chapter 27 Problems

27.1. Determine primitive functions on $\mathbb{R}$ to (a) $(1+x^2)^{-2}2x$, (b) $(1+x)^{-99}$, (c) $(1+(1+x^3)^2)^{-2}2(1+x^3)3x^2$.

27.2. Compute the area under the graph of the function $(1+x)^{-2}$ between $x=1$ and $x=2$.

27.3. A car travels along the x-axis with speed $v(t)=t^{\frac{3}{2}}$ starting at $x=0$ for $t=0$. Compute the position of the car for $t=10$.

27.4. Carry out the proof of the Fundamental Theorem for the variations (27.36) and (27.37).

27.5. Construct a *mechanical integrator* solving the differential equation $u'(x)=f(x)$ for $x>0$, $u(0)=0$ through an analog mechanical devise. Hint: Get hold of a rotating cone and a string.

27.6. Explain the principle behind Thompson's analog integrator.

27.7. Construct a *mechanical* speedometer reporting the speed and traveled distance. Hint: Check the construction of the speedometer of your bike.

27.8. Find the solutions of the initial value problem $u'(x)=f(x)$ for $x>0$, $u(0)=1$, in the following cases: (a) $f(x)=0$, (b) $f(x)=1$, (c) $f(x)=x^r,\ r>0$.

27.9. Find the solution to the second order initial value problem $u''(x)=f(x)$ for $x>0$, $u(0)=u'(0)=1$, in the following cases: (a) $f(x)=0$, (b) $f(x)=1$, (c) $f(x)=x^r,\ r>0$. Explain why two initial conditions are specified.

27.10. Solve initial value problem $u'(x) = f(x)$ for $x \in (0, 2]$, $u(0) = 1$, where $f(x) = 1$ for $x \in [0, 1)$ and $f(x) = 2$ for $x \in [1, 2]$. Draw a graph of the solution and calculate $u(3/2)$. Show that $f(x)$ is not Lipschitz continuous on $[0, 2]$ and determine if $u(x)$ is Lipschitz continuous on $[0, 2]$.

27.11. The time it takes for a light beam to travel through an object is $t = \frac{d}{c/n}$, where c is the speed of light in vacuum, n is the refractive index of the object and d is its thickness. How long does it take for a light beam to travel the shortest way through the center of a glass of water, if the refractive index of the water varies as a certain function $n_w(r)$ with the distance r from the center of glass, the radius of the glass is R and the thickness and that the walls have constant thickness h and constant refractive index equal to n_g.

27.12. Assume that f and g are Lipschitz continuous on $[0, 1]$. Show that $\int_0^1 |f(x) - g(x)|dx = 0$ if and only if $f = g$ on $[0, 1]$. Does this also hold if $\int_0^1 |f(x) - g(x)|dx$ is replaced by $\int_0^1 (f(x) - g(x))dx$?

Fig. 27.10. David Hilbert (1862–1943) at the age of 24: "A mathematical theory is not to be considered complete until you have made it so clear that you can explain it to the first man whom you meet on the street"

28
Properties of the Integral

> For more than two thousand years some familiarity with mathematics has been regarded as an indispensable part of the intellectual equipment of every cultured person. Today the traditional place of mathematics in education is in great danger. Unfortunately, professional representatives of mathematics share the responsibility. The teaching of mathematics has sometimes degenerated into empty drill in problem solving, which may develop formal ability but does not lead to real understanding or to greater intellectual independence... Teachers, students and the general public demand constructive reform, not resignation along the lines of least resistance. (Richard Courant, in Preface to What is Mathematics?, 1941)

28.1 Introduction

In this chapter, we gather together various useful properties of the integral. We may prove these properties in two ways: (i) by using the connection between the integral and the derivative and using properties of the derivative, or (ii) using that the integral is the limit of Riemann sum approximations, that is, using the area interpretation of the integral. We indicate both types of proofs to help the reader getting familiar with different aspects of the integral, and leave some of the work to the problem section.

Throughout the chapter we assume that $f(x)$ and $g(x)$ are Lipschitz continuous on the interval $[a, b]$, and we assume that

$$\sum_{i=1}^{N} f(x_{i-1}^n)h_n \qquad \text{and} \qquad \sum_{i=1}^{N} g(x_{i-1}^n)h_n$$

are Riemann sum approximations of $\int_a^b f(x)\,dx$ and $\int_a^b g(x)\,dx$ with step length $h_n = 2^{-n}(b-a)$ and $x_i^n = a + ih_n$, $i = 0, 1, \ldots, N = 2^n$, as in the previous chapter.

28.2 Reversing the Order of Upper and Lower Limits

So far we have defined the integral $\int_a^b f(x)\,dx$ assuming that $a \le b$, that is that the upper limit of integration b is larger than (or equal to) the lower limit a. It is useful to *extend* the definition to cases with $a > b$ by defining

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx. \tag{28.1}$$

In other words, we decide that switching the limits of integration should change the sign of an integral. As a motivation we may consider the case $f(x) = 1$ and $a > b$, and recall that $\int_b^a 1\,dx = a - b > 0$. Using the same formula with a and b interchanged, we would have $\int_a^b 1\,dx = b - a = -(a-b) = -\int_b^a 1\,dx$, which motivates the sign change under the switch of upper and lower limits. The motivation carries over to the general case using the Riemann sum approximation. Notice that here we do not *prove* anything, we simply introduce a *definition*. Of course we seek a definition which is natural, easy to remember and which allows efficient symbolic computation. The definition we chose fulfills these conditions.

Example 28.1. We have

$$\int_2^1 2x\,dx = -\int_1^2 2x\,dx = -\left[x^2\right]_1^2 = -(4-1) = -3.$$

28.3 The Whole Is Equal to the Sum of the Parts

We shall now prove that if $a \le c \le b$, then

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx. \tag{28.2}$$

One way to do this is to use the area interpretation of the integral and simply notice that the area under $f(x)$ from a to b should be equal to the sum of the area under $f(x)$ from a to c and the area under $f(x)$ from c to b.

We can also give an alternative proof using that that $\int_a^b f(x)\,dx = u(b)$, where $u(x)$ satisfies $u'(x) = f(x)$ for $a \le x \le b$, and $u(a) = 0$. Letting now $w(x)$ satisfy $w'(x) = f(x)$ for $c \le x \le b$, and $w(c) = u(c)$, we have by uniqueness that $w(x) = u(x)$ for $c \le x \le b$, and thus

$$u(b) = w(b) = u(c) + \int_c^b f(y)\,dy = \int_a^c f(y)\,dy + \int_c^b f(y)\,dy,$$

which is the desired result.

Example 28.2. We have

$$\int_0^2 x\,dx = \int_0^1 x\,dx + \int_1^2 x\,dx,$$

which expresses the identity

$$2 = \left(\frac{1}{2}\right) + \left(2 - \frac{1}{2}\right).$$

Note that by the definition (28.1), (28.2) actually holds for any a, b and c.

28.4 Integrating Piecewise Lipschitz Continuous Functions

A functions is said to be *piecewise Lipschitz continuous* on a finite interval $[a, b]$ if $[a, b]$ can be divided up into a finite number of sub-intervals on which the function is Lipschitz continuous, allowing the function to have jumps at the ends of the subintervals, see Fig. 28.1.

We now extend (in the obvious way) the definition of the integral $\int_a^b f(x)\,dx$ to a piecewise Lipschitz continuous function $f(x)$ on an interval $[a, b]$ starting with the case of two subintervals with thus $f(x)$ Lipschitz continuous separately on two adjoining intervals $[a, c]$ and $[c, b]$, where $a \le c \le b$. We define

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx,$$

which obviously fits with (28.2). The extension is analogous for several subintervals with the integral over the whole interval being the sum of the integrals over the subintervals.

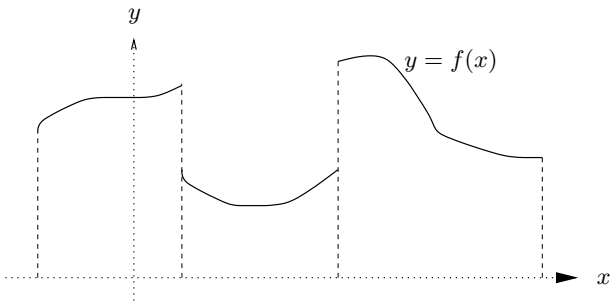

Fig. 28.1. A piecewise Lipschitz continuous function

28.5 Linearity

We shall now prove the following property of *linearity* of the integral: If α and β are real numbers then,

$$\int_a^b (\alpha f(x) + \beta g(x))\,dx = \alpha \int_a^b f(x)\,dx + \beta \int_a^b g(x)\,dx. \tag{28.3}$$

With $\alpha = \beta = 1$ this property expresses that the area (from a to b) underneath the sum of two functions is equal to the sum of the areas underneath each function. Further, with $g(x) = 0$ and $\alpha = 2$ say, the property expresses that the area under the function $2f(x)$ is equal to 2 times the area under the function $f(x)$.

More generally, the linearity of the integral is inherited from the linearity of the Riemann sum approximation, which we may express as

$$\sum_{i=1}^{N} (\alpha f(x_{i-1}^n) + \beta g(x_{i-1}^n) h_n = \alpha \sum_{i=1}^{N} f(x_{i-1}^n) h_n + \beta \sum_{i=1}^{N} g(x_{i-1}^n) h_n, \tag{28.4}$$

and which follows from basic rules for computing with real numbers.

A differential equation proof goes as follows: Define

$$u(x) = \int_a^x f(y)\,dy \quad \text{and} \quad v(x) = \int_a^x g(y)\,dy, \tag{28.5}$$

that is, $u(x)$ is a primitive function of $f(x)$ satisfying $u'(x) = f(x)$ for $a < x \leq b$ and $u(a) = 0$, and $v(x)$ is a primitive function of $g(x)$ satisfying $v'(x) = g(x)$ for $a < x \leq b$ and $v(a) = 0$. Now, the function $w(x) = \alpha u(x) + \beta v(x)$ is a primitive function of the function $\alpha f(x) + \beta g(x)$, since by the linearity of the derivative, $w'(x) = \alpha u'(x) + \beta v'(x) = \alpha f(x) + \beta g(x)$, and $w(a) = \alpha u(a) + \beta v(a) = 0$. Thus, the left hand side of (28.3) is equal to $w(b)$, and since $w(b) = \alpha u(b) + \beta v(b)$, the desired equality follows from setting $x = b$ in (28.5).

Example 28.3. We have

$$\int_0^b (2x + 3x^2)\,dx = 2\int_0^b x\,dx + 3\int_0^b x^2\,dx = 2\frac{b^2}{2} + 3\frac{b^3}{3} = b^2 + b^3.$$

28.6 Monotonicity

The *monotonicity* property of the integral states that if $f(x) \geq g(x)$ for $a \leq x \leq b$, then

$$\int_a^b f(x)\,dx \geq \int_a^b g(x)\,dx. \tag{28.6}$$

This is the same as stating that if $f(x) \geq 0$ for $x \in [a, b]$, then

$$\int_a^b f(x)\,dx \geq 0, \tag{28.7}$$

which follows from the fact that all Riemann sum approximations $\sum_{i=1}^{j} f(x_{i-1}^n)h_n$ of $\int_a^b f(x)\,dx$ are all non-negative if $f(x) \geq 0$ for $x \in [a, b]$.

28.7 The Triangle Inequality for Integrals

We shall now prove the following *triangle inequality for integrals*:

$$\Big|\int_a^b f(x)\,dx\Big| \leq \int_a^b |f(x)|\,dx, \tag{28.8}$$

stating that moving the absolute value inside the integral increases the value (or leaves the value unchanged). This property follows from applying the usual triangle inequality to Riemann sum approximations to get

$$\Big|\sum_{i=1}^{N} f(x_{i-1}^n)h_n\Big| \leq \sum_{i=1}^{N} \Big|f(x_{i-1}^n)\Big|h_n$$

and then passing to the limit. Evidently there may be cancellations on the left hand side if $f(x)$ has changes sign, while on the right hand side we always add nonnegative contributions, making the right hand side at least as big as the left hand side.

Another proof uses the monotonicity as follows: Apply (28.7) to the function $|f| - f \geq 0$ to obtain

$$\int_a^{\bar{x}} f(x)\,dx \leq \int_a^{\bar{x}} |f(x)|\,dx.$$

Replacing f by the function $-f$ we obtain

$$-\int_a^{\bar{x}} f(x)\,dx = \int_a^{\bar{x}} (-f(x))\,dx \leq \int_a^{\bar{x}} |-f(x)|\,dx = \int_a^{\bar{x}} |f(x)|\,dx,$$

which proves the desired result.

28.8 Differentiation and Integration are Inverse Operations

The Fundamental Theorem says that integration and differentiation are *inverse operations* in the sense that first integrating and then differentiating, or first differentiating and then integrating, gives the net result of doing nothing! We make this clear by repeating a part of the proof of the Fundamental Theorem to prove that if $f : [a, b] \to \mathbb{R}$ is Lipschitz continuous then for $x \in [a, b]$,

$$\frac{d}{dx}\int_a^x f(y)\,dy = f(x). \tag{28.9}$$

In other words, integrating a function $f(x)$ and then differentiating the primitive function, gives back the function $f(x)$. Surprise? We illustrate in Fig. 28.2. To properly understand the equality (28.9), it is important to realize that $\int_a^x f(y)\,dy$ is a function of x and thus depends on x. The area under the function f from a to x, of course depends on the upper limit x.

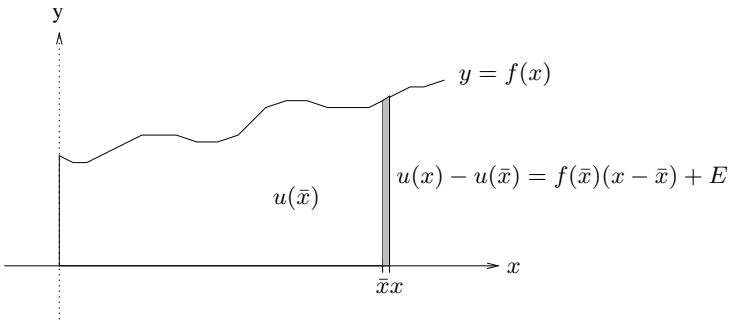

Fig. 28.2. The derivative of $\int_0^x f(y)\,dy$ at $x = \bar{x}$ is $f(\bar{x})$: $|E| \leq \frac{1}{2}L_f|x - \bar{x}|^2$

We may express (28.9) in words as follows: Differentiating an integral with respect to the upper limit of integration gives the value of the integrand at the upper limit of integration.

To prove (28.9) we now choose x and $\bar{x}$ in $[a, b]$ with $x \geq \bar{x}$, and use (28.2) to see that

$$u(x) - u(\bar{x}) = \int_a^x f(z)\,dz - \int_a^{\bar{x}} f(y)dy = \int_{\bar{x}}^x f(y)\,dy$$

so that

$$\begin{aligned}|u(x) - u(\bar{x}) - f(\bar{x})(x-\bar{x})| &= \Big| \int_{\bar{x}}^{x} f(y)\,dy - f(\bar{x})(x-\bar{x}) \Big| \\ &= \Big| \int_{\bar{x}}^{x} (f(y) - f(\bar{x}))\,dy \Big| \\ &\le \int_{\bar{x}}^{x} |f(y) - f(\bar{x})|\,dy \\ &\le \int_{\bar{x}}^{x} L_f|y-\bar{x}|\,dy = \frac{1}{2}L_f(x-\bar{x})^2.\end{aligned}$$

This proves that $u(x)$ is uniformly differentiable on $[a,b]$ with derivative $u'(x) = f(x)$ and constant $K_u \le \frac{1}{2}L_f$.

We also note that (28.1) implies

$$\frac{d}{dx}\int_{x}^{a} f(y)\,dy = -f(x). \tag{28.10}$$

which we may express in words as: Differentiating an integral with respect to the lower limit of integration gives minus the value of the integrand at the lower limit of integration.

Example 28.4. We have

$$\frac{d}{dx}\int_{0}^{x} \frac{1}{1+y^2}\,dy = \frac{1}{1+x^2}.$$

Example 28.5. Note that we can combine (27.30) with the Chain Rule:

$$\frac{d}{dx}\int_{0}^{x^3} \frac{1}{1+y^2}\,dy = \frac{1}{1+(x^3)^2}\frac{d}{dx}\big(x^3\big) = \frac{3x^2}{1+x^6}.$$

28.9 Change of Variables or Substitution

We recall that the Chain rule tells us how to differentiate the composition of two functions. The analogous property of the integral is known as the *change of variables*, or *substitution* formula and plays an important role. For example, it can be used to compute many integrals analytically. The idea is that an integral may be easier to compute analytically if we change scales in the independent variable.

Let now $g : [a,b] \to I$, be uniformly differentiable on an interval $[a,b]$, where I is an interval, and let $f : I \to \mathbb{R}$ be Lipschitz continuous. Typically, g is strictly increasing (or decreasing) and maps $[a,b]$ onto I, so that g :

$[a,b] \to I$ corresponds to a change of scale, but more general situations are allowed. The *change of variables* formula reads

$$\int_a^x f(g(y))g'(y)\,dy = \int_{g(a)}^{g(x)} f(z)\,dz \quad \text{for } x \in [a,b]. \tag{28.11}$$

This is also called *substitution* since the left hand side $L(x)$ is formally obtained by in the right hand hand side $H(x)$ setting $z = g(y)$ and replacing dz by $g'(y)\,dy$ motivated by the relation

$$\frac{dz}{dy} = g'(y),$$

and noting that as y runs from a to x then z runs from $g(a)$ to $g(x)$.

To verify (28.11), we now prove that $H'(x) = L'(x)$ and use the fact that $H(a) = L(a) = 0$ and the uniqueness of the integral. The Chain rule and (27.30) imply that

$$H'(x) = f(g(x))\,g'(x).$$

Further,

$$L'(x) = f(g(x))\,g'(x),$$

which thus proves the desired equality.

We now give a two examples. We will meet many more examples below.

Example 28.6. To integrate

$$\int_0^2 (1+y^2)^{-2} 2y\,dy$$

we make the observation that

$$\frac{d}{dy}(1+y^2) = 2y.$$

Thus, if we set $z = g(y) = 1+y^2$, then applying (28.11) noting that $g(0) = 1$ and $g(2) = 5$ and formally $dz = 2ydy$, we have that

$$\int_0^2 (1+y^2)^{-2} 2y\,dy = \int_0^2 (g(y))^{-2} g'(y)\,dy = \int_1^5 z^{-2}\,dz.$$

Now, the right hand integral can easily be evaluated:

$$\int_1^5 z^{-2}\,dz = [-z^{-1}]_{z=1}^{z=5} = -\left(\frac{1}{5} - 1\right),$$

and thus

$$\int_0^2 (1+y^2)^{-2} 2y\,dy = \frac{4}{5}.$$

Example 28.7. We have setting $y = g(x) = 1+x^4$ noting that then formally $dy = g'(x)dx = 4x^3dx$ and $g(0) = 1$ and $g(1) = 2$, to get

$$\int_0^1 (1+x^4)^{-1/2}x^3\,dx = \frac{1}{4}\int_0^1 (g(x))^{-1/2}g'(x)\,dx = \frac{1}{4}\int_1^2 y^{-1/2}\,dy$$
$$= \frac{1}{2}[y^{1/2}]_1^2 = \frac{\sqrt{2}-1}{2}.$$

28.10 Integration by Parts

We recall that the Product rule is a basic property of the derivative, showing how to compute the derivative of a product of two functions. The corresponding formula for integration is called *integration by parts*. The formula is

$$\int_a^b u'(x)v(x)\,dx = u(b)v(b) - u(a)v(a) - \int_a^b u(x)v'(x)\,dx. \qquad (28.12)$$

The formula follows by applying the Fundamental Theorem to the function $w(x) = u(x)v(x)$, in the form

$$\int_a^b w'(x)\,dx = u(b)v(b) - u(a)v(a),$$

together with the product formula $w'(x) = u'(x)v(x)+u(x)v'(x)$ and (28.3). Below we often write

$$u(b)v(b) - u(a)v(a) = \Big[u(x)v(x)\Big]_{x=a}^{x=b},$$

and we can thus state the formula for integration by parts as

$$\int_a^b u'(x)v(x)\,dx = \Big[u(x)v(x)\Big]_{x=a}^{x=b} - \int_a^b u(x)v'(x)\,dx. \qquad (28.13)$$

This formula is very useful and we will use many times below.

Example 28.8. Computing

$$\int_0^1 4x^3(1+x^2)^{-3}\,dx$$

by guessing at a primitive function for the integrand would be a pretty daunting task. However we can use integration by parts to compute the integral. The trick here is to realize that

$$\frac{d}{dx}(1+x^2)^{-2} = -4x(1+x^2)^{-3}.$$

If we rewrite the integral as

$$\int_0^1 x^2 \times 4x(1+x^2)^{-3}\,dx$$

then we can apply integration by parts with $u(x) = x^2$ and $v'(x) = 4x(1+x^2)^{-3}$, so $u'(x) = 2x$ and $v(x) = -(1+x^2)^{-2}$, to get

$$\begin{aligned}\int_0^1 4x^3(1+x^2)^{-3}\,dx &= \int_0^1 u(x)v'(x)\,dx \\ &= \big[x^2(-(1+x^2)^{-2})\big]_{x=0}^{x=1} - \int_0^1 2x(-(1+x^2)^{-2})\,dx \\ &= -\frac{1}{4} - \int_0^1 (-(1+x^2)^{-2})2x\ dx.\end{aligned}$$

To do the remaining integral, we use the substitution $z = 1 + x^2$ with $dz = 2x\,dx$ to get

$$\begin{aligned}\int_0^1 4x^3(1+x^2)^{-3}\,dx &= -\frac{1}{4} + \int_1^2 z^{-2}\,dz \\ &= -\frac{1}{4} + \big[-z^{-1}\big]_{z=1}^{z=2} = -\frac{1}{4} - \frac{1}{2} + 1 = \frac{1}{4}.\end{aligned}$$

28.11 The Mean Value Theorem

The *Mean Value theorem* states that if $u(x)$ is a differentiable function on $[a,b]$ then there is a point $\bar{x}$ in (a,b) such that the slope $u'(\bar{x})$ of the tangent of the graph of $u(x)$ at $\bar{x}$ is equal to the slope of the secant line, or chord, connecting the points $(a,u(a))$ and $(b,u(b))$. In other words,

$$\frac{u(b)-u(a)}{b-a} = u'(\bar{x}). \tag{28.14}$$

This is geometrically intuitive, see Fig. 28.3, and expresses that the average velocity over $[a,b]$ is equal to momentary velocity $u'(\bar{x})$ at some intermediate point $\bar{x} \in [a,b]$.

To get from the point $(a,u(a))$ to the point $(b,u(b))$, f has to "bend" around in such a way that the tangent becomes parallel to the secant line at least at one point.

Assuming that $u'(x)$ is Lipschitz continuous on $[a,b]$, we shall now prove that there is a real number $\bar{x} \in [a,b]$ such that

$$u(b) - u(a) = (b-a)u'(\bar{x})$$

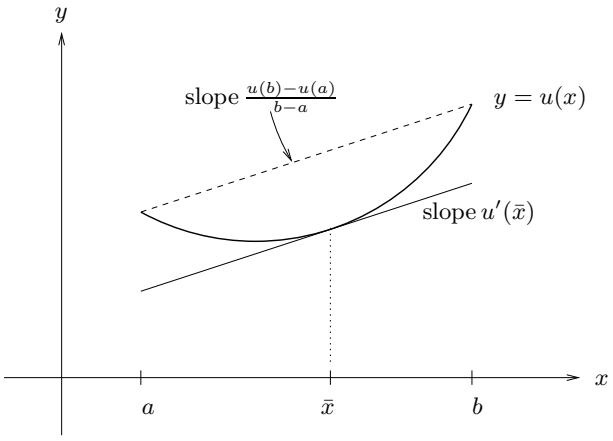

Fig. 28.3. Illustration of the Mean Value theorem

which is equivalent to (28.14). The proof is based on the formula

$$u(b) = u(a) + \int_a^b u'(x)\,dx \tag{28.15}$$

which holds if $u(x)$ is uniformly differentiable on $[a, b]$. If for all $x \in [a, b]$, we would have

$$\frac{u(b) - u(a)}{b - a} > u'(x),$$

then we would have (explain why)

$$u(b) - u(a) = \int_b^a \frac{u(b) - u(a)}{b - a}\,dx > \int_b^a u'(x)\,dx = u(b) - u(a),$$

which is a contradiction. Arguing in the same way, we conclude it is also impossible that

$$\frac{u(b) - u(a)}{b - a} < u'(x)$$

for all $x \in [a, b]$. So there must be numbers c and d in $[a, b]$ such that

$$u'(c) \leq \frac{u(b) - u(a)}{b - a} \leq u'(d).$$

Since $u'(x)$ is Lipschitz continuous for $x \in [a, b]$, it follows by the Intermediate Value theorem that there is a number $\bar{x} \in [a, b]$ such that

$$u'(\bar{x}) = \frac{u(b) - u(a)}{b - a}.$$

We have now proved:

Theorem 28.1 (Mean Value theorem) *If $u(x)$ is uniformly differentiable on $[a, b]$ with Lipschitz continuous derivative $u'(x)$, then there is a (at least one) $\bar{x} \in [a, b]$, such that*

$$u(b) - u(a) = (b - a)u'(\bar{x}). \tag{28.16}$$

The Mean Value Theorem is often written in terms of an integral by setting $f(x) = u'(x)$ in (28.16), which gives

Theorem 28.2 (Mean Value theorem for integrals) *If $f(x)$ is Lipschitz continuous on $[a, b]$, then there is some $\bar{x} \in [a, b]$ such that*

$$\int_a^b f(x)\,dx = (b - a)f(\bar{x}). \tag{28.17}$$

The Mean Value theorem turns out to be very useful in several ways. To illustrate, we discuss two results that can be proved easily using the Mean Value Theorem.

28.12 Monotone Functions and the Sign of the Derivative

The first result says that the sign of the derivative of a function indicates whether the function is increasing or decreasing in value as the input increases. More precisely, the Mean Value theorem implies that if $f'(x) \geq 0$ for all $x \in [a, b]$ then $f(b) \geq f(a)$. Moreover if $x_1 \leq x_2$ are in $[a, b]$, then $f(x_1) \leq f(x_2)$. A function with this property is said to be *non-decreasing* on $[a, b]$. If in fact $f'(x) > 0$ for all $x \in (a, b)$, then $f(x_1) < f(x_2)$ for $x_1 < x_2$ in $[a, b]$ (with strict inequalities), and we say that $f(x)$ is *(strictly) increasing* on the interval $[a, b]$. Corresponding statements hold if $f'(x) \leq 0$ and $f'(x) < 0$, with non-decreasing and (strictly) increasing replaced with *non-increasing* and *(strictly) decreasing*, respectively. Functions that are either (strictly) increasing or (strictly) decreasing on an interval $[a, b]$ are said to be *(strictly) monotone* on $[a, b]$.

28.13 A Function with Zero Derivative is Constant

As a particular consequence of the preceding section, we conclude that if $f'(x) = 0$ for all $x \in [a, b]$, so that $f(x)$ is both non-increasing and non-decreasing on $[a, b]$, then $f(x)$ is constant on $[a, b]$. Thus, a function with derivative vanishing everywhere is a constant function.

28.14 A Bounded Derivative Implies Lipschitz Continuity

As a second consequence of the Mean Value Theorem, we give an alternate, and shorter, proof that a function with a Lipschitz continuous derivative is Lipschitz continuous. Assume that $u : [a, b] \to \mathbb{R}$ has a Lipschitz continuous derivative $u'(x)$ on $[a, b]$ satisfying $|u'(x)| \leq M$ for $x \in [a, b]$. By the Mean Value theorem, we have

$$|u(x) - u(\bar{x})| \leq M|x - \bar{x}| \quad \text{for } x, \bar{x} \in [a, b].$$

We conclude that $u(x)$ is Lipschitz continuous on $[a, b]$ with Lipschitz constant $M = \max_{x\in[a,b]} |u'(x)|$.

28.15 Taylor's Theorem

In earlier chapters, we analyzed a linear approximation to a function u,

$$u(x) \approx u(\bar{x}) + u'(\bar{x})(x - \bar{x}), \tag{28.18}$$

as well as a quadratic approximation

$$u(x) \approx u(\bar{x}) + u'(\bar{x})(x - \bar{x}) + \frac{u''(\bar{x})}{2}(x - \bar{x})^2. \tag{28.19}$$

These approximations are very useful tools for dealing with nonlinear functions. *Taylor's theorem*, invented by Brook Taylor (1685-1731), see Fig. 28.4, generalizes these approximations to any degree.Taylor sided up with Newton in a long scientific fight with associates of Leibniz about "who's best in Calculus?".

Theorem 28.3 (Taylor's theorem) *If $u(x)$ is $n+1$ times differentiable on the interval I with $u^{(n+1)}$ Lipschitz continuous, then for x, $\bar{x} \in I$, we have*

$$u(x) = u(\bar{x}) + u'(\bar{x})(x - \bar{x}) + \cdots + \frac{u^{(n)}(\bar{x})}{n!}(x - \bar{x})^n + \int_{\bar{x}}^{x} \frac{(x - y)^n}{n!}\, u^{(n+1)}(y)\, dy. \tag{28.20}$$

The polynomial

$$P_n(x) = u(\bar{x}) + u'(\bar{x})(x - \bar{x}) + \cdots + \frac{u^{(n)}(\bar{x})}{n!}(x - \bar{x})^n$$

is called the *Taylor polynomial*, or *Taylor expansion*, of $u(x)$ at $\bar{x}$ of degree n,

Fig. 28.4. Brook Taylor, inventor of the Taylor expansion: "I am the best"

The term

$$R_n(x) = \int_{\bar{x}}^{x} \frac{(x-y)^n}{n!}\, u^{(n+1)}(y)\, dy$$

is called the *remainder term* of order n. We have for $x \in I$,

$$u(x) = P_n(x) + R_n(x).$$

It follows directly that

$$\left(\frac{d^k}{dx^k}\right) P_n(\bar{x}) = \left(\frac{d^k}{dx^k}\right) u(\bar{x}) \quad \text{for } k = 0, 1 \cdots, n.$$

Thus Taylor's theorem gives a polynomial approximation $P_n(x)$ of degree n of a given function $u(x)$, such that the derivatives of order $\leq n$ of $P_n(x)$ and $u(x)$ agree at $x = \bar{x}$.

Example 28.9. The Taylor polynomial of order 2 at $x = 0$ for $u(x) = \sqrt{1+x}$ is given by

$$P_2(x) = 1 + \frac{1}{2} - \frac{1}{8}x^2,$$

since $u(0) = 1$, $u'(0) = \frac{1}{2}$, and $u''(0) = -\frac{1}{4}$.

The proof of Taylor's theorem is a wonderful application of integration by parts, discovered by Taylor. We start by noting that Taylor's theorem with $n = 0$ is just the Fundamental Theorem

$$u(x) = u(\bar{x}) + \int_{\bar{x}}^{x} u'(y)\, dy,$$

Using that $\frac{d}{dy}(y-x)=1$, we get integrating by parts

$$\begin{aligned}
u(x) &= u(\bar{x}) + \int_{\bar{x}}^{x} u'(y)\,dy \\
&= u(\bar{x}) + \int_{\bar{x}}^{x} \frac{d}{dy}(y-x)u'(y)\,dy \\
&= u(\bar{x}) + [(y-x)u'(y)]_{y=\bar{x}}^{y=x} - \int_{\bar{x}}^{x} (y-x)u''(y)\,dy \\
&= u(\bar{x}) + (x-\bar{x})u'(\bar{x}) + \int_{\bar{x}}^{x} (x-y)u''(y)\,dy,
\end{aligned}$$

which is Taylor's theorem with $n=1$. Continuing in this manner, integrating by parts, using the notation $k_n(y)=(y-x)^n/n!$, and noting that for $n\geq 1$

$$\frac{d}{dy}k_n(y) = k_{n-1}(y),$$

we get

$$\begin{aligned}
\int_{\bar{x}}^{x} \frac{(x-y)^{n-1}}{(n-1)!}u^{(n)}(y)\,dy &= (-1)^{n-1}\int_{\bar{x}}^{x} k_{n-1}(y)u^{(n)}(y)\,dy \\
&= (-1)^{n-1}\int_{\bar{x}}^{x} \frac{d}{dy}k_n(y)u^{(n)}(y)\,dy \\
&= [(-1)^{n-1}k_n(y)u^{(n)}(y)]_{y=\bar{x}}^{y=x} - (-1)^{n-1}\int_{\bar{x}}^{x} k_n(y)u^{(n+1)}(y)\,dy \\
&= \frac{u^{(n)}(\bar{x})}{n!}(x-\bar{x})^n + \int_{\bar{x}}^{x} \frac{(x-y)^n}{n!}u^{(n+1)}(y)\,dy.
\end{aligned}$$

This proves Taylor's theorem.

Example 28.10. We compute a fourth order polynomial approximation to $f(x)=\frac{1}{1-x}$ near $x=0$. We have

$$\begin{aligned}
f(x) = \frac{1}{1-x} &\implies f(0)=1, \\
f'(x) = \frac{1}{(1-x)^2} &\implies f'(0)=1, \\
f''(x) = \frac{2}{(1-x)^3} &\implies f''(0)=2, \\
f'''(x) = \frac{6}{(1-x)^4} &\implies f'''(0)=6, \\
f''''(x) = \frac{24}{(1-x)^5} &\implies f''''(0)=24,
\end{aligned}$$

and therefore

$$\begin{aligned} P_4(x) &= 1 + 1(x-0)^1 + \frac{2}{2}(x-0)^2 + \frac{6}{6}(x-0)^3 + \frac{24}{24}(x-0)^4 \\ &= 1 + x + x^2 + x^3 + x^4. \end{aligned}$$

We plot the function and the polynomial in Fig. 28.5. Characteristically, the Taylor polynomial is a very accurate approximation near the $\bar{x}$ but the error becomes larger as x moves further away from $\bar{x}$.

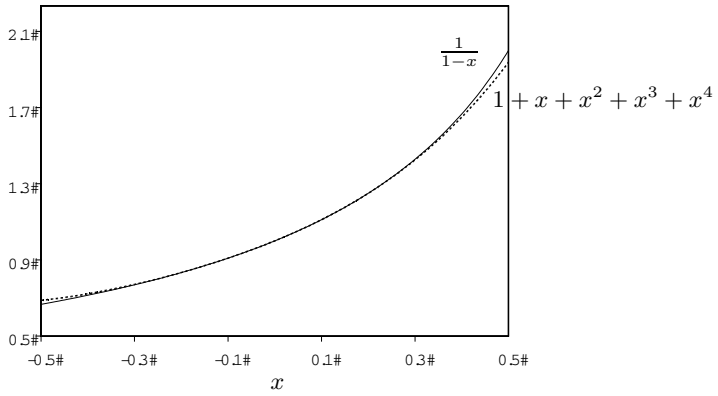

Fig. 28.5. Plots of $f(x) = 1/(1-x)$ and its Taylor polynomial $1+x+x^2+x^3+x^4$

28.16 October 29, 1675

On October 29, 1675, Leibniz got a bright idea while sitting at his desk in Paris. He wrote "Utile erit scribit $\int$ pro omnia", which translates to "It is useful to write $\int$ instead of omnia". This is the moment when the modern notation of calculus was created. Earlier than this date, Leibniz had been working with a notation based on a, l and "omnia" which represented in modern notation dx, dy and $\int$ respectively. This notation resulted in formulas like

$$\text{omn.}l = y, \quad \text{omn.}yl = \frac{y^2}{2}, \quad \text{omn.}xl = x\text{omn.}l - \text{omn.omn.}la,$$

where "omn.", short for omnia, indicated a discrete sum and l and a denoted increments of finite size (often $a = 1$). In the new notation, these formulas became

$$\int dy = y, \quad \int y\,dy = \frac{y^2}{2}, \quad \int x\,dy = xy - \int y\,dx. \tag{28.21}$$

This opened up the possibility of dx and dy being arbitrarily small and the sum being replaced by the "integral".

28.17 The Hodometer

The Romans constructed many roads to keep the Empire together and the need of measuring distances between cities and traveled distance on the road, became very evident. For this purpose the *Hodometer* was constructed by Vitruvius, see Fig. 28.6. For each turn of the wagon wheel, the vertical gear is shifted one step, and for each turn of the vertical gear the horizontal gear is shifted one step. The horizontal gear has a set of holes with one stone in each, and for each shift one stone drops down to a box under the wagon; at the end of the day one computes the number of stones in the box, and the device is so calibrated that this number is equal to the number of traveled miles. Evidently, one may view the hodometer as a kind of simple analog integrator.

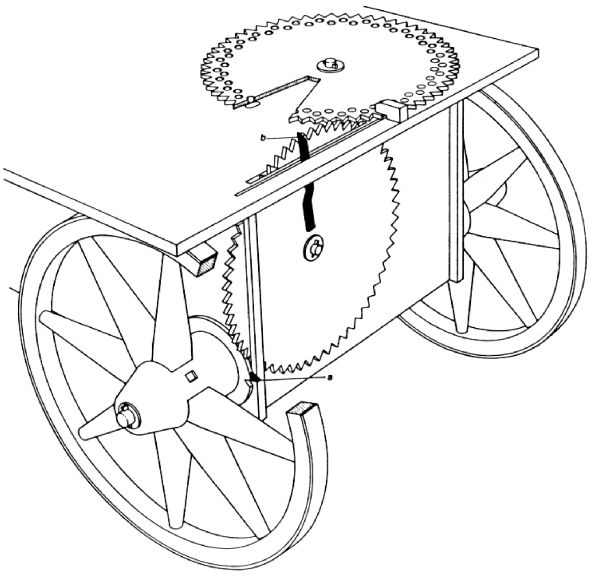

Fig. 28.6. The principle of the Hodometer

Chapter 28 Problems

28.1. Compute the following integrals: a) $\int_0^1 (ax + bx^2)dx$, b) $\int_{-1}^1 |x|dx$, c) $\int_{-1}^1 |x-1|dx$, d) $\int_{-1}^1 |x+a|dx$, e) $\int_{-1}^1 (x-a)^{10}dx$.

28.2. Compute the following integrals by integration by parts. Verify that you get the same result by directly finding the primitive function. a) $\int_0^1 x^2 dx = \int_0^1 x \cdot x dx$, b) $\int_0^1 x^3 dx = \int_0^1 x \cdot x^2 dx$, c) $\int_0^1 x^3 dx = \int_0^1 x^{3/2} \cdot x^{3/2} dx$, d) $\int_0^1 (x^2 - 1)dx = \int_0^1 (x+1) \cdot (x-1)dx$.

28.3. For computing the integral $\int_0^1 x(x-1)^{1000}dx$, what would you rather do; find the primitive function directly or integrate by parts?

28.4. Compute the following integrals: a) $\int_{-1}^2 (2x-1)^7 dx$, b) $\int_0^1 f'(7x)dx$, c) $\int_{-10}^{-7} f'(17x+5)dx$.

28.5. Compute the integral $\int_0^1 x(x^2-1)^{10}dx$ in two ways, first by integration by parts, then by a clever substitution using the chain rule.

28.6. Find Taylor polynomials at $\overline{x}$ of the following functions: a) $f(x) = x$, $\overline{x} = 0$, b) $f(x) = x + x^2 + x^3$, $\overline{x} = 1$, c) $f(x) = \sqrt{\sqrt{x+1}+1}$, $\overline{x} = 0$.

28.7. Find a Taylor expansion of the function $f(x) = x^r - 1$ around a suitable choice of $\overline{x}$, and use the result to compute the limit $\lim_{x\to 1} \frac{x^r-1}{x-1}$. Compare this to using l'Hopital's rule to compute the limit. Can you see the connection between the two methods?

28.8. Motivate the basic properties of linearity and subinterval additivity of the integral using the area interpretation of the integral.

28.9. Prove the basic properties of linearity and subinterval additivity of the integral using that the integral is a limit of discrete sums together with basic properties of discrete sums.

28.10. Make sense out of Leibniz formulas (28.21). Prove, as did Leibniz, the second from a geometrical argument based on computing the area of a right-angled triangle by summing thin slices of variable height y and thickness dy, and the third from computing similarly the area of a rectangle as the sum of the two parts below and above a curve joining two opposite corners of the rectangle.

28.11. Prove the following variant of Taylor's theorem: If $u(x)$ is $n+1$ times differentiable on the interval I, with $u^{(n+1)}(x)$ Lipschitz continuous, then for $\bar{x} \in I$, we have

$$u(x) = u(\bar{x}) + u'(\bar{x})(x-\bar{x}) + \cdots + \frac{u^{(n)}(\bar{x})}{n!}(x-\bar{x})^n + \frac{u^{(n+1)}(\hat{x})}{(n+1)!}(x-\bar{x})^{n+1}$$

where $\hat{x} \in [\bar{x}, x]$. Hint: Use the Mean Value theorem for integrals.

28.12. Prove that if $x = f(y)$ with inverse function $y = f^{-1}(x)$, and $f(0) = 0$, then

$$\int_0^{\bar{y}} f(y)\,dy = \bar{y}\bar{x} - \int_0^{\bar{x}} f^{-1}(x)\,dx.$$

Compare with (28.21) Hint: use integration by parts.

28.13. Show that $x \mapsto F(x) = \int_0^x f(x)dx$ is Lipschitz continuous on $[0, a]$ with Lipschitz constant L_F if $|f(x)| \leq L_F$ for $x \in [0, a]$.

28.14. Why can we think of the primitive function as being "nicer" than the function itself?

28.15. Under what conditions is the following generalized integration by parts formula valid

$$\int_I \frac{d^n f}{dx^n}\varphi dx = (-1)^n \int_I f\frac{d^n \varphi}{dx^n}dx,\ n = 0, 1, 2, \ldots?$$

28.16. Show the following inequality:

$$|\int_I u(x)v(x)\ dx| \leq \sqrt{\int_I u^2(x)\ dx}\ \sqrt{\int_I v^2(x)\ dx},$$

which is referred to as *Cauchy's inequality*. Hint: Let $\overline{u} = u/\sqrt{\int_I u^2(x)\ dx}$, $\overline{v} = v/\sqrt{\int_I v^2(x)\ dx}$, and show that $|\int_I \overline{u}(x)\overline{v}(x)\ dx| \leq 1$ by considering the expression $\int_I(\overline{u}(x) - \int_I \overline{u}(y)\overline{v}(y)\ dy\ \overline{v}(x))\ dx$. Would it be helpful to use the notation $(u, v) = \int_I u(x)v(x)\ dx$, and $\|u\| = \sqrt{\int_I u^2(x)\ dx}$?

28.17. Show that if v is Lipschitz continuous on the bounded interval I and $v = 0$ at one of the endpoints of the interval, then

$$\|v\|_{L^2(I)} \leq C_I\|v'\|_{L^2(I)},$$

for some constant C_I, where the so-called $L^2(I)$ norm of a function v is defined as $\|v\|_{L^2(I)} = \sqrt{\int_I v^2(x)dx}$. What is the value of the constant? Hint: Express v in terms of v' and use the result from the previous problem.

28.18. Check that the inequality from the previous problem holds for the following functions on $I = [0, 1]$: a) $v(x) = x(1 - x)$, b) $v(x) = x^2(1 - x)$, c) $v(x) = x(1 - x)^2$.

28.19. Prove quadratic convergence of Newton's method (25.5) for computing a root $\bar{x}$ of the equation $f(x) = 0$ using Taylor's theorem. Hint: Use the fact that $x_{i+1} - \bar{x} = x_i - \bar{x} + \frac{f(x_i)-f(\bar{x})}{f'(x_i)}$ and Taylor's theorem to see that $f(x_i) - f(\bar{x}) = f'(x_i)(x_i - \bar{x}) + \frac{1}{2}f''(\tilde{x}_i)(x_i - \bar{x})^2$ for some $\tilde{x}_i \approx x_i$.

28.20. Prove (28.3) from (28.4).

29
The Logarithm $\log(x)$

> Nevertheless technicalities and detours should be avoided, and the presentation of mathematics should be just as free from emphasis on routine as from forbidding dogmatism, which refuses to disclose motive or goal and which is an unfair obstacle to honest effort. (R. Courant)

29.1 The Definition of $\log(x)$

We return to the question of the existence of a primitive function of $f(x) = 1/x$ for $x > 0$ posed above. Since the function $f(x) = 1/x$ is Lipschitz continuous on any given interval $[a, b]$ with $0 < a < b$, we know by the Fundamental Theorem that there is a unique function $u(x)$ which satisfies $u'(x) = 1/x$ for $a \leq x \leq b$ and takes on a specific value at some point in $[a, b]$, for example $u(1) = 0$. Since $a > 0$ may be chosen as small as we please and b as large as we please, we may consider the function $u(x)$ to be defined for $x > 0$. We now define the *natural logarithm* $\log(x)$ (or $\ln(x)$) for $x > 0$ as the primitive function $u(x)$ of $1/x$ vanishing for $x = 1$, i.e., $\log(x)$ satisfies

$$\frac{d}{dx}(\log(x)) = \frac{1}{x} \quad \text{for } x > 0, \quad \log(1) = 0. \tag{29.1}$$

Using the definition of the integral, we may express $\log(x)$ as an integral:

$$\log(x) = \int_1^x \frac{1}{y}\, dy \quad \text{for } x > 0. \tag{29.2}$$

In the next chapter we shall use this formula to compute approximations of $\log(x)$ for a given $x > 0$ by computing approximations of the corresponding integral. We plot $\log(x)$ in Fig. 29.1.

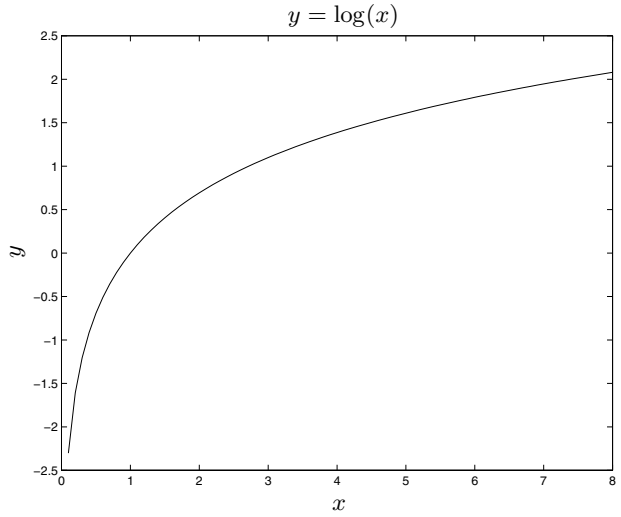

Fig. 29.1. Plot of $\log(x)$

29.2 The Importance of the Logarithm

The logarithm function $\log(x)$ is a basic function in science, simply because it solves a basic differential equation, and thus occurs in many applications. More concretely, the logarithm has some special properties that made previous generations of scientists and engineers use the logarithm intensely, including memorizing long tables of its values. The reason is that one can compute products of real numbers by adding logarithms of real numbers, and thus the operation of multiplication can be replaced by the simpler operation of addition. The *slide rule* is an analog computing device built on this principle, that used to be sign of the engineer visible in the waist-pocket, recall Fig. 1.5. Today the modern computer has replaced the slide rule and does not use logarithms to multiply real numbers. However, the first computer, the mechanical Difference Machine by Babbage from the 1830s, see Fig. 1.2, was used for computing accurate tables of values of the logarithm. The logarithm was discovered by John Napier and presented in *Mirifici logarithmorum canonis descriptio* in 1614. A illuminating citation from the foreword is given in Fig. 29.2.

Fig. 29.2. Napier, Inventor of the Logarithm: "Seeing there is nothing (right well-beloved Students of the Mathematics) that is so troublesome to mathematical practice, nor that doth more molest and hinder calculators, than the multiplications, divisions, square and cubical extractions of great numbers, which besides the tedious expense of time are for the most part subject to many slippery errors, I began therefore to consider in my mind by what certain and ready art I might remove those hindrances. And having thought upon many things to this purpose, I found at length some excellent brief rules to be treated of (perhaps) hereafter. But amongst all, none more profitable than this which together with the hard and tedious multiplications, divisions, and extractions of roots, doth also cast away from the work itself even the very numbers themselves that are to be multiplied, divided and resolved into roots, and putteth other numbers in their place which perform as much as they can do, only by addition and subtraction, division by two or division by three"

29.3 Important Properties of $\log(x)$

We now derive the basic properties of the logarithm function $\log(x)$ using (29.1) or (29.2). We first note that $u(x) = \log(x)$ is *strictly increasing* for $x > 0$, because $u'(x) = 1/x$ is positive for $x > 0$. This can be seen in Fig. 29.1. Recalling (29.1), we conclude that for $a, b > 0$,

$$\int_a^b \frac{dy}{y} = \log(b) - \log(a).$$

Next we note that the Chain rule implies that for any constant $a > 0$

$$\frac{d}{dx}(\log(ax) - \log(x)) = \frac{1}{ax} \cdot a - \frac{1}{x} = 0,$$

and consequently $\log(ax) - \log(x)$ is constant for $x > 0$. Since $\log(1) = 0$, we see by setting $x = 1$ that the constant value is equal to $\log(a)$, and so

$$\log(ax) - \log(x) = \log(a) \quad \text{for } x > 0.$$

Choosing $x = b > 0$, we thus obtain the following fundamental relation satisfied by the logarithm $\log(x)$:

$$\log(ab) = \log(a) + \log(b) \quad \text{for } a, b > 0 \tag{29.3}$$

We can thus compute the logarithm of the product of two numbers by adding the logarithms of the two numbers. We already indicated that this is the principle of the slide rule or using a table of logarithms for multiplying two numbers. More precisely (as first proposed by Napier), to multiply two numbers a and b we first find their logarithms $\log(a)$ and $\log(b)$ from the table, then add them to get $\log(ab)$ using the formula (29.3), and finally we find from the table which real number has the logarithm equal to $\log(ab)$, which is equal to the desired product ab. Clever, right?

The formula (29.3) implies many things. For example, choosing $b = 1/a$, we get

$$\log(a^{-1}) = -\log(a) \quad \text{for } a > 0. \tag{29.4}$$

Choosing $b = a^{n-1}$ with $n = 1, 2, 3, \ldots$, we get

$$\log(a^n) = \log(a) + \log(a^{n-1}),$$

so that by repeating this argument

$$\log(a^n) = n\log(a) \quad \text{for } n = 1, 2, 3, \ldots \tag{29.5}$$

By (29.4) the last equality holds also for $n = -1, -2, \ldots$

More generally, we have for any $r \in \mathbb{R}$ and $a > 0$,

$$\log(a^r) = r\log(a). \tag{29.6}$$

We prove this using the change of variables $x = y^r$ with $dx = ry^{r-1}dy$:

$$\log(a^r) = \int_1^{a^r} \frac{1}{x}\,dx = \int_1^a \frac{ry^{r-1}}{y^r}\,dy = r\int_1^a \frac{1}{y}\,dy = r\log(a).$$

Finally, we note that $1/x$ also has a primitive function for $x < 0$ and for $a, b > 0$, setting $y = -x$,

$$\begin{aligned}\int_{-a}^{-b} \frac{dy}{y} = \int_a^b \frac{-dx}{-x} = \int_a^b \frac{dx}{x} &= \log(b) - \log(a)\\ &= \log(-(-b)) - \log(-(-a)).\end{aligned}$$

Accordingly, we may write for any $a \neq 0$ and $b \neq 0$ that have the same sign,

$$\int_a^b \frac{dx}{x} = \log(|b|) - \log(|a|). \tag{29.7}$$

It is important to understand that (29.7) does *not* hold if a and b have opposite signs.

Chapter 29 Problems

29.1. Prove that $\log(4) > 1$ and $\log(2) \geq 1/2$.

29.2. Prove that

$$\log(x) \to \infty \quad \text{as } x \to \infty,$$
$$\log(x) \to -\infty \quad \text{as } x \to 0^+.$$

Hint: Using that $\log(2) \geq 1/2$ it follows from (29.5) that $\log(2^n)$ tends to infinity as n tends to infinity.

29.3. Give an alternative proof of (29.3) using that

$$\log(ab) = \int_1^{ab} \frac{1}{y}dy = \int_1^a \frac{1}{y}dy + \int_a^{ab} \frac{1}{y}dy = \log(a) + \int_a^{ab} \frac{1}{y}dy,$$

and changing the variable y in the last integral to $z = ay$.

29.4. Prove that $\log(1+x) \leq x$ for $x > 0$, and that $\log(1+x) < x$ for $x \neq 0$ and $x > -1$. Hint: Differentiate.

29.5. Show using the Mean Value theorem, that $\log(1+x) \leq x$ for $x > -1$. Can prove this directly from the definition of the logarithm by sketching the area under the graph?

29.6. Prove that $\log(a) - \log(b) = \log\left(\frac{a}{b}\right)$ for $a, b > 0$.

29.7. Write down the Taylor polynomial of order n for $\log(x)$ at $x = 1$.

29.8. Find a primitive function of $\frac{1}{x^2-1}$. Hint: use that $\frac{1}{x^2-1} = \frac{1}{(x-1)(x+1)} = \frac{1}{2}\left(\frac{1}{x-1} - \frac{1}{x+1}\right)$.

29.9. Prove that $\log(x^r) = r\log(x)$ for $r = \frac{p}{q}$ rational by using (29.5) cleverly.

29.10. Solve the initial value problem $u'(x) = 1/x^a$ for $x > 0$, $u(1) = 0$, for values of the exponent a close to 1. Plot the solutions. Study for which values of a the solution $u(x)$ tends to infinity when x tends to infinity.

29.11. Solve the following equations: (a) $\log(x^2) + \log(3) = \log(\sqrt{x}) + \log(5)$, (b) $\log(7x) - 2\log(x) = \log(3)$, (c) $\log(x^3) - \log(x) = \log(7) - \log(x^2)$.

29.12. Compute the derivatives of the following functions: a) $f(x) = \log(x^3+6x)$, b) $f(x) = \log(\log(x))$, c) $f(x) = \log(x + x^2)$, d) $f(x) = \log(1/x)$, e) $f(x) = x\log(x) - x$.

30
Numerical Quadrature

> "And I know it *seems* easy", said Piglet to himself, "but it isn't *everyone* who could do it". (House at the Pooh Corner, Milne)
>
> Errare humanum est.

30.1 Computing Integrals

In some cases, we can compute a primitive function (or antiderivative or integral) of a given function analytically, that is we can give give a formula for the primitive function in terms of known functions. For example we can give a formula for a primitive function of a polynomial as another polynomial. We will return in Chapter *Techniques of integration* to the question of finding analytical formulas for primitive functions of certain classes of functions. The Fundamental Theorem states that any given Lipschitz continuous function has a primitive function, but does not give any analytical formula for the primitive function. The logarithm,

$$\log(x) = \int_1^x \frac{dy}{y}, \quad \text{where } x > 0,$$

is the first example of this case we have encountered. We know that the logarithm function $\log(x)$ exists for $x > 0$, and we have derived some of its properties indirectly through its defining differential equation, but the question remains how to determine the value of $\log(x)$ for a given $x > 0$. Once we have solved this problem, we may add $\log(x)$ to a list of "elementary" functions that we can play with. Below we will add to this list the

exponential function, the trigonometric functions, and other more exotic functions.

This situation is completely analogous to solving algebraic equations for numbers. Some equations have rational roots and in that case, we feel that we can solve the equation "exactly" by analytic (symbolic) computation. We have a good understanding of rational numbers, even when they have infinite decimal expansions, and we can determine their values, or the pattern in the decimal expansion, with a finite number of arithmetic operations. But most equations have irrational roots with infinite, non-repeating decimal expansions that we can only approximate to a desired level of accuracy in practice. Likewise in a situation in which a function is known only as a primitive function of a given function, the best we can do is to seek to compute its values approximately to a desired level of accuracy. One way to compute values of such a function is through the definition of the integral as a Riemann sum. This is known as *numerical quadrature* or *numerical integration*, and we now explore this possibility.

Suppose thus that we want to compute the integral

$$\int_a^b f(x)\,dx, \tag{30.1}$$

where $f : [a, b] \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant L_f. If we can give a formula for the primitive function $F(x)$ of $f(x)$, then the integral is simply $F(b) - F(a)$. If we cannot give a formula for $F(x)$, then we turn to the Fundamental Theorem and compute an approximate the value of the integral using a Riemann sum approximation

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{N} f(x_{i-1}^n)h_n, \tag{30.2}$$

where $x_i^n = a + ih_n$, $h_n = 2^{-n}(b - a)$, and $N = 2^n$ describes a uniform partition of $[a, b]$, with the *quadrature error*

$$Q_n = \Big| \int_a^b f(x)\,dx - \sum_{i=1}^{N} f(x_{i-1}^n)h_n \Big| \leq \frac{b-a}{2} L_f h_n, \tag{30.3}$$

which tends to zero as we increase the number of steps and $h_n \to 0$. Put another way, if we desire the value of the integral to within a *tolerance* $TOL > 0$ and we know the Lipschitz constant L_f, then we will have $Q_n \leq TOL$ if the mesh size h_n satisfies the *stopping criterion*

$$h_n \leq \frac{2\,TOL}{(b-a)L_f}. \tag{30.4}$$

We refer to the Riemann sum approximation (30.2), compare also with Fig. 30.1, as the *rectangle rule*, which is the simplest method for approximating an integral among many possible methods. The search for more

sophisticated methods for approximating an integral is driven by consideration of the *computational cost* associated to computing the approximation. The cost is typically measured in terms of time because there is a limit on the time we are willing to wait for a solution. In the rectangle rule, the computer spends most of the time evaluating the function f and since each step requires one evaluation of f, the cost is determined ultimately by the number of steps. Considering the cost leads to the optimization problem of trying to compute an approximation of a given accuracy at a relatively low cost.

To reduce the cost, we may construct more sophisticated methods for approximating integrals. But even if we restrict ourselves to the rectangle rule, we can introduce variations that can lower the computational cost of computing an approximation. There are two quantities that we can vary: the point at which we evaluate the function on each interval and the size of the intervals. To understand how these changes could help, consider the illustration of the rectangle rule in Fig. 30.1. Here f varies quite a bit on part of $[a, b]$ and is fairly constant on another part. Consider the approximation to the area under f on the first subinterval on the left. By evaluating f at the left-hand point on the subinterval, we clearly overestimate the area to the maximum degree possible. Choosing to evaluate f at some point inside the subinterval would likely give a better approximation. The same is true of the second subinterval, where choosing the left-hand point clearly leads to an underestimate of the area. On the other hand, consider the approximations to the area in the last four subintervals on the right. Here f is nearly constant and the approximation is very accurate. In fact, we could approximate the area underneath f on this part of $[a, b]$ using one rectangle rather than four. In other words, we would get just as accurate an approximation by using one large subinterval instead of four subintervals. This would cost four times less.

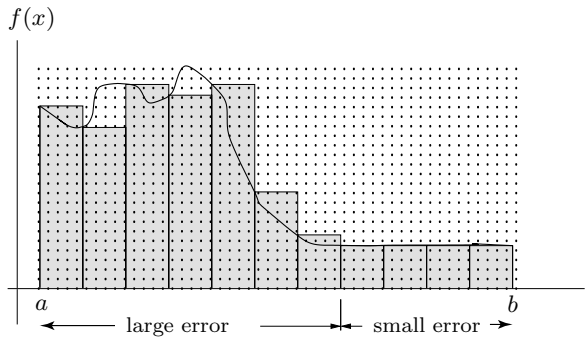

Fig. 30.1. An illustration of the rectangle rule

So we generalize the rectangle rule to allow non-uniform partitions and different points at which to evaluate f. We choose a partition $a = x_0 < x_1 < x_2 \cdots < x_N = b$ of $[a, b]$ into N subintervals $I_j = [x_{j-1}, x_j]$ of lengths $h_j = x_j - x_{j-1}$ for $j = 1, \ldots, N$. Note that N can be any integer and the subintervals may vary in size. By the Mean Value theorem for integrals there is $\bar{x}_j \in I_j$ such that

$$\int_{x_{j-1}}^{x_j} f(x)\,dx = f(\bar{x}_j)h_j, \tag{30.5}$$

and thus we have

$$\int_a^b f(x)\,dx = \sum_{i=1}^{N} \int_{x_{j-1}}^{x_j} f(x)\,dx = \sum_{j=1}^{N} f(\bar{x}_j)h_j.$$

Since the $\bar{x}_j$ are not known in general, we replace $\bar{x}_j$ by a given point $\hat{x}_j \in I_j$. For example, in the original method we use the left end-point $\hat{x}_j = x_{j-1}$, but we could choose the right end-point $\hat{x}_j = x_j$ or the mid-point $\hat{x}_j = \frac{1}{2}(x_{j-1} + x_j)$. We then get the approximation

$$\int_a^b f(x)\,dx \approx \sum_{j=1}^{N} f(\hat{x}_j)h_j. \tag{30.6}$$

We call

$$\sum_{j=1}^{N} f(\hat{x}_j)h_j \tag{30.7}$$

a *quadrature* formula for computing the integral $\int_a^b f(x)\,dx$. We recall that we refer to (30.7) as a *Riemann sum*. The quadrature formula is characterized by the *quadrature points* $\hat{x}_j$ and the *weights* h_j. Note that if $f(x) = 1$ for all x then the quadrature formula is exact and we conclude that $\sum_{j=1}^{N} h_j = b - a$.

We now estimate the *quadrature error*

$$Q_h = \Big| \int_a^b f(x)\,dx - \sum_{j=1}^{N} f(\hat{x}_j)h_j \Big|,$$

where the subscript h refers to the sequence of step sizes h_j. Recalling (30.5), we can do this by estimating the error over each subinterval and then summing. We have

$$\Big| \int_{x_{j-1}}^{x_j} f(x)\,dx - f(\hat{x}_j)h_j \Big| = |f(\bar{x}_j)h_j - f(\hat{x}_j)h_j| = h_j|f(\bar{x}_j) - f(\hat{x}_j)|.$$

We assume that $f'(x)$ is Lipschitz continuous on $[a,b]$. The Mean Value theorem implies that for $x \in [x_{j-1}, x_j]$,

$$f(x) = f(\hat{x}_j) + f'(y)(x - \hat{x}_j),$$

for some $y \in [x_{j-1}, x_j]$. Integrating over $[x_{j-1}, x_j]$ we obtain

$$\Big| \int_{x_{j-1}}^{x_j} f(x)\,dx - f(\hat{x}_j)h_j \Big| \leq \max_{y \in I_j} |f'(y)| \int_{x_{j-1}}^{x_j} |x - \hat{x}_j|\,dx.$$

To simplify the sum on the right, we use the fact that

$$\int_{x_{j-1}}^{x_j} |x - \hat{x}_j|\,dx$$

is maximized if $\hat{x}_j$ is the left (or right) endpoint, in which case

$$\int_{x_{j-1}}^{x_j} (x - x_{j-1})\,dx = \frac{1}{2}h_j^2.$$

We find that

$$\Big| \int_{x_{j-1}}^{x_j} f(x)\,dx - f(\hat{x}_j)h_j \Big| \leq \frac{1}{2} \max_{y \in I_j} |f'(y)| h_j^2.$$

Summing, we conclude

$$Q_h = \Big| \int_a^b f(x)\,dx - \sum_{j=1}^{N} f(\hat{x}_j)h_j \Big| \leq \frac{1}{2} \sum_{j=1}^{N} \left(\max_{y \in I_j} |f'(y)|\, h_j \right) h_j. \quad (30.8)$$

This generalizes the estimate of the Fundamental Theorem to non-uniform partitions. We can see that (30.8) implies that Q_h tends to zero as the maximal step size tends to zero by estimating further:

$$Q_h \leq \frac{1}{2} \max_{[a,b]} |f'| \sum_{j=1}^{N} h_j \max_{1 \leq j \leq N} h_j = \frac{1}{2}(b-a) \max_{[a,b]} |f'| \max_{1 \leq j \leq N} h_j. \quad (30.9)$$

So Q_h tends to zero at the same rate that $\max h_j$ tends to zero.

30.2 The Integral as a Limit of Riemann Sums

We now return to the (subtle) question posed at the end of the Chapter The Integral: Will all limits of Riemann sum approximations (as the maximal subinterval tends to zero) of a certain integral be the same? We recall that we defined the integral using a particular uniform partition and we now

ask if any limit of non-uniform partitions will be the same. The affirmative answer follows from the last statement of the previous section: The quadrature error Q_h tends to zero as $\max h_j$ tends to zero, under the assumption that $\max_{[a,b]} |f'|$ is finite, that is $|f'(x)|$ is bounded on $[a,b]$. This proves the uniqueness of limits of Riemann sum approximations of a certain integral as the maximal subinterval tends to zero, under the assumption that the derivative of the integrand is bounded. This assumption can naturally be relaxed to assuming that the integrand is Lipschitz continuous. We sum up:

Theorem 30.1 *The limit (as the maximal subinterval tends to zero) of Riemann sum approximations of an integral of a Lipschitz continuous function, is unique.*

30.3 The Midpoint Rule

We now analyze the quadrature formula in which the quadrature point is chosen to be the midpoint of each subinterval, $\hat{x}_j = \frac{1}{2}(x_{j-1} + x_j)$. It turns out that this choice gives a formula that is more accurate than any other rectangle rule on a given mesh provided f has a Lipschitz continuous second derivative. Taylor's theorem implies that for $x \in [x_{j-1}, x_j]$,

$$f(x) = f(\hat{x}_j) + f'(\hat{x}_j)(x - \hat{x}_j) + \frac{1}{2}f''(y)(x - \hat{x}_j)^2,$$

for some $y \in [x_{j-1}, x_j]$ if we assume that f'' is Lipschitz continuous. We argue as above by integrating over $[x_{j-1}, x_j]$. Now however we use the fact that

$$\int_{x_{j-1}}^{x_j} (x - \hat{x}_j)\, dx = \int_{x_{j-1}}^{x_j} (x - (x_j + x_{j-1})/2)\, dx = 0$$

which holds only when $\hat{x}_j$ is the midpoint of $[x_{j-1}, x_j]$. This gives

$$\begin{aligned}\Big| \int_{x_{j-1}}^{x_j} f(x)\, dx - f(\hat{x}_j)h_j \Big| &\leq \frac{1}{2} \max_{y \in I_j} |f''(y)| \int_{x_{j-1}}^{x_j} (x - \hat{x}_j)^2\, dx \\ &\leq \frac{1}{24} \max_{y \in I_j} |f''(y)| h_j^3.\end{aligned}$$

Now summing the errors on each subinterval, we obtain the following estimate on the total error

$$Q_h \leq \frac{1}{24} \sum_{j=1}^{N} \left(\max_{y \in I_j} |f''(y)|\, h_j^2 \right) h_j. \tag{30.10}$$

To understand the claim that this formula is more accurate than any other rectangle rule, we estimate further

$$Q_h \leq \frac{1}{24}(b-a) \max_{[a,b]} |f''| \max h_j^2$$

which says that the error decreases as $\max h_j$ decreases like $\max h_j^2$. Compare this to the general result (30.9), which says that the error decreases like $\max h_j$ for general rectangle rules. If we halve the step size $\max h_j$ then in a general rectangle rule the error decreases by a factor of two but in the midpoint rule the error decreases by a factor of *four*. We say that the midpoint rule converges at a *quadratic* rate while the general rectangle rule converges at a *linear* rate.

We illustrate the accuracy of these methods and the error bounds by approximating

$$\log(4) = \int_1^4 \frac{dx}{x} \approx \sum_{j=1}^{N} \frac{h_j}{\hat{x}_j}$$

both with the original rectangle rule with $\hat{x}_j$ equal to the left-hand endpoint x_{j-1} of each subinterval and the midpoint rule. In both cases, we use a constant stepsize $h_i = (4-1)/N$ for $i = 1, 2, \cdots, N$. It is straightforward to evaluate (30.8) and (30.10) because $|f'(x)| = 1/x^2$ and $|f''(x)| = 2/x^3$ are both decreasing functions. We show the results for four different values of N.

	Rectangle rule		Midpoint rule	
N	True error	Error bound	True error	Error bound
25	.046	.049	.00056	.00056
50	.023	.023	.00014	.00014
100	.011	.011	.000035	.000035
200	.0056	.0057	.0000088	.0000088

These results show that the error bounds (30.8) and (30.10) can give quite accurate estimates of the true error. Also note that the midpoint rule is much more accurate than the general rectangle rule on a given mesh and moreover the error in the midpoint rule goes to zero quadratically with the error decreasing by a factor of 4 each time the number of steps is doubled.

30.4 Adaptive Quadrature

In this section, we consider the optimization problem of trying to compute an approximation of an integral to within a given accuracy at a relatively low cost. To simplify the discussion, we use the original rectangle rule with $\hat{x}_j$ equal to the left-hand endpoint x_{j-1} of each subinterval to compute the approximation. The optimization problem becomes to compute an approximation with error smaller than a given tolerance TOL using the least number of steps. Since we do not know the error of the approximation, we use the quadrature estimate (30.8) to estimate the error. The optimization problem is therefore to find a partition $\{x_j\}_{j=0}^N$ using the smallest

number of points N that satisfies the stopping criterion

$$\sum_{j=1}^{N} \left(\max_{y \in I_j} |f'(y)| \, h_j \right) h_j \leq \text{TOL}. \tag{30.11}$$

This equation suggests that we should adjust or *adapt* the stepsizes h_j depending on the size of $\max_{I_j} |f'|$. If $\max_{I_j} |f'|$ is large, then h_j should be small, and vice versa. Trying to find such an optimized partition is referred to as *adaptive* quadrature, because we seek a partition suitably adapted to the nature of the integrand $f(x)$.

There are several possible strategies for finding such a partition and we consider two here.

In the first strategy, or adaptive algorithm, we estimate the sum in (30.11) as follows

$$\sum_{j=1}^{N} \left(\max_{I_j} |f'| \, h_j \right) h_j \leq (b-a) \max_{1 \leq j \leq N} \left(\max_{I_j} |f'| \, h_j \right),$$

where we use the fact that $\sum_{j=1}^{N} h_j = b - a$. It follows that (30.11) is satisfied if the steps are chosen by

$$h_j = \frac{\text{TOL}}{(b-a) \max_{I_j} |f'|} \quad \text{for } j = 1, \ldots, N. \tag{30.12}$$

In general, this corresponds to a nonlinear equation for h_j since $\max_{I_j} |f'|$ depends on h_j.

We apply this adaptive algorithm to the computation of $\log(b)$ and obtain the following results

TOL	b	Steps	Approximate Area	Error
.05	4.077	24	1.36	.046
.005	3.98	226	1.376	.0049
.0005	3.998	2251	1.38528	.0005
.00005	3.9998	22501	1.3861928	.00005

The reason b varies slightly in these results is due to the strategy we use to implement (30.12). Namely, we specify the tolerance and then search for the value of N that gives the closest b to 4.

We plot the sequence of mesh sizes for TOL = .01 in Fig. 30.2, where the adaptivity is plainly visible. In contrast, if we compute with a uniform mesh, we find using (30.11) that we need $N = 9/\text{TOL}$ points to guarantee an accuracy of TOL. For example, this means using 900 points to guarantee an accuracy of .01, which is significantly more than needed for the adapted mesh.

The second adaptive algorithm is based on an *equidistribution of error* in which the steps h_j are chosen so that the contribution to the error from each sub-interval is roughly equal. Intuitively, this should lead to the least number of intervals since the largest error reduction is gained if we subdivide the interval with largest contribution to the error. In this case, we estimate the sum on the left-hand side of (30.11) by

$$\sum_{j=1}^{N} \left(\max_{I_j} |f'| \, h_j \right) h_j \leq N \max_{1 \leq j \leq N} \left(\max_{I_j} |f'| \, h_j^2 \right)$$

and determine the steps h_j by

$$h_j^2 = \frac{\text{TOL}}{N \max_{I_j} |f'|} \quad \text{for } j = 1, \ldots, N. \tag{30.13}$$

As above, we have to solve a nonlinear equation for h_j, now with the additional complication of the explicit presence of the total number of steps N.

We implement (30.13) to compute of $\log(b)$ with $b \approx 4$ and obtain the following results:

TOL	b	Steps	Appr. Area	Error
.05	4.061	21	1.36	.046
.005	4.0063	194	1.383	.005
.0005	3.9997	1923	1.3857	.0005
.00005	4.00007	19220	1.38626	.00005

We plot the sequence of step sizes for TOL = .01 in (30.2). We see that at every tolerance level, the second adaptive strategy (30.13) gives the same accuracy at $x_N \approx 4$ as (30.12) while using fewer steps. It thus seems that the second algorithm is more efficient.

We compare the efficiency of the two adaptive algorithms by estimating the number of steps N required to compute $\log(x)$ to a given tolerance TOL in each case. We begin by noting that the equality

$$N = \frac{h_1}{h_1} + \frac{h_2}{h_2} + \cdots + \frac{h_N}{h_N},$$

implies that, assuming $x_N > 1$,

$$N = \int_1^{x_N} \frac{dy}{h(y)},$$

where $h(y)$ is the piecewise constant *mesh function* with the value $h(s) = h_j$ for $x_{j-1} < s \leq x_j$. In the case of the second algorithm, we substitute the value of h given by (30.13) into the integral to get, recalling that $f(y) = 1/y$ so that $f'(y) = -1/y^2$,

$$N \approx \frac{\sqrt{N}}{\sqrt{\text{TOL}}} \int_1^{x_N} \frac{dy}{y},$$

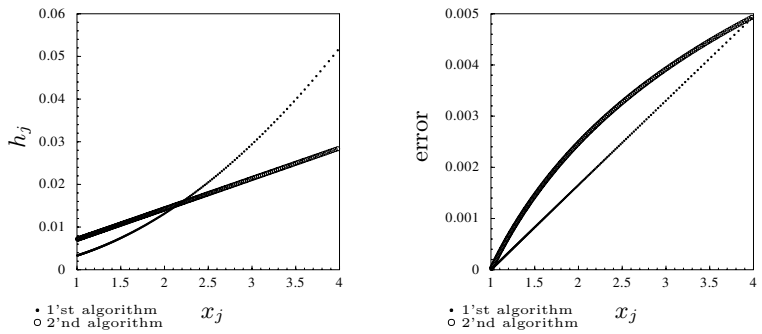

Fig. 30.2. On the *left*, we plot the step sizes generated by two adaptive algorithms for the integration of log(4) using TOL = .01. On the *right*, we plot the errors of the same computations versus x

or

$$N \approx \frac{1}{\text{TOL}} \left(\log(x_N)\right)^2 . \tag{30.14}$$

Making a similar analysis of the first adaptive algorithm, we get

$$N \approx \frac{x_{N-1}}{\text{TOL}} \left(1 - \frac{1}{x_N}\right). \tag{30.15}$$

We see that in both cases, N is inversely proportional to TOL. However, the number of steps needed to reach the desired accuracy using the first adaptive algorithm increases much more quickly as x_N increases than the number needed by the second algorithm, i.e. at a linear rate as opposed to a logarithmic rate. Note that the case $0 < x_N < 1$ may be reduced to the case $x_N > 1$ by replacing x_N by $1/x_N$ since $\log(x) = -\log(1/x)$.

If we use (30.12) or (30.13) to choose the steps h_j over the interval $[a, x_N]$, then of course the quadrature error over any smaller interval $[a, x_i]$ with $i \leq N$, is also smaller than TOL. For the first algorithm (30.12), we can actually show the stronger estimate

$$\left| \int_a^{x_i} f(y)\, dy - \sum_{j=1}^{i} f(x_j)\, h_j \right| \leq \frac{x_i - a}{x_N - a}\, TOL, \quad 1 \leq i \leq N, \tag{30.16}$$

i.e., the error grows at most linearly with x_i as i increases. However, this does not hold in general for the second adaptive algorithm. In Fig. 30.2, we plot the errors versus x_i for $x_i \leq x_N$ resulting from the two adaptive algorithms with TOL = .01. We see the linear growth predicted for the first algorithm (30.12) while the error from the second algorithm (30.13) is larger for $1 < x_i < x_N$.

Chapter 30 Problems

30.1. Estimate the error using endpoint and midpoint quadrature for the following integrals: (a) $\int_0^2 2s\,ds$, (b) $\int_0^2 s^3\,ds$, and (c) $\int_0^2 \exp(-s)\,ds$ using $h = .1$, .01, .001 and .0001. Discuss the results.

30.2. Compute approximations of the following integrals using adaptive quadrature (a) $\int_0^2 2s\,ds$, (b) $\int_0^2 s^3\,ds$, and (c) $\int_0^2 \exp(-s)\,ds$. Discuss the results.

30.3. Compare theoretically and experimentally the number of steps of (30.12) and (30.13) for the computation of integrals of the form $\int_x^1 f(y)\,dy$ for $x > 0$, where $f(y) \sim y^{-\alpha}$ with $\alpha > 1$.

30.4. The *trapezoidal rule* takes the form

$$\int_{x_{j-1}}^{x_j} f(x)dx \approx (x_j - x_{j-1})(f(x_{j-1}) + f(x_j))/2. \tag{30.17}$$

Show that the quadrature is exact if $f(x)$ is a first order polynomial, and give an estimate of the quadrature error analogous to that of the midpoint rule. Compare the the midpoint and the trapezoidal method.

30.5. Design different adaptive quadrature algorithms based on the midpoint rule and make comparisons.

30.6. Consider a quadrature formula of the form

$$\int_a^b f(x)dx \approx (b-a)(f(\hat{x}_1) + f(\hat{x}_2))/2. \tag{30.18}$$

Determine the quadrature points $\hat{x}_1$ and $\hat{x}_2$, so that the quadrature formula is exact for $f(x)$ a second order polynomial. This quadrature rule is called the two-point Gauss rule. Check for which order of polynomials the resulting quadrature formula is exact.

30.7. Compute the value of $\displaystyle\int_0^1 \frac{1}{1+x^2}dx$ by quadrature. Multiply the result by 4. Do you recognize this number?

31
The Exponential Function $\exp(x) = e^x$

> The need for mathematical skills is greater than ever, but it is widely recognized that, as a consequence of computer developments, there is a need for a shift in emphasis in the teaching of mathematics to students studying engineering. This shift is away from the simple mastery of solution techniques and towards development of a greater understanding of mathematical ideas and processes together with efficiency in applying this understanding to the formulation and analysis of physical phenomena and engineering systems. (Glyn James, in Preface to Modern Engineering Mathematics, 1992)
>
> Because of the limitations of human imagination, one ought to say: everything is possible - and a bit more. (Horace Engdahl)

31.1 Introduction

In this chapter we return to study of the *exponential function* $\exp(x)$, which we have met above in Chapter *A very short course in Calculus* and Chapter *Galileo, Newton, Hooke, Malthus and Fourier*, and which is one of the basic functions of Calculus, see Fig. 31.1. We have said that $\exp(x)$ for $x > 0$ is the solution to the following initial value problem: Find a function $u(x)$ such that

$$\begin{aligned} u'(x) &= u(x) \quad \text{for } x > 0, \\ u(0) &= 1. \end{aligned} \tag{31.1}$$

Evidently, the equation $u'(x) = u(x)$ states that the rate of growth $u'(x)$ is equal to the quantity $u(x)$ itself, that is, the exponential function $\exp(x)=e^x$

is characterized by the property that its derivative is equal to itself: $D\exp(x) = \exp(x)$. What a wonderful almost divine property! We also denote the exponential function by e^x, that is, $e^x = \exp(x)$ and $De^x = e^x$.

In this chapter, we give a constructive proof of the *existence* of a unique solution to the initial value problem (31.1), that is, we *prove* the existence of the exponential function $\exp(x) = e^x$ for $x > 0$. Note that above, we just *claimed* the existence of solutions. As always, a constructive proof also shows how we may *compute* $\exp(x)$ for different values of x.

Below we extend $\exp(x)$ to $x < 0$ by setting $\exp(x) = (\exp(-x))^{-1}$ for $x < 0$, and show that $\exp(-x)$ solves the initial value problem $u'(x) = -u(x)$ for $x > 0$, $u(0) = 1$. We plot the functions $\exp(x)$ and $\exp(-x)$ for $x \geq 0$ in Fig. 31.1. We notice that $\exp(x)$ is increasing and $\exp(-x)$ is decreasing with increasing x, and that $\exp(x)$ is positive for all x. Combining $\exp(x)$ and $\exp(-x)$ for $x \geq 0$ defines $\exp(x)$ for $-\infty < x < \infty$. Below we show that $D\exp(x) = \exp(x)$ for $-\infty < x < \infty$.

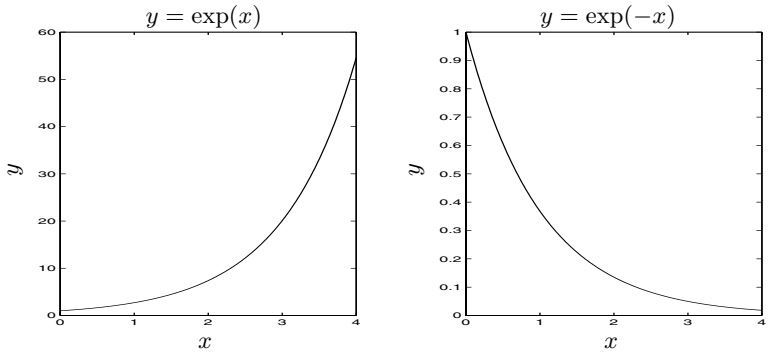

Fig. 31.1. The exponential functions $\exp(x)$ and $\exp(-x)$ for $x \geq 0$

The problem (31.1) is a special case of the Malthus population model (26.17), which also models a large variety of phenomena in e.g. physics and economy:

$$\begin{cases} u'(x) = \lambda u(x) & \text{for } x > 0, \\ u(0) = u_0. \end{cases} \tag{31.2}$$

where λ is a constant and u_0 is a given initial value. The solution of this problem can be expressed in terms of the exponential function as

$$u(x) = \exp(\lambda x)u_0 \quad \text{for } x \geq 0. \tag{31.3}$$

This follows directly from the fact that by the Chain rule, $D\exp(\lambda x) = \exp(\lambda x)\lambda$, where we used that $D\exp(x) = \exp(x)$. Assuming $u_0 > 0$ so that $u(x) > 0$, evidently the sign of λ determines if u decreases ($\lambda < 0$) or increases ($\lambda > 0$). In Fig. 31.1, we plotted the solutions of (31.2) with $\lambda = \pm 1$ and $u_0 = 1$.

Before going into the construction of the exponential function exp(x), we recall two of the key applications of (31.2): population dynamics and banking. Here x represents time and we change notation, replacing x by t.

Example 31.1. We consider a population with constant birth and death rates β and δ, which are the numbers of births and deaths per individual creature per unit time. With $u(t)$ denoting the population at time t, there will be during the time interval from t to $t+\Delta t$ with Δt a small increment, approximately $\beta u(t)\Delta t$ births and $\delta u(t)\Delta t$ deaths. Hence the change in population over the time interval is approximately

$$u(t+\Delta t) - u(t) \approx \beta u(t)\Delta t - \delta u(t)\Delta t$$

and therefore

$$\frac{u(t+\Delta t) - u(t)}{\Delta t} \approx (\beta - \delta)u(t),$$

where the approximation improves as we decrease Δt. Taking the limit as $\Delta t \to 0$, assuming $u(t)$ is a differentiable function, we obtain the model $u'(t) = (\beta - \delta)u(t)$. Assuming the initial population at $t = 0$ is equal to u_0, leads to the model (31.2) with $\lambda = \beta - \delta$, with solution $u(x) = \exp(\lambda x)u_0$.

Example 31.2. An investment u in a saving account earning 5% interest compounded continuously and beginning with \$2000 at time $t = 0$, satisfies

$$\begin{cases} u' = 1.05u, & t > 0, \\ u(0) = 2000, \end{cases}$$

and thus $u(t) = \exp(1.05t)2000$ for $t \geq 0$.

31.2 Construction of the Exponential exp(x) for $x \geq 0$

In the proof of the Fundamental Theorem, we constructed the solution $u(x)$ of the initial value problem

$$\begin{cases} u'(x) = f(u(x), x) & \text{for } 0 < x \leq 1, \\ u(0) = u_0, \end{cases} \tag{31.4}$$

in the case that $f(u(x), x) = f(x)$ depends only on x and not on $u(x)$. We constructed the solution $u(x)$ as the limit of a sequence of functions $\{U^n(x)\}_{n=1}^{\infty}$, where $U^n(x)$ is a piecewise linear function defined at a set of nodes $x_i^n = ih_n$, $i = 0, 1, 2, \ldots, N = 2^n$, $h_n = 2^{-n}$, by the relations

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n f(x_{i-1}^n) \quad \text{for } i = 1, 2, \ldots, N, \quad U^n(0) = u_0. \tag{31.5}$$

We shall now apply the same technique to construct the solution of (31.1), which has the form (31.4) with $f(u(x), x) = u(x)$ and $u_0 = 1$. We carry out the proof in a form which generalizes in a straight-forward way to any system of equations of the form (31.4), which really includes a very wide range of problems. We hope this will motivate the reader to carefully follow every step of the proof, to get properly prepared for the highlight Chapter *The general initial value problem*.

We construct the solution $u(x)$ of (31.1) for $x \in [0, 1]$ as the limit of a sequence of piecewise linear functions $\{U^n(x)\}_{n=1}^{\infty}$ defined at the nodes by the formula

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n U^n(x_{i-1}^n) \quad \text{for } i = 1, 2, \ldots, N, \tag{31.6}$$

with $U^n(0) = 1$, which is an analog of (31.5) obtained by replacing $f(x_{i-1}^n)$ by $U^n(x_{i-1}^n)$ corresponding to replacing $f(x)$ by $f(x, u(x)) = u(x)$. Using the formula we can compute the values $U^n(x_i^n)$ one after the other for $i = 1, 2, 3, \ldots$, starting from the initial value $U^n(0) = 1$, that is marching forward in time with x representing time.

We can write (31.6) in the form

$$U^n(x_i^n) = (1 + h_n)U^n(x_{i-1}^n) \quad \text{for } i = 1, 2, \ldots, N, \tag{31.7}$$

and conclude since $U^n(x_i^n) = (1 + h_n)U^n(x_{i-1}^n) = (1 + h_n)^2 U^n(x_{i-2}^n) = (1 + h_n)^3 U^n(x_{i-3}^n)$ and so on, that the nodal values of $U^n(x)$ are given by the formula

$$U^n(x_i^n) = (1 + h_n)^i, \quad \text{for } i = 0, 1, 2, \ldots, N, \tag{31.8}$$

where we also used that $U^n(0) = 1$. We illustrate in Fig. 31.2. We may view $U^n(x_i^n)$ as the capital obtained at time $x_i^n = ih_n$ starting with a unit capital at time zero, if the interest rate at each capitalization is equal to h_n.

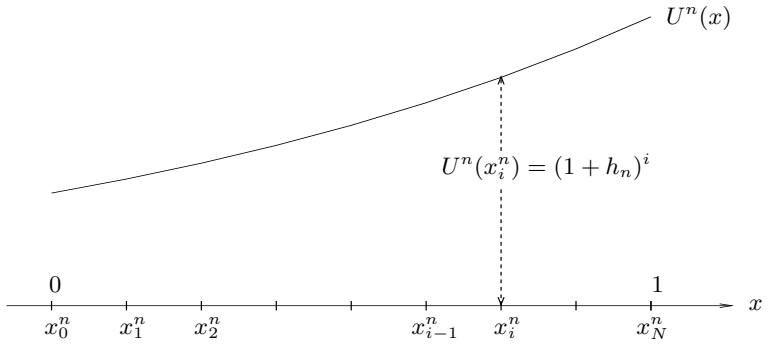

Fig. 31.2. The piecewise linear approximate solution $U^n(x) = (1 + h_n)^i$

To analyze the convergence of $U^n(x)$ as $n \to \infty$, we first prove a bound on the nodal values $U^n(x_i^n)$, by taking the logarithm of (31.8) and using the inequality $\log(1+x) \leq x$ for $x > 0$ from Problem 29.4, to obtain

$$\log(U^n(x_i^n) = i\log(1+h_n) \leq ih_n = x_i^n \leq 1 \quad \text{for } i = 1, 2, \ldots, N.$$

It follows that

$$U^n(x_i^n) = (1+h_n)^i \leq 4 \quad \text{for } i = 1, 2, \ldots, N, \tag{31.9}$$

since $\log(4) > 1$ according to Problem 29.1, and $\log(x)$ is increasing. Since $U^n(x)$ is linear between the nodes, and obviously $U^n(x) \geq 1$, we find that $1 \leq U^n(x) \leq 4$ for all $x \in [0,1]$.

We now show that $\{U^n(x)\}_{n=1}^{\infty}$ is a Cauchy sequence for each fixed $x \in [0,1]$. To see this, we first estimate $|U^n(x) - U^{n+1}(x)|$ at the node points $x = x_i^n = ih_n = 2ih_{n+1} = x_{2i}^{n+1}$ for $i = 0, 1, \ldots, N$, see Fig. 31.3. Notice that $h_{n+1} = h_n/2$ so that two steps with mesh size h_{n+1} corresponds to one step with mesh size h_n. We start by subtracting

$$U^{n+1}(x_{2i}^{n+1}) = (1+h_{n+1})U^{n+1}(x_{2i-1}^{n+1}) = (1+h_{n+1})^2 U^{n+1}(x_{2i-2}^{n+1}),$$

from (31.6), using that $x_i^n = x_{2i}^{n+1}$, and setting $e_i^n = U^n(x_i^n) - U^{n+1}(x_i^n)$, to get

$$e_i^n = (1+h_n)U^n(x_{i-1}^n) - (1+h_{n+1})^2 U^{n+1}(x_{i-1}^n),$$

which we may rewrite using that $(1+h_{n+1})^2 = 1 + 2h_{n+1} + h_{n+1}^2$ and $2h_{n+1} = h_n$, as

$$e_i^n = (1+h_n)e_{i-1}^n - h_{n+1}^2 U^{n+1}(x_{i-1}^n).$$

It follows using the bound $1 \leq U^{n+1}(x) \leq 4$ for $x \in [0,1]$, that

$$|e_i^n| \leq (1+h_n)|e_{i-1}^n| + 4h_{n+1}^2.$$

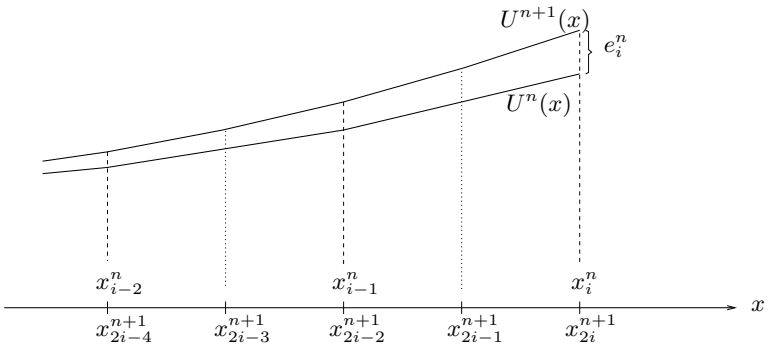

Fig. 31.3. $U^n(x)$ and $U^{n+1}(x)$

Inserting the corresponding estimate for e^n_{i-1}, we get

$$\begin{aligned}|e^n_i| \leq &(1+h_n)\left((1+h_n)|e^n_{i-2}| + 4h^2_{n+1}\right) + 4h^2_{n+1}\\ =&(1+h_n)^2|e^n_{i-2}| + 4h^2_{n+1}\big(1+(1+h_n)\big).\end{aligned}$$

Continuing this way and using that $e^n_0 = 0$, we obtain for $i = 1, \ldots, N$,

$$|e^n_i| \leq 4h^2_{n+1}\sum_{k=0}^{i-1}(1+h_n)^k = h^2_n\sum_{k=0}^{i-1}(1+h_n)^k.$$

Using now the fact that

$$\sum_{k=0}^{i-1} z^k = \frac{z^i - 1}{z-1} \tag{31.10}$$

with $z = 1 + h_n$, we thus obtain for $i = 1, \ldots, N$,

$$|e^n_i| \leq h^2_n\frac{(1+h_n)^i - 1}{h_n} = h_n((1+h_n)^i - 1) \leq 3h_n,$$

where we again used that $(1+h_n)^i = U^n(x^n_i) \leq 4$. We have thus proved that for $\bar{x} = x^n_j$, $j = 1, \ldots, N$,

$$|U^n(\bar{x}) - U^{n+1}(\bar{x})| = |e^n_j| \leq 3h_n,$$

which is analogous to the central estimate (27.24) in the proof of the Fundamental Theorem.

Iterating this estimate over n as in the proof of (27.25), we get for $m > n$,

$$|U^n(\bar{x}) - U^m(\bar{x})| \leq 6h_n, \tag{31.11}$$

which shows that $\{U^n(\bar{x})\}_{n=1}^{\infty}$ is a Cauchy sequence and thus converges to a real number $u(\bar{x})$, which we choose to denote by $\exp(\bar{x}) = e^{\bar{x}}$. As in the proof of the Fundamental Theorem we can extend to a function $u(x) = \exp(x) = e^x$ defined for $x \in [0,1]$. Letting m tend to infinity in (31.11), we see that

$$|U^n(x) - \exp(x)| \leq 6h_n \quad \text{for } x \in [0,1]. \tag{31.12}$$

By the construction, we have if $\bar{x} = jh_n$ so that $h_n = \frac{\bar{x}}{j}$, noting that $j \to \infty$ as $n \to \infty$:

$$\exp(\bar{x}) = \lim_{n\to\infty}(1+h_n)^j = \lim_{j\to\infty}\left(1 + \frac{\bar{x}}{j}\right)^j,$$

that is,

$$\exp(x) = \lim_{j\to\infty}\left(1 + \frac{x}{j}\right)^j \qquad \text{for } x \in [0,1]. \tag{31.13}$$

In particular, we define the number e by

$$e \equiv \exp(1) = \lim_{j\to\infty} \left(1 + \frac{1}{j}\right)^j. \tag{31.14}$$

We refer to e as the *base of the exponential function*. We will prove below that $\log(e) = 1$.

It remains to verify that the function $u(x) = \exp(x) = e^x$ constructed above, indeed satisfies (31.1) for $0 < x \leq 1$. We note that choosing $\bar{x} = jh_n$ and summing over i in (31.6), we get

$$U^n(\bar{x}) = \sum_{i=1}^{j} U^n(x_{i-1}^n)h_n + 1,$$

which we can write as

$$U^n(\bar{x}) = \sum_{i=1}^{j} u(x_{i-1}^n)h_n + 1 + E_n,$$

where $u(x) = \exp(x)$, and using (31.12),

$$|E_n| = \left|\sum_{i=1}^{j}(U^n(x_{i-1}^n) - u(x_{i-1}^n))h_n\right| \leq 6h_n \sum_{i=1}^{j} h_n \leq 6h_n,$$

since obviously $\sum_{i=1}^{j} h_n \leq 1$. Letting n tend to infinity and using $\lim_{n\to\infty} E_n = 0$, we see that $u(\bar{x}) = \exp(\bar{x})$ satisfies

$$u(\bar{x}) = \int_0^{\bar{x}} u(x)\,dx + 1.$$

Differentiating this equality with respect to $\bar{x}$, we get $u'(\bar{x}) = u(\bar{x})$ for $\bar{x} \in [0, 1]$, and we have now proved that the constructed function $u(x)$ indeed solves the given initial value problem.

We conclude the proof by showing uniqueness. Thus, assume that we have two uniformly differentiable functions $u(x)$ and $v(x)$ such that $u'(x) = u(x)$ and $v'(x) = v(x)$ for $x \in (0, 1]$, and $u(0) = v(0) = 1$. The $w = u - v$ satisfies $w'(x) = w(x)$ and $w(0) = 0$, and thus by the Fundamental Theorem,

$$w(x) = \int_0^x w'(y)\,dy = \int_0^x w(y)\,dy \quad \text{for } x \in [0, 1].$$

Setting $a = \max_{0\leq x\leq 0.5} |w(x)|$, we thus have

$$a \leq \int_0^{0.5} a\,dy = 0.5a$$

which is possible only if $a = 0$ showing uniqueness for $0 \le x \le 0.5$. Repeating the argument on $[0.5, 1]$ proves that $w(x) = 0$ for $x \in [0, 1]$ and the uniqueness follows.

The proof immediately generalizes to $x \in [0, b]$ where b is any positive real number. We now summarize:

Theorem 31.1 *The initial value problem $u'(x) = u(x)$ for $x > 0$, and $u(0) = 1$, has a unique solution $u(x) = \exp(x)$ given by* (31.13).

31.3 Extension of the Exponential $\exp(x)$ to $x < 0$

If we define

$$\exp(-x) = \frac{1}{\exp(x)} \quad \text{for } x \ge 0,$$

then we find that

$$D\exp(-x) = D\frac{1}{\exp(x)} = -\frac{D\exp(x)}{(\exp(x))^2} = -\frac{1}{\exp(x)} = -\exp(-x). \quad (31.15)$$

We conclude that $\exp(-x)$ solves the initial value problem

$$u'(x) = -u(x) \quad \text{for } x > 0,\ u(0) = 1.$$

31.4 The Exponential Function $\exp(x)$ for $x \in \mathbb{R}$

Piecing together the functions $\exp(x)$ and $\exp(-x)$ with $x \ge 0$, we obtain the function $u(x) = \exp(x)$ defined for $x \in \mathbb{R}$, which satisfies $u'(x) = u(x)$ for $x \in \mathbb{R}$ and $u(0) = 1$, see Fig. 31.4 and Fig. 31.5.

To see that $\frac{d}{dx}\exp(x)$ for $x < 0$, we set $y = -x > 0$ and compute $\frac{d}{dx}\exp(x) = \frac{d}{dy}\exp(-y)\frac{dy}{dx} = -\exp(-y)(-1) = \exp(x)$, where we used (31.15).

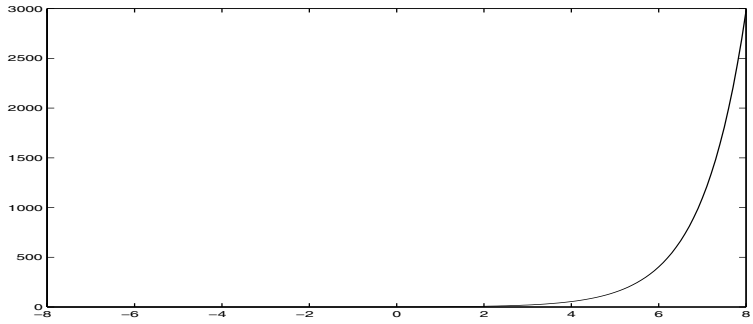

Fig. 31.4. The exponential $\exp(x)$ for $x \in [-2.5, 2.5]$

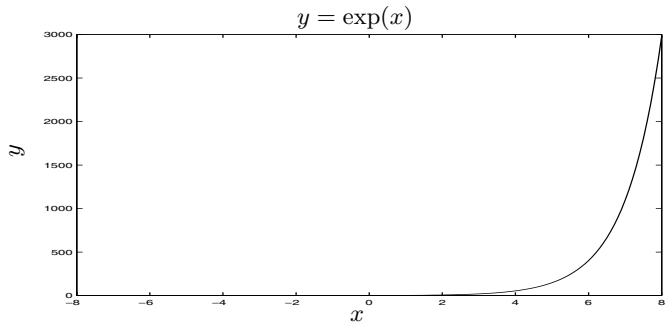

Fig. 31.5. The exponential $\exp(x)$ for $x \in [-8, 8]$

31.5 An Important Property of $\exp(x)$

We now prove the basic property of the exponential function $\exp(x)$ using the fact that $\exp(x)$ satisfies the differential equation $D\exp(x) = \exp(x)$. We start considering the initial value problem

$$u'(x) = u(x) \quad \text{for } x > a, \quad u(a) = u_a, \tag{31.16}$$

with initial value at some point a other than zero. Setting $x = y + a$ and $v(y) = u(y + a) = u(x)$, we obtain by the Chain rule

$$v'(y) = \frac{d}{dy}u(y + a) = u'(y + a)\frac{d}{dy}(y + a) = u'(x),$$

and thus $v(y)$ satisfies the differential equation

$$v'(y) = v(y) \quad \text{for } y > 0, \quad v(0) = u_a.$$

This means that

$$v(y) = \exp(y)u_a \quad \text{for } y > 0.$$

Going back to the original variables, using that $y = x - a$, we find that the solution of (31.16) is given by

$$u(x) = \exp(x - a)u_a \quad \text{for } x \geq a. \tag{31.17}$$

We now prove that for $a, b \in \mathbb{R}$,

$$\exp(a + b) = \exp(a)\exp(b) \quad \text{or } e^{a+b} = e^a e^b, \tag{31.18}$$

which is the basic property of the exponential function. We do this by using the fact that $u(x) = \exp(x)$ satisfies the differential equation $u'(x) = u(x)$ and $\exp(0) = 1$. We have on the one hand that $u(a + b) = \exp(a + b)$ is the value of the solution $u(x)$ for $x = a+b$. We may reach to $x = a+b$, assuming

$0 < a, b$ to start with, by first computing the solution $u(x) = \exp(x)$ from $x = 0$ up to $x = a$, which gives $u(a) = \exp(a)$. We next consider the following problem

$$v'(x) = v(x) \quad \text{for } x > a, \quad v(a) = \exp(a)$$

with solution $v(x) = \exp(x - a)\exp(a)$ for $x \geq a$. We have $v(x) = u(x)$ for $x \geq a$, since $u(x)$ also solves $u'(x) = u(x)$ for $x > a$, and $u(a) = \exp(a)$. Thus $v(b+a) = u(a+b)$, which translates into the desired equality $\exp(b)\exp(a) = \exp(a+b)$. The proof extends to any $a, b \in \mathbb{R}$.

31.6 The Inverse of the Exponential is the Logarithm

We shall now prove that

$$\log(\exp(x)) = x \quad \text{for } x \in \mathbb{R}, \tag{31.19}$$

and conclude that

$$y = \exp(x) \quad \text{if and only if } x = \log(y), \tag{31.20}$$

which states that the inverse of the exponential is the logarithm.

We prove (31.19) by differentiation to get by the Chain rule for $x \in \mathbb{R}$,

$$\frac{d}{dx}(\log(\exp(x)) = \frac{1}{\exp(x)}\,\frac{d}{dx}(\exp(x)) = \frac{1}{\exp(x)}\,\exp(x) = 1,$$

and noting that $\log(\exp(0)) = \log(1) = 0$, which gives (31.19). Setting $x = \log(y)$ in (31.19), we have $\log\big(\exp(\log(y))\big) = \log(y)$, that is

$$\exp(\log(y)) = y \quad \text{for } y > 0. \tag{31.21}$$

We note in particular that

$$\exp(0) = 1 \text{ and } \log(e) = 1 \tag{31.22}$$

since $0 = \log(1)$ and $e = \exp(1)$ respectively.

In many Calculus books the exponential function $\exp(x)$ is defined as the inverse of the logarithm $\log(x)$ (which is defined as an integral). However, we prefer to directly prove the existence of $\exp(x)$ through its defining initial value problem, since this prepares the construction of solutions to general initial value problems.

31.7 The Function a^x with $a > 0$ and $x \in \mathbb{R}$

We recall that in Chapter *The function* $y = x^r$ we defined the function x^r for $r = p/q$ rational with p and $q \neq 0$ integers, and x is a positive real number, as the solution y to the equation $y^q = x^p$.

We thus are familiar with a^x with $a > 0$ and x rational, and we may extend to $x \in \mathbb{R}$ by defining:

$$a^x = \exp(x \log(a)). \tag{31.23}$$

We now prove the basic properties of a^x with $x \in \mathbb{R}$, that is, the positive number a raised to the power $x \in \mathbb{R}$, extending our previous experience with x rational. We note that by the Chain rule the function $u(x) = a^x$ satisfies the differential equation

$$u'(x) = \log(a)u(x)$$

and $u(0) = 1$. In particular, choosing $a = e = \exp(1)$, we find that $a^x = e^x = \exp(x)$, and we thus conclude that the exponential function $\exp(x)$ indeed equals the number e raised to the power x. Note that before we just used e^x just as a another notation for $\exp(x)$.

Using now the exponential law (31.18) for $\exp(x)$, we obtain with a direct computation using the definition (31.23) the following analog for a^x:

$$a^{x+y} = a^x a^y. \tag{31.24}$$

The other basic rule for a^x reads:

$$(a^x)^y = a^{xy}, \tag{31.25}$$

which follows from the following computation:

$$(a^x)^y = \exp(y \log(a^x)) = \exp(y \log(\exp(x \log(a)))) = \exp(yx \log(a)) = a^{xy}.$$

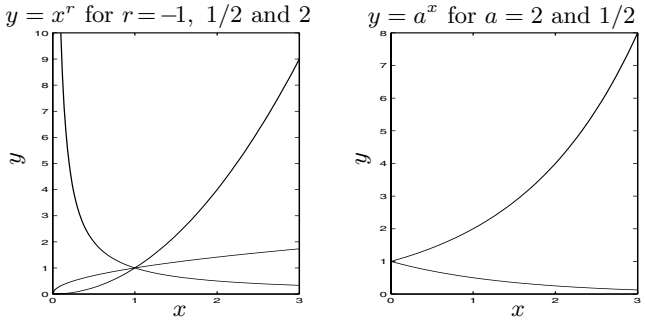

Fig. 31.6. Examples of functions x^r and a^x

As indicated, the rules (31.24) and (31.25) generalize the corresponding rules with x and y rational proved above.

We conclude computing the derivative of the function a^x from the definition (31.23) using the Chain rule:

$$\frac{d}{dx}a^x = \log(a)a^x. \tag{31.26}$$

Chapter 31 Problems

31.1. Define $U^n(x_i^n)$ alternatively by $U^n(x_i^n) = U^n(x_{i-1}^n) \pm h_n U^n(x_i^n)$, and prove that the corresponding sequence $\{U^n(x)\}$ converges to $\exp(\pm x)$.

31.2. Prove that for $x > 0$

$$\left(1 + \frac{x}{n}\right)^n < \exp(x) \quad \text{for } n = 1, 2, 3 \ldots. \tag{31.27}$$

Hint: Take logarithm and use that $\log(1+x) < x$ for $x > 0$, and that the logarithm is increasing.

31.3. Prove directly the existence of a unique solution of $u'(x) = -u(x)$ for $x > 0$, $u(0) = 1$, that is construct $\exp(-x)$ for $x \geq 0$.

31.4. Show that the more often the bank capitalizes your interest, the better off you are, that is verify that

$$\left(1 + \frac{a}{n}\right)^n \leq \left(1 + \frac{a}{n+1}\right)^{n+1}. \tag{31.28}$$

Hint: Use the Binomial theorem.

31.5. Assume a bank offers "continuous capitalization" of the interest, corresponding to the (annual) interest rate a. What is then the "effective annual interest rate"?

31.6. Prove the differentiation formula $\frac{d}{dx}x^r = rx^{r-1}$ for $r \in \mathbb{R}$.

31.7. Prove the basic properties of the exponential function using that it is the inverse of the logarithm and use properties of the logarithm.

31.8. Given that the equation $u'(x) = u(x)$ has a solution for $x \in [0, 1]$ with $u(0) = 1$, construct a solution for all $x \geq 0$. Hint: use that if $u(x)$ satisfies $u'(x) = u(x)$ for $0 < x \leq 1$, then $v(x) = u(x-1)$ satisfies $v'(x) = v(x)$ for $1 < x \leq 2$ and $v(1) = u(0)$.

31.9. Give the Taylor polynomial of order n with error term for $\exp(x)$ at $x = 0$.

31.10. Find a primitive function to (a) $x\exp(-x^2)$, (b) $x^3\exp(-x^2)$.

31.11. Compute the derivatives of the following functions: a) $f(x) = a^x$, $a > 0$, b) $f(x) = \exp(x+1)$, c) $f(x) = x\exp(x^2)$, d) $f(x) = x^3\exp(x^2)$, e) $f(x) = \exp(-x^2)$.

31.12. Compute the integrals $\int_0^1 f(x)dx$ of the functions in the previous exercise, except for the one in e), $f(x) = \exp(-x^2)$. Why do you think we left this one out?

31.13. Try to find the value of $\int_{-\infty}^{\infty} \exp(-x^2)dx$ numerically by quadrature. Square the result. Do you recognize this number?

31.14. Show that $\exp(x) \geq 1 + x$ for all x, not just for $x > -1$.

31.15. Show, by induction, that

$$\frac{d^n}{dx^n}\left(\mathrm{e}^x f(x)\right) = \mathrm{e}^x \left(1 + \frac{d}{dx}\right)^n f(x).$$

31.16. Prove (31.24) using the basic property (31.18) of the exponential and the definition (31.23).

31.17. Construct directly, without using the exponential function, the solution to the initial value problem $u'(x) = au(x)$ for $x \geq 0$ with $u(0) = 1$, where a is a real constant. Call the solution $\mathrm{aexp}(x)$. Prove that the function $\mathrm{aexp}(x)$ satisfies $\mathrm{aexp}(x+y) = \mathrm{aexp}(x)\mathrm{aexp}(y)$ for $x, y \geq 0$.

31.18. Define with $a > 0$ given, the function $y = \log_a(x)$ for $x > 0$ as the solution y to the equation $a^y = x$. With $a = e$ we get $\log_e(x) = \log(x)$, the natural logarithm. With $a = 10$ we get the so-called 10-logarithm. Prove that (i) $\log_a(xy) = \log_a(x) + \log_a(y)$ for $x, y > 0$, (ii) $\log_a(x^r) = r\log_a(x)$ for $x > 0$ and $r \in \mathbb{R}$, and (iii) $\log_a(x)\log(a) = \log(x)$ for $x > 0$.

31.19. Give the details of the proof of (31.26).

32
Trigonometric Functions

> When I get to the bottom, I go back to the top of the slide where I stop and I turn and I go for a ride 'til I get to the bottom and I see you again. (Helter Skelter, Lennon-McCartney, 1968)

32.1 The Defining Differential Equation

In this chapter, we shall study the following *initial value problem for a second order differential equation*: Find a function $u(x)$ defined for $x \geq 0$ satisfying

$$u''(x) = -u(x) \quad \text{for } x > 0, \quad u(0) = u_0,\ u'(0) = u_1, \tag{32.1}$$

where u_0 and u_1 are given *initial data*. We here require two initial conditions because the problem involves a second order derivative. We may compare with the first order initial value problem: $u'(x) = -u(x)$ for $x > 0$, $u(0) = u_0$, with the solution $u(x) = \exp(-x)$, which we studied in the previous chapter.

We shall demonstrate below, in Chapter *The general initial value problem*, that (32.1) has a unique solution for any given values of u_0 and u_1, and we shall in this chapter show that the solution with initial data $u_0 = 0$ and $u_1 = 1$ is an old friend, namely, $u(x) = \sin(x)$, and the solution with $u_0 = 1$ and $u_1 = 0$ is $u(x) = \cos(x)$. Here $\sin(x)$ and $\cos(x)$ are the usual trigonometric functions defined geometrically in Chapter *Pythagoras and Euclid*, with the change that we measure the angle x in the unit of *radians*

instead of degrees, with one radian being equal to $\frac{180}{\pi}$ degrees. In particular, we shall explain why one radian equals $\frac{180}{\pi}$ degrees.

We may thus define the trigonometric functions $\sin(x)$ and $\cos(x)$ as the solutions of (32.1) with certain initial data if we measure angles in the unit of radian. This opens a fresh route to understanding properties of the trigonometric functions by studying properties of solutions the differential equation (32.1), and we shall now explore this possibility.

We start by rewriting (32.1) changing the independent variable from x to t, since to aid our intuition we will use a mechanical interpretation of (32.1), where the independent variable represents time. We denote the derivative with respect to t with a dot, so that $\dot{u} = \frac{du}{dt}$, and $\ddot{u} = \frac{d^2u}{dt^2}$. We thus rewrite (32.1) as

$$\ddot{u}(t) = -u(t) \quad \text{for } t > 0, \quad u(0) = 0,\ \dot{u}(0) = 1, \tag{32.2}$$

where we chose $u_0 = 0$ and $u_1 = 1$ anticipating that we are looking for $\sin(t)$.

We now recall that (32.2) is a model of the motion of unit mass along a friction-less horizontal x-axis with the mass connected to one end of a Hookean spring with spring constant equal to 1 and with the other end connected to the origin, see Fig. 26.3. We let $u(t)$ denotes the position ($x-$coordinate) of the mass at time t, and we assume that the mass is started at time $t = 0$ at the origin with speed $\dot{u}(0) = 1$, that is, $u_0 = 0$ and $u_1 = 1$. The spring exerts a force on the mass directed towards the origin, which is proportional to the length of the spring, since the spring constant is equal to 1, and the equation (32.2) expresses Newton's law: the acceleration $\ddot{u}(t)$ is equal to the spring force $-u(t)$. Because there is no friction, we would expect the mass to oscillate back and forth across the equilibrium position at the origin. We plot the solution $u(t)$ to (32.2) in Fig. 32.1, which clearly resembles the plot of the $\sin(t)$ function.

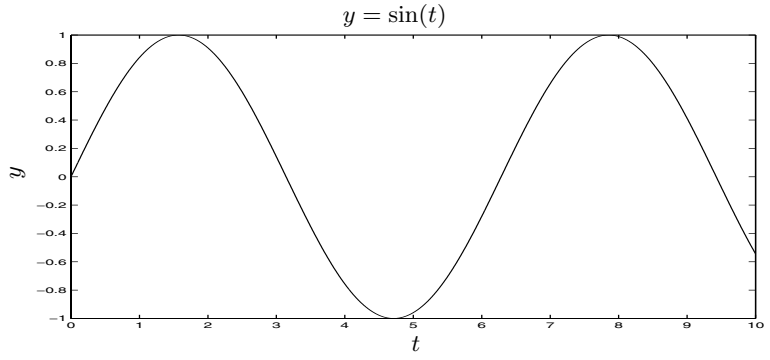

Fig. 32.1. The solution of (32.2). Is it the function $\sin(t)$?

Let's now prove that our intuitive feeling indeed is correct, that is, let us prove that the solution of (32.2) indeed is our old friend $\sin(t)$. The key step then turns out to be to multiply our equation $\ddot{u} + u = 0$ by $\dot{u}$, to get

$$\frac{d}{dt}(\dot{u}^2 + u^2) = 2\dot{u}\ddot{u} + 2u\dot{u} = 2\dot{u}(\ddot{u} + u) = 0.$$

We conclude that $\dot{u}^2(t)+u^2(t)$ is constant for all t, and since $\dot{u}^2(0)+u^2(0) = 1 + 0 = 1$, we have found that the solution $u(t)$ of (32.2) satisfies the *conservation property*

$$\dot{u}^2(t) + u^2(t) = 1 \quad \text{for } t > 0, \tag{32.3}$$

which states that the point $(\dot{u}(t), u(t)) \in \mathbb{R}^2$ lies on the unit circle in $\mathbb{R}^2$, see Fig. 32.2.

We remark that in mechanical terms, the relation (32.3) expresses that the *total energy*

$$E(t) \equiv \frac{1}{2}\dot{u}^2(t))\text{TS}^1 + \frac{1}{2}u^2(t), \tag{32.4}$$

is preserved ($= 1/2$) during the motion. The total energy at time t is the sum of the *kinetic energy* $\dot{u}^2(t)/2$, and the *potential energy* $u^2(t)/2$. The potential energy is the energy stored in the spring, which is equal to the *work* $W(u(t))$ to stretch the spring the distance $u(t)$:

$$W(u(t)) = \int_0^{u(t)} v\,dv = \frac{1}{2}u^2(t),$$

where we used the principle that to stretch the spring from v to $v+\Delta v$, the work is $v\Delta v$ since the spring force is v. At the extreme points with $\dot{u}(t) = 0$, the kinetic energy is zero and all energy occurs as potential energy, while all energy occurs as kinetic energy when the body passes the origin with $u(t) = 0$. During the motion of the body, the energy is thus periodically transferred from kinetic energy to potential energy an back again.

Going now back to (32.3), we thus see that the point $(\dot{u}(t), u(t)) \in \mathbb{R}^2$ moves on the unit circle and the *velocity* of the motion is given by $(\ddot{u}(t), \dot{u}(t))$, which we obtain by differentiating each coordinate function with respect to t. We will return to this issue in Chapter *Curves* below. Using the differential equation $\ddot{u} + u = 0$, we see that

$$(\ddot{u}(t), \dot{u}(t)) = (-u(t), \dot{u}(t)),$$

and conclude recalling (32.3) that the modulus of the velocity is equal to 1 for all t. We conclude that the point $(\dot{u}(t), u(t))$ moves around the unit circle with unit velocity and at time $t = 0$ the point is at position $(1, 0)$. But this directly connects with the usual geometrical definition of $(\cos(t), \sin(t))$ as the coordinates of a point on the unit circle at the angle t, see Fig. 32.2, so that we should have $(\dot{u}(t), u(t)) = (\cos(t), \sin(t))$. To make this connection

TS[1] Is there an opening parenthesis missing here?

Editor's or typesetter's annotations (will be removed before the final TEX run)

straight, we of course need to measure angles properly, and the proper measure is *radians* with 2π radians corresponding to 360 degrees. This is because the time for one revolution with speed 1 should be equal to 2π, that is the length of the circumference of the unit circle.

In fact, we can use the solution $\sin(t)$ of the initial value problem (32.2) to define the number π as the smallest positive root $\bar{t}$ of $\sin(t)$, corresponding to one half revolution with $u(\bar{t}) = 0$ and $\dot{u}(\bar{t}) = -1$.

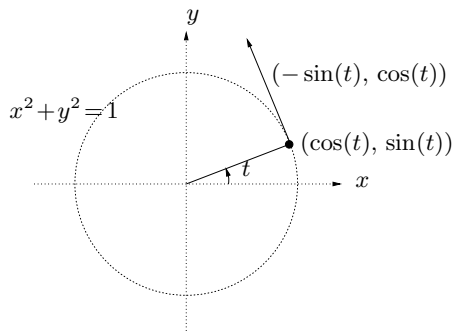

Fig. 32.2. Energy conservation

We may now conclude that the solution $u(t)$ of (32.2) satisfies $(\dot{u}(t), u(t)) = (\cos(t), \sin(t))$, so that in particular $u(t) = \sin(t)$ and $\frac{d}{dt}\sin(t) = \cos(t)$, where $(\cos(t), \sin(t))$ is defined geometrically as the point on the unit circle of angle t radians.

We can now turn the argument around, and simply define $\sin(t)$ as the solution $u(t)$ to (32.2) with $u_0 = 0$ and $u_1 = 1$, and then define $\cos(t) = \frac{d}{dt}\sin(t)$. Alternatively, we can define $\cos(t)$ as the solution of the problem

$$\ddot{v}(t) = -v(t) \quad \text{for } t > 0, \quad v(0) = 1,\ \dot{v}(0) = 0, \tag{32.5}$$

which we obtain by differentiation of (32.2) with respect to t and using the initial conditions for $\sin(t)$. Differentiating once more, we see that $\frac{d}{dt}\cos(t) = -\sin(t)$.

Both $\sin(t)$ and $\cos(t)$ will be *periodic with period* 2π, because the point $(\dot{u}(t), u(t))$ moves around the unit circle with velocity one and comes back the same point after a time period of 2π. As we said, we may in particular define π as the first value of $t > 0$ for which $\sin(t) = 0$, which corresponds the point $(\dot{u}, u) = (-1, 0)$, and 2π will then be time it takes for the point $(\dot{u}, u)$ to make one complete revolution starting at $(1, 0)$, moving to $(-1, 0)$ following the upper semi-circle and then returning to $(1, 0)$ following the lower semi-circle. The periodicity of $u(t)$ with period 2π is expressed as

$$u(t + 2n\pi) = u(t) \quad \text{for } t \in \mathbb{R},\ n = 0, \pm 1, \pm 2, \ldots \tag{32.6}$$

The energy conservation (32.3) translates into the most well known of all trigonometric formulas:

$$\sin^2(t) + \cos^2(t) = 1 \quad \text{for } t > 0. \tag{32.7}$$

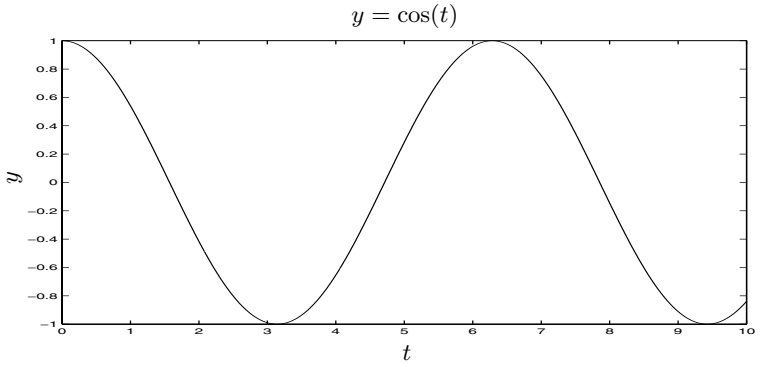

Fig. 32.3. The function $\cos(t)$!

To compute the values of $\sin(t)$ and $\cos(t)$ for a given t, we may compute the solution to the corresponding defining differential initial value problem. We return to this topic below.

We summarize:

Theorem 32.1 *The initial value problem $u''(x) + u(x) = 0$ for $x > 0$ with $u_0 = 0$ and $u_1 = 1$, has a unique solution, which is denoted by $\sin(x)$. The initial value problem $u''(x) + u(x) = 0$ for $x > 0$ with $u_0 = 1$ and $u_1 = 0$, has a unique solution, which is denoted by $\cos(x)$. The functions $\sin(x)$ and $\cos(x)$ extend to $x < 0$ as solutions of $u''(x) + u(x) = 0$ and are periodic with period 2π, and $\sin(\pi) = 0$, $\cos(\frac{\pi}{2}) = 0$. We have $\frac{d}{dx}\sin(x) = \cos(x)$ and $\frac{d}{dx}\cos(x) = -\sin(x)$. Further $\cos(-x) = \cos(x)$, $\cos(\pi - x) = -\cos(x)$, $sin(\pi - x) = \sin(x)$, $sin(-x) = -\sin(x)$, $\cos(x) = \sin(\frac{\pi}{2} - x)$, $\sin(x) = \cos(\frac{\pi}{2} - x)$, $\sin(\frac{\pi}{2} + x) = \cos(x)$, and $\cos(\frac{\pi}{2} + x) = -\sin(x)$.*

32.2 Trigonometric Identities

Using the defining differential equation $u''(x) + u(x) = 0$, we can verify the following basic trigonometric identities for $x, y \in \mathbb{R}$:

$$\sin(x + y) = \sin(x)\cos(y) + \cos(x)\sin(y) \tag{32.8}$$

$$\sin(x - y) = \sin(x)\cos(y) - \cos(x)\sin(y) \tag{32.9}$$

$$\cos(x + y) = \cos(x)\cos(y) - \sin(x)\sin(y) \tag{32.10}$$

$$\cos(x - y) = \cos(x)\cos(y) + \sin(x)\sin(y). \tag{32.11}$$

TS[m] In the hardcopy, this equation is labelled (41.12), please check.

Editor's or typesetter's annotations (will be removed before the final TEX run)

For example, to prove (32.8),[TS m] we note that both the right hand and left hand side satisfy the equation $u''(x) + u(x) = 0$, and the initial conditions $u(0) = \sin(y)$, $u'(0) = \cos(y)$, with y acting as a parameter, and thus are equal.

We note the particular special cases:

$$\sin(2x) = 2\sin(x)\cos(x) \tag{32.12}$$

$$\cos(2x) = \cos^2(x) - \sin^2(x) = 2\cos^2(x) - 1 = 1 - 2\sin^2(x). \tag{32.13}$$

Adding (32.8)[TS m] and (32.9)[TS m], we obtain

$$\sin(x+y) + \sin(x-y) = 2\sin(x)\cos(y).$$

Setting $\bar{x} = x + y$ and $\bar{y} = x - y$ we obtain the first of the following set of formulas, all proved similarly,

$$\sin(\bar{x}) + \sin(\bar{y}) = 2\sin\left(\frac{\bar{x}+\bar{y}}{2}\right)\cos\left(\frac{\bar{x}-\bar{y}}{2}\right) \tag{32.14}$$

$$\sin(\bar{x}) - \sin(\bar{y}) = 2\cos\left(\frac{\bar{x}+\bar{y}}{2}\right)\sin\left(\frac{\bar{x}-\bar{y}}{2}\right) \tag{32.15}$$

$$\cos(\bar{x}) + \cos(\bar{y}) = 2\cos\left(\frac{\bar{x}+\bar{y}}{2}\right)\cos\left(\frac{\bar{x}-\bar{y}}{2}\right) \tag{32.16}$$

$$\cos(\bar{x}) - \cos(\bar{y}) = -2\sin\left(\frac{\bar{x}+\bar{y}}{2}\right)\sin\left(\frac{\bar{x}-\bar{y}}{2}\right). \tag{32.17}$$

32.3 The Functions $\tan(x)$ and $\cot(x)$ and Their Derivatives

We define

$$\tan(x) = \frac{\sin(x)}{\cos(x)}, \quad \cot(x) = \frac{\cos(x)}{\sin(x)}, \tag{32.18}$$

for values of x such that the denominator is different from zero. We compute the derivatives

$$\frac{d}{dx}\tan(x) = \frac{\cos(x)\cos(x) - \sin(x)(-\sin(x))}{\cos^2(x)} = \frac{1}{\cos^2(x)}, \tag{32.19}$$

and similarly

$$\frac{d}{dx}\cot(x) = -\frac{1}{\sin^2(x)}. \tag{32.20}$$

Dividing (32.8)[TS m] by (32.10)[TS m], and dividing both numerator and denominator by $\cos(x)\cos(y)$, we obtain

$$\tan(x+y) = \frac{\tan(x) + \tan(y)}{1 - \tan(x)\tan(y)}, \tag{32.21}$$

and similarly,

$$\tan(x-y) = \frac{\tan(x) - \tan(y)}{1 + \tan(x)\tan(y)}. \tag{32.22}$$

32.4 Inverses of Trigonometric Functions

Inverses of the basic trigonometric functions $\sin(x)$, $\cos(x)$, $\tan(x)$ and $\cot(x)$, are useful in applications. We now introduce and give names to these inverses and derive their basic properties.

The function $f(x) = \sin(x)$ is strictly increasing from -1 to 1 on $[\frac{-\pi}{2}, \frac{\pi}{2}]$, because the derivative $f'(x) = \cos(x)$ is positive on $(\frac{-\pi}{2}, \frac{\pi}{2})$. Thus the function $y = f(x) = \sin(x)$ with $D(f) = [\frac{-\pi}{2}, \frac{\pi}{2}]$ and $R(f) = [-1, 1]$, therefore has an inverse $x = f^{-1}(y)$, which we denote by

$$x = f^{-1}(y) = \arcsin(y), \tag{32.23}$$

and $D(f^{-1}) = D(\arcsin) = [-1, 1]$ and $R(f^{-1}) = R(\arcsin) = [\frac{-\pi}{2}, \frac{\pi}{2}]$, see Fig. 32.4.

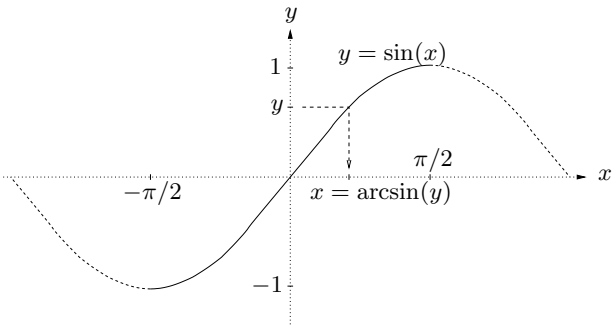

Fig. 32.4. The function $x = \arcsin(y)$

We thus have

$$\sin(\arcsin(y)) = y \quad \arcsin(\sin(x)) = x \quad \text{for } x \in \left[\frac{-\pi}{2}, \frac{\pi}{2}\right], y \in [-1, 1]. \tag{32.24}$$

We next compute the derivative of $\arcsin(y)$ with respect to y:

$$\frac{d}{dy}\arcsin(y) = \frac{1}{\frac{d}{dx}\sin(x)} = \frac{1}{\cos(x)} = \frac{1}{\sqrt{1-\sin^2(x)}} = \frac{1}{\sqrt{1-y^2}}.$$

Similarly, the function $y = f(x) = \tan(x)$ is strictly increasing on $D(f) = (\frac{-\pi}{2}, \frac{\pi}{2})$ and $R(f) = \mathbb{R}$, and thus has an inverse, which we denote by

$$x = f^{-1}(y) = \arctan(y),$$

with $D(\arctan) = \mathbb{R}$ and $R(\arctan) = (\frac{-\pi}{2}, \frac{\pi}{2})$, see Fig. 32.5.

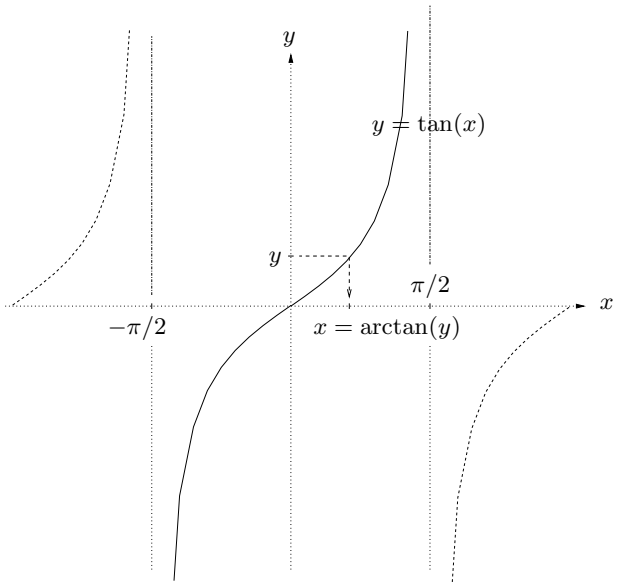

Fig. 32.5. The function $x = \arctan(y)$

We compute the derivative of $\arctan(y)$:

$$\frac{d}{dy}\arctan(y) = \frac{1}{\frac{d}{dx}\tan(x)} = \cos^2(x)$$
$$= \frac{\cos^2(x)}{\cos^2(x) + \sin^2(x)} = \frac{1}{1 + \tan^2(x)} = \frac{1}{1 + y^2}.$$

We define similarly the inverse of $y = f(x) = \cos(x)$ with $D(f) = [0, \pi]$ and denote the inverse by $x = f^{-1}(y) = \arccos(y)$ with $D(\arccos) = [-1, 1]$ and $R(\arccos) = [0, \pi]$. We have

$$\frac{d}{dy}\arccos(y) = \frac{1}{\frac{d}{dx}\cos(x)} = -\frac{1}{\sin(x)} = -\frac{1}{\sqrt{1 - \cos^2(x)}} = -\frac{1}{\sqrt{1 - y^2}}.$$

Finally, we define the inverse of $y = f(x) = \cot(x)$ with $D(f) = (0, \pi)$ and denote the inverse by $x = f^{-1}(y) = \operatorname{arccot}(y)$ with $D(\operatorname{arccot}) = \mathbb{R}$ and $R(\operatorname{arccot}) = (0, \pi)$. We have

$$\frac{d}{dy}\operatorname{arccot}(y) = \frac{1}{\frac{d}{dx}\cot(x)} = -\sin^2(x) = -\frac{\sin^2(x)}{\cos^2(x) + \sin^2(x)}$$
$$= -\frac{1}{1 + \cot^2(x)} = -\frac{1}{1 + y^2}.$$

We summarize:

$$\begin{aligned}
\frac{d}{dx}\arcsin(x) &= \frac{1}{\sqrt{1-x^2}} \quad \text{for } x \in (-1,1)\\
\frac{d}{dx}\arccos(x) &= -\frac{1}{\sqrt{1-x^2}} \quad \text{for } x \in (-1,1)\\
\frac{d}{dx}\arctan(x) &= \frac{1}{1+x^2} \quad \text{for } x \in \mathbb{R}\\
\frac{d}{dx}\text{arccot}(x) &= -\frac{1}{1+x^2} \quad \text{for } x \in \mathbb{R}.
\end{aligned} \tag{32.25}$$

In other words,

$$\begin{aligned}
\arcsin(x) &= \int_0^x \frac{1}{\sqrt{1-y^2}}\,dy \quad \text{for } x \in (-1,1)\\
\arccos(x) &= \frac{\pi}{2} - \int_0^x \frac{1}{\sqrt{1-y^2}}\,dy \quad \text{for } x \in (-1,1)\\
\arctan(x) &= \int_0^x \frac{1}{1+y^2}\,dy \quad \text{for } x \in \mathbb{R}\\
\text{arccot}(x) &= \frac{\pi}{2} - \int_0^x \frac{1}{1+y^2}\,dy \quad \text{for } x \in \mathbb{R}.
\end{aligned} \tag{32.26}$$

We also note the following analog of (32.21) obtained by setting $x = \arctan(u)$ and $y = \arctan(v)$, so that $u = \tan(x)$ and $v = \tan(y)$, and assuming that $x + y \in (-\frac{\pi}{2}, \frac{\pi}{2})$:

$$\arctan(u) + \arctan(v) = \arctan\left(\frac{u+v}{1-uv}\right). \tag{32.27}$$

32.5 The Functions sinh(x) and cosh(x)

We define for $x \in \mathbb{R}$

$$\sinh(x) = \frac{e^x - e^{-x}}{2} \quad \text{and} \quad \cosh(x) = \frac{e^x + e^{-x}}{2}. \tag{32.28}$$

We note that

$$D\sinh(x) = \cosh(x) \quad \text{and} \quad D\cosh(x) = \sinh(x). \tag{32.29}$$

We have $y = f(x) = \sinh(x)$ is strictly increasing and thus has an inverse $x = f^{-1}(y) = \text{arcsinh}(y)$ with $D(\text{arcsinh}) = \mathbb{R}$ and $R(\text{arcsinh}) = \mathbb{R}$. Further, $y = f(x) = \cosh(x)$ is strictly increasing on $[0,\infty)$, and thus has an inverse $x = f^{-1}(y) = \text{arccosh}(y)$ with $D(\text{arccosh}) = [1,\infty)$ and $R(\text{arccosh}) = [0,\infty)$. We have

$$\frac{d}{dy}\text{arcsinh}(y) = \frac{1}{\sqrt{y^2+1}}, \quad \frac{d}{dy}\text{arccosh}(y) = \frac{1}{\sqrt{y^2-1}}. \tag{32.30}$$

32.6 The Hanging Chain

Consider a hanging chain fixed at $(-1,0)$ and $(1,0)$ in a coordinate system with the x-axis horizontal and y-axis vertical. Let us seek the curve $y = y(x)$ described by the chain. Let $(F_h(x), F_v(x))$ be the two components of the *force* in the chain at x. Vertical and horizontal equilibrium of the element of the chain between x and $x + \Delta x$ gives

$$F_h(x + \Delta x) = F_h(x), \quad F_v(x) + m\Delta s = F_v(x + \Delta x),$$

where $\Delta s \approx \sqrt{(\Delta x)^2 + (\Delta y)^2} \approx \sqrt{1 + (y'(x))^2}\Delta x$, and m is the weight of the chain per unit length. We conclude that $F_h(x) = F_h$ is constant, and

$$F_v'(x) = m\sqrt{1 + (y'(x))^2}.$$

Momentum equilibrium around the midpoint of the element of the chain between x and $x + \Delta x$, gives

$$F_h\Delta y = \frac{1}{2}F_v(x + \Delta x)\Delta x + \frac{1}{2}F_v(x)\Delta x \approx F_v(x)\Delta x,$$

which leads to

$$y'(x) = \frac{F_v(x)}{F_h}. \tag{32.31}$$

Assuming that $F_h = 1$, we are thus led to the differential equation

$$F_v'(x) = m\sqrt{1 + (F_v(x))^2}.$$

We can check by direct differentiation that this differential equation is satisfied if $F_v(x)$ solves the equation

$$\text{arcsinh}(F_v(x)) = mx,$$

and we also have $F_v(0) = 0$. Therefore

$$F_v(x) = \sinh(mx),$$

and thus by (32.31),

$$y(x) = \frac{1}{m}\cosh(mx) + c$$

with the constant c to be chosen so that $y(\pm 1) = 0$. We thus obtain the following solution

$$y(x) = \frac{1}{m}(\cosh(mx) - \cosh(m)). \tag{32.32}$$

The curve $y(x) = \cosh(mx)+c$ with m and c constants, is called the *hanging chain curve*, or the *catenaria*.

32.7 Comparing $u'' + k^2u(x) = 0$ and $u'' - k^2u(x) = 0$

We summarize some experience from above. The solutions of the equation $u'' + k^2u(x) = 0$ are linear combinations of $\sin(kx)$ and $\cos(kx)$. The solutions of $u'' - k^2u(x) = 0$ are linear combinations of $\sinh(kx)$ and $\cosh(kx)$.

Chapter 32 Problems

32.1. Show that the solution of $\ddot{u}(t) + u(t) = 0$ for $t > 0$ with $u(0) = \sin(\alpha)$ and $u'(0) = \cos(\alpha)$ is given by $u(t) = \cos(t)\sin(\alpha) + \sin(t)\cos(\alpha) = \sin(t + \alpha)$.

32.2. Show that the solution of $\ddot{u}(t)+u(t) = 0$ for $t > 0$ with $u(0) = r\cos(\alpha)$ and $u'(0) = r\sin(\alpha)$ is given by $u(t) = r(\cos(t)\cos(\alpha) + \sin(t)\sin(\alpha)) = r\cos(t - \alpha)$.

32.3. Show that the solution to $\ddot{u}(t) + ku(t) = 0$ for $t > 0$ with $u(0) = r\cos(\alpha)$ and $u'(0) = r\sin(\alpha)$, where k is a given positive constant, is given by $r\cos(\sqrt{k}(t - \alpha))$. Give a mechanical interpretation of this model.

32.4. Show that the function $\sin(nx)$ solves the boundary value problem $u''(x)+n^2u(x) = 0$ for $0 < x < \pi$, $u(0) = u(\pi) = 0$.

32.5. Solve $u'(x) = \sin(x)$, $\quad x > \pi/4$, $u(\pi/4) = 2/3$.

32.6. Show that (a) $\sin(x) < x$ for $x > 0$, (b) $x < \tan(x)$ for $0 < x < \pi/2$.

32.7. Show that $\lim_{x\to 0} \dfrac{\sin(x)}{x} = 1$.

32.8. Show the following relations from the definition, i.e. from the differential equation defining $\sin(x)$ and $\cos(x)$: (a) $\sin(-x) = -\sin(x)$, (b) $\cos(-x) = \cos(x)$, (c) $\sin(\pi - x) = \sin(x)$, (d) $\cos(\pi - x) = -\cos(x)$, (e) $\sin(\pi/2 - x) = \cos(x)$, (f) $\cos(\pi/2 - x) = \sin(x)$.

32.9. Prove the product formulas show that

$$\sin(x)\sin(y) = \frac{1}{2}\left(\cos(x - y) - \cos(x + y)\right),$$
$$\cos(x)\cos(y) = \frac{1}{2}\left(\cos(x - y) + \cos(x + y)\right),$$
$$\sin(x)\cos(y) = \frac{1}{2}\left(\sin(x - y) + \sin(x + y)\right).$$

32.10. Compute the following integrals by integrating by parts:
(a) $\int_0^1 x^3\sin(x)dx$, (b) $\int_0^1 \exp(x)\sin(x)dx$, (c) $\int_0^1 x^2\cos(x)dx$.

32.11. Determine Taylor's formula for $\arctan(x)$ at $x = 0$ and use your result to calculate approximations of π. Hint: $\arctan(1) = \pi/4$.

32.12. Show that $\arctan(1) = \arctan(1/2) + \arctan(1/3)$. Try to find other rational numbers a and b such that $\arctan(1) = \arctan(a) + \arctan(b)$. In particular seek to find a and b as small as possible.

32.13. Combine your results from the previous two exercises to construct a better algorithm for computing π. Even more efficient methods may be obtained using the identity $\pi/4 = 4\arctan(1/5) - \arctan(1/239)$. Compare the two algorithms and explain why the second is more efficient.

32.14. Show that: (a) $\arcsin(-x) = -\arcsin(x)$, (b) $\arccos(-x) = \pi - \arccos(x)$, (c) $\arctan(-x) = -\arctan(x)$, (d) $\operatorname{arccot}(-x) = \pi - \operatorname{arccot}(x)$, (e) $\arcsin(x) + \arccos(x) = \pi/2$, (f) $\arctan(x) + \operatorname{arccot}(x) = \pi/2$.

32.15. Calculate analytically: (a) $\arctan(\sqrt{2}-1)$, (b) $\arctan(1/8)+\arctan(7/9)$, (c) $\arcsin(1/7) + \arcsin(11/4)$, (d) $\tan(\arcsin(3/5)/2)$, (e) $\sin(2\arcsin(0.8))$, (f) $\arctan(2) + \arcsin(3/\sqrt{10})$.

32.16. Solve the equation: (a) $\arccos(2x) = \arctan(x)$, (b) $\arcsin(\cos(x)) = x\sqrt{3}$.

32.17. Calculate the derivative, if possible, of (a) $\arctan(\sqrt{x} - x^5)$, (b) $\arcsin(1/x^2)\arcsin(x^2)$, (c) $\tan(\arcsin(x^2))$, (d) $1/\arctan(\sqrt{x})$.

32.18. Compute numerically for different values of x, (a) $\arcsin(x)$, (b) $\arccos(x)$, (c) $\arctan(x)$, (d) $\operatorname{arccot}(x)$.

32.19. Prove (32.30).

32.20. Verify that $\cosh^2(x) - \sinh^2(x) = 1$.

32.21. (a) Find the inverse $x = \operatorname{arcsinh}(y)$ of $y = sinh(x) = \frac{1}{2}(e^x - e^{-x})$ by solving for x in terms of y. Hint: Multiply by e^x and solve for $z = e^x$. Then take logarithms. (b) Find a similar formula for $\operatorname{arccosh}(y)$.

32.22. Compute analytically the area of a disc of radius 1 by computing the integral

$$\int_{-1}^{1} \sqrt{1 - x^2}\, dx.$$

How do you handle the fact that $\sqrt{1-x^2}$ is not Lipschitz continuous on $[-1, 1]$? Hint: Use the substitution $x = \sin(y)$ and the fact the $\cos^2(y) = \frac{1}{2}(1 + \cos(2y))$.

33

The Functions $\exp(z)$, $\log(z)$, $\sin(z)$ and $\cos(z)$ for $z \in \mathbb{C}$

> The shortest path between two truths in the real domain passes through the complex domain. (Hadamard 1865-1963)

33.1 Introduction

In this chapter we extend some of the elementary functions to complex arguments. We recall that we can write a complex number z in the form $z = |z|(\cos(\theta) + i\sin(\theta))$ with $\theta = \arg z$ the argument of z, and $0 \leq \theta = \text{Arg } z < 2\pi$ the principal argument of z.

33.2 Definition of $\exp(z)$

We define, writing $z = x + iy$ with $x, y \in \mathbb{R}$,

$$\exp(z) = e^z = e^x(\cos(y) + i\sin(y)), \tag{33.1}$$

which extends the definition of e^z with $z \in \mathbb{R}$ to $z \in \mathbb{C}$. We note that in particular for $y \in \mathbb{R}$,

$$e^{iy} = \cos(y) + i\sin(y), \tag{33.2}$$

which is also referred to as *Euler's formula*. We note that

$$\sin(y) = \frac{e^{iy} - e^{-iy}}{2i}, \quad \cos(y) = \frac{e^{iy} + e^{-iy}}{2}, \quad \text{for } y \in \mathbb{R}, \tag{33.3}$$

and

$$|e^{iy}| = 1 \quad \text{for } y \in \mathbb{R}. \tag{33.4}$$

We can now express a complex number $z = r(\cos(\theta) + i\sin(\theta))$ in the form

$$z = re^{i\theta} \tag{33.5}$$

with $\theta = \arg z$ and $r = |z|$.

One can prove (using the basic trigonometric formulas) that $\exp(z)$ satisfies the usual law for exponentials so that in particular for $z, \zeta \in \mathbb{C}$,

$$e^z e^\zeta = e^{z+\zeta}. \tag{33.6}$$

In particular, the rule for multiplication of two complex numbers $z = |z|e^{i\theta}$ and $\zeta = |\zeta|e^{i\varphi}$ can be expressed as follows:

$$z\zeta = |z|e^{i\theta}|\zeta|e^{i\varphi} = |z||\zeta|e^{i(\theta+\varphi)}. \tag{33.7}$$

33.3 Definition of sin(z) and cos(z)

We define for $z \in \mathbb{C}$

$$\sin(z) = \frac{e^{iz} - e^{-iz}}{2i}, \quad \cos(z) = \frac{e^{iz} + e^{-iz}}{2}, \tag{33.8}$$

which extends (33.3) to $\mathbb{C}$.

33.4 de Moivres Formula

We have for $\theta \in \mathbb{R}$ and n an integer

$$(e^{i\theta})^n = e^{in\theta},$$

that is,

$$(\cos(\theta) + i\sin(\theta))^n = \cos(n\theta) + i\sin(n\theta), \tag{33.9}$$

which is referred to as *de Moivres formula*. In particular,

$$(\cos(\theta) + i\sin(\theta))^2 = \cos(2\theta) + i\sin(2\theta),$$

from which follows separating into real and complex parts

$$\cos(2\theta) = \cos^2(\theta) - \sin^2(\theta), \quad \sin(2\theta) = 2\cos(\theta)\sin(\theta).$$

Using de Moivres formula gives a quick way of deriving some of the basic[TS^n] trigonometric formulas (in case one has forgotten these formulas).

[TS^n] Please check it.

Editor's or typesetter's annotations (will be removed before the final TEX run)

33.5 Definition of log(z)

We have defined above log(x) for $x > 0$ and we now pose the problem of defining log(z) for $z \in \mathbb{C}$. We recall that $w = \log(x)$ can be viewed as the unique solution to the equation $e^w = x$, where $x > 0$. We consider therefore the equation

$$e^w = z,$$

with $z = |z|(\cos(\theta) + i\sin(\theta)) \in \mathbb{C}$ being given assuming $z \neq 0$, and we seek $w = \text{Re } w + i\text{Im } w \in \mathbb{C}$, with the intention to call a solution $w = \log(z)$. Here Re w and Im w denote the real and imaginary parts of w. Equating the modulus of both sides of the equation $e^w = z$, we get

$$e^{\text{Re } w} = |z|,$$

and thus

$$\text{Re } w = \log(|z|).$$

Further, equating the argument of both sides, we get

$$\text{Im } w = \theta = \arg z,$$

and thus

$$w = \log(|z|) + i \arg z.$$

We are thus led to define

$$\log(z) = \log(|z|) + i \arg z, \tag{33.10}$$

which extends the definition of the natural logarithm from the positive real numbers to non-zero complex numbers. We see that the imaginary part log(z) is not uniquely defined up to multiples of 2π, since arg z is not, and thus log(z) is *multi-valued*: the imaginary part of log(z) is not uniquely defined up to multiples of 2π. Choosing $\theta = \text{Arg } z$ with $0 \leq \text{Arg } z < 2\pi$, we obtain the *principal branch* of log(z) denoted by

$$\text{Log}(z) = \log(|z|) + i\text{Arg } z.$$

We see that if we let arg z increase from 0 beyond 2π, the function Log (z) will be discontinuous at Im $z = 2\pi$. We thus have to remember that the imaginary part of log(z) is not uniquely defined.

Chapter 33 Problems

33.1. Describe in geometrical terms the mappings $f : \mathbb{C} \to \mathbb{C}$ given by (a) $f(z) = \exp(z)$, (b) $f(z) = \text{Log}(z)$, (c) $\sin(z)$.

34
Techniques of Integration

> A poor head, having subsidiary advantages, ... can beat the best, just as a child can draw a line with a ruler better than the greatest master by hand. (Leibniz)

34.1 Introduction

It is not generally possible to find an explicit formula for a primitive function of a given arbitrary function in terms of known *elementary functions*, by which we mean the polynomials, rational functions, root functions, exponentials and trigonometric functions along with their inverses and combinations. It is not even true that the primitive function of an elementary function is another elementary function. A famous example is given by the function $f(x) = \exp(-x^2)$, whose primitive function $F(x)$ (with $F(0) = 0$), which exists by the Fundamental Theorem, is known *not* to be an elementary function (by a tricky proof by contradiction). To compute values of $F(x) = \int_0^x \exp(y)\,dy$ for different values of x we therefore have to use numerical quadrature just as in the case of the logarithm. Of course we can give $F(x)$ a *name*, for example we may agree to call it the *error function* $F(x) = \mathrm{erf}(x)$ and add it to our list of known functions that we can use. Nevertheless there will be other functions (such as $\frac{\sin(x)}{x}$) whose primitive function cannot be expressed in the known functions.

The question of how to handle such functions (including also $\log(x)$, $\exp(x)$, $\sin(x)\ldots$) of course arises: should we pre-compute long tables of values of these functions and print them in thick books or store them in

the computer, or should we compute each required value from scratch using numerical quadrature? The first option was favored in earlier times when computing power was sparse, and the second one is favored today (even in the pocket calculator).

Despite the impossibility to reach generality, it is it is possible (and useful) to compute primitive functions analytically in certain cases, and in this chapter, we collect some tricks that have proved useful for doing this. The tricks we present are basically various clever substitutions together with integration by parts. We have no ambition to be encyclopedic. We refer to Mathematics Handbook for Science and Engineering for further development.

We start with rational functions, and then proceed to various combinations of polynomials, logarithms, exponentials and trigonometric functions.

34.2 Rational Functions: The Simple Cases

Integration of rational functions depends on three basic formulas

$$\int_{x_0}^{x} \frac{1}{s-c}\,ds = \log|x-c| - \log|x_0-c|, \quad c \neq 0 \tag{34.1}$$

$$\int_{x_0}^{x} \frac{s-a}{(s-a)^2+b^2}\,dx = \frac{1}{2}\log((x-a)^2+b^2) - \frac{1}{2}\log((x_0-a)^2+b^2) \tag{34.2}$$

and

$$\int_{x_0}^{x} \frac{1}{(s-a)^2+b^2}\,ds = \left[\frac{1}{b}\arctan\left(\frac{x-a}{b}\right)\right] - \left[\frac{1}{b}\arctan\left(\frac{x_0-a}{b}\right)\right], b \neq 0. \tag{34.3}$$

These formulas can be verified by differentiation. Using the formulas can be straightforward as in

Example 34.1.

$$\int_6^8 \frac{ds}{s-4} = \log 4 - \log 2 = \log 2.$$

Or more complicated as in

Example 34.2.

$$\begin{aligned}\int_2^4 \frac{ds}{2(s-2)^2+6} &= \frac{1}{2}\int_2^4 \frac{ds}{(s-2)^2+3} \\ &= \frac{1}{2}\int_2^4 \frac{ds}{(s-2)^2+(\sqrt{3})^2} \\ &= \frac{1}{2}\left(\frac{1}{\sqrt{3}}\arctan\left(\frac{4-2}{\sqrt{3}}\right) - \frac{1}{\sqrt{3}}\arctan\left(\frac{2-2}{\sqrt{3}}\right)\right).\end{aligned}$$

Of course we may combine these formulas with substitution:

Example 34.3.

$$\int_0^x \frac{\cos(s)\,ds}{\sin(s)+2} = \int_0^{\sin(x)} \frac{du}{u+2} = \log|\sin(x)+2| - \log 2.$$

Using (34.2) and (34.3) may require *completing the square*, as we now show in

Example 34.4. For example, consider

$$\int_0^3 \frac{ds}{s^2-2s+5}.$$

We want to get s^2-2s+5 into the form $(s-a)^2+b^2$ if possible. We set

$$(s-a)^2+b^2 = s^2-2as+a^2+b^2 = s^2-2s+5.$$

Equating the coefficients of s on both sides gives $a=1$. Equating the constant terms on both sides gives $b^2 = 5-1 = 4$ and therefore we may take $b=2$. After a little practice with completing the square, we can often argue directly, as

$$s^2-2s+5 = s^2-2s+1^2-1^2+5 = (s-1)^2+2^2.$$

Returning to the integral, we have

$$\begin{aligned}\int_0^3 \frac{ds}{s^2-2s+5} &= \int_0^3 \frac{ds}{(s-1)^2+2^2} \\ &= \frac{1}{2}\arctan\left(\frac{3-2}{2}\right) - \frac{1}{2}\arctan\left(\frac{0-2}{2}\right).\end{aligned}$$

34.3 Rational Functions: Partial Fractions

We now investigate a systematic method for computing integrals of *rational* functions $f(x)$, i.e. functions of the form $f(x) = p(x)/q(x)$, where $p(x)$ and $q(x)$ are polynomials. The method is based manipulating the integrand so that the basic formulas (34.1)–(34.3) can be used. The manipulation is based on the observation that it is possible to write a complicated rational function as a sum of relatively simple rational functions.

Example 34.5. Consider the integral

$$\int_4^5 \frac{s^2+s-2}{s^3-3s^2+s-3}\,ds.$$

The integrand can be expanded

$$\frac{s^2+s-2}{s^3-3s^2+s-3} = \frac{1}{s^2+1} + \frac{1}{s-3}$$

which we can verify by adding the two fractions on the right after computing a common denominator,

$$\begin{aligned}\frac{1}{s^2+1} + \frac{1}{s-3} &= \frac{s-3}{s-3} \times \frac{1}{s^2+1} + \frac{s^2+1}{s^2+1} \times \frac{1}{s-3} \\ &= \frac{s-3+s^2+1}{(s^2+1)(s-3)} = \frac{s^2+s-2}{s^3-3s^2+s-3}.\end{aligned}$$

Therefore we can integrate

$$\begin{aligned}\int_4^5 \frac{s^2+s-2}{s^3-3s^2+s-3}\,ds &= \int_4^5 \frac{1}{s^2+1}\,ds + \int_4^5 \frac{1}{s-3}\,ds \\ &= (\arctan(5) - \arctan(4)) + (\log(5-3) - \log(4-3)).\end{aligned}$$

The general technique of *partial fractions* is based on a systematic method for writing a rational function as a sum of simple rational functions that can be integrated with the basic formulas (34.1)–(34.3). The method is analogous to "reversing" the addition of rational functions by finding a common denominator.

Applying the technique of partial fractions to a general rational function has several steps, which we explain in "reverse" order. So we begin by assuming that the numerator $p(x)$ of the rational function $p(x)/q(x)$ has smaller degree than the denominator $q(x)$, i.e. $\deg p(x) < \deg q(x)$, and that $q(x)$ has the form

$$\frac{p(x)}{q(x)} = \frac{p(x)}{k(x-c_1)\cdots(x-c_n)((x-a_1)^2+b_1^2)\cdots((x-a_m)^2+b_m^2)}, \quad (34.4)$$

where k is a number, the c_i are the real roots of $q(x)$, and the second degree factors $(x-a_j)^2+b_j^2$ correspond to the complex roots $a_j \pm ib_j$ of $q(x)$ that necessarily come in pairs of complex conjugates. We call polynomials of the form $(x-a_j)^2+b_j^2$ *irreducible* because we cannot factor them as a product of linear polynomials with *real* coefficients.

In the first instance, we assume that the zeroes $\{c_i\}$ and $\{a_j \pm ib_j\}$ are distinct. In this case, we rewrite $p(x)/q(x)$ as the sum of partial fractions

$$\begin{aligned}\frac{p(x)}{q(x)} =& \frac{C_1}{x-c_1} + \cdots + \frac{C_n}{x-c_n} \\ &+ \frac{A_1(x-a_1)+B_1}{(x-a_1)^2+b_1^2} + \cdots + \frac{A_m(x-a_m)+B_m}{(x-a_m)^2+b_m^2}, \end{aligned} \quad (34.5)$$

for some constants C_i, $1 \le j \le n$, and A_j, B_j, $1 \le j \le m$ that we have to determine. The motivation to rewrite $p(x)/q(x)$ in this way is that we can then compute an integral of $p(x)/q(x)$ by applying the formulas (34.1)–(34.3) to integrate the individual terms on the right-hand side of (34.5) as in the example above.

Example 34.6. For $p(x) = q(x) = (x-1)/(x^2 - x - 2)$ with $q(x) = (x-2)(x+1)$ we have

$$\frac{x-1}{x^2-x-2} = \frac{x-1}{(x-2)(x+1)} = \frac{1/3}{x-2} + \frac{2/3}{x+1},$$

and thus

$$\begin{aligned}\int_{x_0}^{x} \frac{s-1}{s^2-s-2}\,ds &= \frac{1}{3}\int_{x_0}^{x} \frac{1}{s-2}\,ds + \frac{2}{3}\int_{x_0}^{x} \frac{1}{s+1}\,ds \\ &= \frac{1}{3}[\log(s-2)]_{s=x_0}^{s=x} + \frac{2}{3}[\log(s+1)]_{s=x_0}^{s=x}.\end{aligned}$$

The rationale for the expansion (34.5) is simply that if we ask for the most general sum of rational functions with denominators of degrees 1 and 2 that can yield $p(x)/q(x)$, where $q(x)$ is the common denominator for the sum, then we get precisely the right-hand side of (34.5). In particular if the terms on the right had numerators of any higher degree, then $p(x)$ would have to have degree greater than $q(x)$.

The constants C_i, A_j and B_j in (34.5) can be found by rewriting the right-hand side of (34.5) with a common denominator.

Example 34.7. In the last example with $q(x) = (x-2)(x+1)$, we find that

$$\frac{C_1}{x-2} + \frac{C_2}{x+1} = \frac{C_1(x+1) + C_2(x-2)}{(x-2)(x+1)} = \frac{(C_1+C_2)x + (C_1 - 2C_2)}{(x-2)(x+1)},$$

which equals

$$\frac{x-1}{(x-2)(x+1)}$$

if and only if

$$C_1 + C_2 = 1 \quad \text{and} \quad C_1 - 2C_2 = -1,$$

that is if $C_1 = 1/3$ and $C_2 = 2/3$.

Since it is cumbersome to compute the constants by dealing with the fractions, we usually rewrite the problem by multiplying both sides of (34.5) by the common denominator.

Example 34.8. We multiply both sides of

$$\frac{x-1}{(x-2)(x+1)} = \frac{C_1}{x-2} + \frac{C_2}{x+1}$$

by $(x-2)(x+1)$ to get

$$x-1 = C_1(x+1)C_2(x-2) = (C_1+C_2)x + (C_1-2C_2).$$

Equating coefficients, we find $C_1+C_2=1$ and $C_1-2C_2=-1$, which yields $C_1=1/3$ and $C_2=2/3$.

Example 34.9. To integrate $f(x) = (5x^2-3x+6)/((x-2)(\text{TS}^{\text{O}}(x+1)^2+2^2))$, we begin by writing the partial fraction expansion

$$\frac{5x^2-3x+6}{(x-2)((x+1)^2+2^2)} = \frac{C}{x-2} + \frac{A(x+1)+B}{(x+1)^2+2^2}.$$

To determine the constants, we multiply both sides by $(x-2)((x+1)^2+2^2))$ to obtain

$$\begin{aligned} 5x^2-3x+6 &= C((x+1)^2+2^2) + (A(x+1)+B)(x-2) \\ &= (C+A)x^2 + (2C-2A+B)x + (4C-2A-2B). \end{aligned}$$

Equating coefficients, we find that $C+A=0$, $2C-2A+B=1$ and $5C-2A-2B=0$, that is $C=2$, $A=3$ and $B=-1$. Therefore we find that

$$\begin{aligned} &\int_{x_0}^{x} \frac{5s^2-3s+6}{(s-2)((s+1)^2+2^2))\text{TS}^{\text{P}}}\,ds \\ &\quad = 2\int_{x_0}^{x} \frac{1}{s-2}\,ds + \int_{x_0}^{x} \frac{3(s+1)-1}{(s+1)^2+2^2}\,ds \\ &\quad = 2\int_{x_0}^{x} \frac{1}{s-2}\,ds + 3\int_{x_0}^{x} \frac{s+1}{(s+1)^2+2^2}\,ds - \int_{x_0}^{x} \frac{1}{(s+1)^2+2^2}\,ds \\ &\quad = 2\big(\log|x-2| - \log|x_0-2|\big) \\ &\qquad\qquad + \frac{3}{2}\big(\log((x+1)^2+4) - \log((x_0+1)^2+4)\big) \\ &\qquad\qquad - \frac{1}{2}\left(\arctan\left(\frac{x+1}{2}\right) - \arctan\left(\frac{x_0+1}{2}\right)\right). \end{aligned}$$

In the case that some of the factors in the factorization of the denominator (34.4) are repeated, i.e. some of the roots have multiplicity greater than one, then we have to modify the partial sum expansion (34.5). We do not write out a general case because it is a mess and nearly unreadable, we just note that the principle for determining the correct partial fractions is always to write down the most general sum that can give the indicated common denominator.

Example 34.10. The general partial fraction expansion of $f(x) = x^2/((x-2)(x+1)^2)$ has the form

$$\frac{x^2}{(x-2)(x+1)^2} = \frac{C_1}{x-2} + \frac{C_{2,1}}{x+1} + \frac{C_{2,2}}{(x+1)^2},$$

TS^O Please check this left parenthesis.

TS^P Please check this right parenthesis.

Editor's or typesetter's annotations (will be removed before the final TEX run)

for constants C_1, $C_{2,1}$ and $C_{2,2}$ because all of the terms on the right-hand will yield the common denominator $(x-2)(x+1)^2$. Multiplying both sides by the common denominator and equating coefficients as usual, we find that $C_1 = 4/9$, $C_{2,1} = 5/9$ and $C_{2,2} = -3/9$.

In general if $q(x)$ has the multiple factor $(x-c_i)^L$ the term $\frac{C_i}{x-c_i}$ in the partial fraction expansion (34.5) should be replaced by the *sum* of fractions $\sum_{l=1}^{l=L} \frac{C_{i,l}}{(x-c)^l}$. There is a corresponding procedure for multiple factors of the form $((x-a)^2+b^2)^L$.

We have discussed how to integrate rational functions $p(x)/q(x)$ where $\deg p < \deg q$ and q is factored into a product of linear and irreducible quadratic polynomials. Now we discuss removing these restrictions. First we deal with the factorization of the denominator $q(x)$. The Fundamental Theorem of Algebra says that a polynomial q of degree n with real coefficients has exactly n roots and hence it can be factored into a product of n linear polynomials with possibly complex coefficients. However, because the polynomial q has real coefficients, the complex roots always come in complex conjugate pairs, i.e. if r is a root of q then so is $\bar{r}$. This means that there are an even number of linear factors of q corresponding to complex roots and furthermore we can combine the factors corresponding to conjugate roots to get quadratic factors with real coefficients. For example, $(x-3+i)(x-3-i) = (x-3)^2+1$. Therefore every polynomial $q(x)$ can theoretically be factored into a product $k(x-c_1)\cdots(x-c_n)((x-a_1)^2+b_1^2)\cdots((x-a_m)^2+b_m^2)$.

However, we caution that this theoretical result does not carry over in practice to situations in which the degree of q is large. To determine the factorization of q, we must determine the roots of q. In the problems and examples, we stick to cases in which the roots are simple, relatively small integers. But in general we know that the roots can be any kind of algebraic number which we can only approximate. Unfortunately it turns out that it is extremely difficult to determine the roots of a polynomial of high degree, even using Newton's method. So the method of partial fractions is used only for low degree polynomials in practice, though it is a very useful theoretical tool.

Finally we remove the restriction that $\deg p < \deg q$. When the degree of the numerator polynomial $p(x)$ is $\geq$ the degree of the denominator polynomial $q(x)$, we first use polynomial division to rewrite $f(x)$ as the sum of a polynomial $s(x)$ and a rational function $\frac{r(x)}{q(x)}$ for which the degree of the numerator $r(x)$ is *less* than the degree of the denominator $q(x)$.

Example 34.11. For $f(x) = (x^3-x)/(x^2+x+1)$, we divide to get $f(x) = x-1+(1-x)/(x^2+x+1)$, so that

$$\int_0^{\bar{x}} \frac{x^3}{x^2+x+1}\,dx = \left[\frac{1}{2}x^2 - x\right]_{x=0}^{x=\bar{x}} + \int_0^{\bar{x}} \frac{1-x}{x^2+x+1}\,dx.$$

34.4 Products of Polynomial and Trigonometric or Exponential Functions

To integrate the product of a polynomial and a trigonometric or exponential function, we use integration by parts repeatedly to reduce the polynomial to an constant.

Example 34.12. To compute a primitive function of $x\cos(x)$, we inte- grate by parts once

$$\int_0^x y\cos(y)\,dy = [y\sin(y)]_{y=0}^{y=x} - \int_0^x \sin(y)\,dy = x\sin(x) + \cos(x) + 1.$$

To handle higher order polynomials, we use integration by parts several times.

Example 34.13. We have

$$\begin{aligned}\int_0^x s^2e^s\,ds &= s^2(e^s)_{s=0}^{s=x} - 2\int_0^x se^s\,ds \\ &= [s^2e^s]_{s=0}^{s=x} - 2\left([se^s]_{s=0}^{s=x} - \int_0^x e^s\,ds\right) \\ &= [s^2e^s)_{s=0}^{s=x} - 2\big([se^s]_{s=0}^{s=x} - [e^s]_{s=0}^{s=x}\big) \\ &= x^2e^x - 2xe^x + 2e^x - 2.\end{aligned}$$

34.5 Combinations of Trigonometric and Root Functions

To compute a primitive function of $\sin(\sqrt{y})$ for $x > 0$, we set $y = t^2$ and obtain by using partial integration

$$\begin{aligned}\int_0^x \sin(\sqrt{y})\,dy &= \int_0^{\sqrt{x}} 2t\sin(t)\,dt = [-2t\cos(t)]_{t=0}^{t=\sqrt{x}} + 2\int_0^{\sqrt{x}} \cos(t)\,dt \\ &= -2\sqrt{x}\cos(\sqrt{x}) + 2\sin(\sqrt{x}).\end{aligned}$$

TS[q] Please check this right parenthesis.

Editor's or typesetter's annotations (will be removed before the final TEX run)

34.6 Products of Exponential and Trigonometric Functions

To compute a primitive function of $e^y \sin(y)$, we use repeated integration by parts as follows

$$\begin{aligned}\int_0^x e^y \sin(y)\,dy &= [e^y \sin(y)]_{y=0}^{y=x} - \int_0^x e^y \cos(y)\,dy \\ &= e^x \sin(x) - [e^y \cos(y)]_{y=0}^{y=x} - \int_0^x e^y \sin(y)\,dy,\end{aligned}$$

which shows that

$$\int_0^x e^y \sin(y)\,dy = \frac{1}{2}(e^x \sin(x) - e^x \cos(x) + 1)$$

34.7 Products of Polynomials and Logarithm Functions

To compute a primitive function of $x^2 \log(x)$, we integrate by parts:

$$\int_1^x y^2 \log(y)\,dy = \left[\frac{y^3}{3}\log(y)\right]_{y=1}^{y=x} - \int_1^x \frac{y^3}{3}\frac{1}{y}\,dy = \frac{x^3}{3}\log(x) - \frac{x^3}{9} + \frac{1}{9}.$$

Chapter 34 Problems

34.1. Compute (a) $\int_0^x t\sin(2t)\,dt$ (b) $\int_0^x t^2\cos(t)\,dt$ (c) $int_0^x t\exp(-2t)\,dt$. Hint: Integrate by parts.

34.2. Compute (a) $\int_1^x y\log(y)\,dy$ (b) $\int_1^x \log(y)\,dy$ (c) $\int_0^x \arctan(t)\,dt$ (d) $\int_0^x \exp(-t)\cos(2t)\,dt$. Hint: Integrate by parts.

34.3. Compute using the formula $\int_0^x \frac{g'(y)}{g(y)}\,dy = \log(g(x)) - \log(g(0))$ the following integrals. (a) $\int_0^x \frac{y}{y^2+1}\,dy$ (b) $\int_0^x \frac{e^t}{e^t+1}\,dt$.

34.4. Compute by a suitable change of variable (a) $\int_0^x y\exp(y^2)\,dy$ (b) $\int_0^x y\sqrt{y-1}\,dy$ (c) $\int_0^x \sin(t)\cos^2(t)\,dt$.

34.5. Compute (a) $\int_0^x \frac{dy}{y^2-y-2}\,dy$ (b) $\int_0^x \frac{y^3}{y^2+2y-3}\,dy$ (c) $\int_0^x \frac{dy}{y^2+2y+5}\,dy$ (d) $\int_0^x \frac{x-x^2}{(y-1)(y^2+2y+5)}\,dy$ (e) $\int_0^x \frac{x^4}{(x-1)(x^2+x-6)}\,dy$.

34.6. Recalling that a function is called *even* if $f(-x) = f(x)$ and *odd* if $f(-x) = -f(x)$ for all x, (a) give examples of even and odd functions (b) sketch their graphs, and (c) show that

$$\int_{-a}^{a} f(x)\,dx = 2\int_{0}^{a} f(x)\,dx \text{ if } f \text{ is even}, \quad \int_{-a}^{a} f(x)\,dx = 0 \text{ if } f \text{ is odd}. \quad (34.6)$$

34.7. Compute (a) $\int_{-\pi}^{\pi} |x|\cos(x)\,dx$ (b) $\int_{-\pi}^{\pi} \sin^2(x)\,dx$ (c) $\int_{-\pi}^{\pi} x\sin^2(x)\,dx$ (d) $\int_{-\pi}^{\pi} \arctan(x+3x^3)\,dx$.

35

Solving Differential Equations Using the Exponential

> ... he climbed a little further... and further... and then just a little further. (Winnie-the-Pooh)

35.1 Introduction

The exponential function plays a fundamental role in modeling and analysis because of its basic properties. In particular it can be used to solve a variety of differential equations analytically as we show in this chapter. We start with generalizations of the initial value problem (31.2) from Chapter *The exponential function*:

$$u'(x) = \lambda u(x) \quad \text{for } x > a, \quad u(a) = u_a, \tag{35.1}$$

where $\lambda \in \mathbb{R}$ is a constant, with solution

$$u(x) = \exp(\lambda(x-a))u_a \quad \text{for } x \geq a. \tag{35.2}$$

Analytic solutions formulas may give very important information and help the intuitive understanding of different aspects of a mathematical model, and should therefore be kept as valuable gems in the scientist and engineer's tool-bag. However, useful analytical formulas are relatively sparse and must be complemented by numerical solutions techniques. In the Chapter *The General Initial Value Problem* we extend the constructive numerical method for solving (35.1) to construct solutions of general initial value problems for systems of differential equations, capable of modeling a very

large variety of phenomena. We can thus numerically compute the solution to just about any initial value problem, with more or less computational work, but we are limited to computing one solution for each specific choice of data, and getting qualitative information for a variety of different data may be costly. On the other hand, an analytical solution formula, when available, may contain this qualitative information for direct information.

An analytical solution formula for a differential equation may thus be viewed as a (smart and beautiful) short-cut to the solution, like evaluating an integral of a function by just evaluating two values of a corresponding primitive function. On the other hand, numerical solution of a differential equation is like a walk along a winding mountain road from point A to point B, without any short-cuts, similar to computing an integral by numerical quadrature. It is useful to be able to use both approaches.

35.2 Generalization to $u'(x) = \lambda(x)u(x) + f(x)$

The first problem we consider is a model in which the rate of change of a quantity $u(x)$ is proportional to the quantity with a variable factor of proportionality $\lambda(x)$, and moreover in which there is an external "forcing" function $f(x)$. The problem reads:

$$u'(x) = \lambda(x)u(x) + f(x) \quad \text{for } x > a, \quad u(a) = u_a, \tag{35.3}$$

where $\lambda(x)$ and $f(x)$ are given functions of x, and u_a is a given initial value. We first describe a couple physical situations being modeled by (35.3).

Example 35.1. Consider for time $t > 0$ the population $u(t)$ of rabbits in West Virginia with inital value $u(0) = u_0$ given, which we assume has time dependent known birth rate $\beta(t)$ and death rate $\delta(t)$. In general, we would expect that rabbits will migrate quite freely back and forth across the state border and that the rates of the migration would vary with the season, i.e. with time t. We let $f_i(t)$ and $f_o(t)$ denote the rate of migration into and out of the state respectively at time t, which we assume to be known (realistic?). Then the population $u(t)$ will satisfy

$$\dot{u}(t) = \lambda(t)u(t) + f(t), \quad \text{for } t > a, \quad u(a) = u_a, \tag{35.4}$$

with $\lambda(t) = \beta(t) - \delta(t)$ and $f(t) = f_i(t) - f_o(t)$, which is of the form (35.3). Recall that $\dot{u} = \frac{du}{dt}$.

Example 35.2. We model the amount of solute such as salt in a solvent such as water in a tank in which there is both inflow and outflow, see Fig. 35.1. We let $u(t)$ denote the amount of solute in the tank at time t and suppose that we know the initial amount u_0 at $t = 0$. We suppose that a mixture of solute/solvent, of concentration C_i in say grams per liter, flows into the

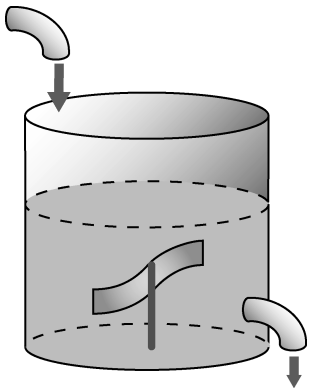

Fig. 35.1. An illustration of a chemical mixing tank

tank at a rate σ_i liters per second. We assume there is also outflow at a rate of σ_o liters per second, and we assume that the mixture in the tank is well mixed with a uniform concentration $C(t)$ at any time t.

To get a differential equation for $u(t)$, we compute the change $u(t+\Delta t) - u(t)$ during the interval $[t, t+\Delta t]$. The amount of solute that flows into the tank during that time interval is $\sigma_i C_i \Delta t$, while the amount of solute that flows out of the tank during that time equals $\sigma_o C(t)\Delta t$, and thus

$$u(t+\Delta t) - u(t) \approx \sigma_i C_i \Delta t - \sigma_o C(t)\Delta t, \tag{35.5}$$

where the approximation improves when we decrease Δt. Now the concentration at time t will be $C(t) = u(t)/V(t)$ where $V(t)$ is the volume of fluid in the tank at time t. Substituting this into (35.5) and dividing by Δt gives

$$\frac{u(t+\Delta t) - u(t)}{\Delta t} \approx \sigma_i C_i - \sigma_o \frac{u(t)}{V(t)}$$

and taking the limit $\Delta t \to 0$ assuming $u(t)$ is differentiable gives the following differential equation for u,

$$\dot{u}(t) = -\frac{\sigma_o}{V(t)} u(t) + \sigma_i C_i.$$

The volume $V(t)$ is determined simply by the flow rates of fluid in and out of the tank. If there is initially V_0 liters in the tank then at time t, $V(t) = V_0 + (\sigma_i - \sigma_o)t$ because the flow rates are assumed to be constant. This gives again a model of the form (35.3):

$$\dot{u}(t) = -\frac{\sigma_o}{V_0 + (\sigma_i - \sigma_o)t} u(t) + \sigma_i C_i \quad \text{for } t > 0, \quad u(0) = u_0. \tag{35.6}$$

The Method of Integrating Factor

We now return to derive an analytical solution formula for (35.3), using the method of *integrating factor*. To work out the solution formula, we begin with the special case

$$u'(x) = \lambda(x)u(x) \quad \text{for } x > a, \quad u(a) = u_a, \tag{35.7}$$

where $\lambda(x)$ is a given function of x. We let $\Lambda(x)$ be a primitive function of $\lambda(x)$ such that $\Lambda(a) = 0$, assuming that $\lambda(x)$ is Lipschitz continuous on $[a, \infty)$. We now multiply the equation $0 = u'(x) - \lambda(x)u(x)$ by $\exp(-\Lambda(x))$, and we get

$$0 = u'(x)\exp(-\Lambda(x)) - u(x)\exp(-(x))\lambda(x) = \frac{d}{dx}(u(x)\exp(-\Lambda(x))),$$

where we refer to $\exp(-\Lambda(x))$ as an integrating factor because it brought the given equation to the form $\frac{d}{dx}$ of something, namely $u(x)\exp(-\Lambda(x))$, equal to zero. We conclude that $u(x)\exp(-\Lambda(x))$ is constant and is therefore equal to u_a since $u(a)\exp(-\Lambda(a)) = u(a) = u_a$. In other words, the solution to (35.7) is given by the formula

$$u(x) = \exp(\Lambda(x))u_a = e^{\Lambda(x)}u_a \quad \text{for } x \geq a. \tag{35.8}$$

We can check by differentiation that this function satisfies (35.7), and thus by uniqueness is the solution. To sum up, we have derived a solution formula for (35.7) in terms of the exponential function and a primitive function $\Lambda(x)$ of the coefficient $\lambda(x)$.

Example 35.3. If $\lambda(x) = \frac{r}{x}$ and $a = 1$ then $\Lambda(x) = r\log(x) = \log(x^r)$, and the solution of

$$u'(x) = \frac{r}{x}u(x) \quad \text{for } x \neq 1, \quad u(1) = 1, \tag{35.9}$$

is according to (35.8) given by $u(x) = \exp(r\log(x)) = x^r$. We may define x^r for

Duhamel's Principle

We now continue with the general problem to (35.3). We multiply by $e^{-\Lambda(x)}$, where again $\Lambda(x)$ is the primitive function of $\lambda(x)$ satisfying $\Lambda(a) = 0$, and get

$$\frac{d}{dx}\left(u(x)e^{-\Lambda(x)}\right) = f(x)e^{-\Lambda(x)}.$$

Integrating both sides, we see that the solution $u(x)$ satisfying $u(a) = u_a$ can be expressed as

$$u(x) = e^{\Lambda(x)}u_a + e^{\Lambda(x)}\int_a^x e^{-\Lambda(y)}f(y)\,dy. \tag{35.10}$$

This formula for the solution $u(x)$ of (35.3), expressing $u(x)$ in terms of the given data u_a and the primitive function $\Lambda(x)$ of $\lambda(x)$ satisfying $\Lambda(a) = 0$, is referred to as *Duhamel's principle* or the *variation of constants formula*.

We can check the validity of (35.10) by directly computing the derivative of $u(x)$:

$$\begin{aligned} u'(x) &= \lambda e^{\Lambda(x)} u_a + f(x) + \int_0^x (\lambda(x) e^{\Lambda(x)-\Lambda(y)} f(y)\, dy \\ &= \lambda(x) \left(e^{\Lambda(x)} u_a + \int_0^x e^{\Lambda(x)-\Lambda(y)} f(y)\, dy \right) + f(x). \end{aligned}$$

Example 35.4. If $\lambda(x) = \lambda$ is constant, $f(x) = x$, $a = 0$ and $u_0 = 0$, the solution of (35.3) is given by

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

Example 35.5. In the model of the rabbit population (35.4), consider a situation with an initial population of 100, the death rate is greater than the birth rate by a constant factor 4, so $\lambda(t) = \beta(t) - \delta(t) = -4$, and there is a increasing migration into the state, so $f(t) = f_i(t) - f_o(t) = t$. Then (35.10) gives

$$\begin{aligned} u(t) &= e^{-4t} 100 + e^{-4t} \int_0^t e^{4s} s\, ds \\ &= e^{-4t} 100 + e^{-4t} \left(\frac{1}{4} s e^{4s} |_0^t - \frac{1}{4} \int_0^t e^{4s}\, ds \right) \\ &= e^{-4t} 100 + e^{-4t} \left(\frac{1}{4} t e^{4t} - \frac{1}{16} e^{4t} + \frac{1}{16} \right) \\ &= 100.0625 e^{-4t} + \frac{t}{4} - \frac{1}{16}. \end{aligned}$$

Without the migration into the state, the population would decrease exponentially, but in this situation the population decreases only for a short time before beginning to increase at a linear rate.

Example 35.6. Consider a mixing tank in which the input flow at a rate of $\sigma_i = 3$ liters/sec has a concentration of $C_i = 1$ grams/liter, and the outflow is at a rate of $\sigma_o = 2$ liters/sec, the initial volume is $V_0 = 100$ liters with no solute dissolved, so $u_0 = 0$. The equation is

$$\dot{u}(t) = -\frac{2}{100 + t} u(t) + 3.$$

TS[F] Please check this left parenthesis.

Editor's or typesetter's annotations (will be removed before the final TEX run)

We find $\Lambda(t) = 2\ln(100+t)$ and so

$$\begin{aligned} u(t) &= 0 + e^{2\ln(100+t)} \int_0^t e^{-2\ln(100+s)} 3\, ds \\ &= (100+t)^2 \int_0^t (100+s)^{-2} 3\, ds \\ &= (100+t)^2 \left(\frac{-3}{100+t} + \frac{3}{100} \right) \\ &= \frac{3}{100} t(100+t). \end{aligned}$$

As expected from the conditions, the concentration increases steadily until the tank is full.

35.3 The Differential Equation $u''(x) - u(x) = 0$

Consider the second order initial value problem

$$u''(x) - u(x) = 0 \quad \text{for } x > 0,\ u(0) = u_0,\ u'(0) = u_1, \tag{35.11}$$

with two initial conditions. We can write the differential equation $u''(x) - u(x) = 0$ formally as

$$(D+1)(D-1)u = 0,$$

where $D = \frac{d}{dx}$, since $(D+1)(D-1)u = D^2u - Du + Du - u = D^2u - u$. Setting $w = (D-1)u$, we thus have $(D+1)w = 0$, which gives $w(x) = ae^{-x}$ with $a = u_1 - u_0$, since $w(0) = u'(0) - u(0)$. Thus, $(D-1)u = (u_1 - u_0)e^{-x}$, so that by Duhamel's principle

$$\begin{aligned} u(x) &= e^x u_0 + \int_0^x e^{x-y}(u_1 - u_0)e^{-y}\, dy \\ &= \frac{1}{2}(u_0 + u_1)e^x + \frac{1}{2}(u_0 - u_1)e^{-x}. \end{aligned}$$

We conclude that the solution $u(x)$ of $u''(x) - u(x) = 0$ is a linear combination of e^x and e^{-x} with coefficients determined by the initial conditions. The technique of "factoring" the differential equation $(D^2 - 1)u = 0$ into $(D+1)(D-1)u = 0$, is very powerful and we now proceed to follow up this idea.

35.4 The Differential Equation $\sum_{k=0}^{n} a_k D^k u(x) = 0$

In this section, we look for solutions of the *linear differential equation with constant coefficients*:

$$\sum_{k=0}^{n} a_k D^k u(x) = 0 \quad \text{for } x \in I, \tag{35.12}$$

where the coefficients a_k are given real numbers, and I is a given interval. Corresponding to the *differential operator* $\sum_{k=0}^{n} a_k D^k$, we define the polynomial $p(x) = \sum_{k=0}^{n} a_k x^k$ in x of degree n with the same coefficients a_k as the differential equation. This is called the *characteristic polynomial* of the differential equation. We can now express the differential operator formally as

$$p(D)u(x) = \sum_{k=0}^{n} a_k D^k u(x).$$

For example, if $p(x) = x^2 - 1$ then $p(D)u = D^2 u - u$.

The technique for finding solutions is based on the observation that the exponential function $\exp(\lambda x)$ has the following property:

$$p(D)\exp(\lambda x) = p(\lambda)\exp(\lambda x), \tag{35.13}$$

which follows from repeated use of the Chain rule. This translates the differential operator $p(D)$ acting on $\exp(\lambda x)$ into the simple operation of multiplication by $p(\lambda)$. Ingenious, right?

We now seek solutions of the differential equation $p(D)u(x) = 0$ on an interval I of the form $u(x) = \exp(\lambda x)$. This leads to the equation

$$p(D)\exp(\lambda x) = p(\lambda)\exp(\lambda x) = 0, \quad \text{for } x \in I,$$

that is, λ should be a root of the polynomial equation

$$p(\lambda) = 0. \tag{35.14}$$

This algebraic equation is called the *characteristic equation* of the differential equation $p(D)u = 0$. To find the solutions of a differential equation $p(D)u = 0$ on the interval I, we are thus led to search for the roots $\lambda_1, \ldots \lambda_n$, of the algebraic equation $p(\lambda) = 0$ with corresponding solutions $\exp(\lambda_1 x), \ldots, \exp(\lambda_n x)$. Any linear combination

$$u(x) = \alpha_1 \exp(\lambda_1 x) + \ldots + \alpha_n \exp(\lambda_n x), \tag{35.15}$$

with α_i real (or complex) constants, will then be a solution of the differential equation $p(D)u = 0$ on I. If there are n distinct roots $\lambda_1, \ldots, \lambda_n$, then the *general solution* of $p(D)u = 0$ has this form. The constants α_i will be determined from initial or boundary conditions in a specific situation.

If the equation $p(\lambda) = 0$ has a multiple roots λ_i of multiplicity r_i, then the situation is more complicated. It can be shown that the solution is a sum of terms of the form $q(x)\exp(\lambda_i x)$, where $q(x)$ is a polynomial of degree at most $r_i - 1$. For example, if $p(D) = (D-1)^2$, then the general solution of $p(D)u = 0$ has the form $u(x) = (\alpha_0 + \alpha_1 x)\exp(x)$. In the Chapter *N-body systems* below we study the the constant coefficient linear second order equation $a_0 + a_1 Du + a_2 D^2 u = 0$ in detail, with interesting results!

The translation from a differential equation $p(D)u = 0$ to an algebraic equation $p(\lambda) = 0$ is very powerful, but requires the coefficients a_k of $p(D)$ to be independent of x and is thus not very general. The whole branch of *Fourier analysis* is based on the formula (35.13).

Example 35.7. The characteristic equation for $p(D) = D^2 - 1$ is $\lambda^2 - 1 = 0$ with roots $\lambda_1 = 1, \lambda_2 = -1$, and the corresponding general solution is given by $\alpha_1 \exp(x) + \alpha_2 \exp(-x)$. We already met this example just above.

Example 35.8. The characteristic equation for $p(D) = D^2 + 1$ is $\lambda^2 + 1 = 0$ with roots $\lambda_1 = i, \lambda_2 = -i$, and the corresponding general solution is given by

$$\alpha_1 \exp(ix) + \alpha_2 \exp(-ix).$$

with the α_i complex constants. Taking the real part, we get solutions of the form

$$\beta_1 \cos(x) + \beta_2 \sin(x)$$

with the β_i real constants.

35.5 The Differential Equation $\sum_{k=0}^{n} a_k D^k u(x) = f(x)$

Consider now the nonhomogeneous differential equation

$$p(D)u(x) = \sum_{k=0}^{n} a_k D^k u(x) = f(x), \tag{35.16}$$

with given constant coefficients a_k, and a given right hand side $f(x)$. Suppose $u_p(x)$ is any solution of this equation, which we refer to as a *particular solution*. Then any other solution $u(x)$ of $p(D)u(x) = f(x)$ can be written

$$u(x) = u_p(x) + v(x)$$

where $v(x)$ is a solution of the corresponding homogeneous differential equation $p(D)v = 0$. This follows from linearity and uniqueness since $p(D)(u - u_p) = f - f = 0$.

Example 35.9. Consider the equation $(D^2 - 1)u = f(x)$ with $f(x) = x^2$. A particular solution is given by $u_p(x) = -x^2 - 2$, and thus the general solution is given by

$$u(x) = -x^2 - 2 + \alpha_1 \exp(x) + \alpha_2 \exp(-x).$$

35.6 Euler's Differential Equation

In this section, we consider Euler's equation

$$a_0 u(x) + a_1 x u'(x) + a_2 x^2 u''(x) = 0, \tag{35.17}$$

which has variable coefficients $a_i x^i$ of a very particular form. Following a grand mathematical tradition, we guess, or make an *Ansatz* on the form of the solution, and assume that $u(x) = x^m$ for some m to be determined. Substituting into the differential equation, we get

$$a_0 x^m + a_1 x (x^m)' + a_2 x^2 (x^m)'' = (a_0 + (a_1 - 1)m + a_2 m^2) x^m,$$

and we are thus led to the auxiliary algebraic equation

$$a_0 + (a_1 - 1)m + a_2 m^2 = 0$$

in m. Letting the roots of this equation be m_1 and m_2, assuming the roots are real, any linear combination

$$\alpha_1 x^{m_1} + \alpha_2 x^{m_2}$$

Fig. 35.2. Leonard Euler: "...I soon found an opportunity to be introduced to a famous professor Johann Bernoulli...True, he was very busy and so refused flatly to give me private lessons; but he gave me much more valuable advice to start reading more difficult mathematical books on my own and to study them as diligently as I could; if I came across some obstacle or difficulty, I was given permission to visit him freely every Sunday afternoon and he kindly explained to me everything I could not understand..."

is a solution of (35.17). In fact the general solution of (35.17) has this form if m_1 and m_2 are distinct and real.

Example 35.10. The auxiliary equation for the differential equation $x^2u'' - \frac{3}{2}xu' - 2u = 0$ is $m^2 - \frac{7}{2}m - 2 = 0$ with roots $m_1 = -\frac{1}{2}$ and $m_2 = 4$ and thus the general solution takes the form

$$u(x) = \alpha_1 \frac{1}{\sqrt{x}} + \alpha_2 x^4.$$

Leonard Euler (1707-83) is the mathematical genius of the 18th century, with an incredible production of more than 800 scientific articles half of them written after he became completely blind in 1766, see Fig. 35.2.

Chapter 35 Problems

35.1. Solve the initial value problem (35.7) with $\lambda(x) = x^r$, where $r \in \mathbb{R}$, and $a = 0$.

35.2. Solve the following initial value problems: a) $u'(x) = 8xu(x), u(0) = 1,\ x > 0$, b) $\frac{(15x+1)u(x)}{u'(x)} = 3x,\ u(1) = \mathrm{e},\ x > 1$, c) $u'(x) + \frac{x}{(1-x)(1+x)}u = 0,\ u(0) = 1,\ x > 0$.

35.3. Make sure that you got the correct answer in the previous problem, part c). Will your solution hold for $x > 1$ as well as $x < 1$?

35.4. Solve the following initial value problems: a) $xu'(x) + u(x) = x,\ u(1) = \frac{3}{2},\ x > 1$, b) $u'(x) + 2xu = x,\ u(0) = 1,\ x > 0$, c) $u'(x) = \frac{x+u}{2},\ u(0) = 0,\ x > 0$.

35.5. Describe the behavior of the population of rabbits in West Virginia in which the birth rate exceeds the death rate by 5, the initial population is 10000 rabbits, and (a) there is a net migration out of the state at a rate of $5t$ (b) there is a net migration out of the state at a rate of $\exp(6t)$.

35.6. Describe the concentration in a mixing tank with an initial volume of 50 liters in which 20 grams of solute are dissolved, there is an inflow of 6 liters/sec with a concentration of 10 grams/liter and an outflow of 7 liters/sec.

36
Improper Integrals

All sorts of funny thoughts, run around my head. (When We Were Very Young, Milne)

36.1 Introduction

In some applications, it is necessary to compute integrals of functions that are unbounded at isolated points or to compute integrals of functions over unbounded intervals. We call such integrals *improper*, or sometimes (more properly) *generalized* integrals. We compute these integrals using the basic results on convergence of sequences that we have already developed.

We now consider these two kinds of improper integrals: integrals over unbounded intervals and integrals of unbounded functions.

36.2 Integrals Over Unbounded Intervals

We start considering the following example of an integral over the unbounded interval $[0, \infty)$:

$$\int_0^\infty \frac{1}{1+x^2}\,dx.$$

The integrand $f(x) = (1 + x^2)^{-1}$ is a smooth (positive) function that we can integrate over any finite interval $[0, n]$ to get,

$$\int_0^n \frac{1}{1+x^2}\,dx = \arctan(n). \tag{36.1}$$

Now we consider what happens as n increases, that is we integrate f over increasingly longer intervals. Since $\lim_{n\to\infty} \arctan(n) = \pi/2$, we may write

$$\lim_{n\to\infty} \int_0^n \frac{1}{1+x^2}\,dx = \frac{\pi}{2},$$

and we are thus led to *define*

$$\int_0^\infty \frac{1}{1+x^2}\,dx = \lim_{n\to\infty} \int_0^n \frac{1}{1+x^2}\,dx = \frac{\pi}{2}.$$

We generalize in the obvious way to an arbitrary (Lipschitz continuous) function $f(x)$ defined for $x > a$, and thus define

$$\int_a^\infty f(x)\,dx = \lim_{n\to\infty} \int_a^n f(x)\,dx \tag{36.2}$$

granted that the limit is defined and is finite. In this case, we say the improper integral is *convergent* (or is *defined*) and that the function $f(x)$ is *integrable* over $[a, \infty)$. Otherwise, we say the integral is *divergent* (or is *undefined*), and that $f(x)$ is *not* integrable over $[a, \infty)$.

If the function $f(x)$ is positive, then in order for the integral $\int_a^\infty f(x)\,dx$ to be convergent, the integrand $f(x)$ has to get sufficiently small for large values of x, since otherwise $\lim_{n\to\infty} \int_a^n f(x)\,dx = \infty$ and the integral is divergent. We saw above that the function $\frac{1}{1+x^2}$ was decaying to zero sufficiently quickly for large values of x to be integrable over $[a, \infty)$.

Consider now the function $\frac{1}{1+x}$ with a less quick decay as $x \to \infty$. Is it integrable on $[0, \infty)$? Well, we have

$$\int_0^n \frac{1}{1+x}\,dx = \Big[\log(1+x)\Big]_0^n = \log(1+n),$$

and since

$$\log(1+n) \to \infty \quad \text{as } n \to \infty$$

although the divergence is slow, we understand that $\int_0^\infty \frac{1}{1+x}\,dx$ is divergent.

Example 36.1. The improper integral

$$\int_1^\infty \frac{dx}{x^\alpha}$$

is convergent for $\alpha > 1$, since

$$\lim_{n\to\infty} \int_1^n \frac{dx}{x^{-\alpha}} = \lim_{n\to\infty} \left[-\frac{x^{-(\alpha-1)}}{\alpha-1}\right]_1^n = \frac{1}{\alpha-1}.$$

We can sometimes show that an improper integral exists even when we can not compute its value.

Example 36.2. Consider the improper integral

$$\int_1^\infty \frac{e^{-x}}{x}\,dx.$$

Since $f(x) = \frac{e^{-x}}{x} > 0$ for $x > 1$, we see that the sequence $\{I_n\}_{n=1}^\infty$, with

$$I_n = \int_1^n \frac{e^{-x}}{x}\,dx$$

is increasing. By Chapter *Optimization* we know that $\{I_n\}_{n=1}^\infty$ will have a limit if we only can show that $\{I_n\}_{n=1}^\infty$ is bounded above. Since trivially $1/x \leq 1$ if $x \leq 1$, we have for all $n \geq 1$

$$I_n \leq \int_1^n e^{-x}\,dx = e^{-1} - e^{-n} \leq e^{-1}.$$

We conclude that $\int_1^\infty \frac{e^{-x}}{x}\,dx$ converges. Note that we may restrict n to take integer values because the integrand e^{-x}/x tends to zero as x tends to infinity.

We may also compute integrals of the form

$$\int_{-\infty}^\infty f(x)\,dx.$$

We do this by choosing an arbitrary point $-\infty < a < \infty$ and defining

$$\begin{aligned}\int_{-\infty}^\infty f(x)\,dx &= \int_{-\infty}^a f(x)\,dx + \int_a^\infty f(x)\,dx \\ &= \lim_{m\to-\infty}\int_m^a f(x)\,dx + \lim_{n\to\infty}\int_a^n f(x)\,dx,\end{aligned} \tag{36.3}$$

where we compute the two limits independently and both must be defined and finite for the integral to exist.

36.3 Integrals of Unbounded Functions

We begin this section by considering the integral

$$\int_a^b f(x)\,dx,$$

where $f(x)$ is unbounded at a, i.e. $\lim_{x\downarrow a} f(x) = \pm\infty$. We consider the following example:

$$\int_0^1 \frac{1}{\sqrt{x}}\,dx.$$

The function $\frac{1}{\sqrt{x}}$ is unbounded on $(0,1]$, but bounded and Lipschitz continuous on $[\epsilon,1]$ for any $1 \geq \epsilon > 0$. This means that the integrals

$$I_\epsilon = \int_\epsilon^1 \frac{1}{\sqrt{x}}\,dx = 2 - 2\sqrt{\epsilon} \tag{36.4}$$

are defined for any $1 \geq \epsilon > 0$, and evidently

$$\lim_{\epsilon\downarrow 0 \to\infty} I_\epsilon = 2,$$

where we recall that $\epsilon \downarrow 0$ means that ϵ tends to zero through positive values. It is thus natural to define

$$\int_0^1 \frac{1}{\sqrt{x}}\,dx = \lim_{\epsilon\downarrow 0} \int_\epsilon^1 \frac{1}{\sqrt{x}}\,dx = 2.$$

In general if $f(x)$ is unbounded close to a, then we define

$$\int_a^b f(x)\,dx = \lim_{s\downarrow a} \int_s^b f(x)\,dx, \tag{36.5}$$

and if $f(x)$ is unbounded at b then we define

$$\int_a^b f(x)\,dx = \lim_{s\uparrow a} \int_a^s f(x)\,dx \tag{36.6}$$

when these limits are defined and finite. As above, we say the improper integrals are convergent and defined if the limits exist and are finite, and otherwise say the integrals are divergent and not defined.

We may naturally extend this definition to the case when $f(x)$ is unbounded at a point $a < c < b$ by defining

$$\begin{aligned}\int_a^b f(x)\,dx &= \lim \int_a^c f(x)\,dx + \lim \int_c^b f(x)\,dx \\ &= \lim_{s\uparrow c} \int_a^s f(x)\,dx + \lim_{t\downarrow c} \int_t^b f(x)\,dx\end{aligned} \tag{36.7}$$

where the two limits are computed independently and must both be defined and finite for the integral to converge.

Chapter 36 Problems

36.1. If possible, compute the following integrals

1. $\displaystyle\int_0^\infty \frac{x}{(1+x^2)^2}\,dx$
2. $\displaystyle\int_{-\infty}^\infty xe^{-x^2}\,dx$
3. $\displaystyle\int_0^1 \frac{1}{\sqrt{1-x}}\,dx$
4. $\displaystyle\int_0^\pi \frac{\cos(x)}{(1-\sin(x))^{1/3}}\,dx$

36.2. Prove that if $\int_0^\infty |f(x)|\,dx$ is convergent, then so is $\int_0^\infty f(x)\,dx$, that is, absolute convergence implies convergence.

36.3. Prove that $\int_B \|x\|^{-\alpha}\,dx$, where $B = \{x \in \mathbb{R}^d : \|x\| < 1\}$, is convergent if $\alpha < d$ for $d = 1, 2, 3$.

37
Series

> If you disregard the very simplest cases, there is in all of mathematics not a single series whose sum has been rigorously determined. In other words, the most important part of mathematics stand without a foundation. (Abel 1802–1829)

37.1 Introduction

In this chapter we consider the concept of *series*, which is a sum of numbers. We distinguish between a *finite* series, where the sum has a finite number of terms, and an *infinite series* with an infinite number of terms. A finite series does not pose any mysteries; we can, at least in principle, compute the sum of a finite series by adding the terms one-by-one, given enough time. The concept of an infinite series requires some explanation, since we cannot actually add an infinite number of terms one-by-one, and we thus need to define what we mean by an "infinite sum".

The concept of infinite series has a central role in Calculus, because a basic idea has been to seek to express "arbitrary" functions in terms of series as sums of simple terms. This was the grand idea of Fourier who thought of representing general functions as sums of trigonometric functions in the form of Fourier series, and Weierstrass who tried to do the same with monomials or polynomials in the form of power series. There are limitations to both Fourier and power series and the role of such series is today largely being taken over by computational methods. We therefore do not go into

any excessive treatment of series, but we do present some important basic facts, which are useful to know.

We recall that we already met one infinite series, namely the *geometric series*

$$\sum_{i=0}^{\infty} a^i = 1 + a + a^2 + a^3 + \cdots,$$

where a is a real number. We determined the sum of this infinite series in the case $|a| < 1$ by first computing the *partial sum of order* n:

$$s_n = \sum_{i=0}^{n} a^i = 1 + a + a^2 + \cdots + a^n = \frac{1 - a^{n+1}}{1 - a}.$$

by summing the terms a^i with $i \leq n$. We then made the observation that if $|a| < 1$, then

$$\lim_{n\to\infty} s_n = \lim_{n\to\infty} \frac{1 - a^{n+1}}{1 - a} = \frac{1}{1 - a},$$

and so we defined for $|a| < 1$ the sum of the infinite geometric series to be

$$\sum_{i=0}^{\infty} a^i = \lim_{n\to\infty} \sum_{i=0}^{n} a^i = \frac{1}{1 - a}.$$

We note that if $|a| \geq 1$, then we had to leave the sum of the geometric series $\sum_{i=0}^{\infty} a^i$ undefined. If $|a| \geq 1$, then $|s_n - s_{n-1}| = |a^n| \geq 1$, and therefore $\{s_n\}_{n=0}^{\infty}$ is not a Cauchy sequence, and thus $\lim_{n\to\infty} s_n = \lim_{n\to\infty} \sum_{i=0}^{n} a^i$ does not exist. Evidently, a necessary condition for convergence is that the terms a^i tend to zero as i tends to infinity.

37.2 Definition of Convergent Infinite Series

We now generalize these ideas to arbitrary infinite series. Thus let $\{a_n\}_{n=0}^{\infty}$ denote a sequence of real numbers and consider the *sequence of partial sums* $\{s_n\}_{n=0}^{\infty}$, where

$$s_n = \sum_{i=0}^{n} a_i = a_0 + a_1 + \cdots + a_n \tag{37.1}$$

is the *partial sum of order* n. We now say that the series $\sum_{i=0}^{\infty} a_i$ is *convergent* if the corresponding sequence of partial sums $\{s_n\}_{n=0}^{\infty}$ converges, and we then write

$$\sum_{i=0}^{\infty} a_i = \lim_{n\to\infty} s_n = \lim_{n\to\infty} \sum_{i=0}^{n} a_i, \tag{37.2}$$

which we refer to as the sum of the series. The convergence of a series $\sum_{i=1}^{\infty} a_i$ is thus reduced to the convergence of the sequence of it's partial

sums. All convergence issues for a series are handled in this way by reduction to convergence of sequences. This chapter therefore may be viewed as a direct gives a direct continuation of Chapters *Sequences and limits* and *Real numbers*. In particular, we understand as in the case of a geometric series, that a necessary condition for convergence of a series $\sum_{i=0}^{\infty} a_i$ is that the terms a_i tend to zero as i tends to infinity. However, this condition is not sufficient, as we should know from our previous experience with sequences, and as we will see again below.

Note that we can similarly consider series of the form $\sum_{i=1}^{\infty} a_i$ or $\sum_{i=m}^{\infty} a_i$ for any integer m.

Note that in a few special cases like the geometric series, we can actually find an analytic formula for the sum of the series. However, for most series $\sum_{i=0}^{\infty} a_i$ this is not possible, or may be so be tricky that we can't make it. Of course, we can then usually compute an approximation by directly computing a partial sum $s_n = \sum_{i=0}^{n} a_i$ for some appropriate n, that is, if n is not too big and the terms a_i not too difficult to evaluate. To then estimate the error, we are led to estimate the remainder $\sum_{i=n+1}^{\infty} a_i$. Thus we see a need to be able to analytically estimate the sum of a series, which may be easier than to analytically compute the exact sum.

In particular, such estimation may be used to decide if a series is convergent or not, which of course is an important issue because playing around with divergent series cannot have any meaning. In this pursuit, it is natural to distinguish between series in which all of the terms have the same sign and those in which the terms can have different signs. It may be more difficult to determine convergence for a series in which the terms can have different signs because of the possibility of cancellation between the terms.

Further, if we bound a series remainder $\sum_{i=n+1}^{\infty} a_i$ by using the triangle inequality, we get

$$|\sum_{i=n+1}^{\infty} a_i| \leq \sum_{i=n+1}^{\infty} |a_i|,$$

where the series on the right hand side is positive. So, positive series are of prime importance and we now turn to this topic.

37.3 Positive Series

A series $\sum_{i=1}^{\infty} a_i$ is said to be a *positive series*, if $a_i \geq 0$ for $i = 1, 2, \ldots$. The important point about a positive series is that the sequence of partial sums is non-decreasing, because

$$s_{n+1} - s_n = \sum_{i=1}^{n+1} a_i - \sum_{i=1}^{n} a_i = a_{n+1} \geq 0. \tag{37.3}$$

In Chapter *Optimization* below we shall prove that a nondecreasing sequence converges if and only if the sequence is bounded above. If we accept this as a fact, we understand that a positive series is convergent if and only if the sequence of partial sums is bounded above, that is there is a constant C such that

$$\sum_{i=1}^{n} a_i \leq C \quad \text{for } n = 1, 2, \ldots, . \tag{37.4}$$

This gives a definite way to check convergence, which we state as a theorem:

Theorem 37.1 *A positive series converges if and only if the sequence of partial sums is bounded above.*

This result does not apply if the series has terms with different signs. For example, the series $\sum_{i=0}^{\infty}(-1)^i = 1 - 1 + 1 - 1 + 1 \ldots$ has bounded partial sums, but is not convergent since $(-1)^i$ does not tend to zero as i tends to infinity.

Example 37.1. We can sometimes use an integral to bound the partial sums of a positive series and thus to prove convergence or estimate remainders. As an example, consider the positive series $\sum_{i=2}^{\infty} \frac{1}{i^2}$. The partial sum

$$s_n = \sum_{i=2}^{n} \frac{1}{i^2}$$

may be viewed as a quadrature formula for the integral of $\int_1^n x^{-2}\,dx$, see Fig. 37.1.

More precisely, we see that

$$\begin{aligned}\int_1^n x^{-2}\,dx &= \int_1^2 x^{-2}\,dx + \int_2^3 x^{-2}\,dx + \cdots + \int_{n-1}^n x^{-2}\,dx \\ &\geq \int_1^2 \frac{1}{2^2}\,dx + \int_2^3 \frac{1}{3^2}\,dx + \cdots + \int_{n-1}^n \frac{1}{n^2}\,dx \\ &\geq \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{n^2} = s_n.\end{aligned}$$

Since

$$\int_1^n x^{-2}\,dx = \left(1 - \frac{1}{n}\right) \leq 1,$$

we conclude that $s_n \leq 1$ for all n and therefore the series $\sum_{i=2}^{\infty} \frac{1}{i^2}$ is convergent. To compute an approximation of the sum of the series, we of course compute a partial sum s_n with n sufficiently large. To estimate the remainder we may of course use a similar comparison, see Problem 37.5.

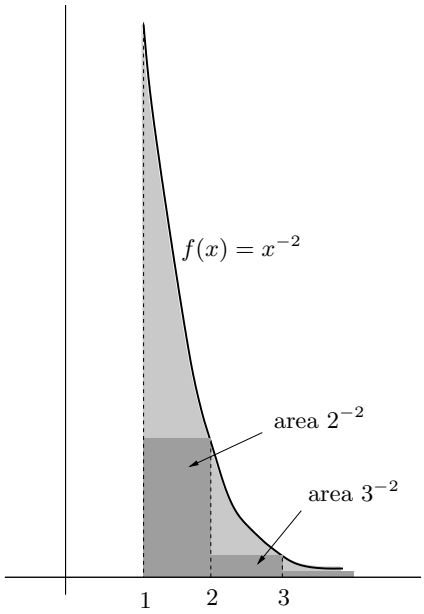

Fig. 37.1. The relation between $\int_1^n x^{-2}\,dx$ and $\sum_{i=2}^n i^{-2}$

Example 37.2. The positive series $\sum_{i=1}^{\infty} \frac{1}{i+i^2}$ converges because for all n

$$s_n = \sum_{i=1}^{n} \frac{1}{i+i^2} \leq \sum_{i=1}^{n} \frac{1}{i^2} \leq 2$$

by the previous example.

Similarly, a *negative series* with all terms non-positive, converges if and only if its partial sums are bounded *below*.

Example 37.3. For the *alternating* series

$$\sum_{i=1}^{\infty} \frac{(-1)^i}{i},$$

we have that the difference between two successive partial sums

$$s_n - s_{n-1} = \frac{(-1)^n}{n}$$

alternates in sign, and thus the sequence of partial sums is not monotone, and therefore we cannot decide convergence or not from the above theorem. We shall return to this series below and prove that it is in fact convergent.

37.4 Absolutely Convergent Series

Now we turn to series with terms of different signs. We begin by first considering series that converge regardless of any cancellation between the terms. We are motivated by the convergence results for positive series. A series $\sum_{i=1}^{\infty} a_i$ is said to be *absolutely convergent* if the series

$$\sum_{i=1}^{\infty} |a_i|$$

converges. By the previous result we know that a series $\sum_{i=1}^{\infty} a_i$ is absolutely convergent if and only if the sequence $\{\hat{s}_n\}$ with

$$\hat{s}_n = \sum_{i=1}^{n} |a_i|, \tag{37.5}$$

is bounded above.

We shall now prove that an absolutely convergent series $\sum_{i=1}^{\infty} a_i$ is convergent. By the triangle inequality we have for $m > n$,

$$|s_m - s_n| = \left|\sum_{n}^{m} a_i\right| \leq \sum_{n}^{m} |a_i| = |\hat{s}_m - \hat{s}_n|. \tag{37.6}$$

Now, since we can make $|\hat{s}_m - \hat{s}_n|$ arbitrarily small by taking m and n large, because $\sum_{i=1}^{\infty} a_i$ is absolutely convergent and thus $\{\hat{s}_n\}_{n=1}^{\infty}$ is a Cauchy sequence, we conclude that $\{s_n\}_{n=1}^{\infty}$ is a Cauchy sequence and therefore converges and thus the series $\sum_{i=1}^{\infty} a_i$ is convergent. We state this fundamental result as a theorem:

Theorem 37.2 *An absolutely convergent series is convergent.*

Example 37.4. The series $\sum_{i=1}^{\infty} \frac{(-1)^i}{i^2}$ is convergent because $\sum_{i=1}^{\infty} \frac{1}{i^2}$ is convergent.

37.5 Alternating Series

The convergence of a general series with terms of "random" sign may be very difficult to analyze because of cancellation of terms. We now consider a special case with a regular pattern to the signs of the terms:

$$\sum_{i=0}^{\infty} (-1)^i a_i \tag{37.7}$$

where $a_i \geq 0$ for all i. This is called an *alternating series* since the signs of the terms alternate. We shall now prove that if $a_{i+1} \leq a_i$ for $i = 0, 1, 2 \ldots$

and $\lim_{i\to\infty} a_i = 0$, then the alternating series converges. The key observation is that the sequence $\{s_n\}$ of partial sums satisfies

$$s_1 \leq s_3 \leq s_5 \leq \ldots s_{2j+1} \leq s_{2i} \leq \ldots \leq s_4 \leq s_2 \leq s_0, \tag{37.8}$$

which shows that both limits $\lim_{j\to\infty} s_{2j+1}$ and $\lim_{i\to\infty} s_{2i}$ exist. Since $a_i \to 0$ as i tends to infinity, $\lim_{j\to\infty} s_{2j+1} = \lim_{i\to\infty} s_{2i}$, and thus $\lim_{n\to\infty} s_n$ exists and convergence of the series $\sum_{i=0}^{\infty}(-1)^i a_i$ follows. We summarize in the following theorem first stated and proved by Leibniz:

Theorem 37.3 *An alternating series with the property that the modulus of its terms tends monotonically to zero, converges.*

Example 37.5. The *harmonic series*

$$\sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots$$

converges. We now proceed to show that this series is not absolutely convergent.

37.6 The Series $\sum_{i=1}^{\infty} \frac{1}{i}$ Theoretically Diverges!

We shall now show that the *harmonic series* $\sum_{i=1}^{\infty} \frac{(-1)^i}{i}$ is **not** absolutely convergent, i.e. we shall prove that the series

$$\sum_{i=1}^{\infty} \frac{1}{i} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$$

diverges. We do this by proving that the sequence $\{s_n\}_{n=1}^{\infty}$ of partial sums

$$s_n = \sum_{i=1}^{n} \frac{1}{i}$$

can become arbitrarily large if n is large enough. To see this we group the terms of a partial sum as follows:

$$1 + \overline{\frac{1}{2} + \overline{\frac{1}{3} + \frac{1}{4}} + \overline{\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}}} + \overline{\frac{1}{9} + \frac{1}{10} + \frac{1}{11} + \frac{1}{12} + \frac{1}{13} + \frac{1}{14} + \frac{1}{15} + \frac{1}{16}} + \overline{\frac{1}{17} + \cdots + \frac{1}{32}} + \cdots$$

The first "group" is 1/2. The second group is

$$\frac{1}{3}+\frac{1}{4}\geq\frac{1}{4}+\frac{1}{4}=\frac{1}{2}.$$

The third group is

$$\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}\geq\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}=\frac{1}{2}.$$

The fourth group

$$\frac{1}{9}+\frac{1}{10}+\frac{1}{11}+\frac{1}{12}+\frac{1}{13}+\frac{1}{14}+\frac{1}{15}+\frac{1}{16}$$

has 8 terms that are larger than 1/16, so it also gives a sum larger than $8/16 = 1/2$. We can continue in this way, taking the next 16 terms, all of which are larger than 1/32, then the next 32 terms, all of which are larger than 1/64, and so on. Each time we take a group, we get a contribution to the overall sum that is larger than 1/2.

When we take n larger and larger, we can combine more and more terms in this way, making the sum larger in increments of 1/2 each time. The partial sums therefore just become larger and larger as n increases, which means the partial sums diverge to infinity.

Note that by the arithmetic rules, the partial sum s_n should be the same whether we compute the sum in the "forward" direction

$$s_n = 1+\frac{1}{2}+\frac{1}{3}+\cdots\frac{1}{n-1}+\frac{1}{n}$$

or the "backward" direction

$$s_n = \frac{1}{n}+\frac{1}{n-1}+\cdots+\frac{1}{3}+\frac{1}{2}+1.$$

In Fig. 37.2, we list various partial sums in both the forward and backward directions computed using FORTRAN with single precision variables with about 7 digits of accuracy. Note two things about these results:

First, the computed partial sums s_n all become equal when n is large enough, even though theoretically they should keep increasing to infinity as n increases. This is because in finite precision the new terms we add eventually get so small that they effectively give zero contribution. Thus, although in principle the series diverges, in practice the series appears to converge on the computer. This gives an illustration of idealism vs realism in mathematics!

Second, the backward sum is strictly larger than the forward sum! This is because in the summation a term effectively adds zero when the term is sufficiently small compared to the current partial sum, and the size of the partial sums is vastly different if we add in a forward or backward manner.

n	forward sum	backward sum
10000	9.787612915039062	9.787604331970214
100000	12.090850830078120	12.090151786804200
1000000	14.357357978820800	14.392651557922360
2000000	15.311032295227050	15.086579322814940
3000000	15.403682708740240	15.491910934448240
5000000	15.403682708740240	16.007854461669920
10000000	15.403682708740240	16.686031341552740
20000000		17.390090942382810
30000000		17.743585586547850
40000000		18.257812500000000
50000000		18.807918548583980
100000000	15.403682708740240	18.807918548583980
200000000		18.807918548583980
1000000000		18.807918548583980

Fig. 37.2. Forward $1 + \frac{1}{2} + \cdots + \frac{1}{n}$ and backward $\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{2} + 1$ partial harmonic sums for various n computed with double precision

37.7 Abel

Niels Henrik Abel (1802–1829), the great mathematical genius of Norway, is today world famous for his half-page proof from 1824 of the impossibility of solving polynomial equations of degree larger or equal to five by root-extraction. This settled a famous problem which had haunted many generations of mathematicians. However, Abel's life was short and tragic and his fame came only after his sudden death at the age of 27. Gauss in Göttingen

Fig. 37.3. Niels Henrik Abel (1802–1829):" The divergent series are the invention of the devil, and it is a shame to base on them any demonstration whatsoever. By using them, one may draw any conclusion he pleases and that is why these series have produced so many fallacies and so many paradoxes. . . "

was indifferent to the proof when it was first presented, based on his view expressed in his thesis of 1801 that the algebraic solution of an equation was no better than devising a symbol for the root of the equation and then saying that the equation had a root equal to the symbol (compare the square root of two).

Abel tried also unsuccessfully to convince Cauchy on a trip to Paris 1825, which ended in misery, and he then left for Berlin on borrowed money but succeeded to produce another master-piece now on so called elliptic integrals. After returning to a modest position in Christiania he continued to pour out high quality mathematics while his health was deteriorating. After a sled journey to visit his girl friend for Christmas 1828 he became seriously ill and died quickly after.

37.8 Galois

Abel is contemporary with Evariste Galois (1811–32), who independently 1830 proved the same fifth order equation result as Abel, again with no reaction from Cauchy. Galois was refused twice in the entrance exam to Ecole Polytechnique apparently after accusing the examiner for posing questions incorrectly. Galois was imprisoned for a revolutionary speech against King

Fig. 37.4. Evariste Galois: (1811–1832):“ Since the beginning of the century, computational procedures have become so complicated that any progress by those means has become impossible, without the elegance which modern mathematicians have brought to bear on their research, and by means of which the spirit comprehends quickly and in one step a great many computations. It is clear that elegance, so vaunted and so aptly named, can have no other purpose... Go to the roots, of these calculations! Group the operations. Classify them according to their complexities rather than their appearances! This, I believe, is the mission of future mathematicians. This is the road on which I am embarking in this work” (from the preface to Galois' final manuscript)

Louis Philippe 1830, was released in 1832 but soon died after wounds from a duel about his girl friend, at the age of 21.

Chapter 37 Problems

37.1. Prove that the series $\sum_{i=1}^{\infty} i^{-\alpha}$ converges if and only if $\alpha > 1$. Hint: Compare with a primitive function of $x^{-\alpha}$.

37.2. Prove that the series $\sum_{i=1}^{\infty} (-1)^i i^{-\alpha}$ converges if and only if $\alpha > 0$.

37.3. Prove that the following series converges: (a) $\sum_{i=1}^{\infty} e^{-i}$. (b) $\sum_{i=1}^{\infty} \frac{1+(-1)^i}{i^2}$. (c) $\sum_{i=1}^{\infty} \frac{e^{-i}}{i}$. (d) $\sum_{i=1}^{\infty} \frac{1}{(i+1)(i+4)}$.

37.4. Prove that $\sum_{i=1}^{\infty} \frac{1}{i^2 - i}$ converges. Hint: first show that $\frac{1}{2}i^2 - i \geq 0$ for $i \geq 2$.

37.5. Estimate the remainder $\sum_{i=n}^{\infty} \frac{1}{i^2}$ for different values of n.

37.6. Prove that $\sum_{i=1}^{\infty} (-1)^i \sin(1/i)$ converges. More difficult: prove that it is **not** absolutely convergent.

37.7. Explain in detail why the backward partial sum of the series $\sum_{i=1}^{\infty} \frac{1}{i}$ is larger than the forward sum.

38

Scalar Autonomous Initial Value Problems

He doesn't[TS] use long, difficult words, like Owl. (The House at Pooh Corner, Milne)

38.1 Introduction

In this chapter, we consider the initial value problem for a *scalar autonomous non-linear differential equation*: Find a function $u : [0,1] \to \mathbb{R}$ such that

$$u'(x) = f(u(x)) \quad \text{for } 0 < x \leq 1,\ u(0) = u_0, \tag{38.1}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a given function and u_0 a given initial value. We assume that $f : \mathbb{R} \to \mathbb{R}$ is bounded and Lipschitz continuous, that is, there are constants L_f and M_f such that for all $v, w \in \mathbb{R}$,

$$|f(v) - f(w)| \leq L_f|v - w|, \quad \text{and} \quad |f(v)| \leq M_f. \tag{38.2}$$

For definiteness, we choose the interval $[0,1]$, and we may of course generalize to any interval $[a,b]$.

The problem (38.1) is in general *non-linear*, since $f(v)$ in general is non-linear in v, that is, $f(u(x))$ depends non-linearly on $u(x)$. We have already in Chapter *The exponential function* considered the basic case with f linear, which is the case $f(u(x)) = u(x)$ or $f(v) = v$. Now we pass on to nonlinear functions such as $f(v) = v^2$ and others.

Further, we call (38.1) *autonomous* because $f(u(x))$ depends on the value of the solution $u(x)$, but not directly on the independent variable x. A *non-autonomous* differential equation has the form $u'(x) = f(u(x), x)$, where

$f(u(x), x)$ depends on both $u(x)$ and x. The differential equation $u'(x) = xu^2(x)$ is non-autonomous and non-linear with $f(v, x) = xv^2$, while the equation $u'(x) = u(x)$ defining the exponential is autonomous and linear with $f(v) = v$.

Finally, we refer to (38.1) as a *scalar* problem since $f : \mathbb{R} \to \mathbb{R}$ is a real valued function of one real variable, that is, $v \in \mathbb{R}$ and $f(v) \in \mathbb{R}$, and thus $u(x)$ takes real values or $u : [0, 1] \to \mathbb{R}$. Below we shall consider *systems* of equations with $f : \mathbb{R}^d \to \mathbb{R}^d$ and $u : [0, 1] \to \mathbb{R}^d$, where $d > 1$, which models a very large range of phenomena.

We hope the reader (like Owl) is now at ease with the terminology: In this chapter we thus focus on scalar autonomous non-linear differential equations.

The initial value problem for a scalar autonomous differential equation is the simplest of all initial value problems and the solution (when it exists) can be expressed analytically in terms of a primitive function $F(v)$ of the function $1/f(v)$. In the next chapter we present an extension of this solution formula to a certain class of scalar non-autonomous differential equations referred to as *separable* differential equations. The analytical solution formula does not generalize to an initial value problems for a system of differential equations, and is thus of very very limited use. However, the solution formula is really a beautiful application of Calculus, which may give valuable information in compact form in the special cases when it is applicable.

We also present a direct constructive proof of existence of a solution to the scalar autonomous problem, which generalizes to the very general case of a initial value problems for (autonomous and non-autonomous) systems of differential equations, as presented in Chapter *The general initial value problem* below.

38.2 An Analytical Solution Formula

To derive the analytical solution formula, we let $F(v)$ be a primitive function of the function $1/f(v)$, assuming v takes values so that zeros of $f(v)$ are avoided. Observe that here $F(v)$ is a primitive function of the function $1/f(v)$, and not of $f(v)$. We can then write the equation $u'(x) = f(u(x))$ as

$$\frac{d}{dx}F(u(x)) = 1,$$

since by the Chain rule $\frac{d}{dx}F(u(x)) = F'(u(x))u'(x) = \frac{u'(x)}{f(u(x))}$. We conclude that

$$F(u(x)) = x + C,$$

TS^S I changed doesn' to doesn't, please check it.

Editor's or typesetter's annotations (will be removed before the final TEX run)

where the constant C is to be determined by the initial condition by setting $F(u_0) = C$ at $x = 0$. Formally, we can carry out the calculus as follows: We write the differential equation $\frac{du}{dx} = f(u)$ in the form

$$\frac{du}{f(u)} = dx$$

and integrate to get

$$F(u) = x + C,$$

which gives the solution formula

$$u(x) = F^{-1}(x + F(u_0)), \tag{38.3}$$

where F^{-1} is the inverse of F.

The Model $u' = u^n$ for $n > 1$

We use this example to show that the nonlinear nature of (38.1) allows the interesting behavior of *finite-time-blow-up* of the solution. First consider the case $n = 2$, that is, the initial value problem

$$u'(x) = u^2(x) \quad \text{for } x > 0, \quad u(0) = u_0 > 0, \tag{38.4}$$

with $f(v) = v^2$. In this case $F(v) = -1/v$ with $F^{-1}(w) = -1/w$, and we obtain the solution formula

$$u(x) = \frac{1}{u_0^{-1} - x} = \frac{u_0}{1 - u_0 x}.$$

We see that that $u(x) \to \infty$ as $x \to u_0^{-1}$, that is, the solution $u(x)$ of (38.1) with $f(u) = u^2$ tends to infinity as x increases to u_0^{-1} and the solution does not exist beyond this point, see Fig. 38.1. We say that the solution u *blows up* in finite time or exhibits *finite time blow-up*.

If we consider $u'(x) = u^2(x)$ as a model for the growth of a quantity $u(x)$ with time x in which the rate of growth is proportional to $u^2(x)$ and compare with the model $u'(x) = u(x)$ with solution $u_0 \exp(x)$ showing exponential growth. In the model $u'(x) = u^2(x)$ the growth is eventually much quicker than exponential growth since $u^2(x) > u(x)$ as soon as $u(x) > 1$.

We now generalize to

$$u'(x) = u^n(x) \quad \text{for } x > 0, \quad u(0) = u_0,$$

where $n > 1$. In this case $f(v) = v^{-n}$ and $F(v) = -\frac{1}{n-1} v^{-(n-1)}$, and we find the solution formula

$$u(x) = \frac{1}{(u_0^{-n+1} - (n-1)x)^{1/(n-1)}}.$$

Again the solution exhibits finite time blow-up.

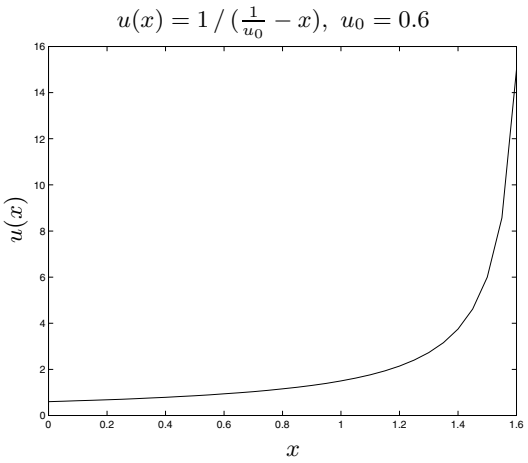

Fig. 38.1. Solution of the equation $u' = u^2$

The Logistic Equation $u' = u(1-u)$

We now consider the initial value problem for the *logistic equation*

$$u'(x) = u(x)(1 - u(x)) \quad \text{for } x > 0, \quad u(0) = u_0,$$

which was derived by the mathematician and biologist Verhulst as a model of a population with the *growth rate* decreasing with the factor $(1-u)$, as compared with the basic model $u' = u$, as the population approaches the value 1. Typically we assume $0 < u_0 < 1$ and expect to have $0 \le u(x) \le 1$.

In this case we have $f(u) = \frac{1}{u(1-u)}$ and using that $f(u) = \frac{1}{u} + \frac{1}{1-u}$, we find that

$$F(u) = \log(u) - \log(1-u) = \log\left(\frac{u}{1-u}\right),$$

so that

$$\log\left(\frac{u}{1-u}\right) = x + C,$$

or

$$\frac{u}{1-u} = \exp(C)\exp(x).$$

Solving for u and using the initial condition we find that

$$u(x) = \frac{1}{\frac{1-u_0}{u_0}\exp(-x) + 1}.$$

We see that the solution $u(x)$ increases from $u_0 < 1$ to 1 as x increases to infinity, see Fig. 38.2, which gives the famous logistic S-curve modeling growth with decreasing growth rate.

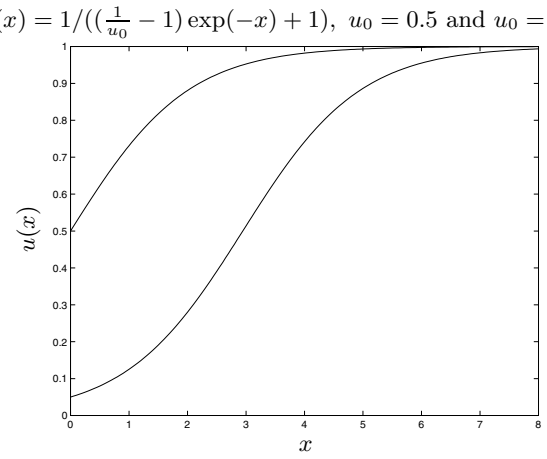

Fig. 38.2. Solution of the logistic equation

38.3 Construction of the Solution

For the direct construction of a solution of (38.1), we shall use the same technique as that used for the linear problem $f(u(x)) = u(x)$ considered in Chapter *The exponential function*. Of course, one may ask why we should worry about constructing the solution, when we already have the solution formula (38.3). We may reply that the solution formula involves the (inverse of) the primitive function $F(v)$ of $1/f(v)$, which we may have to construct anyway, and then a direct construction of the solution may in fact be preferable. In general, a solution formula when available may give valuable information about qualitative properties of the solution such as dependence of parameters of the problem, even if it is not necessarily the most effective way of actually computing the solution.

To construct the solution we introduce meshes with nodes $x_i^n = ih_n$ for $i = 1, \cdots, N$, where $h_n = 2^{-n}$ and $N = 2^n$, and for $n = 1, 2, \cdots$. we then define an approximate continuous piecewise linear solution $U^n(x)$ for $0 < x \leq 1$ by the formula

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n f(U^n(x_{i-1}^n)) \quad \text{for } i = 1, \cdots, N, \tag{38.5}$$

with $U^n(0) = u_0$.

We want to prove that $\{U^n(x)\}$ is a Cauchy sequence for $x \in [0, 1]$ and we start by estimating $U^n(x_i^n) - U^{n+1}(x_i^n)$ for $i = 1, \cdots, N$. Taking two steps with step size $h_{n+1} = \frac{1}{2}h_n$ to go from time $x_{i-1}^n = x_{2i-2}^{n+1}$ to $x_i^n = x_{2i}^{n+1}$, we get

$$U^{n+1}(x_{2i-1}^{n+1}) = U^{n+1}(x_{2i-2}^{n+1}) + h_{n+1} f(U^{n+1}(x_{2i-2}^{n+1})),$$
$$U^{n+1}(x_{2i}^{n+1}) = U^{n+1}(x_{2i-1}^{n+1}) + h_{n+1} f(U^{n+1}(x_{2i-1}^{n+1})).$$

Inserting now the value of $U^{n+1}(x_{2i-1}^{n+1})$ at the intermediate step x_{2i-1}^{n+1} from the first equation into the second equation gives

$$U^{n+1}(x_{2i}^{n+1}) = U^{n+1}(x_{2i-2}^{n+1}) + h_{n+1}f(U^{n+1}(x_{2i-2}^{n+1})) + h_{n+1}f\big(U^{n+1}(x_{2i-2}^{n+1}) + h_{n+1}f(U^{n+1}(x_{2i-2}^{n+1}))\big). \quad (38.6)$$

Setting $e_i^n \equiv U^n(x_i^n) - U^{n+1}(x_{2i}^{n+1})$ and subtracting (38.6) from (38.5), we get

$$\begin{aligned} e_i^n &= e_{i-1}^n + h_n\big(f(U^n(x_{i-1}^n)) - f(U^{n+1}(x_{2i-2}^{n+1}))\big) \\ &\quad + h_{n+1}\Big(f(U^{n+1}(x_{2i-2}^{n+1})) - f\big(U^{n+1}(x_{2i-2}^{n+1}) + h_{n+1}f(U^{n+1}(x_{2i-2}^{n+1}))\big)\Big) \\ &\equiv e_{i-1}^n + F_{1,n} + F_{2,n}, \end{aligned}$$

with the obvious definition of $F_{1,n}$ and $F_{2,n}$. Using the Lipschitz continuity and boundedness (38.2), we have

$$\begin{aligned} |F_{1,n}| &\leq L_f h_n |e_{i-1}^n|, \\ |F_{2,n}| &\leq L_f h_{n+1}^2 |f(U^{n+1}(x_{2i-2}^{n+1}))| \leq L_f M_f h_{n+1}^2. \end{aligned}$$

Thus for $i = 1, \cdots, 2^N$,

$$|e_i^n| \leq (1 + L_f h_n)|e_{i-1}^n| + L_f M_f h_{n+1}^2.$$

Iterating this inequality over i and using that $e_0^n = 0$, we get

$$|e_i^n| \leq L_f M_f h_{n+1}^2 \sum_{k=0}^{i-1} (1 + L_f h_n)^k \quad \text{for } i = 1, \cdots, N.$$

Now recalling (31.10) and (31.27), we have

$$\sum_{k=0}^{i-1} (1 + L_f h_n)^k \leq \frac{\exp(L_f) - 1}{L_f h_n},$$

and thus we have proved that for $i = 1, \cdots, N$,

$$|e_i^n| \leq \frac{1}{2} M_f \exp(L_f) h_{n+1},$$

that is, for $\bar{x} = i h_n$ with $i = 0, \ldots, N$,

$$|U^n(\bar{x}) - U^{n+1}(\bar{x})| \leq \frac{1}{2} M_f \exp(L_f) h_{n+1}.$$

Iterating this inequality as in the proof of the Fundamental Theorem, we get for $m > n$ and $\bar{x} = ih_n$ with $i = 0, \ldots, N$,

$$|U^n(\bar{x}) - U^m(\bar{x})| \leq \frac{1}{2} M_f \exp(L_f) h_n.$$

Again as in the proof of the Fundamental Theorem, we conclude that $\{U^n(x)\}$ is a Cauchy sequence for each $x \in [0,1]$, and thus converges to a function $u(x)$, which by the construction satisfies the differential equation $u'(x) = f(u(x))$ for $x \in (0,1]$ and $u(0) = u_0$, and thus the limit $u(x)$ is a solution of the initial value problem (38.1).

It remains to prove uniqueness. Assume that $v(x)$ satisfies $v'(x) = f(v(x))$ for $x \in (0,1]$ and $v(0) = u_0$, and consider the function $w = u - v$. Since $w(0) = 0$,

$$\begin{aligned} |w(x)| &= \left| \int_0^x w'(y)\, dy \right| = \left| \int_0^x f(u(y)) - f(v(y))\, dy \right| \\ &\leq \int_0^x |f(u(y)) - f(v(y))|\, dy \leq \int_0^x L_f |w(y)|\, dy. \end{aligned}$$

Setting $a = \max_{0 \leq x \leq (2L_f)^{-1}} |w(x)|$, we have

$$a \leq \int_0^{(2L_f)^{-1}} L_f a\, dy \leq \frac{1}{2} a$$

which proves that $w(x) = 0$ for $0 \leq x \leq (2L_f)^{-1}$. We now repeat the argument for $x \geq (2L_f)^{-1}$ to get uniqueness for $0 \leq x \leq 1$.

We have now proved:

Theorem 38.1 *The initial value problem* (38.1) *with* $f : \mathbb{R} \to \mathbb{R}$ *Lipschitz continuous and bounded has a unique solution* $u : [0,1] \to \mathbb{R}$*, which is the limit of the sequence of continuous piecewise linear functions* $\{U^n(x)\}$ *constructed from* (38.5) *and satisfying* $|u(x) - U^n(x)| \leq \frac{1}{2} M_f \exp(L_f) h_n$ *for* $x \in [0,1]$.

The attentive reader will note that the existence proof does not seem to apply to e.g. the initial value problem (38.4), because the function $f(v) = v^2$ is not Lipschitz continuous and bounded on $\mathbb{R}$. In fact, the solution $u(x) = \frac{u_0}{1 - u_0 x}$ only exists on the interval $[0, u_0^{-1}]$ and blows up at $x = u_0^{-1}$. However, we can argue that *before* blow-up with say $|u(x)| \leq M$ for some (large) constant M, it suffices to consider the function $f(v) = v^2$ on the interval $[-M, M]$ where the assumption of Lipschitz continuity and boundedness is satisfied. We conclude that for functions $f(v)$ which are Lipschitz continuous and bounded on bounded intervals of $\mathbb{R}$, the constructive existence proof applies as longs as the solution does not blow up.

Chapter 38 Problems

38.1. Solve the following initial value problem analytically: $u'(x) = f(u(x))$ for $x > 0$, $u(0) = u_0$, with (a) $f(u) = -u^2$, (b) $f(u) = \sqrt{u}$, (c) $f(u) = u\log(u)$, (d) $f(u) = 1 + u^2$, (e) $f(u) = \sin(u)$, (f) $f(u) = (1 + u)^{-1}$, (g) $f(u) = \sqrt{u^2 + 4}$.

38.2. Verify that the constructed function $u(x)$ satisfies (38.1). Hint: Use that by the construction we have $u(x) = u_0 + \int_0^x f(u(y))\,dy$ for $x \in [0, 1]$.

38.3. Find the velocity of a parachute jumper assuming that the air resistance is proportional to the square of the velocity.

38.4. Let $u(t)$ be the position of a body sliding along x-axis with the velocity $\dot{u}(t)$ satisfying $\dot{u}(t) = -\exp(-u)$. How long time does it take for the body to reach the position $u = 0$ starting from $u(0) = 5$

39

Separable Scalar Initial Value Problems

> The search for general methods for integrating ordinary differential equations ended about 1755. (Mathematical Thought, from Ancient to Modern Times, Kline)

39.1 Introduction

We now consider the initial value problem for a scalar non-autonomous differential equation:

$$u'(x) = f(u(x), x) \quad \text{for } 0 < x \leq 1,\ u(0) = u_0, \tag{39.1}$$

in the special case when $f(u(x), x)$ has the form

$$f(u(x), x) = \frac{h(x)}{g(u(x))}, \tag{39.2}$$

where $h : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$. We thus consider the initial value problem

$$u'(x) = \frac{h(x)}{g(u(x))} \quad \text{for } 0 < x \leq 1,\ u(0) = u_0, \tag{39.3}$$

where $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ are given functions, which we refer to as a *separable* problem, because the right hand side $f(u(x), x)$ separates into the quotient of one function $h(x)$ of x only, and one function $g(u(x))$ of $u(x)$ only according to (39.2).

39.2 An Analytical Solution Formula

We shall now derive an analytical solution formula that generalizes the solution formula (38.3) for a scalar autonomous problem (corresponding to the case $h(x) = 1$). Let then $G(v)$ and $H(x)$ be primitive functions of $g(v)$ and $h(x)$ so that $\frac{dG}{dv} = g$ and $\frac{dH}{dx} = h$, and suppose that the function $u(x)$ solves the equation

$$G(u(x)) = H(x) + C, \tag{39.4}$$

for $x \in [0, 1]$, where C is a constant. Differentiating with respect to x using the Chain rule on the left hand side, we then find that $g(u(x))u'(x) = h(x)$, that is $u(x)$ solves the differential equation $u'(x) = h(x)/g(u(x)) = f(u(x), x)$ as desired. Choosing the constant C so that $u(0) = u_0$, we thus obtain a solution $u(x)$ of (39.3), that is the problem (39.1) with $f(u(x), x)$ of the separable form (39.2).

Note that (39.4) is an algebraic equation for the value of the solution $u(x)$ for each value of x. We have thus rewritten the differential equation (39.3) as an algebraic equation (39.4) with x acting as a parameter, and involving primitive functions of $g(y)$ and $h(x)$.

Of course, we may consider (39.1) with x in an interval $[a, b]$ or $[a, \infty)$ with $a, b \in \mathbb{R}$.

Example 39.1. Consider the separable initial value problem

$$u'(x) = xu(x), \quad x > 0,\ u(0) = u_0, \tag{39.5}$$

where $f(u(x), x) = h(x)/g(u(x))$ with $g(v) = 1/v$ and $h(x) = x$. The equation $G(u(x)) = H(x) + C$ takes the form

$$\log(u(x)) = \frac{x^2}{2} + C, \tag{39.6}$$

and thus the solution $u(x)$ of (39.5) is given by the formula

$$u(x) = \exp\left(\frac{x^2}{2} + C\right) = u_0 \exp\left(\frac{x^2}{2}\right),$$

with $\exp(C) = u_0$ chosen so that the initial condition $u(0) = u_0$ is satisfied. We check by differentiation using the Chain rule that indeed $u_0 \exp(\frac{x^2}{2})$ satisfies $u'(x) = xu(x)$ for $x > 0$.

Formally ("multiplying by dx"), we can rewrite (39.5) as

$$\frac{du}{u} = x\,dx$$

and integrate to get

$$\log(u) = \frac{x^2}{2} + C,$$

which corresponds to the equation (39.6).

Example 39.2. On the rainy evening of November 11 1675 Leibniz successfully solved the following problem as a first (crucial) test of the power of the Calculus he had discovered on October 29: Find a curve $y = y(x)$ such that the *subnormal* p, see Fig. 39.1, is inversely proportional to y. Leibniz argued as follows: By similarity, see again Fig. 39.1, we have

$$\frac{dy}{dx} = \frac{p}{y},$$

and assuming the subnormal p to be inversely proportional to y, that is,

$$p = \frac{\alpha}{y}$$

with α a positive constant, we get the differential equation

$$\frac{dy}{dx} = \frac{\alpha}{y^2} = \frac{h(x)}{g(y)}, \tag{39.7}$$

which is separable with $h(x) = \alpha$ and $g(y) = y^2$. The solution $y = y(x)$ with $y(0) = 0$ thus is given by, see Fig. 39.1,

$$\frac{y^3}{3} = \alpha x, \quad \text{that is } y = (3\alpha x)^{\frac{1}{3}}, \tag{39.8}$$

The next morning Leibniz presented his solution to a stunned audience of colleagues in Paris, and rocketed to fame as a leading mathematician and Inventor of Calculus.

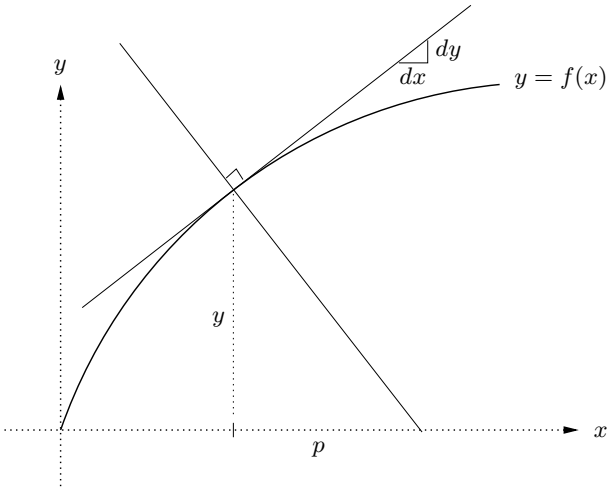

Fig. 39.1. Leibniz' subnormal problem (change from $y = f(x)$ to $y = y(x)$)

39.3 Volterra-Lotka's Predator-Prey Model

We now consider a biological system consisting of prey and predators like rabbits and foxes which interact. Let $x(t)$ be the density of the prey and $y(t)$ that of the predators at time t and consider *Volterra-Lotka's predator-prey model* for their interaction:

$$\begin{aligned} \dot{x}(t) &= ax(t) - bx(t)y(t), \\ \dot{y}(t) &= -\alpha y(t) + \beta x(t)y(t) \end{aligned} \tag{39.9}$$

where a, b, α and β are positive constants, and $\dot{x} = \frac{dx}{dt}$ and $\dot{y} = \frac{dy}{dt}$. The model includes a growth term $ax(t)$ for the prey corresponding to births and a decay term $bx(t)y(t)$ proportional to the density of prey and predators corresponding to the consumption of prey by the predators, together with corresponding terms for the predators with different signs.

This is a system of two differential equations in two unknowns $x(t)$ and $y(t)$ for which analytical solutions are unknown in general. However, we can derive an equation satisfied by the points $(x(t), y(t))$ in an $x-y$ plane, referred to as the $x-y$ *phase plane*, as follows: Dividing the two equations, we get

$$\frac{\dot{y}}{\dot{x}} = \frac{-\alpha y + \beta xy}{ax - bxy}$$

and formally replacing $\frac{\dot{y}}{\dot{x}}$ (by formally dividing out the common dt), we are led to the equation

$$y'(x) = \frac{-\alpha y + \beta xy}{ax - bxy} = \frac{y(-\alpha + \beta x)}{(a - by)x},$$

where $y' = \frac{dy}{dx}$, which is a separable equation with solution $y = y(x)$ satisfying

$$a\log(y) - by = -\alpha\log(x) + \beta x + C,$$

or

$$y^a \exp(-by) = \exp(C)x^{-\alpha}\exp(\beta x)$$

where C is a constant determined by the initial conditions. We plot pairs of (x, y) satisfying this equation in Fig. 39.2 as we let the prey x vary, which traces a *phase plane curve* of the solution $(x(t), y(t)$ of Fig. 39.2 as t varies. We see that the solution is periodic with a variation from (many rabbits, many foxes) to (few rabbits, many foxes) to (few rabbits, few foxes) to (many rabbits, few foxes) and back to (many rabbits, many foxes). Note that the phase plane curve shows the different combinations of rabbits and foxes (x, y), but does *not* give the time evolution $(x(t), y(t))$ of their interaction as a function of time t. We know that for a given t, the point $(x(t), y(t))$ lies on the phase plane curve, but not where.

$(x(t), y(t))$ for 0<t<25.5 with (a, b, c, d)=(.5, 1, .2, 1), (x_0, y_0)=(.5, .3)

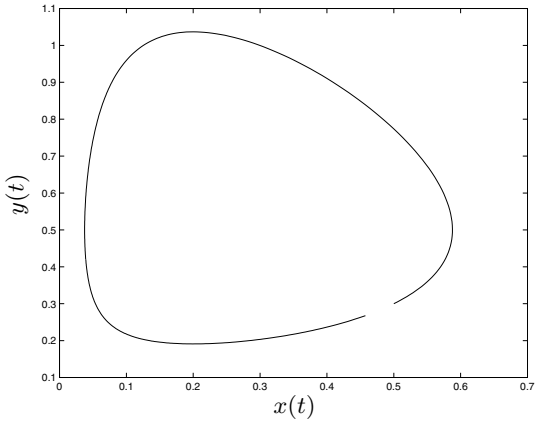

Fig. 39.2. Phase plane plot of a solution of Volterra-Lotka's equation

39.4 A Generalization

We now consider a generalization of the separable differential equation (39.3) with solution $u(x)$ satisfying an equation of the form $G(u(x)) - H(x) = C$, to a differential equation with solution satisfying a more general equation of the form $F(x, u(x)) = C$. This closely couples to Chapter *Potential fields* below, and uses a generalization of the Chain rule, which can be accepted right now by a willing reader, and which we will meet again in Chapter *Vector-valued functions of several variables* below.

We thus consider the scalar initial value problem

$$u'(x) = f(u(x), x) \quad \text{for } 0 < x \leq 1,\ u(0) = u_0, \tag{39.10}$$

in the case $f(u(x), x)$ has the form

$$f(u(x), x) = \frac{h(u(x), x)}{g(u(x)), x)}, \tag{39.11}$$

where $h(v, x)$ and $g(v, x)$ are functions of v and x with the special property that

$$g(v, x) = \frac{\partial F}{\partial v}(v, x), \qquad h(v, x) = -\frac{\partial F}{\partial x}(v, x), \tag{39.12}$$

where $F(v, x)$ is a given function of v and x. Above we considered the case when $g(v, x) = g(v)$ is a function of v only and $h(v, x) = h(x)$ is a function of x only, and $F(v, x) = G(v) - H(x)$ with $G(v)$ and $H(x)$ primitive functions of $g(v)$ and $h(x)$, respectively. Now we allow $F(v, x)$ to have a more general form.

Assume now that $u(x)$ satisfies the equation

$$F(u(x), x) = C \qquad \text{for } 0 < x \leq 1.$$

Differentiating both sides with respect to x, using a generalization of the Chain rule, we then get

$$\frac{\partial F}{\partial u}\frac{du}{dx} + \frac{\partial F}{\partial x}\frac{dx}{dx} = g(x,u(x))u'(x) - h(x,u(x)) = 0,$$

and thus $u'(x)$ solves (39.10) with $f(u(x),x)$ of the form (39.11). Again, we thus have rewritten a differential equation as an algebraic equation $F(x,u(x)) = C$ with x acting as a parameter. We give an example. The reader can construct many other similar examples.

Example 39.3. Let $F(v,x) = \frac{x^3}{3} + xv + \frac{v^3}{3}$ so that $g(v,x) = \frac{\partial F}{\partial v} = x + v^2$ and $h(v,x) = -\frac{\partial F}{\partial x} = -x^2 - v$. If $u(x)$ satisfies the algebraic equation $\frac{x^3}{3} + xu(x) + \frac{u^3(x)}{3} = C$ for $x \in [0,1]$, then $u(x)$ solves the differential equation

$$u'(x) = -\frac{x^2 + u(x)}{x + u^2(x)} \qquad \text{for } 0 < x < 1.$$

To sum up: In this chapter we have given analytical solution formula for some special cases of the scalar initial value problem (39.1), but we were not able to give a solution formula in the case of a general non-autonomous scalar equation.

Chapter 39 Problems

39.1. Prove that solutions $(x(t), y(t)$ of the Volterra-Lotka model satisfies

$$\bar{x} = \frac{1}{T}\int_0^T x(t)\,dt = \frac{c}{d}, \quad \bar{y} = \frac{1}{T}\int_0^T y(t)\,dt = \frac{a}{b},$$

where T is the period of periodic solutions. Investigate the effect on the mean values $\bar{x}$ and $\bar{y}$ of hunting of both predator and prey corresponding to including dissipative terms $-\epsilon x$ and $-\epsilon y$ with $\epsilon > 0$. Hint: Consider the integral of $\dot{x}/x$ over a period.

39.2. Extend the Volterra-Lotka model to the model

$$\begin{aligned} \dot{x}(t) &= ax(t) - bx(t)y(t) - ex^2(t), \\ \dot{y}(t) &= -cy(t) + dx(t)y(t) - fy^2(t), \end{aligned} \tag{39.13}$$

where e and f are positive constants, with the additional terms modeling negative influences from competition within the species as the populations densities increase. Compare the solutions of the two models numerically. Is the extended system separable?

39.3. Consider the spread of an infection modeled by

$$\begin{aligned} \dot{u} &= -auv, \\ \dot{v} &= auv - bv, \end{aligned}$$

where $u(t)$ is the density of the susceptibles and $v(t)$ is that of the infectives at time t, and a and b are positive constants. The term $\pm auv$ models the transfer of susceptibles to infectives at a rate proportional to auv, and $-bv$ models the decay of infectives by death or immunity. Study the qualitative behavior of phase plane curves.

39.4. Extend the previous model by changing the first equation to $\dot{u} = -auv + \mu$, with μ a positive constant modeling a constant growth of the susceptibles. Find the equilibrium point, and study the linearized model linearized at the equilibrium point.

39.5. Motivate the following model for a national economy:

$$\dot{u} = u - av, \quad \dot{v} = b(u - v - w),$$

where u is the national income, v the rate of consumer spending and w the rate of government spending, and $a > 0$ and $b \geq 1$ are constants. Show that if w is constant, then there is an equilibrium state, that is a solution independent of time satisfying $u - av = b(u - v - w) = 0$. Show that the economy oscillates if $b = 1$. Study the stability of solutions. Study a model with $w = w_0 + cu$ with w_0 a constant. Show that there is no equilibrium state in this model if $c \geq (a-1)/a$. Draw some conclusion. Study a model with $w = w_0 + cu^2$.

39.6. Consider a boat being rowed across a river occupying the strip $\{(x, y) : 0 \leq x \leq 1, y \in \mathbb{R}\}$, in such a way that the boat always points in the direction of $(0, 0)$. Assume that the boat moves with the constant speed u relative to the water and that the river flows with constant speed v in the positive y-direction. Show that the equations of motion are

$$\dot{x} = -\frac{ux}{\sqrt{x^2 + y^2}}, \ \dot{y} = -\frac{uy}{\sqrt{x^2 + y^2}}.$$

Show that the phase-plane curves are given by

$$y = \sqrt{x^2 + y^2} = Ax^{1-\alpha}, \quad \text{where } \alpha - \frac{v}{u}.$$

What happens if $v > u$?. TS^t Compute solutions.

TS^t Please check the punctuation mark.

Editor's or typesetter's annotations (will be removed before the final TEX run)

40
The General Initial Value Problem

Things are as they are because they were as they were.
(Thomas Gold)

40.1 Introduction

We now consider the Initial Value Problem or IVP for a *system* of nonlinear differential equations of the form: Find $u : [0,1] \to \mathbb{R}^d$ such that

$$u'(x) = f(u(x), x) \quad \text{for } 0 < x \leq 1,\ u(0) = u^0, \tag{40.1}$$

where $f : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ is a given bounded and Lipschitz continuous function, $u^0 \in \mathbb{R}^d$ is a given initial value, and $d \geq 1$ is the dimension of the system. The reader may assume $d = 2$ or $d = 3$, recalling the chapters on analytic geometry in $\mathbb{R}^2$ and $\mathbb{R}^3$, and extend to the case $d > 3$ after having read the chapter on analytic geometry in $\mathbb{R}^n$ below. The material in Chapter *Vector-valued functions of several real variables* is largely motivated from the need of studying problems of the form (40.1), and there is thus a close connection between this chapter and the present one. We keep this chapter abstract (and a bit philosophical), and present many concrete examples below. Note that for notational convenience we here use superscript index in the initial value u^0 (instead of u_0).

The IVP (40.1) is the non-autonomous vector version of the scalar initial value problem (38.1), and reads as follows in component form: Find

functions $u_i : [0,1] \to \mathbb{R}$, $i = 1, \ldots, d$, such that

$$\begin{aligned} u_1'(x) &= f_1(u_1(x), u_2(x), \ldots, u_d(x), x) \quad \text{for } 0 < x \leq 1, \\ u_2'(x) &= f_2(u_1(x), u_2(x), \ldots, u_d(x), x) \quad \text{for } 0 < x \leq 1, \\ &\cdots\cdots\cdots \\ u_d'(x) &= f_d(u_1(x), u_2(x), \ldots, u_d(x), x) \quad \text{for } 0 < x \leq 1, \\ u_1(0) &= u_{10},\ u_2(0) = u_{20},\ u_d(0) = u_d^0, \end{aligned} \tag{40.2}$$

where $f_i : \mathbb{R}^d \times [0,1] \to \mathbb{R}$, $i = 1, \ldots, d$, are given functions and u_i^0, $i = 1, \ldots, d$, are given initial values. With vector notation writing $u = (u_1, \ldots, u_d)$, $f = (f_1, \ldots, f_d)$ and $u^0 = (u_1^0, \ldots, u_d^0)$, we may write (40.2) in the compact form (40.1). Of course, writing $f : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$, means that for each vector $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$ and $x \in [0,1]$ there is assigned a vector $f(v,x) = (f_1(v,x), \ldots, f_d(v,x)) \in \mathbb{R}^d$, where $f_i(v,x) = f_i(v_1, \ldots, v_d, x)$.

We assume Lipschitz continuity and boundedness of $f : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ in the form: There are constants L_f and M_f such that for all $v, w \in \mathbb{R}^d$ and $x, y \in [0,1]$,

$$|f(v,x) - f(w,y)| \leq L_f(|v - w| + |x - y|) \quad \text{and} \quad |f(v,x)| \leq M_f, \tag{40.3}$$

where $|v| = (\sum_{i=1}^d v_i^2)^{1/2}$ is the Euclidean norm of $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$.

In short, everything looks the same as in the scalar case of (38.1) with the natural extension to a non-autonomous problem, but the vector interpretation makes the actual content of this chapter vastly different from that of Chapter *Scalar autonomous initial value problems*. In particular, there is in general no analytical solution formula if $d > 1$, since the solution formula for $d = 1$ based on the existence of a primitive function of $1/f(v)$, does not generalize to $d > 1$.

We prove the existence of a unique solution of the IVP (40.1) by using a constructive process which is a direct generalization of the method used for the scalar problem (38.1), which was a direct generalization of method used to construct the integral. The result of this chapter is definitely one of the highlights of mathematics (or at least of this book), because of its generality and simplicity: $f : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ can be *any* bounded Lipschitz continuous function with the dimension d arbitrarily large, and the proof looks exactly the same as in the scalar case. Therefore this chapter has a central role in the book and couples closely to several other chapters below including *Analytic geometry in* $\mathbb{R}^n$, *Solving systems of linear equations*, *Linearization and stability of IVP*, *Adaptive IVP-solvers*, *Vector-valued functions of several real variales* and various chapters on applications including mechanical systems, electrical circuits, chemical reactions, and other phenomena. This means that a full appreciation of this chapter can only be made after digesting all this material. Nevertheless, it should be possible to go through this chapter and understand that the general IVP (40.1) can be solved through a constructive process requiring more or less

work. This chapter also may be used as a basis for a bit of philosophical discussion on constructive aspects of the World, as we now proceed to do (for the interested reader).

40.2 Determinism and Materialism

Before we present the existence proof (which we thus have already seen), we pause to reflect a little on the related *mechanistic/deterministic* view of science and philosophy going back to Descartes and Newton and forming the basis the industrial society leading into our own time. With this view the World is like a big mechanical Clock governed by laws of mechanics, which may be modeled as an initial value problem of the form (40.1) with a certain function f and initial value u^0 at time $x = 0$. The state of this system for positive time is, according to the existence proof, uniquely determined by the function f and u^0, which would support a *deterministic* or *materialistic* view of the World including the mental processes in human beings: everything that will happen in the future is in principle determined by the present state (assuming no blow-up). Of course, this view is in serious conflict with massive everyday experience of unpredictability and our firm belief in the existence of a *free will*, and considerable efforts have gone into resolving this paradox through the centuries without complete success.

Let's see if we can approach this paradox from a mathematical point of view. After all, the determinitic/materialistic view is founded on a proof of existence of a unique solution of an initial value problem of the form (40.1), and thus the roots of the paradox may be hidden in the mathematical proof itself. We will argue that the resolution of the paradox must be coupled to aspects of *predictability* and *computability* of the problem (40.1), which we will now briefly touch upon and return to in more detail below. We hope the reader is open for this type of discussion, seldom met in a Calculus text. We try to point to the necessity of a proper understanding of a mathematical result, which may appear to be very simple and clear, like the existence proof to be presented, but which in fact may require a lot of explanation and qualification to avoid misunderstanding.

40.3 Predictability and Computability

The *predictability* of the problem (40.1) concerns the *sensitivity* of the solution to the given *data*, that is, the function f and the initial value u^0. The sensitivity is a measure of the change of the solution under changes of the data f and u^0. If the solution changes very much even for very small changes of data, then the sensitivity is very high. In such a case we need to know the data with very high precision to accurately predict the solution.

We shall see below that solutions of certain initial value problems are highly sensitive to changes in data and in these problems accurate prediction will be impossible. An example is given by the familiar process of tossing a coin, which can be modeled as an initial value problem. In principle, by repeatedly choosing the same initial value, the person tossing the coin should be able to always get heads, for example. However, we all know that this is impossible in practice, because the process is too sensitive to small changes in the initial value (and the corresponding function f). To handle this type of unpredictability the scientific field of *statistics* has been developed.

Similarly, the *computability* of the problem (40.1) concerns (i) the sensitivity of the solution to errors made in constructing the solution according to the existence proof, and (ii) the amount of computational work needed to construct the solution. Usually, (i) and (ii) go hand in hand: if the sensitivity is high, then a lot of work is required and of course the work also increases with the dimension d. A highly sensitive problem with d very large is thus a computational night-mare. To construct the solution of the initial value problem for even a small part of the Universe will thus be practically impossible with any kind of computer, and claiming that in principle the solution is determined would make little sense.

We will meet this problem with painstaking evidence when we turn into numerical methods. We will see that most systems of the form (40.1) with d small ($d \leq 10$ say) may be solved within fractions of a second on a PC, while some systems (like the famous Lorenz system with $d = 3$ to be studied below) quickly will exhaust even supercomputers because of very high sensitivity. We will further see that many systems of practical interest with d large ($d \approx 10^6 - 10^7$) can be solved within minutes/hours on a PC, while accurate modeling of e.g. turbulent flow requires $d \geq 10^{10}$ and supercomputer power. The most powerful super-computer in sight, the *Blue Gene* consisting of 10^6 connected PCs to appear in a couple of years, is designed for initial value problems of molecular dynamics of protein folding for the purpose of medical drug design. A landmark in computing was set in 1997 when the chess computer *Deep Blue* put the the world-champion Gary Kasparov chess mate.

The computational work required to solve (40.1) may thus vary considerably. Below we shall successively uncover a bit of this mystery and identify basic features of problems requiring different amounts of computational work.

We will return to the concepts of predictability and computability of differential equations below. Here we just wanted to give some perspective on the constructive existence proof to be given showing some limits of mathematics as a human activity.

40.4 Construction of the Solution

The construction of the solution $u(x)$ of (40.1) looks identical to the construction of the solution of (38.1), after we interpret $u(x)$ and $f(u(x))$ as vectors instead of scalars and make the natural extension to a non-autonomous problem.

We begin by discretizing $[0,1]$ using a mesh with nodes $x_i^n = ih_n$ for $i = 1, \cdots, N$, where $h_n = 2^{-n}$ and $N = 2^n$. For $n = 1, \cdots, N$, we define an approximate piecewise linear solution $U^n : [0,1] \to \mathbb{R}^d$ by the formula

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n f(U^n(x_{i-1}^n), x_{i-1}^n), \quad \text{for } i = 1, \cdots, N, \quad (40.4)$$

and setting $U^n(0) = u^0$. Note that $U^n(x)$ is linear on each subinterval $[x_{n-1}^n, x_i^n]$.

We want to prove that for $x \in [0,1]$, $\{U^n(x)\}_{n=1}^{\infty}$ is a Cauchy sequence in $\mathbb{R}^d$. We start by estimating $U^n(x_i^n) - U^{n+1}(x_i^n)$ for $i = 1, \cdots, N$. Taking two steps with step size $h_{n+1} = \frac{1}{2}h_n$ to go from time $x_{i-1}^n = x_{2i-2}^{n+1}$ to $x_i^n = x_{2i}^{n+1}$, we have

$$\begin{aligned} U^{n+1}(x_{2i-1}^{n+1}) &= U^{n+1}(x_{2i-2}^{n+1}) + h_{n+1} f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n), \\ U^{n+1}(x_{2i}^{n+1}) &= U^{n+1}(x_{2i-1}^{n+1}) + h_{n+1} f(U^{n+1}(x_{2i-1}^{n+1}), x_{2i-1}^{n+1}). \end{aligned}$$

Inserting now the value of $U^{n+1}(x_{2i-1}^{n+1})$ at the intermediate step x_{2i-1}^{n+1} from the first equation into the second equation, we get

$$\begin{aligned} U^{n+1}(x_{2i}^{n+1}) = U^{n+1}(x_{2i-2}^{n+1}) &+ h_{n+1} f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n) \\ &+ h_{n+1} f\big(U^{n+1}(x_{2i-2}^{n+1}) + h_{n+1} f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n), x_{2i-1}^{n+1}\big). \end{aligned} \quad (40.5)$$

Setting $e_i^n \equiv U^n(x_i^n) - U^{n+1}(x_{2i}^{n+1})$ and subtracting (40.5) from (40.4) gives

$$\begin{aligned} e_i^n = e_{i-1}^n &+ h_n \big(f(U^n(x_{i-1}^n), x_{i-1}^n) - f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n)\big) \\ &+ h_{n+1} \Big(f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n) - f\big(U^{n+1}(x_{2i-2}^{n+1}) \\ &+ h_{n+1} f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n), x_{2i-1}^{n+1}\big) \Big) \equiv e_{i-1}^n + F_{1,n} + F_{2,n}, \end{aligned}$$

with the obvious definitions of $F_{1,n}$ and $F_{2,n}$. Using (40.3), we have

$$\begin{aligned} &|F_{1,n}| \le L_f h_n |e_{i-1}^n|, \\ &|F_{2,n}| \le L_f h_{n+1}^2 (|f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n)| + 1) \le L_f \bar{M}_f h_{n+1}^2, \end{aligned}$$

where $\bar{M}_f = M_f + 1$, and so for $i = 1, \cdots, N$,

$$|e_i^n| \le (1 + L_f h_n)|e_{i-1}^n| + L_f \bar{M}_f h_{n+1}^2.$$

Iterating this inequality over i and using that $e_0^n = 0$, we get

$$|e_i^n| \le L_f \bar{M}_f h_{n+1}^2 \sum_{k=0}^{i-1} (1 + L_f h_n)^i \quad \text{for } i = 1, \cdots, N.$$

Recalling (31.10) and (31.27), we have

$$\sum_{k=0}^{i-1} (1 + L_f h_n)^k = \frac{(1 + L_f h_n)^i - 1}{L_f h_n} \le \frac{\exp(L_f) - 1}{L_f h_n},$$

and thus we have proved that for $i = 1, \cdots, N$,

$$|e_i^n| \le \frac{1}{2} \bar{M}_f \exp(L_f) h_{n+1},$$

that is, for $\bar{x} = i h_n$ with $i = 0, \ldots, N$,

$$|U^n(\bar{x}) - U^{n+1}(\bar{x})| \le \frac{1}{2} \bar{M}_f \exp(L_f) h_{n+1}.$$

Iterating this inequality as in the proof of the Fundamental Theorem, we get for $m > n$ and $\bar{x} = i h_n$ with $i = 0, \ldots, N$,

$$|U^n(\bar{x}) - U^m(\bar{x})| \le \frac{1}{2} \bar{M}_f \exp(L_f) h_n. \tag{40.6}$$

Again as in the proof of the Fundamental Theorem, we conclude that $\{U^n(x)\}$ is a Cauchy sequence for each $x \in [0, 1]$, and thus converges to a function $u(x)$, which by the construction satisfies the differential equation $u'(x) = f(u(x))$ for $x \in (0, 1]$ and $u(0) = u^0$, and thus the limit $u(x)$ is a solution of the initial value problem (40.1). Uniqueness of a solution follows as in the scalar case considered in Chapter *Scalar autonomous initial value problems*. We have now proved the following basic result:

Theorem 40.1 *The initial value problem* (40.1) *with* $f : \mathbb{R}^d \times [0, 1] \to \mathbb{R}^d$ *bounded and Lipschitz continuous, has a unique solution* $u(x)$*, which is the limit of the sequence of continuous piecewise linear functions* $\{U^n(x)\}$ *constructed from* (40.4) *and satisfying*

$$|u(x) - U^n(x)| \le (M_f + 1) \exp(L_f) h_n \qquad \text{for } x \in [0, 1]. \tag{40.7}$$

40.5 Computational Work

The convergence estimate (40.7) indicates that the work required to compute a solution $u(x)$ of (40.1) to a given accuracy is proportional to $\exp(L_f)$ and to $\exp(L_f T)$ if we consider a time interval $[0, T]$ instead of $[0, 1]$. With

$L_f = 10$ and $T = 10$, which would seem to be a very innocent case, we would have $\exp(L_f T) = \exp(10^2)$ and we would thus have to choose h_n smaller than $\exp(-10^2) \approx 10^{-30}$, and the number of computational operations would be of the order 10^{30} which would be at the limit of any practical possibility. Already moderately large constants such as $L_f = 100$ and $T = 100$, would give an exponential factor $\exp(10^4)$ way beyond any comprehension. We conclude that the appearance of the exponential factor $\exp(L_f T)$, which corresponds to a worst possible case, seems to limit the interest of the existence proof. Of course, the worst possible case does not necessarily have to occur always. Below we will present problems with special features for which the error is actually smaller than worst possible, including the important class of *stiff problems* where large Lipschitz constants cause quick exponential decay instead of exponential growth, and the Lorenz system where the error growth turns out to be of order $\exp(T)$ instead of $\exp(L_f T)$ with $L_f = 100$.

40.6 Extension to Second Order Initial Value Problems

Consider a second order initial value problem

$$\ddot{v}(t) = g(v(t), \dot{v}(t)) \text{ for } 0 < t \leq 1,\ v(0) = v_0,\ \dot{v}(0) = \dot{v}_0, \tag{40.8}$$

with initial conditions for $v(0)$ and $\dot{v}(0)$, where $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous, $v : [0,1] \to \mathbb{R}^d$ and $\dot{v} = \frac{dv}{dt}$. In mechanics, initial value problems often come in such second order form as they express Newton's Law with $\ddot{v}(t)$ representing acceleration and $g(v(t), \dot{v}(t))$ force. This problem can be reduced to a first order system of the form (40.1) by introducing the new variable $w(t) = \dot{v}(t)$ and writing (40.8) as

$$\begin{aligned} \dot{w}(t) &= g(v(t), w(t)) \quad \text{for } 0 < t \leq 1, \\ \dot{v}(t) &= w(t) \quad \text{for } 0 < t \leq 1, \\ v(0) &= v_0,\ w(0) = \dot{v}_0. \end{aligned} \tag{40.9}$$

Setting $u = (u_1, \ldots, u_{2d}) = (v_1, \ldots, v_d, w_1, \ldots, w_d)$ and $f(u) = (g_1(u), \cdots, g_d(u), u_{d+1}, \ldots, u_{2d})$, the system (40.9) takes the form $\dot{u}(t) = f(u(t))$ for $0 < t \leq 1$, and $u(0) = (v_0, \dot{v}_0)$.

In particular, we can rewrite the second order scalar equations $\ddot{v} + v = 0$ as a first order system and obtain existence of the trigonometric functions via the general existence result for first order systems as solutions of the corresponding initial value problem with appropriate data.

40.7 Numerical Methods

The computational solution of differential equations is an important subject with many aspects. The overall objective may be viewed to be to compute approximate solutions with as little work as possible per digit of accuracy. So far we have discussed only the simplest method for constructing approximate solutions. In this section, we give a brief glimpse of other methods. In Chapter *Adaptive IVP solvers*, we continue this study.

The computational method we have used so far, in which

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n f(U^n(x_{i-1}^n), x_{i-1}^n), \quad \text{for } i = 1, \cdots, N, \quad (40.10)$$

with $U^n(0) = u^0$, is called the *forward Euler* method. The forward Euler method is an *explicit* method because we can directly compute $U^n(x_i^n)$ from $U^n(x_{i-1}^n)$ without solving a system of equations.

In contrast, the *backward Euler* method in which the approximate solution is computed via the equation

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n f(U^n(x_i^n), x_i^n), \quad \text{for } i = 1, \cdots, N, \quad (40.11)$$

with $U^n(0) = u^0$, is an *implicit* method. At each step we need to solve the system

$$V = U^n(x_{i-1}^n) + h_n f(V, x_i^n), \quad (40.12)$$

to compute $U^n(x_i^n)$ from $U^n(x_{i-1}^n)$. Another implicit method is the *midpoint method*

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n f\left(\frac{1}{2}(U^n(x_{i-1}^n) + U^n(x_i^n)), \bar{x}_{i-1}^n\right), \; i = 1, \cdots, N, \quad (40.13)$$

with $\bar{x}_{i-1}^n = \frac{1}{2}(x_{i-1}^n + x_i^n)$, where we have to solve the system

$$V = U^n(x_{i-1}^n) + h_n f\left(\frac{1}{2}(U^n(x_{i-1}^n) + V), \bar{x}_{i-1}^n\right) \quad (40.14)$$

at each step. Note that both (40.12) and (40.14) are nonlinear equations when f is nonlinear. We may use Fixed Point Iteration or Newton's method to solve them, see Chapter *Vector-valued functions of several real variables* below.

We also present the following variant of the midpoint method, which we call the cG(1), *continuous Galerkin method with trial functions of order* 1:.[TS u] The approximate solution is computed via

$$U^n(x_i^n) = U^n(x_{i-1}^n) + \int_{x_{i-1}^n}^{x_i^n} f(U(x), x)\, dx, \quad i = 1, \cdots, N, \quad (40.15)$$

and $U^n(0) = u^0$, where $U^n(x)$ is continuous piecewise linear function with the values $U^n(x_i^n)$ at the nodes x_i^n. If we evaluate the integral in (40.15)

[TS u] Please check this punctuation mark.
[TS v] Please check it.

Editor's or typesetter's annotations (will be removed before the final TEX run)

with the midpoint quadrature rule, we obtain the midpoint method. We can of course use other quadrature formulas to get different methods.

We shall see that cG(1) is the first in a family of methods cG(q)TSY with $q = 1, 2, \ldots$, where the solution is approximated by continuous piecewise polynomials of order q. The Galerkin feature of cG(1) is expressed by the fact that the method can be formulated as

$$\int_{x_{i-1}^n}^{x_i^n} \left(\frac{dU^n}{dx}(x) - f(U^n(x), x) \right) dx = 0,$$

stating that the mean-value over each subinterval of the *residual* $\frac{dU^n}{dx}(x) - f(U(x), x)$ of the continuous piecewise linear approximate solution $U^n(x)$, is equal to zero (or that the residual is orthogonal to the set of constant functions on each subinterval with a terminology to be used below).

We can prove convergence of the backward Euler and midpoint methods in the same way as for the forward Euler method. The forward and backward Euler methods are *first order accurate* methods in the sense that the error $|u(x) - U^n(x)|$ is proportional to the step size h_n, while the midpoint method is *second order accurate* with the error proportional to h_n^2 and thus in general is more accurate. The computational work per step is generally smaller for an explicit method than for an implicit method, since no system of equations has to be solved at each step. For so-called stiff problems, explicit methods may require very small time steps compared to implicit methods, and then implicit methods can give a smaller total cost. We will return to these issues in Chapter *Adaptive IVP solvers* below.

Note that all of the methods discussed so far generalize to allow non-uniform meshes $0 = x_0 < x_1 < x_2 < \ldots < x_N = 1$ with possibly varying steps $x_i - x_{i-1}$. We will below return to the problem of *automatic step-size control* with the purpose of keeping the error $|u(x_i) - U(x_i)| \leq TOL$ for $i = 1, \cdots, N$, where TOL is a given tolerance, while using as few time steps as possible by varying the mesh steps, cf. the Chapter *Numerical Quadrature*.

Chapter 40 Problems

40.1. Prove existence of a solution of the initial value problem (40.1) using the backward Euler method or the midpoint method.

40.2. Complete the proof of existence for (40.1) by proving that the constructed limit function $u(x)$ solves the initial value problem. Hint: use that $u_i(x) = \int_0^x f_i(u(y))\, dy$ for $x \in [0, 1]$, $i = 1, \ldots, d$.

40.3. Give examples of problems of the form (40.1).

41
Calculus Tool Bag I

> After experience had taught me that all the usual surroundings of social life are vain and futile; seeing that none of the objects of my fears contained in themselves anything either good or bad, except in so far as the mind is affected by them, I finally resolved to inquire whether there might be some real good having power to communicate itself, which would affect the mind singly, to the exclusion of all else: whether, in fact, there might be anything of which the discovery and attainment would enable me to enjoy continuous, supreme, and unending happiness. (Spinoza)
>
> Sapiens nihil affirmat quod non probat.

41.1 Introduction

We present a *Calculus Tool Bag I* containing a minimal set of important tools and concepts of Calculus for functions $f : \mathbb{R} \to \mathbb{R}$. Below, we present a *Calculus Tool Bag II* containing the corresponding of tools and concepts of Calculus for functions $f : \mathbb{R}^m \to \mathbb{R}^n$.

41.2 Rational Numbers

We start with the set of integers $\mathbb{Z} = \{\cdots, -3, -2, -1, 0, 1, 2, 3, \cdots\}$ together with the usual operations of addition, subtraction and multiplication. We define the set of rational numbers $\mathbb{Q}$ as the set of pairs (p, q) with

p and $q \neq 0$ integers, and we write $(p,q) = \frac{p}{q}$ along with the arithmetic operations of addition

$$\frac{p}{q} + \frac{r}{s} = \frac{ps+qr}{qs},$$

multiplication

$$\frac{p}{q} \times \frac{r}{s} = \frac{pr}{qs},$$

and division

$$(p,q)/(r,s) = \frac{(p,q)}{(r,s)} = (ps, qr),$$

assuming $r \neq 0$. With the operation of division, we can solve the equation $ax = b$ to get $x = b/a$ for $a, b \in \mathbb{Q}$ with $a \neq 0$.

Rational numbers have periodic decimal expansions. There is no rational number x such that $x^2 = 2$.

41.3 Real Numbers. Sequences and Limits

Definitions: A real number is specified by an *infinite decimal expansion* of the form

$$\pm p_m \cdots p_0.q_1 q_2 q_3 \cdots$$

with a never ending list of decimals q_1, q_2, ..., where each of the p_i and q_j are one of the 10 digits $0, 1, \ldots, 9$. The set of (all possible) real numbers is denoted by $\mathbb{R}$.

A sequence $\{x_i\}_{i=1}^{\infty}$ of real numbers converges to a real number x if for any $\epsilon > 0$ there is a natural number N such that $|x_i - x| < \epsilon$ for $i \geq N$ and we then write $x = \lim_{i\to\infty} x_i$.

A sequence $\{x_i\}_{i=1}^{\infty}$ of real numbers is a Cauchy sequence if for all $\epsilon > 0$ there is a natural number N such that

$$|x_i - x_j| \leq \epsilon \quad \text{for} \quad i, j \geq N.$$

Basic properties: A convergent sequence of real numbers is a Cauchy sequence. A Cauchy sequence of real numbers converges to a unique real number. We have $\lim_{i\to\infty} x_i = x$, where $\{x_i\}_{i=1}^{\infty}$ is the sequence of truncated decimal expansions of x.

41.4 Polynomials and Rational Functions

A polynomial function $f : \mathbb{R} \to \mathbb{R}$ of degree n has the form $f(x) = a_0 + a_1 x + \cdots + a_n x^n$ with coefficients $a_i \in \mathbb{R}$. A rational function $h(x)$ has the form $h(x) = f(x)/g(x)$, where $f(x)$ and $g(x)$ are polynomials.

41.5 Lipschitz Continuity

Definition: A function $f : I \to \mathbb{R}$, where I is an interval of real numbers, is Lipschitz continuous on I with Lipschitz constant $L_f \geq 0$ if

$$|f(x_1) - f(x_2)| \leq L_f|x_1 - x_2| \quad \text{for all } x_1, x_2 \in I.$$

Basic facts: Polynomial functions are Lipschitz continuous on bounded intervals. Sums, products and composition of Lipschitz continuous functions are Lipschitz continuous. Quotients of Lipschitz continuous functions are Lipschitz continuous on intervals where the denominator is bounded away from zero. A Lipschitz continuous function $f : I \to \mathbb{R}$, where I is an interval of real numbers, satisfies:

$$f(\lim_{i\to\infty} x_i) = \lim_{i\to\infty} f(x_i),$$

for any convergent sequence $\{x_i\}$ in I with $\lim_{i\to\infty} x_i \in I$.

41.6 Derivatives

Definition: The function $f : (a, b) \to \mathbb{R}$ is differentiable at $\bar{x} \in (a, b)$ with derivative $f'(\bar{x}) = \frac{df}{dx}(\bar{x})$ if there are real numbers $f'(\bar{x})$ and $K_f(\bar{x})$ such that for $x \in (a, b)$ close to $\bar{x}$,

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + E_f(x, \bar{x}),$$
$$\text{with } |E_f(x, \bar{x})| \leq K_f(\bar{x})|x - \bar{x}|^2.$$

If the constant $K_f(\bar{x})$ can be chosen independently of $\bar{x} \in (a, b)$, then $f : (a, b) \to \mathbb{R}$ is said to be uniformly differentiable on (a, b).

Derivative of x^α with $\alpha \neq 0$: The derivative of $f(x) = x^\alpha$ is $f'(x) = \alpha x^{\alpha-1}$ for $\alpha \neq 0$, and $x \neq 0$ for $\alpha < 1$.

Bounded derivative implies Lipschitz continuity: If $f(x)$ is uniformly differentiable on the interval $I = (a, b)$ and there is a constant L such that

$$|f'(x)| \leq L, \quad \text{for } x \in I,$$

then $f(x)$ is Lipschitz continuous on I with Lipschitz constant L.

41.7 Differentiation Rules

Linear Combination rule:

$$(f + g)'(x) = f'(x) + g'(x),$$
$$(cf)'(x) = cf'(x),$$

where c is a constant.

Product rule:

$$(fg)'(x) = f(x)g'(x) + f'(x)g(x).$$

Chain rule:

$$(f \circ g)'(x) = f'(g(x))g'(x), \quad \text{or}$$
$$\frac{dh}{dx} = \frac{df}{dy}\frac{dy}{dx},$$

where $h(x) = f(y)$ and $y = g(x)$, that is $h(x) = f(g(x)) = (f \circ g)(x)$.

Quotient rule:

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2},$$

provided $g(x) \neq 0$.

The derivative of an inverse function:

$$\frac{d}{dy}f^{-1}(y) = \frac{1}{\frac{d}{dx}f(x)}\,.$$

where $y = f(x)$ and $x = f^{-1}(y)$.

41.8 Solving $f(x) = 0$ with $f : \mathbb{R} \to \mathbb{R}$

Bisection: If $f : [a, b] \to \mathbb{R}$ is Lipschitz continuous on $[a, b]$ and $f(a)f(b) < 0$, then the Bisection algorithm converges to a root $\bar{x} \in [a, b]$ of $f(x) = 0$.

Fixed Point Iteration: A Lipschitz continuous function $g : \mathbb{R} \to \mathbb{R}$ with Lipschitz constant $L < 1$ is said to be a contraction mapping. A contraction mapping $g : \mathbb{R} \to \mathbb{R}$ has a unique fixed point $\bar{x} \in \mathbb{R}$ satisfying $\bar{x} = g(\bar{x})$ and any sequence $\{x_i\}_{i=1}^{\infty}$ generated by Fixed Point Iteration $x_i = g(x_{i-1})$ converges to $\bar{x}$.

Bolzano's theorem: If $f : [a, b] \to \mathbb{R}$ is Lipschitz continuous and $f(a)f(b) < 0$, then there is a real number $\bar{x} \in [a, b]$ such that $f(\bar{x}) = 0$ (consequence of Bisection above).

Newton's method: Newton's method $x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$ for computing a root $\bar{x}$ of $f : \mathbb{R} \to \mathbb{R}$ converges quadratically if $f'(x)$ is bounded away from zero for x close to $\bar{x}$ and the initial approximation is sufficiently close to the root $\bar{x}$.

41.9 Integrals

The Fundamental Theorem of Calclulus: If $f : [a, b]$ is Lipschitz continuous, then there is a unique uniformly differentiable function $u : [a, b] \to \mathbb{R}$, that solves the initial value problem

$$\begin{cases} u'(x) = f(x) & \text{for } x \in (a, b], \\ u(a) = u_a, \end{cases}$$

where $u_a \in \mathbb{R}$ is given. The function $u : [a, b] \to \mathbb{R}$ can be expressed as

$$u(\bar{x}) = u_a + \int_a^{\bar{x}} f(x)\,dx \quad \text{for } \bar{x} \in [a, b],$$

where

$$\int_0^{\bar{x}} f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{j} f(x_{i-1}^n)h_n,$$

with $\bar{x} = x_j^n$, $x_i^n = a + ih_n$, $h_n = 2^{-n}(b-a)$. More precisely, if the Lipschitz constant of f is L_f then for $n = 1, 2, \ldots,$

$$\Big| \int_a^{\bar{x}} f(x)\,dx - \sum_{i=1}^{j} f(x_{i-1}^n)h_n \Big| \le \frac{1}{2}(\bar{x} - a)L_f h_n.$$

Furthermore, if $|f(x)| \le M_f$ for $x \in [a, b]$, then $u : [a, b] \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant M_f and $K_u \le \frac{1}{2}L_f$, where K_u is the constant of uniform differentiability of u.

Additivity:

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

Linearity: If α and β are real numbers then,

$$\int_a^b (\alpha f(x) + \beta g(x))\,dx = \alpha \int_a^b f(x)\,dx + \beta \int_a^b g(x)\,dx.$$

Monotonicity: If $f(x) \ge g(x)$ for $a \le x \le b$, then

$$\int_a^b f(x)\,dx \ge \int_a^b g(x)\,dx.$$

Differentiation and integration are inverse operations:

$$\frac{d}{dx}\int_a^x f(y)\,dy = f(x).$$

Change of variables: Setting $y = g(x)$, we have with formally $dy = g'(x)\,dx$,

$$\int_a^b f(g(x))g'(x)\,dx = \int_{g(a)}^{g(b)} f(y)\,dy.$$

Integration by parts:

$$\int_a^b u'(x)v(x)\,dx = u(b)v(b) - u(a)v(a) - \int_a^b u(x)v'(x)\,dx.$$

The Mean Value theorem: If $u(x)$ is uniformly differentiable on $[a, b]$ with Lipschitz continuous derivative $u'(x)$, then there is a (at least one) $\bar{x} \in [a, b]$, such that

$$u(b) - u(a) = u'(\bar{x})(b - a).$$

Taylor's theorem:

$$u(x) = u(\bar{x}) + u'(\bar{x})(x - \bar{x}) + \cdots + \frac{u^{(n)}(\bar{x})}{n!}(x - \bar{x})^n + \int_{\bar{x}}^x \frac{(x-y)^n}{n!}\,u^{(n+1)}(y)\,dy.$$

41.10 The Logarithm

Definition:

$$\log(x) = \int_1^x \frac{1}{y}\,dy \quad \text{for } x > 0.$$

Basic properties:

$$\begin{aligned}\frac{d}{dx}\log(x) &= \frac{1}{x} \quad \text{for } x > 0,\\ \log(ab) &= \log(a) + \log(b) \quad \text{for } a, b > 0,\\ \log(a^r) &= r\log(a), \quad \text{for } r \in \mathbb{R}, a > 0.\end{aligned}$$

41.11 The Exponential

Definition: $\exp(x) = e^x$ is the unique solution of the differential equation $u'(x) = u(x)$ for $x \in \mathbb{R}$ and $u(0) = 1$.

Basic properties:

$$\frac{d}{dx}\exp(x) = \exp(x),$$
$$\exp(a+b) = \exp(a)\exp(b) \quad \text{or } e^{a+b} = e^a e^b,$$
$$\exp(x) = \lim_{j\to\infty}\left(1 + \frac{x}{j}\right)^j.$$

The inverse of the exponential is the logarithm:

$$y = \exp(x) \quad \text{if and only if } x = \log(y).$$

The function a^x with $a > 0$:

$$a^x = \exp(x\log(a)), \quad \frac{d}{dx}a^x = \log(a)a^x.$$

41.12 The Trigonometric Functions

Definition of $\sin(x)$ **and** $\cos(x)$: The initial value problem $u''(x)+u(x)=0$ for $x > 0$ with $u_0 = 0$ and $u_1 = 1$, has a unique solution, which is denoted by $\sin(x)$. The initial value problem $u''(x) + u(x) = 0$ for $x > 0$ with $u_0 = 1$ and $u_1 = 0$, has a unique solution, which is denoted by $\cos(x)$. The functions $\sin(x)$ and $\cos(x)$ extend to $x < 0$ as solutions of $u''(x)+u(x) = 0$ and are periodic with period 2π, and $\sin(\pi) = 0$, $\cos(\frac{\pi}{2}) = 0$.

Properties:

$$\frac{d}{dx}\sin(x) = \cos(x),$$
$$\frac{d}{dx}\cos(x) = -\sin(x), \cos(-x) = \cos(x),$$
$$\sin(-x) = -\sin(x),$$
$$\cos(\pi - x) = -\cos(x),$$
$$\sin(\pi - x) = \sin(x),$$
$$\cos(x) = \sin\left(\frac{\pi}{2} - x\right),$$
$$\sin(x) = \cos\left(\frac{\pi}{2} - x\right),$$
$$\sin\left(\frac{\pi}{2} + x\right) = \cos(x),$$
$$\cos\left(\frac{\pi}{2} + x\right) = -\sin(x).$$

Definition of $\tan(x)$ **and** $\cot(x)$:

$$\tan(x) = \frac{\sin(x)}{\cos(x)}, \quad \cot(x) = \frac{\cos(x)}{\sin(x)}.$$

Derivatives of $\tan(x)$ **and** $\cot(x)$:

$$\frac{d}{dx}\tan(x) = \frac{1}{\cos^2(x)}, \quad \frac{d}{dx}\cot(x) = -\frac{1}{\sin^2(x)}.$$

Trigonometric formulas:

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

$$\cos(x) - \cos(y) = -2\sin\left(\frac{x+y}{2}\right)\sin\left(\frac{x-y}{2}\right),$$

$$\tan(x+y) = \frac{\tan(x)+\tan(y)}{1-\tan(x)\tan(y)},$$

$$\tan(x-y) = \frac{\tan(x)-\tan(y)}{1+\tan(x)\tan(y)},$$

$$\sin(x) + \sin(y) = 2\sin\left(\frac{x+y}{2}\right)\cos\left(\frac{x-y}{2}\right),$$

$$\sin(x) - \sin(y) = 2\cos\left(\frac{x+y}{2}\right)\sin\left(\frac{x-y}{2}\right),$$

$$\cos(x) + \cos(y) = 2\cos\left(\frac{x+y}{2}\right)\cos\left(\frac{x-y}{2}\right).$$

Inverses of trigonometric functions: The inverse of $f(x) = \sin(x)$ with $D(f) = [\frac{-\pi}{2}, \frac{\pi}{2}]$ is $f^{-1}(y) = \arcsin(y)$ with $D(\arcsin) = [-1, 1]$. The inverse of $f(x) = \tan(x)$ with $D(f) = (\frac{-\pi}{2}, \frac{\pi}{2})$ is $f^{-1}(y) = \arctan(y)$ with $D(\arctan) = \mathbb{R}$. The inverse of $y = f(x) = \cos(x)$ with $D(f) = [0, \pi]$ is $f^{-1}(y) = \arccos(y)$ with $D(\arccos) = [-1, 1]$. The inverse of $f(x) = \cot(x)$ with $D(f) = (0, \pi)$ is $f^{-1}(y) = \operatorname{arccot}(y)$ with $D(\operatorname{arccot}) = \mathbb{R}$. We have

$$\frac{d}{dy}\arcsin(y) = \frac{1}{\sqrt{1-y^2}}$$

$$\frac{d}{dy}\arctan(y) = \frac{1}{1+y^2}$$

$$\frac{d}{dy}\arccos(y) = -\frac{1}{\sqrt{1-y^2}}$$

$$\frac{d}{dy}\operatorname{arccot}(y) = -\frac{1}{1+y^2},$$

$$\arctan(u) + \arctan(v) = \arctan\left(\frac{u+v}{1-uv}\right).$$

Definition of sinh(x) and cosh(x):

$$\sinh(x) = \frac{e^x - e^{-x}}{2} \quad \text{and} \quad \cosh(x) = \frac{e^x + e^{-x}}{2} \quad \text{for } x \in \mathbb{R}.$$

Derivatives of of sinh(x) and cosh(x):

$$D\sinh(x) = \cosh(x) \quad \text{and} \quad D\cosh(x) = \sinh(x).$$

Inverses of of sinh(x) and cosh(x): the inverse of $y = f(x) = \sinh(x)$ with $D(f) = \mathbb{R}$ is $f^{-1}(y) = \operatorname{arcsinh}(y)$ with $D(\operatorname{arcsinh}) = \mathbb{R}$. The inverse of $y = f(x) = \cosh(x)$ with $D([0, \infty))$, is $f^{-1}(y) = \operatorname{arccosh}(y)$ with $D(\operatorname{arccosh}) = [1, \infty)$. We have

$$\frac{d}{dy}\operatorname{arcsinh}(y) = \frac{1}{\sqrt{y^2+1}}, \quad \frac{d}{dy}\operatorname{arccosh}(y) = \frac{1}{\sqrt{y^2-1}}.$$

41.13 List of Primitive Functions

$$\int_{x_0}^{x} \frac{1}{s-c}\,ds = \log|x-c| - \log|x_0-c|, \quad c \neq 0,$$

$$\int_{x_0}^{x} \frac{s-a}{(s-a)^2+b^2}\,dx = \frac{1}{2}\log((x-a)^2+b^2) - \frac{1}{2}\log((x_0-a)^2+b^2),$$

$$\int_{x_0}^{x} \frac{1}{(s-a)^2+b^2}\,ds = \left[\frac{1}{b}\arctan\left(\frac{x-a}{b}\right)\right] - \left[\frac{1}{b}\arctan\left(\frac{x_0-a}{b}\right)\right], \quad b \neq 0,$$

$$\int_{0}^{x} y\cos(y)\,dy = x\sin(x) + \cos(x) + 1,$$

$$\int_{0}^{x} \sin(\sqrt{y})\,dy = -2\sqrt{x}\cos(\sqrt{x}) + 2\sin(\sqrt{x}),$$

$$\int_{1}^{x} y^2\log(y)\,dy = \frac{x^3}{3}\log(x) - \frac{x^3}{9} + \frac{1}{9}$$

$$\int_{0}^{x} \frac{1}{\sqrt{1-y^2}}\,dy = \arcsin(x) \quad \text{for } x \in (-1,1)$$

$$\int_{0}^{x} \frac{1}{\sqrt{1-y^2}} = \frac{\pi}{2} - \arccos(x)\,dy \quad \text{for } x \in (-1,1)$$

$$\int_{0}^{x} \frac{1}{1+y^2}\,dy = \arctan(x) \quad \text{for } x \in \mathbb{R}$$

$$\int_{0}^{x} \frac{1}{1+y^2}\,dy = \frac{\pi}{2} - \operatorname{arccot}(x) \quad \text{for } x \in \mathbb{R}.$$

41.14 Series

Definition of convergence: A series $\sum_{i=1}^{\infty} a_i$ converges if and only if the sequence $\{s_n\}_{n=1}^{\infty}$ of partial sums $s_n = \sum_{i=1}^{n} a_i$ converges.

Geometric series: $\sum_{i=0}^{\infty} a^i = \frac{1}{1-a}$ if $|a| < 1$.

Basic facts: A positive series $\sum_{i=1}^{\infty} a_i$ converges if and only if the sequence of partial sums is bounded above.

The series $\sum_{i=1}^{\infty} i^{-\alpha}$ converges if and only if $\alpha > 1$.

An absolutely convergent series is convergent.

An alternating series with the property that the modulus of its terms tends monotonically to zero, converges. Example: $\sum_{i=1}^{\infty}(-i)^{-1}$ converges.

41.15 The Differential Equation $\dot{u} + \lambda(x)u(x) = f(x)$

The solution to the initial-value problem $\dot{u} + \lambda(x)u(x) = f(x)$ for $x > 0$, $u(0) = u^0$, is given by

$$u(x) = \exp(-\Lambda(x))u^0 + \exp(-\Lambda(x)) \int_0^x \exp(\Lambda(y))f(y)\, dy,$$

where $\Lambda(x)$ is a primitive function of $\lambda(x)$ satisfying $\Lambda(0) = 0$.

41.16 Separable Scalar Initial Value Problems

The solution of the separable scalar initial value problem

$$u'(x) = \frac{h(x)}{g(u(x))} \quad \text{for } 0 < x \leq 1,\ u(0) = u_0,$$

where $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ are given functions, satisfies for $0 \leq x \leq 1$ the algebraic equation

$$G(u(x)) = H(x) + C,$$

where $G(v)$ and $H(x)$ are primitive functions of $g(v)$, and $C = G(u_0) - H(0)$.

42
Analytic Geometry in $\mathbb{R}^n$

> I also think that the (mathematical) mine has become too deep and sooner or later it will be necessary to abandon it if new ore-bearing veins shall not be discovered. Physics and Chemistry display now treasures much more brilliant and easily exploitable, thus, apparently, everybody has turned completely in this direction, and possibly posts in Geometry in the Academy of Sciences will some day be like chairs in Arabic Language in universities at present. (Lagrange, 1781)

42.1 Introduction and Survey of Basic Objectives

We now generalize the discussion of analytic geometry to $\mathbb{R}^n$, where n is an arbitrary natural number. Following the pattern set above for $\mathbb{R}^2$ and $\mathbb{R}^3$, we define $\mathbb{R}^n$ to be the set of all possible ordered n-tuples of the form $(x_1, x_2, \ldots, x_n)$ with $x_i \in \mathbb{R}$ for $i = 1, \ldots, n$. We refer to $\mathbb{R}^n$ as *n-dimensional Euclidean space*.

We all have a direct concrete experience of $\mathbb{R}^3$ as the three-dimensional space of the real World, and we may think of $\mathbb{R}^2$ as an infinite flat surface, but we don't have a similar experience with for example $\mathbb{R}^4$, except possibly from some science fiction novel with space ships travelling in four-dimensional space-time. Actually, Einstein in his theory of relativity used $\mathbb{R}^4$ as the set of space-time coordinates (x_1, x_2, x_3, x_4) with $x_4 = t$ representing time, but of course had the same difficulty as we all have of

"seeing" an object in $\mathbb{R}^4$. In Fig. 42.1, we show a projection into $\mathbb{R}^3$ of a 4-cube in $\mathbb{R}^4$, and we hope the clever reader can "see" the 4-cube.

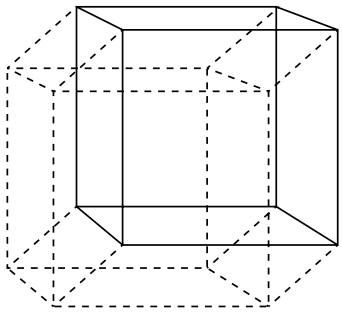

Fig. 42.1. A cube in $\mathbb{R}^4$

More generally, the need of using $\mathbb{R}^n$ arises as soon as we have n different variables to deal with, which occurs all the time in applications, and $\mathbb{R}^n$ is thus one of the most useful concepts in mathematics. Fortunately, we can work with $\mathbb{R}^n$ purely algebraically without having to draw geometric pictures, that is we can use the tools of analytic geometry in $\mathbb{R}^n$ in pretty much the same way as we have done in $\mathbb{R}^2$ and $\mathbb{R}^3$.

Most of this chapter is one way or the other connected to systems of m linear equations in n unknowns $x_1, \ldots, x_n$, of the form

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \quad \text{for } i = 1, \ldots, m, \tag{42.1}$$

that is,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2, \\ &\cdots\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_m, \end{aligned} \tag{42.2}$$

where the a_{ij} are given (real) coefficients and $(b_1, \ldots, b_m) \in \mathbb{R}^m$ is a given right-hand side. We will write this system in matrix form as

$$Ax = b, \tag{42.3}$$

that is

$$\begin{pmatrix} a_{11} & a_{12} & .. & a_{1n} \\ . & . & .. & . \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ . \\ . \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ . \\ . \\ b_m \end{pmatrix}, \tag{42.4}$$

where $A = (a_{ij})$ is a $m \times n$ matrix with rows $(a_{i1}, \ldots, a_{in})$, $i = 1, \ldots, m$, and columns $(a_{1j}, \ldots, a_{mj})$, $j = 1, \ldots, n$, and we view $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$

and $b = (b_1, \ldots, b_m) \in \mathbb{R}^m$ as column vectors. We will also write the system in the form

$$x_1 a_1 + \cdots + x_n a_n = b, \tag{42.5}$$

expressing the given column vector $b \in \mathbb{R}^m$ as a linear combination of the column vectors $a_j = (a_{1j}, a_{2j}, \ldots, a_{mj})$, $j = 1, 2, \ldots, n$, with coefficients $(x_1, \ldots, x_n)$. Notice that we use both (column) vectors in $\mathbb{R}^m$ (such as the columns of the matrix A and the right hand side b) and (column) vectors in $\mathbb{R}^n$ such as the solution vector x.

We shall view $f(x) = Ax$ as a function or transformation $f : \mathbb{R}^n \to \mathbb{R}^m$, and we thus focus on a particular case of our general problem of solving systems of equations of the form $f(x) = b$, where $f : \mathbb{R}^n \to \mathbb{R}^m$ is the *linear transformation* $f(x) = Ax$. We shall denote by $R(A)$ the *range* of $f(x) = Ax$, that is

$$R(A) = \{Ax \in \mathbb{R}^m : x \in \mathbb{R}^n\} = \left\{ \sum_{j=1}^{n} x_j a_j : x_j \in \mathbb{R} \right\},$$

and by $N(A)$ the *null space* of $f(x) = Ax$ that is

$$N(A) = \{x \in \mathbb{R}^n : Ax = 0\} = \left\{ x \in \mathbb{R}^n : \sum_{j=1}^{n} x_j a_j = 0 \right\}.$$

We are interested in the question of existence and/or uniqueness of solutions $x \in \mathbb{R}^n$ to the problem $Ax = b$ for a given $m \times n$ matrix A and right hand side $b \in \mathbb{R}^m$. Of particular interest is the case $m = n$ with as many equations as unknowns.

Existence of a solution x to $Ax = b$ is of course the same as saying that $b \in R(A)$, which is the same as saying that b is a linear combination of the columns of A. Uniqueness is the same as saying that $N(A) = 0$, because if x and $\hat{x}$ satisfy $Ax = b$ and $A\hat{x} = b$, then by linearity, $A(x - \hat{x}) = 0$, and if $N(A) = 0$ then $x - \hat{x} = 0$ that is $x = \hat{x}$. Further, the non-uniqueness of solutions of $Ax = b$ is described by $N(A)$: If $A\hat{x} = b$ and $Ax = b$, then $x - \hat{x} \in N(A)$.

We may thus formulate the following prime objectives of our study of the linear transformation $f(x) = Ax$ given by the matrix A:

- Determine $R(A)$.
- Determine $N(A)$.
- Solve $Ax = b$ for given b.

We state here the following partial answer given by the *Fundamental Theorem of Linear Algebra*, which we will prove in a couple of different ways below: Let $m = n$ and suppose that $N(A) = 0$. Then $Ax = b$ has a unique

solution for any $b \in \mathbb{R}^m$, that is, $R(A) = \mathbb{R}^m$. In other words, if $m = n$, then uniqueness implies existence.

In our study we will be led to concepts such as: linear combination, linear span, linear space, vector space, subspace, linear independence, basis, determinant, linear transformation and projection, which we have already met in the chapters on analytic geometry in $\mathbb{R}^2$ and $\mathbb{R}^3$ above.

This chapter focusses mostly on theoretical issues while the computational methods such as Gaussian elimination and iterative methods are considered in more detail in Chapter *Solving systems of linear equations* below.

42.2 Body/Soul and Artificial Intelligence

Before plunging into the geometry of $\mathbb{R}^n$, we take a brake and return to the story of Body and Soul which continues into our time with new questions: Is it possible to create computer programs for Artificial Intelligence AI, that is, can we give the computer some more or less advanced capability of acting like an intelligent organism with some ability of "thinking"? It appears that this question does not yet have a clear positive answer, despite many dreams in that direction during the development of the computer. In seeking an answer, Spencer's principle of adaptivity of course plays an important role: an intelligent system must be able to adapt to changes in its environment. Further, the presence of a goal or final cause according to Leibniz, seems to be an important feature of intelligence, to judge if an action of a system is stupid or not. Below we will design adaptive IVP-solvers, which are computer programs for solving systems of differential equations, with features of adaptive feed-back from the computational process towards the goal of error control. These IVP-solvers thus show some kind of rudiments of intelligence, and at any rate are infinitely much more "clever" than traditional non-adaptive IVP-solvers with no feed-back.

42.3 The Vector Space Structure of $\mathbb{R}^n$

We view $\mathbb{R}^n$ as a *vector space* consisting of *vectors* which are ordered n-tuples, $x = (x_1, \ldots, x_n)$ with components $x_i \in \mathbb{R}$, $i = 1, \ldots, n$. We write $x = (x_1, \ldots, x_n)$ for short, and refer to $x \in \mathbb{R}^n$ as a vector with component x_i in position i.

We may *add* two vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ in $\mathbb{R}^n$ by componentwise addition to get a new vector $x + y$ in $\mathbb{R}^n$ defined by

$$x + y = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n). \tag{42.6}$$

Further, we may multiply a vector $x = (x_1, \ldots, x_n)$ by a real number λ by componentwise multiplication with λ, to get a new vector λx in $\mathbb{R}^n$ defined

by

$$\lambda x = (\lambda x_1, \ldots, \lambda x_n). \tag{42.7}$$

The operations of adding two vectors in $\mathbb{R}^n$ and multiplying a vector in $\mathbb{R}^n$ with a real number, of course directly generalize the corresponding operations from the cases $n = 2$ and $n = 3$ considered above. The generalization helps us to deal with $\mathbb{R}^n$ using concepts and tools which we have found useful in $\mathbb{R}^2$ and $\mathbb{R}^3$.

We may thus add vectors in $\mathbb{R}^n$ and multiply them by real numbers (scalars), the usual commutative and distributive rules hold for these operations, and $\mathbb{R}^n$ is thus a vector space. We say that $(0, 0, \ldots, 0)$ is the *zero vector* in $\mathbb{R}^n$ and write $0 = (0, 0, \ldots, 0)$.

Linear algebra concerns vectors in vector spaces, also refereed to as *linear spaces*, and linear functions of vectors, that is *linear transformations* of vectors. As we just saw, $\mathbb{R}^n$ is a vector space, but there are also many other types of vector spaces, where the vectors have a different nature. In particular, we will below meet vector spaces consisting of vectors which are functions. In this chapter we focus on $\mathbb{R}^n$, the most basic of all vector spaces. We know that linear transformations in $\mathbb{R}^2$ and $\mathbb{R}^3$ lead to 2×2 and 3×3 matrices, and we shall now generalize to linear transformations from $\mathbb{R}^n$ into $\mathbb{R}^m$ which can be represented by $m \times n$ *matrices*.

We give in this chapter a condensed (and dry) presentation of some basic facts of linear algebra in $\mathbb{R}^n$. Many applications of the theoretical results presented will appear in the rest of the book.

42.4 The Scalar Product and Orthogonality

We define the *scalar product* $x \cdot y = (x, y)$ of two vectors x and y in $\mathbb{R}^n$, by

$$x \cdot y = (x, y) = \sum_{i=1}^{n} x_i y_i. \tag{42.8}$$

This generalizes the scalar product in $\mathbb{R}^2$ and $\mathbb{R}^3$. Note that here we introduce a new notation for the scalar product of two vectors x and y, namely (x, y), as an alternative to the "dot product" $x \cdot y$ used in $\mathbb{R}^2$ and $\mathbb{R}^3$. We should be ready to use both notations.

The scalar product is *bilinear* in the sense that $(x+y, z) = (x, z)+(y, z)$, $(\lambda x, z) = \lambda(x, z)$, $(x, y + z) = (x, y) + (x, z)$ and $(x, \lambda y) = \lambda(x, y)$, and *symmetric* in the sense that $(x, y) = (y, x)$, for all vectors $x, y, z \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$.

We say that two vectors x and y in $\mathbb{R}^n$ are *orthogonal* if $(x, y) = 0$. We define

$$|x| = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2} = (x, x)^{1/2} \tag{42.9}$$

to be the Euclidean *length* or *norm* of the vector x. Note that this definition of *length* is a direct generalization of the natural length $|x|$ of a vector x in $\mathbb{R}^n$, $n = 1, 2, 3$.

Example 42.1. Let $x = (2, -4, 5, 1, 3)$ and $y = (1, 4, 6, -1, 2)$ be two vectors in $\mathbb{R}^5$. We compute $(x, y) = 2 \times 1 + (-4) \times 4 + 5 \times 6 + 1 \times (-1) + 3 \times 2 = 21$.

42.5 Cauchy's Inequality

Cauchy's inequality states that for $x, y \in \mathbb{R}^n$,

$$|(x, y)| \leq |x|\,|y|.$$

In words: the absolute value of the scalar product of two vectors is bounded by the product of the norms of the vectors. We prove Cauchy's inequality by noting that for all $s \in \mathbb{R}$,

$$0 \leq |x + sy|^2 = (x + sy, x + sy) = |x|^2 + 2s(x, y) + s^2|y|^2,$$

and then assuming that $y \neq 0$, choosing $s = -(x, y)/|y|^2$ (which minimizes the right-hand side), to get

$$0 \leq |x|^2 - 2\frac{(x, y)^2}{|y|^2} + \frac{(x, y)^2}{|y|^2} = |x|^2 - \frac{(x, y)^2}{|y|^2},$$

which proves the desired result.

We recall that for $n = 2, 3$,

$$(x, y) = x \cdot y = \cos(\theta)|x||y|,$$

where θ is the angle between x and y, from which of course Cauchy's inequality follows directly using the fact that $|\cos(\theta)| \leq 1$.

We define the *angle* $\theta \in [0, 2\pi)$ between two non-zero vectors x and y in $\mathbb{R}^n$ by

$$\cos(\theta) = \frac{(x, y)}{|x||y|}, \tag{42.10}$$

which generalizes the corresponding notion for $n = 2, 3$.

Example 42.2. The angle between the vectors $x = (1, 2, 3, 4)$ and $y = (4, 3, 2, 1)$ in $\mathbb{R}^4$ is equal to $\arccos\frac{2}{3} \approx 0.8411 \approx 48^\circ$ since $(x, y) = 20$ and $|x| = |y| = \sqrt{30}$.

42.6 The Linear Combinations of a Set of Vectors

We know that two non-parallel vectors a_1 and a_2 in $\mathbb{R}^3$ define a plane in $\mathbb{R}^3$ through the origin consisting of all the linear combinations $\lambda_1 a_1 + \lambda_2 a_2$ with coefficients λ_1 and λ_2 in $\mathbb{R}$. The normal to the plane is given by $a_1 \times a_2$. A plane through the origin is an example of *subspace* of $\mathbb{R}^3$, which is a subset of $\mathbb{R}$ with the property that vector addition and scalar multiplication does not lead outside the set. So, a subset S of $\mathbb{R}^3$ is a subspace if the sum of any two vectors in S belongs to S and scalar multiplication of a vector in S gives a vector in S. Clearly, a plane through the origin is a subspace of $\mathbb{R}^3$. Similarly, a line through the origin defined as the scalar multiples $\lambda_1 a_1$ with coefficients $\lambda_1 \in \mathbb{R}$ and a_1 a given vector in $\mathbb{R}^3$, is a subspace of $\mathbb{R}^3$. The subspaces of $\mathbb{R}^3$ consist of lines and planes through the origin. Notice that a plane or line in $\mathbb{R}^3$ not passing through the origin, is not a subspace.

More generally, we use the concept of a *vector space* to denote a set of vectors for which the operations of vector addition and scalar multiplication does not lead outside the set. Of course, $\mathbb{R}^3$ is a vector space. A subspace of $\mathbb{R}^3$ is a vector space. A plane or line in $\mathbb{R}^3$ through the origin is a vector space. The concept of vector space is fundamental in mathematics and we will meet this term many times below.

We will now generalize to $\mathbb{R}^m$ with $m > 3$ and we will then meet new examples of vector spaces and subspaces of vector spaces. Let a_1, a_2,...,a_n, be n non-zero vectors in $\mathbb{R}^m$. A vector b in $\mathbb{R}^m$ of the form

$$b = \lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_n a_n, \tag{42.11}$$

where the $\lambda_i \in \mathbb{R}$, is said to be a *linear combination* of the set of vectors $\{a_1, \ldots, a_n\}$ with *coefficients* $\lambda_1, \ldots, \lambda_n$. If

$$c = \mu_1 a_1 + \mu_2 a_2 + \cdots + \mu_n a_n, \tag{42.12}$$

is another linear combination of $\{a_1, \ldots, a_n\}$ with coefficients $\mu_j \in \mathbb{R}$, then the vector

$$b + c = (\lambda_1 + \mu_1)a_1 + (\lambda_2 + \mu_2)a_2 + \cdots + (\lambda_+ \mu_n)a_n, \tag{42.13}$$

is again a linear combination of $\{a_1, \ldots, a_m\}$ now with coefficients $\lambda_j + \mu_j$. Further, for any $\alpha \in \mathbb{R}$ the vector

$$\alpha b = \alpha\lambda_1 a_1 + \alpha\lambda_2 a_2 + \cdots + \alpha\lambda_m a_m \tag{42.14}$$

is also a linear combination of $\{a_1, \ldots, a_m\}$ with coefficients $\alpha\lambda_j$. This means that if we let $S(a_1, \ldots a_n)$ denote the set of all linear combinations

$$\lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_n a_n, \tag{42.15}$$

of $\{a_1, \ldots, a_n\}$, where the coefficients $\lambda_j \in \mathbb{R}$, then $S(a_1, \ldots a_n)$ is indeed a vector space, since vector addition and multiplication by scalars do not

lead outside the set. The sum of two linear combinations of $\{a_1, \ldots, a_n\}$ is also a linear combination of $\{a_1, \ldots, a_n\}$, and a linear combination of $\{a_1, \ldots, a_n\}$ multiplied by a real number is also a linear combination of $\{a_1, \ldots, a_n\}$.

We refer to the vector space $S(a_1, \ldots a_n)$ of all linear combinations of the form (42.15) of the vectors $\{a_1, \ldots, a_n\}$ in $\mathbb{R}^m$ as the *subspace of* $\mathbb{R}^m$ *spanned by the vectors* $\{a_1, \ldots, a_n\}$, or simply just the *span of* $\{a_1, \ldots, a_n\}$, which we may describe as:

$$S(a_1, \ldots a_n) = \left\{ \sum_{i=1}^{n} \lambda_j a_j : \ \lambda_j \in \mathbb{R}, j = 1, \ldots, n \right\}.$$

If $m = 2$ and $n = 1$, then the subspace $S(a_1)$ is a line in $\mathbb{R}^2$ through the origin with direction a_1. If $m = 3$ and $n = 2$, then $S(a_1, a_2)$ corresponds to the plane in $\mathbb{R}^3$ through the origin spanned by a_1 and a_2 (assuming a_1 and a_2 are non-parallel), that is, the plane through the origin with normal given by $a_1 \times a_2$.

Note that for any $\mu \in \mathbb{R}$, we have

$$S(a_1, a_2, \ldots, a_n) = S(a_1, a_2 - \mu a_1, a_3, \ldots, a_n), \tag{42.16}$$

since we can replace each occurrence of a_2 by the linear combination $(a_2 - \mu a_1) + \mu a_1$ of $a_2 - \mu a_1$ and a_1. More generally, we can add any multiple of one vector to one of the other vectors without changing the span of the vectors! Of course we may also replace any vector a_j with a μa_j where μ is a non-zero real number without changing the span. We shall return to these operations below.

42.7 The Standard Basis

The set of vectors in $\mathbb{R}^n$:

$$\{(1, 0, 0, \cdots, 0, 0), (0, 1, 0, \cdots, 0, 0), \cdots, (0, 0, 0, \cdots, 0, 1)\},$$

commonly denoted by $\{e_1, \ldots, e_n\}$, where $e_i = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$ with a single coefficient 1 at position i, is called the *standard basis* for $\mathbb{R}^n$. Any vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ can be written as a linear combination of the basis vectors $\{e_1, \ldots, e_n\}$:

$$x = x_1 e_1 + x_2 e_2 + \cdots + x_n e_n, \tag{42.17}$$

with the coefficients x_j of x appearing as coefficients of the basis vectors e_j.

We note that $(e_j, e_k) = e_j \cdot e_k = 0$ for $j \neq k$, that is the standard basis vectors are *pairwise orthogonal*, and of length one since $(e_j, e_j) = |e_j|^2 = 1$. We may thus express the coefficients x_i of a given vector $x = (x_1, \ldots, x_n)$ with respect to the standard basis $\{e_1, \ldots, e_n\}$ as follows:

$$x_i = (e_i, x) = e_i \cdot x. \tag{42.18}$$

42.8 Linear Independence

We recall that to specify a plane in $\mathbb{R}^3$ as the set of linear combinations of two given vectors a_1 and a_2, we assume that a_1 and a_2 are non-parallel. The generalization of this condition to a set $\{a_1, \ldots, a_m\}$ of m vectors in $\mathbb{R}^n$, is referred to as *linear independence*, which we now proceed to define. Eventually, this will lead us to the concept of *basis* of a vector space, which is one of the most basic(!) concepts of linear algebra.

A set $\{a_1, \ldots, a_n\}$ of vectors in $\mathbb{R}^m$ is said to be *linearly independent* if none of the vectors a_i can be expressed as a linear combination of the others. Conversely, if some of the vectors a_i can be expressed as a linear combination of the others, for example if

$$a_1 = \lambda_2 a_2 + \ldots + \lambda_n a_n \tag{42.19}$$

for some numbers $\lambda_2, \ldots, \lambda_n$, we say that the set $\{a_1, a_2, \ldots, a_n\}$ is *linearly dependent*. As a test of linear independence of $\{a_1, a_2, \ldots, a_n\}$, we can use: if

$$\lambda_1 a_1 + \lambda_2 a_2 + \ldots + \lambda_n a_n = 0 \tag{42.20}$$

implies that $\lambda_1 = \lambda_2 = \ldots = \lambda_n = 0$, then $\{a_1, a_2, \ldots, a_n\}$ is linearly independent. This is because if (42.20) holds with some of the λ_j different from 0, for example, $\lambda_1 \neq 0$, then we could divide by λ_1 and express a_1 as a linear combination of $\{a_2, \ldots, a_m\}$:

$$a_1 = -\frac{\lambda_2}{\lambda_1} a_2 + \ldots + -\frac{\lambda_n}{\lambda_1} a_n. \tag{42.21}$$

The standard basis $\{e_1, \ldots, e_n\}$ is (of course) a linearly independent set, since if

$$\lambda_1 e_1 + \ldots + \lambda_n e_n = 0,$$

then $\lambda_i = 0$ for $i = 1, \ldots, n$, because $0 = (0, 0, \ldots, 0) = \lambda_1 e_1 + \ldots + \lambda_n e_n = (\lambda_1, \ldots, \lambda_n)$.

42.9 Reducing a Set of Vectors to Get a Basis

Consider the subspace $S(a_1, \ldots, a_n)$ spanned by the set of vectors $\{a_1, a_2, \ldots, a_n\}$. If the set $\{a_1, a_2, \ldots, a_n\}$ is linearly dependent, say that a_n can be expressed as a linear combination of $\{a_1, \ldots, a_{n-1}\}$, then $S(a_1, \ldots, a_n)$ is in fact spanned by $\{a_1, \ldots, a_{n-1}\}$ and thus $S(a_1, \ldots, a_n) = S(a_1, \ldots, a_{n-1})$. This follows simply by replacing all occurrences of a_n by its linear combination of $\{a_1, \ldots, a_{n-1}\}$. Continuing this way, eliminating linearly dependent vectors, we may express $S(a_1, \ldots, a_n)$ as the span of $\{a_1, a_2, \ldots, a_k\}$ (with a suitable enumeration), that is, $S(a_1, \ldots, a_n) = S(a_1, a_2, \ldots, a_k\}$, where $k \leq n$, and the set $\{a_1, a_2, \ldots, a_k\}$ is linearly independent. This means that $\{a_1, a_2, \ldots, a_k\}$ is a *basis* for the vector space $S = S(a_1, \ldots, a_n)$ in the sense that the following two conditions are fulfilled:

- any vector in S can be expressed as a linear combination of $\{a_1, a_2, \ldots, a_k\}$,
- the set $\{a_1, a_2, \ldots, a_k\}$ is linearly independent.

Note that by the linear independence the coefficients in the linear combination are uniquely determined: if two linear combinations $\sum_{j=1}^k \lambda_j a_j$ and $\sum_{j=1}^k \mu_j a_j$ are equal, then $\lambda_j = \mu_j$ for $j = 1, \ldots, k$.

Each vector $b \in S$ can thus be expresses as a unique linear combination of the basis vectors $\{a_1, a_2, \ldots, a_k\}$:

$$b = \sum_{j=1}^{k} \lambda_j a_j,$$

and we refer to $(\lambda_1, \ldots, \lambda_k)$ as the *coefficients* of b with respect to the basis $\{a_1, a_2, \ldots, a_k\}$.

The *dimension* of a vector space S is equal to the number of basis vectors in a basis for S. We prove below that the dimension is uniquely defined so that two sets of basis vectors always have the same number of elements.

Example 42.3. Consider the three vectors $a_1 = (1, 2, 3, 4)$, $a_2 = (1, 1, 1, 1)$, and $a_3 = (3, 3, 5, 6)$ in $\mathbb{R}^4$. We see that $a_3 = a_1 + 2a_2$, and thus the set $\{a_1, a_2, a_3\}$ is linearly dependent. The span of $\{a_1, a_2, a_3\}$ thus equals the span of $\{a_1, a_2\}$, since each occurrence of a_3 can be replaced by a_1+2a_2. The vector a_3 is thus redundant, since it can be replaced by a linear combination of a_1 and a_2. Evidently, $\{a_1, a_2\}$ is linearly independent, since a_1 and a_2 are non-parallel. Thus, $\{a_1, a_2\}$ is a linearly independent set spanning the same subset as $\{a_1, a_2, a_3\}$. We can also express a_2 in terms of a_1 and a_3, or a_1 in terms of a_2 and a_3, and thus any set of two vectors $\{a_1, a_2\}$, $\{a_1, a_3\}$ or $\{a_2, a_3\}$, can serve as a basis for the subspace spanned by $\{a_1, a_2, a_3\}$.

42.10 Using Column Echelon Form to Obtain a Basis

We now present a constructive process for determining a basis for the vector space $S(a_1,\ldots,a_n)$ spanned by the set of vectors $\{a_1,a_2,\ldots,a_n\}$, where $a_j=(a_{1j},\ldots,a_{mj})\in\mathbb{R}^m$ for $j=1,\ldots,n$ which we view as column vectors. We refer to this process as *reduction to column echelon form*. It is of fundamental importance and we shall return to it below in several different contexts. Assume then first that $a_{11}=1$ and choose $\mu\in\mathbb{R}$ so that $\mu a_{11}=a_{12}$, and note that $S(a_1,\ldots,a_n)=S(a_1,a_2-\mu a_1,a_3,\ldots,a_n)$, where now the first component of $a_2-\mu a_1$ is zero. We here used the fact that we can add one vector multiplied by a scalar to another vector without changing the span of the vectors. Continuing in the same way we obtain $S(a_1,\ldots,a_n)=S(a_1,\hat{a}_2,\hat{a}_3,\ldots\hat{a}_n)$ where $\hat{a}_{1j}=0$ for $j>1$. In matrix form with the $a_j\in\mathbb{R}^m$ as column vectors, we may express this as follows:

$$S\begin{pmatrix} a_{11} & a_{12} & .. & a_{1n}\\ a_{21} & a_{22} & .. & a_{2n}\\ . & . & .. & .\\ a_{m1} & a_{m2} & .. & a_{mn}\end{pmatrix}=S\begin{pmatrix} 1 & 0 & .. & 0\\ a_{21} & \hat{a}_{22} & .. & \hat{a}_{22}\\ . & . & .. & .\\ a_{m1} & \hat{a}_{m2} & .. & \hat{a}_{mn}\end{pmatrix}$$

We can now repeat the process by cutting out the first row and first column and reduce to a set of $n-1$ vectors in $\mathbb{R}^{m-1}$. Before doing this we take care of the case $a_{11}\neq 1$. If $a_{11}\neq 0$, then we transform to the case $a_{11}=1$ by replacing a_1 by μa_1 with $\mu=1/a_{11}$, noting that we can multiply any column with a non-zero real number without changing the span. By renumbering the vectors we may then assume that either $a_{11}\neq 0$, which thus led to the above construction, or $a_{1j}=0$ for $j=1,\ldots,n$, in which case we seek to compute a basis for

$$S\begin{pmatrix} 0 & 0 & .. & 0\\ a_{21} & a_{22} & .. & a_{22}\\ . & . & .. & .\\ a_{m1} & a_{m2} & .. & a_{mn}\end{pmatrix}$$

with only zeros in the first row. We may then effectively cut out the first row and reduce to a set of n vectors in $\mathbb{R}^{m-1}$.

Repeating now the indicated process, we obtain with $k\leq\min(n,m)$,

$$S\begin{pmatrix} a_{11} & a_{12} & .. & a_{1n}\\ . & . & .. & .\\ . & . & .. & .\\ a_{m1} & a_{m2} & .. & a_{mn}\end{pmatrix}=S\begin{pmatrix} 1 & 0 & 0 & .. & 0 & .. & 0\\ \hat{a}_{21} & 1 & 0 & .. & 0 & .. & 0\\ . & . & . & 0 & 0 & .. & 0\\ . & . & .. & 1 & 0 & .. & 0\\ . & . & .. & . & . & .. & .\\ \hat{a}_{m1} & \hat{a}_{m2} & .. & \hat{a}_{mk} & 0 & .. & 0\end{pmatrix}$$

where we refer to the matrix on the right as the *column echelon form* of the matrix to the left, and k is the number of non-zero columns. We

see that each non-zero column $\hat{a}_j$, $j = 1, \ldots, k$, in the echelon form has a coefficient equal to 1 and that all matrix elements to the right and above that coefficient is equal to zero. Further, the ones appear in a staircase form descending to the right on or below the diagonal. The set of non-zero columns $\{\hat{a}_1, \ldots, \hat{a}_k\}$ is linearly independent, because if

$$\sum_{j=1}^{k} \hat{x}_j \hat{a}_j = 0,$$

then we get successively $\hat{x}_1 = 0$, $\hat{x}_2 = 0$,..., $\hat{x}_k = 0$, and thus $\{\hat{a}_1, \ldots, \hat{a}_k\}$ forms a basis for $S(a_1, \ldots, a_n)$. The dimension of $S(a_1, \ldots, a_n)$ is equal to k. If zero columns appear in the echelon form, then the original set $\{a_1, \ldots, a_n\}$ is linearly dependent.

We note that, because of the construction, zero columns must appear if $n > m$, and we thus understand that a set of n vectors in $\mathbb{R}^m$ is linearly dependent if $n > m$. We may also understand that if $n < m$, then the set $\{a_1, \ldots, a_n\}$ cannot span $\mathbb{R}^m$, because if $k < m$, then there are vectors $b \in \mathbb{R}^m$ which cannot be expressed as linear combinations of $\{\hat{a}_1, \ldots, \hat{a}_k\}$ as we now show: if

$$b = \sum_{j=1}^{k} \hat{x}_j \hat{a}_j,$$

then the coefficients $\hat{x}_1, \ldots, \hat{x}_k$ are determined by the coefficients $b_1, \ldots, b_k$, of b occurring in the rows with the coefficient 1. For example, in the case the 1s appear on the diagonal, we first compute $\hat{x}_1 = b_1$, then $\hat{x}_2 = b_1 - \hat{a}_{21}\hat{x}_1$ etc, and thus the remaining coefficients $b_{k+1}, \ldots, b_m$ of b cannot be arbitrary.

42.11 Using Column Echelon Form to Obtain $R(A)$

By reduction to column echelon form we can construct a basis for $R(A)$ for a given $m \times n$ matrix A with column vectors $a_1, \ldots, a_n$ because

$$Ax = \sum_{j=1}^{n} x_j a_j$$

and thus $R(A) = S(a_1, \ldots, a_n)$ expressing that the range $R(A) = \{Ax : x \in \mathbb{R}^n\}$ is equal to the vector space $S(a_1, \ldots, a_n)$ of all linear combinations of the set of column vectors $\{a_1, \ldots, a_n\}$. Setting now

$$A = \begin{pmatrix} a_{11} & a_{12} & .. & a_{1n} \\ . & . & .. & . \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix}, \qquad \hat{A} = \begin{pmatrix} 1 & 0 & 0 & .. & 0 & .. & 0 \\ \hat{a}_{21} & 1 & 0 & .. & 0 & .. & 0 \\ . & . & . & 0 & 0 & .. & 0 \\ . & . & .. & 1 & 0 & .. & 0 \\ . & . & .. & . & . & .. & . \\ \hat{a}_{m1} & \hat{a}_{m2} & .. & \hat{a}_{mk} & 0 & .. & 0 \end{pmatrix}$$

with $\hat{A}$ obtained from A by reduction to column echelon form, we have

$$R(A) = R(\hat{A}) = S(\hat{a}_1, \ldots, \hat{a}_k),$$

and thus $\{\hat{a}_1, \ldots, \hat{a}_k\}$ forms a basis for $R(A)$. In particular we can easily check if a given vector $b \in \mathbb{R}^m$ belongs to $R(A)$, by using the echelon form. By reduction to column echelon form we can thus give an answer to the basic problem of determining $R(A)$ for a given matrix A. Not bad. For example, in the case $m = n$ we have that $R(A) = \mathbb{R}^m$ if and only if $k = n = m$, in which case the echelon form $\hat{A}$ has 1s all along the diagonal.

We give an example showing the sequence of matrices appearing in reduction to column echelon form:

Example 42.4. We have

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 7 \\ 1 & 3 & 4 & 5 & 8 \\ 1 & 4 & 5 & 6 & 9 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 2 & 3 & 6 \\ 1 & 2 & 3 & 4 & 7 \\ 1 & 3 & 4 & 5 & 8 \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & -1 & -2 & -5 \\ 1 & 3 & -2 & -4 & -10 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & -2 & -5 \\ 1 & 3 & 2 & -4 & -10 \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 1 & 3 & 2 & 0 & 0 \end{pmatrix} = \hat{A}.$$

We conclude that $R(A)$ is spanned by the 3 non-zero columns of $\hat{A}$ and thus in particular that the dimension of $R(A)$ is equal to 3. In this example, A is a 4×5 matrix and $R(A)$ does not span $\mathbb{R}^4$. Solving the system

$$\hat{A}\hat{x} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 1 & 3 & 2 & 0 & 0 \end{pmatrix} \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \\ \hat{x}_4 \\ \hat{x}_5 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix}$$

we compute uniquely $\hat{x}_1$, $\hat{x}_2$ and $\hat{x}_3$ from the first three equations, and to have the fourth equation satisfied, we must have $b_4 = \hat{x}_1 + 3\hat{x}_2 + 2\hat{x}_3$ and thus b_4 can not be chosen freely.

42.12 Using Row Echelon Form to Obtain $N(A)$

We take the chance to solve the other basic problem of determining $N(A)$ by *reduction to row echelon form*, which is analogous to reduction to column echelon form working now with the rows instead of the columns. We thus consider a $m \times n$ matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & .. & a_{1n} \\ . & . & .. & . \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix},$$

and perform the operations of (i) multiplying one row with a real number and (ii) multiplying one row with a real number and subtracting it from another row. We then obtain the *row echelon form* of A (possibly by reordering rows):

$$\hat{A} = \begin{pmatrix} 1 & \hat{a}_{12} & . & . & .. & . & \hat{a}_{1n} \\ 0 & 1 & . & .. & . & .. & \hat{a}_{2n} \\ . & . & . & . & . & .. & . \\ 0 & 0 & .. & 1 & . & .. & \hat{a}_{kn} \\ 0 & 0 & .. & 0 & 0 & .. & 0 \\ . & . & .. & . & . & .. & . \\ 0 & 0 & .. & 0 & 0 & .. & 0 \end{pmatrix}$$

Each non-zero row of the row echelon matrix $\hat{A}$ has one element equal to 1 and all elements to the left and below are equal to zero, and the 1s appear in a staircase form on or to the right of the diagonal from the upper left corner.

We notice that the row operations do not change the null space $N(A) = \{x : Ax = 0\}$, because we may perform the row operations in the system of equations $Ax = 0$, that is[TS a]

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= 0, \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= 0, \\ &\cdots\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= 0, \end{aligned}$$

to reduce it to the echelon form system $\hat{A}x = 0$ without changing the vector $x = (x_1, \ldots, x_n)$. We conclude that

$$N(A) = N(\hat{A})$$

and we may thus determine $N(A)$ by using that we can directly determine $N(\hat{A})$ from the echelon form of A. It is easy to see that the dimension of $N(A) = N(\hat{A})$ is equal to $n - k$, as illustrated in the following example. In

[TS a] The following equation was numbered with (42.22) in the hard copy, please check it.

Editor's or typesetter's annotations (will be removed before the final TEX run)

the case $n = m$, we have that $N(A) = 0$ if and only if $k = m = n$ in which all diagonal elements of $\hat{A}$ are equal to 1.

We give an example showing the sequence of matrices appearing in reduction to row echelon form:

Example 42.5. We have

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 7 \\ 1 & 3 & 4 & 5 & 8 \\ 1 & 4 & 5 & 6 & 9 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 6 \\ 0 & 2 & 3 & 4 & 7 \\ 0 & 3 & 4 & 5 & 8 \end{pmatrix} \rightarrow$$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 6 \\ 0 & 0 & -1 & -2 & -5 \\ 0 & 0 & -2 & -4 & -10 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 6 \\ 0 & 0 & 1 & 2 & 5 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} = \hat{A}.$$

We now determine $N(A)$ by determining $N(\hat{A}) = N(A)$ by seeking the solutions $x = (x_1, \ldots, x_5)$ of the system $\hat{A}x = 0$, that is

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 6 \\ 0 & 0 & 1 & 2 & 5 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

We see that we can freely choose x_4 and x_5 and then solve for x_3, x_2 and x_1 to get the solution in the form

$$x = \lambda_1 \begin{pmatrix} 0 \\ 1 \\ -2 \\ 1 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} 0 \\ 4 \\ -5 \\ 0 \\ 1 \end{pmatrix}$$

where λ_1 and λ_2 are any real numbers. We have now computed a basis for $N(A)$ and we see in particular that the dimension of $N(A)$ is equal to 2. We recall that the dimension of $R(A)$ is equal to 3 and we note that the sum of the dimensions of $R(A)$ and $N(A)$ happens to be equal to 5 that is the number of columns of A. This is a general fact which we prove in the Fundamental Theorem below.

42.13 Gaussian Elimination

Gaussian elimination to compute solutions to the system

$$\begin{pmatrix} a_{11} & a_{12} & .. & a_{1n} \\ . & . & .. & . \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ . \\ . \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ . \\ . \\ b_m \end{pmatrix}$$

closely couples to reduction to row echelon form. Performing row operations we may reduce to a system of the form

$$\hat{A} = \begin{pmatrix} 1 & \hat{a}_{12} & . & . & .. & . & \hat{a}_{1n} \\ 0 & 1 & . & .. & . & .. & \hat{a}_{2n} \\ . & . & . & . & . & .. & . \\ 0 & 0 & .. & 1 & . & .. & \hat{a}_{kn} \\ 0 & 0 & .. & 0 & 0 & .. & 0 \\ . & . & .. & . & . & .. & . \\ 0 & 0 & .. & 0 & 0 & .. & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ . \\ . \\ x_n \end{pmatrix} = \begin{pmatrix} \hat{b}_1 \\ . \\ . \\ \hat{b}_m \end{pmatrix}$$

with the same solution vector x. We may assume, by possibly by renumbering the components of x, that the 1s appear on the diagonal. We see that solvability is equivalent to having $\hat{b}_j = 0$ for $j = k+1, \ldots, m$, and the non-uniqueness is expressed by $N(A)$ as explained above. In the case $m = n$ we have that $N(A) = 0$ if and only if $k = m = n$ in which case all diagonal elements of $\hat{A}$ are equal to 1, and the system $\hat{A}x = \hat{b}$ is uniquely solvable for all $\hat{b} \in \mathbb{R}^m$, and thus $Ax = b$ is uniquely solvable for all $b \in \mathbb{R}^m$. We conclude that if $m = n$, then uniqueness implies existence. We may thus say that by Gaussian elimination or reduction to row echelon form, we may solve our basic problems of existence and uniqueness of solutions to the system $Ax = b$. We shall add more information on these problems in the Fundamental Theorem of Linear Algebra below. For more information on Gaussian elimination, we refer to Chapter *Solving Linear Algebraic Systems* below.

42.14 A Basis for $\mathbb{R}^n$ Contains n Vectors

Let us now prove that if $\{a_1, \ldots, a_m\}$ is a basis for $\mathbb{R}^n$, then $m = n$, that is any basis for $\mathbb{R}^n$ has exactly n elements, no more no less. We already deduced this fact from the column echelon form above, but we here give a "coordinate-free" proof which applies to more general situations.

We recall that a set $\{a_1, \ldots, a_m\}$ of vectors in $\mathbb{R}^n$ is a *basis* for $\mathbb{R}^n$ if the following two conditions are fulfilled:

- $\{a_1, \ldots, a_m\}$ is linearly independent,
- any vector $x \in \mathbb{R}^n$ can be expressed as a linear combination $x = \sum_{j=1}^m \lambda_j a_j$ of $\{a_1, \ldots, a_m\}$ with coefficients λ_j.

Of course, $\{e_1, \ldots, e_n\}$ is a basis for $\mathbb{R}^n$ in this sense.

To prove that $m = n$, we consider the set $\{e_1, a_1, a_2, \ldots, a_m\}$. Since $\{a_1, \ldots, a_m\}$ is a basis for $\mathbb{R}^n$, that is spans $\mathbb{R}^n$, the vector e_1 can be expressed as a linear combination of $\{a_1, \ldots, a_m\}$:

$$e_1 = \sum_{j=1}^m \lambda_j a_j,$$

with some $\lambda_j \neq 0$. Suppose $\lambda_1 \neq 0$. Then, dividing by λ_1 expresses a_1 as a linear combination of $\{e_1, a_2, \ldots, a_m\}$. This means that $\{e_1, a_2, \ldots, a_m\}$ spans $\mathbb{R}^n$. Consider now the set $\{e_1, e_2, a_2, \ldots, a_m\}$. The vector e_2 can be expressed as a linear combination of $\{e_1, a_2, \ldots, a_m\}$ and some of the coefficients of the a_j must be non-zero, since $\{e_1, e_2\}$ are linearly independent. Supposing the coefficient of a_2 is non-zero, we can eliminate a_2 and thus the set $\{e_1, e_2, a_3, \ldots, a_m\}$ now spans $\mathbb{R}^n$. Continuing this way we get the set $\{e_1, e_2, \ldots, e_n, a_{n+1}, \ldots, a_m\}$ if $m > n$ and the set $\{e_1, e_2, \ldots, e_n\}$ if $m = n$, which both span $\mathbb{R}^n$. We conclude that $m \geq n$, since if e.g. $m = n - 1$, we would end up with the set $\{e_1, e_2, \ldots, e_{n-1}\}$ which does not span $\mathbb{R}^n$ contrary to the assumption.

Repeating this argument with the roles of the basis $\{e_1, e_2, \ldots, e_n\}$ and $\{a_1, a_2, \ldots, a_m\}$ interchanged, we get the reverse inequality $n \geq m$ and thus $n = m$. Of course, intuitively, there are n independent directions in $\mathbb{R}^n$ and thus a basis of $\mathbb{R}^n$ has n elements, no more no less.

We also note that if $\{a_1, \ldots, a_m\}$ is a linearly independent set in $\mathbb{R}^n$, then it can be extended to a basis $\{a_1, \ldots, a_m, a_{m+1}, \ldots, a_n\}$ by adding suitable elements $a_{m+1}, \ldots, a_n$. The extension starts by adding a_{m+1} as any vector which cannot be expressed as a linear combination of the set $\{a_1, \ldots, a_m\}$. Then $\{a_1, \ldots, a_m, a_{m+1}\}$ is linearly independent, and if $m + 1 < n$, the process may be continued.

We summarize as follows:

Theorem 42.1 *Any basis of $\mathbb{R}^n$ has n elements. Further, a set of n vectors in $\mathbb{R}^n$ span $\mathbb{R}^n$ if and only if it is linearly independent, that is a set of n vectors in $\mathbb{R}^n$ that spans $\mathbb{R}^n$ or is independent, must be a basis. Also, a set of fewer than n vectors in $\mathbb{R}^n$ cannot span $\mathbb{R}^n$, and a set of more than n vectors in $\mathbb{R}^n$ must be linearly dependent.*

The argument used to prove this result can also be used to prove that the dimension of a vector space S is well defined in the sense that any two bases have the same number of elements.

42.15 Coordinates in Different Bases

There are many different bases in $\mathbb{R}^n$ if $n > 1$ and the coordinates of a vector with respect to one basis are not equal to the coordinates with respect to another basis.

Suppose $\{a_1, \ldots, a_n\}$ is a basis for $\mathbb{R}^n$ and let us seek the connection between the coordinates of one and the same vector in the standard basis $\{e_1, \ldots, e_n\}$ and the basis $\{a_1, \ldots, a_n\}$. Assume then that the coordinates of the basis vectors a_j in the standard basis $\{e_1, \ldots, e_n\}$ are given by $a_j = (a_{1j}, \ldots a_{nj})$ for $j = 1, \ldots n$, that is

$$a_j = \sum_{i=1}^{n} a_{ij} e_i.$$

Denoting the coordinates of a vector x with respect to $\{e_1, \ldots, e_n\}$ by x_j and the coordinates with respect to $\{a_1, \ldots, a_n\}$ by $\hat{x}_j$, we have

$$x = \sum_{j=1}^{n} \hat{x}_j a_j = \sum_{j=1}^{n} \hat{x}_j \sum_{i=1}^{n} a_{ij} e_i = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} a_{ij} \hat{x}_j \right) e_i, \tag{42.22}$$

that is since also $x = \sum_{i=1}^{n} x_i e_i$ and the coefficients x_i of x are unique,

$$x_i = \sum_{j=1}^{n} a_{ij} \hat{x}_j \quad \text{for } i = 1, \ldots n. \tag{42.23}$$

This relation expresses the connection between the coordinates $\hat{x}_j$ with respect to the basis $\{a_1, \ldots, a_n\}$, and the coordinates x_i with respect to the standard basis $\{e_1, \ldots, e_n\}$, in terms of the coordinates a_{ij} of the basis vectors a_j with respect to $\{e_1, \ldots, e_n\}$. This is a basic connection, which will play a central role in the development to come.

Using the scalar product we can express the coordinates a_{ij} of the basis vector a_j as $a_{ij} = (e_i, a_j)$. To find the connection (42.23) between the coordinates $\hat{x}_j$ with respect to the basis $\{a_1, \ldots, a_n\}$, and the coordinates x_i with respect to the standard basis $\{e_1, \ldots, e_n\}$, we may start from the equality $\sum_{j=1}^{n} x_j e_j = x = \sum_{j=1}^{n} \hat{x}_j a_j$ and take the scalar product of both sides with e_i, to get

$$x_i = \sum_{j=1}^{n} \hat{x}_j (e_i, a_j) = \sum_{j=1}^{n} a_{ij} \hat{x}_j, \tag{42.24}$$

where $a_{ij} = (e_i, a_j)$.

Example 42.6. The set $\{a_1, a_2, a_3\}$ with $a_1 = (1, 0, 0)$, $a_2 = (1, 1, 0)$, $a_3 = (1, 1, 1)$ in the standard basis, forms a basis for $\mathbb{R}^3$ since the set $\{a_1, a_2, a_3\}$ is

linearly independent. This is because if $\lambda_1 a_1 + \lambda_2 a_2 + \lambda_3 a_3 = 0$, then $\lambda_3 = 0$ and thus also $\lambda_2 = 0$ and thus also $\lambda_1 = 0$. If (x_1, x_2, x_3) are the coordinates with respect to the standard basis and $(\hat{x}_1, \hat{x}_2, \hat{x}_3)$ are the coordinates with respect to $\{a_1, a_2, a_3\}$ of a certain vector, then the connection between the coordinates is given by $(x_1, x_2, x_3) = \hat{x}_1 a_1 + \hat{x}_2 a_2 + \hat{x}_3 a_3 = (\hat{x}_1 + \hat{x}_2 + \hat{x}_3, \hat{x}_2 + \hat{x}_3, \hat{x}_3)$. Solving for the $\hat{x}_j$ in terms of the x_i, we get $(\hat{x}_1, \hat{x}_2, \hat{x}_3) = (x_1 - x_2, x_2 - x_3, x_3)$.

42.16 Linear Functions $f : \mathbb{R}^n \to \mathbb{R}$

A *linear function* $f : \mathbb{R}^n \to \mathbb{R}$ satisfies

$$f(x + y) = f(x) + f(y),\ f(\alpha x) = \alpha f(x) \quad \text{for all } x, y \in \mathbb{R}^n,\ \alpha \in \mathbb{R}. \tag{42.25}$$

We say that $f(x)$ is a *scalar linear function* since $f(x) \in \mathbb{R}$. Expressing $x = x_1 e_1 + \ldots + x_n e_n$ in the standard basis $\{e_1, \ldots, e_n\}$, and using the linearity of $f(x)$, we find that

$$f(x) = x_1 f(e_1) + \ldots + x_n f(e_n), \tag{42.26}$$

and thus $f(x)$ has the form

$$f(x) = f(x_1, \ldots, x_n) \quad = \quad a_1 x_1 + a_2 x_2 + \ldots + a_n x_n, \tag{42.27}$$

where the $a_j = f(e_j)$ are real numbers. We can write $f(x)$ as

$$f(x) = (a, x) = a \cdot x, \tag{42.28}$$

where $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$, that is $f(x)$ can be expressed as the scalar product of x with the vector $a \in \mathbb{R}^n$ with components a_j given by $a_j = f(e_j)$.

The set of scalar linear functions is the mother of all other functions. We now generalize to systems of scalar linear functions. Linear algebra is the study of systems of linear functions.

Example 42.7. $f(x) = 2x_1 + 3x_2 - 7x_3$ defines a linear function $f : \mathbb{R}^3 \to \mathbb{R}$ with coefficients $f(e_1) = a_1 = 2$, $f(e_2) = a_2 = 3$ and $f(e_3) = a_3 = -7$.

42.17 Linear Transformations $f : \mathbb{R}^n \to \mathbb{R}^m$

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be *linear* if

$$f(x+y) = f(x)+f(y),\ f(\alpha x) = \alpha f(x) \quad \text{for all } x, y \in \mathbb{R}^n,\ \alpha \in \mathbb{R}. \tag{42.29}$$

We also refer to a linear function $f : \mathbb{R}^n \to \mathbb{R}^m$ as a *linear transformation* of $\mathbb{R}^n$ into $\mathbb{R}^m$.

The image $f(x)$ of $x \in \mathbb{R}^n$ is a vector in $\mathbb{R}^m$ with components which we denote by $f_i(x)$, $i = 1, 2, \ldots, m$, so that $f(x) = (f_1(x), \ldots, f_m(x))$. Each *coordinate function* $f_i(x)$ is a linear scalar function $f_i : \mathbb{R}^n \to \mathbb{R}$ if $f : \mathbb{R}^n \to \mathbb{R}^m$ is linear. We can thus represent a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ as

$$\begin{array}{rcl} f_1(x) & = & a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ f_2(x) & = & a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \\ \ldots\ldots & & \\ f_m(x) & = & a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \end{array} \tag{42.30}$$

with the coefficients $a_{ij} = f_i(e_j) = (e_i, f(e_j)) \in \mathbb{R}$.

We can write (42.30) in condensed form as

$$f_i(x) = \sum_{j=1}^{n} a_{ij}x_j \quad \text{for } i = 1, \ldots, m. \tag{42.31}$$

Example 42.8. $f(x) = (2x_1 + 3x_2 - 7x_3, x_1 + x_3)$ defines a linear function $f : \mathbb{R}^3 \to \mathbb{R}^2$ with coefficients $f_1(e_1) = a_{11} = 2$, $f_1(e_2) = a_{12} = 3$ and $f_1(e_3) = a_{13} = -7$, $f_2(e_1)a_{21} = 1$, $f_2(e_2)a_{22} = 0$ and $f_2(e_3) = a_{23} = 1$.

42.18 Matrices

We now return to the notion of a *matrix* and develop a matric calculus. The connection to linear transformations is very important. We define the $m \times n$ *matrix* $A = (a_{ij})$ as the rectangular array

$$\begin{pmatrix} a_{11} & a_{12} & .. & a_{1n} \\ . & . & .. & . \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix} \tag{42.32}$$

with rows $(a_{i1}, \ldots, a_{in})$, $i = 1, \ldots, m$, and columns $(a_{1j}, \ldots, a_{mj})$, $j = 1, \ldots, n$, where $a_{ij} \in \mathbb{R}$.

We may view each row $(a_{i1}, \ldots, a_{in})$ as a n-row vector or as a $1 \times n$ matrix, and each column $(a_{1j}, \ldots, a_{mj})$ as an m-column vector or a $m \times 1$ matrix. We can thus view the $m \times n$ matrix $A = (a_{ij})$ with elements a_{ij}, as consisting of m row vectors $(a_{i1}, \ldots, a_{in})$, $i = 1, \ldots, m$ or n column vectors $(a_{1j}, \ldots, a_{mj})$, $j = 1, \ldots, n$.

42.19 Matrix Calculus

Let $A = (a_{ij})$ and $B = (b_{ij})$ be two $m \times n$ matrices. We define $C = A + B$ as the $m \times n$ matrix $C = (c_{ij})$ with elements

$$c_{ij} = a_{ij} + b_{ij}, \quad i = 1, \ldots, n,\ j = 1, \ldots, m. \tag{42.33}$$

We may thus add two $m \times n$ matrices by adding the corresponding elements.

Similarly, we define for λ a real number the matrix λA with elements (λa_{ij}), corresponding to multiplying all elements of A by the real number λ.

We shall now define matrix multiplication and we start by defining the product Ax of an $m \times n$ matrix $A = (a_{ij})$ with a $n \times 1$ column vector $x = (x_j)$ as the $m \times 1$ column vector $y = Ax$ with elements $y_i = (Ax)_i$ given by

$$(Ax)_i = \sum_{j=1}^{n} a_{ij} x_j, \tag{42.34}$$

or with matrix notation

$$\begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_m \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \ldots & & & \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix}.$$

The element $y_i = (Ax)_i$ of the matrix-vector product Ax is thus obtained by taking the scalar product of row i of A with the vector x, as expressed by (42.34).

We can now express a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ as a matrix-vector product

$$f(x) = Ax,$$

where $A = (a_{ij})$ is an $m \times n$ matrix with elements $a_{ij} = f_i(e_j) = (e_i, f(e_j))$, where $f(x) = (f_1(x), \ldots, f_m(x))$. This is a restatement of (42.31).

We now proceed to define the product of an $m \times p$ matrix $A = (a_{ij})$ with a $p \times n$ matrix $B = (b_{ij})$. We do this by connecting the matrix product to the composition $f \circ g : \mathbb{R}^n \to \mathbb{R}^m$ given by

$$f \circ g(x) = f(g(x)) = f(Bx) = A(Bx), \tag{42.35}$$

where $f : \mathbb{R}^p \to \mathbb{R}^m$ is the linear transformation given by $f(y) = Ay$, where $A = (a_{ij})$ and $a_{ik} = f_i(e_k)$, and $g : \mathbb{R}^n \to \mathbb{R}^p$ is the linear transformation given by $g(x) = Bx$, where $B = (b_{kj})$ and $b_{kj} = g_k(e_j)$. Here e_k denote the standard basis vectors in $\mathbb{R}^p$, and e_j the corresponding basis vectors in $\mathbb{R}^n$. Clearly, $f \circ g : \mathbb{R}^n \to \mathbb{R}^m$ is linear and thus can be represented by an $m \times n$ matrix. Letting $(f \circ g)_i(x))$ denote the components of $(f \circ g)(x)$, we have

$$(f \circ g)_i(e_j) = f_i(g(e_j)) = f_i\left(\sum_{k=1}^{p} b_{kj} e_k\right) = \sum_{k=1}^{p} b_{kj} f_i(e_k) = \sum_{k=1}^{p} a_{ik} b_{kj},$$

which shows that $f \circ g(x) = Cx$, where $C = (c_{ij})$ is the $m \times n$ matrix with elements c_{ij} given by the formula

$$c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}, \quad i = 1, \ldots, m,\ j = 1, \ldots, n. \tag{42.36}$$

We conclude that $A(Bx) = Cx$, and we are thus led to define the matrix product $AB = C$ by (42.36), where thus A is an $m \times p$ matrix and B is a $p \times n$ matrix, and the product AB is a $m \times n$ matrix. We can then write

$$A(Bx) = ABx$$

as a reflection of $f(g(x)) = f \circ g(x)$.

Formally, to get the $m \times n$ format of the product AB, we cancel the p in the $m \times p$ format of A and the $p \times n$ format of B. We see that the formula (42.36) may be expressed as follows: the element c_{ij} in row i and column j of AB is obtained by taking the scalar product of row i of A with column j of B.

We may write the formula for matrix multiplication as follows:

$$(AB)_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}, \quad \text{for } i = 1, \ldots, n, j = 1, \ldots, m, \tag{42.37}$$

or with matrix notation

$$AB = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1p} \\ a_{21} & a_{22} & \ldots & a_{2p} \\ \ldots & & & \\ a_{m1} & a_{m2} & \ldots & a_{mp} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \ldots & b_{1n} \\ b_{21} & b_{22} & \ldots & b_{2n} \\ \ldots & & & \\ b_{p1} & b_{p2} & \ldots & b_{pn} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{k=1}^{p} a_{1k} b_{k1} & \sum_{k=1}^{p} a_{1k} b_{k2} & \ldots & \sum_{k=1}^{p} a_{1k} b_{kn} \\ \sum_{k=1}^{p} a_{2k} b_{k1} & \sum_{k=1}^{p} a_{2k} b_{k2} & \ldots & \sum_{k=1}^{p} a_{2k} b_{kn} \\ \ldots & & & \\ \sum_{k=1}^{p} a_{mk} b_{k1} & \sum_{k=1}^{p} a_{mk} b_{k2} & \ldots & \sum_{k=1}^{p} a_{mk} b_{kn} \end{pmatrix}.$$

Matrix multiplication is not commutative, that is $AB \neq BA$ in general. In particular, BA is defined only if $n = m$.

As a special case, we have that the product Ax of an $m \times n$ matrix A with an $n \times 1$ matrix x is given by (42.34). We may thus view the matrix-vector product Ax defined by (42.34) as a special case of the matrix product (42.36) with the $n \times 1$ matrix x being a column vector. The vector Ax is obtained taking the scalar product of the rows of A with the column vector x.

We sum up in the following theorem.

Theorem 42.2 *A linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ can be expressed as*

$$f(x) = Ax, \tag{42.38}$$

where $A = (a_{ij})$ is an $m \times n$ matrix with elements $a_{ij} = f_i(e_j) = (e_i, f(e_j))$, where $f(x) = (f_1(x), \ldots, f_m(x))$. If $g : \mathbb{R}^n \to \mathbb{R}^p$ and $f : \mathbb{R}^p \to \mathbb{R}^m$ are two linear transformations with corresponding matrices A and B, then the matrix of $f \circ g : \mathbb{R}^n \to \mathbb{R}^m$ is given by AB.

42.20 The Transpose of a Linear Transformation

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation defined by $f(x) = Ax$, where $A = (a_{ij})$ is an $m \times n$-matrix. We now define another linear transformation $f^\top : \mathbb{R}^m \to \mathbb{R}^n$, which we refer to as the *transpose* of f, by the relation

$$(x, f^\top(y)) = (f(x), y) \quad \text{for all } x \in \mathbb{R}^n,\ y \in \mathbb{R}^m. \tag{42.39}$$

Using that $f(x) = Ax$, we have

$$(f(x), y) = (Ax, y) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}x_j y_i, \tag{42.40}$$

and thus setting $x = e_j$, we see that

$$(f^\top(y))_j = \sum_{i=1}^{m} a_{ij}y_i. \tag{42.41}$$

This shows that $f^\top(y) = A^\top y$, where $A^\top$ is the $n \times m$ matrix with elements $(a_{ji}^\top)$ given by $a_{ji}^\top = a_{ij}$. In other words, the columns of $A^\top$ are the rows of A and vice versa. For example, if

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}, \quad \text{then} \quad A^\top = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}. \tag{42.42}$$

Summing up we have:

Theorem 42.3 *If $A = (a_{ij})$ is a $m \times n$ matrix, then the transpose $A^\top$ is an $n \times m$ matrix with elements $a_{ji}^\top = a_{ij}$, and*

$$(Ax, y) = (x, A^\top y) \quad \textit{for all } x \in \mathbb{R}^n,\ y \in \mathbb{R}^m. \tag{42.43}$$

An $n \times n$ matrix such that $A^\top = A$, that is $a_{ij} = a_{ji}$ for $i, j = 1, \ldots n$, is said to be a *symmetric* matrix.

42.21 Matrix Norms

In many situations we need to estimate the "size" of a $m \times n$ matrix $A = (a_{ij})$. We may use this information to estimate the "length" of $y = Ax$ in terms of the "length" of x. We observe that

$$\sum_{i=1}^m |y_i| \le \sum_{i=1}^m \sum_{j=1}^n |a_{ij}||x_j| = \sum_{j=1}^n \sum_{i=1}^m |a_{ij}||x_j| \le \max_{j=1,\ldots,n} \sum_{i=1}^m |a_{ij}| \sum_{j=1}^n |x_j|,$$

which shows that if we define $\|x\|_1 = \sum |x_j|$ and $\|y\|_1 = \sum |y_i|$, then

$$\|y\|_1 \le \|A\|_1 \|x\|_1$$

if we define

$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^m |a_{ij}|$$

Similarly, we have

$$\max_i |y_i| \le \max_i \sum_{j=1}^n |a_{ij}|x_j| \le \max_i^n \sum_{j=1}^n |a_{ij}| \max_j |x_j|$$

which shows that if we define $\|x\|_\infty = \max_j |x_j|$ and $\|y\|_\infty = \max_i |y_i|$, then

$$\|y\|_\infty \le \|A\|_\infty \|x\|_\infty$$

if we define

$$\|A\|_\infty = \max_{i+1,\ldots,m} \sum_{j=1}^n |a_{ij}|.$$

We may also define the *Euclidean norm* $\|A\|$ by

$$\|A\| = \max_{x \in \mathbb{R}^n} \frac{\|Ax\|}{\|x\|}, \tag{42.44}$$

where we maximize over $x \neq 0$, and $\|\cdot\|$ denotes the Euclidean norm. We thus define $\|A\|$ to be the smallest constant C such that $\|Ax\| \le C\|x\|$ for all $x \in \mathbb{R}^n$. We shall return in Chapter *The Spectral theorem* below to the problem of giving a formula for $\|A\|$ in terms of the coefficients of A in the case A is symmetric (with in particular $m = n$). By definition, we clearly have

$$\|Ax\| \le \|A\| \, \|x\|. \tag{42.45}$$

If $A = (\lambda_i)$ is a diagonal $n \times n$ matrix with diagonal elements $a_{ii} = \lambda_i$, then

$$\|A\| = \max_{i=1,\ldots,n} |\lambda_i|. \tag{42.46}$$

TS[a] The following equation was numbered with (42.48), please check it.

Editor's or typesetter's annotations (will be removed before the final TEX run)

42.22 The Lipschitz Constant of a Linear Transformation

Consider a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ given by a $m \times n$ matrix $A = (a_{ij})$, that is[TS a]

$$f(x) = Ax, \qquad \text{for } x \in \mathbb{R}^n.$$

By linearity we have

$$\|f(x) - f(y)\| = \|Ax - Ay\| = \|A(x-y)\| \leq \|A\|\|x-y\|.$$

We may thus say that the Lipschitz constant of $f : \mathbb{R}^n \to \mathbb{R}^m$ is equal to $\|A\|$. Alternatively, working with the norms $\|\cdot\|_1$ or $\|\cdot\|_\infty$, we may view the Lipschitz constant to be equal to $\|A\|_1$ or $\|A\|_\infty$.

42.23 Volume in $\mathbb{R}^n$: Determinants and Permutations

Let $\{a_1, a_2, \ldots, a_n\}$ be a set of n vectors in $\mathbb{R}^n$. We shall now generalize the concept of *volume* $V(a_1, \ldots, a_n)$ spanned by $\{a_1, a_2, \ldots, a_n\}$, which we have met above in the case $n = 2$ and $n = 3$. In particular, the volume will give us a tool to determine whether the set of vectors $\{a_1, \ldots, a_n\}$ is linearly independent or not. Using the determinant we shall also develop Cramer's solution formula for an $n \times n$ system $Ax = b$, which generalizes the solution formulas for 2×2 and 3×3 which we have already met. The determinant is a quite complicated object, and we try to make the presentation as accessible as possible. When it comes to computing determinants. we shall return to the column echelon form.

Before actually giving a formula for the volume $V(a_1, \ldots, a_n)$ in terms of the coordinates $(a_{1j}, \ldots, a_{nj})$ of the vectors a_j, $j = 1, 2, \ldots, n$, we note that from our experience in $\mathbb{R}^2$ and $\mathbb{R}^3$, we expect $V(a_1, \ldots, a_n)$ to be a *multilinear alternating form*, that is

$$\begin{gathered} V(a_1, \ldots, a_n) \in \mathbb{R}, \\ V(a_1, \ldots, a_n) \quad \text{is linear in each argument } a_j, \\ V(a_1, \ldots, a_n) = -V(\hat{a}_1, \ldots, \hat{a}_n), \end{gathered}$$

where $\hat{a}_1, \ldots, \hat{a}_n$ is a listing of $a_1, \ldots, a_n$ with two of the a_j interchanged. For example $\hat{a}_1 = a_2$, $\hat{a}_2 = a_1$ and $\hat{a}_j = a_j$ for $j = 3, \ldots, n$. We note that if two of the arguments in an alternating form is the same, for example $a_1 = a_2$, then $V(a_1, a_2, a_3, \ldots, a_n) = 0$. This follows at once from the fact that $V(a_1, a_2, a_3, \ldots, a_n) = -V(a_2, a_1, a_3, \ldots, a_n)$. We are familiar with these properties in the case $n = 2, 3$.

We also need a little preliminary work on permutations. A *permutation* of the ordered list $\{1, 2, 3, 4, \ldots, n\}$ is a reordering of the list. For example $\{2, 1, 3, 4, \ldots, n\}$ is a permutation corresponding to interchanging the elements 1 and 2. Another permutation is $\{n, n-1, \ldots, 2, 1\}$ corresponding to reversing the order.

We can also describe a permutation as a one-to-one mapping of the set $\{1, 2, \ldots, n\}$ onto itself. We may denote the mapping by $\pi : \{1, 2, \ldots, n\} \to \{1, 2, , n\}$, that is $\pi(j)$ is one of the numbers $1, 2, \ldots, n$ for each $j = 1, 2, \ldots, n$ and $\pi(i) \neq \pi(j)$ if $i \neq j$. We can then talk about the *product* $\sigma\tau$ of two permutations σ and τ defined as the composition of τ and σ:

$$\sigma\tau(j) = \sigma(\tau(j)), \quad \text{for } j = 1, \ldots, n, \tag{42.47}$$

which is readily seen to be a permutation. Note that the order may be important: in general the permutation $\sigma\tau$ is different from the permutation $\tau\sigma$. In other words, multiplication of permutations is not commutative. However, multiplication is associative:

$$(\pi\sigma)\tau = \pi(\sigma\tau), \tag{42.48}$$

which directly follows from the definition by composition of functions.

A permutation corresponding to interchanging two elements, is called a *transposition*. More precisely, if π is a transposition then there are two elements p and q of the elements $\{1, 2, \ldots, n\}$, such that

$$\begin{aligned} \pi(p) &= q \\ \pi(q) &= p \\ \pi(j) &= j \quad \text{for } j \neq p,\ j \neq q. \end{aligned}$$

The permutation π with $\pi(j) = j$ for $j = 1, \ldots, n$ is called the identity permutation.

We shall use the following basic fact concerning permutations (a proof will be given in the Appendix).

Theorem 42.4 *Every permutation can be written as a product of transpositions. The representation is not unique, but for a given permutation the number of transpositions in such a representation cannot be odd in one case and even in another case; it is odd for all representations or even for all representations.*

We call a permutation *even* if it contains an even number of transposition factors, and *odd* if it contains an odd number of transpositions. The number of even perturbations is equal to the number of odd perturbations, and thus the total number of perturbations, including the identity, is even.

42.24 Definition of the Volume $V(a_1,\ldots,a_n)$

Assuming that $V(a_1,\ldots,a_n)$ is multilinear and alternating and that $V(e_1,e_2,\ldots,e_n)=1$, we get the following relation

$$\begin{aligned} V(a_1,\ldots,a_n) &= V(\sum_j a_{j1}e_j,\sum_j a_{j2}e_j,\ldots,\sum_j a_{jn}e_j) \\ &= \sum_\pi \pm a_{\pi(1)\,1}a_{\pi(2)\,2}\cdots a_{\pi(n)\,n}, \end{aligned} \tag{42.49}$$

where we sum over all permutations π of the set $\{1,\ldots,n\}$, and the sign indicates if the permutation is even (+) or odd (-). Note that we give the identity permutation, which is included among the permutations, the sign +. We recall that $a_j=(a_{1j},\ldots,a_{nj})$ for $j=1,\ldots,n$.

We now turn around in this game, and simply take (42.49) as a definition of the volume $V(a_1,\ldots,a_n)$ spanned by the set of vectors $\{a_1,\ldots,a_n\}$. By this definition it follows that $V(a_1,\ldots,a_n)$ is indeed a multilinear alternating form on $\mathbb{R}^n$. Further, $V(e_1,\ldots,e_n)=1$, since the only non-zero term in the sum (42.49) in this case corresponds to the identity perturbation.

We can transform the definition $V(a_1,\ldots,a_n)$ to matrix language as follows. Let $A=(a_{ij})$ be the $n\times n$ matrix

$$\begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ . & . & \ldots & . \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix} \tag{42.50}$$

formed by the column vectors $a_1,\ldots,a_n$ with coefficients $a_j=(a_{1j},\ldots a_{nj})$. We define the *determinant* $\det A$ of A, by

$$\det A = V(a_1,\ldots,a_n) = \sum_\pi \pm a_{\pi(1)\,1}a_{\pi(2)\,2}\cdots a_{\pi(n)\,n},$$

where we sum over all permutations π of the set $\{1,\ldots,n\}$, and the sign indicates if the permutation is even (+) or odd (-).

We note that since the unit vectors e_j in $\mathbb{R}^n$ are mapped by A into the column vectors a_j, that is since $Ae_j=a_j$, we have that A maps the unit n-cube in $\mathbb{R}^n$ onto the parallelepiped in $\mathbb{R}^n$ spanned by $a_1,\ldots,a_n$. Since the volume of the n-cube is one and the volume of the the parallelepiped spanned by $a_1,\ldots,a_n$ is $V(a_1,\ldots,a_n)$, the *volume scale* of the mapping $x\to Ax$ is equal to $V(a_1,\ldots,a_n)$.

42.25 The Volume $V(a_1, a_2)$ in $\mathbb{R}^2$

If A is the 2×2-matrix

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$

then $\det A = V(a_1, a_2)$ is given

$$\det A = V(a_1, a_2) = a_{11}a_{22} - a_{21}a_{12}. \tag{42.51}$$

of course, $a_1 = (a_{11}, a_{21})$ and $a_2 = (a_{12}, a_{22})$ are the column vectors of A.

42.26 The Volume $V(a_1, a_2, a_3)$ in $\mathbb{R}^3$

If A is the 3×3-matrix

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},$$

then $\det A = V(a_1, a_2, a_3)$ is given by

$$\begin{aligned} &\det A = V(a_1, a_2, a_3) = a_1 \cdot a_2 \times a_3 \\ &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}). \end{aligned} \tag{42.52}$$

We see that we can express $\det A$ as

$$\begin{aligned} \det A &= a_{11} \det A_{11} - a_{12} \det A_{12} + a_{13}A_{13} \\ &= a_{11}V(\hat{a}_2, \hat{a}_3) - a_{12}V(\hat{a}_1, \hat{a}_3) + a_{13}V(\hat{a}_1, \hat{a}_2) \end{aligned} \tag{42.53}$$

where the A_{1j} are 2×2 matrices formed by cutting out the first row and j:th column of A, explicitly given by

$$A_{11} = \begin{pmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix} \quad A_{12} = \begin{pmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{pmatrix} \quad A_{13} = \begin{pmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix}$$

and $\hat{a}_1 = (a_{21}, a_{31})$, $\hat{a}_2 = (a_{22}, a_{32})$, $\hat{a}_3 = (a_{23}, a_{33})$ are the 2-column vectors formed by cutting out the first element of the 3-columns a_j. We say that (42.53) is an *expansion* of the 3×3 matrix A in terms of the elements of the first row of A and the corresponding 2×2 matrices. The expansion formula follows by collecting all the terms with a_{11} as a factor, and all the terms with a_{12} as a factor and all the terms with a_{13} as a factor.

42.27 The Volume $V(a_1, a_2, a_3, a_4)$ in $\mathbb{R}^4$

Using the expansion formula we can compute the determinant $\det A = V(a_1, \ldots, a_4)$ of a 4×4 matrix $A = (a_{ij})$ with column vectors $a_j = (a_{1j}, \ldots, a_{4j})$ for $j = 1, 2, 3, 4$. We have

$$\det A = V(a_1, a_2, a_3, a_4) = a_{11}V(\hat{a}_2, \hat{a}_3, \hat{a}_4) - a_{12}V(\hat{a}_1, \hat{a}_3, \hat{a}_4) + a_{13}V(\hat{a}_1, \hat{a}_2, \hat{a}_4) - a_{14}V(\hat{a}_1, \hat{a}_2, \hat{a}_3),$$

where the $\hat{a}_j$, $j = 1, 2, 3, 4$ are the 3-column vectors corresponding to cutting out the first coefficient of the a_j. We have now expressed the determinant of the 4×4 matrix A as a sum of determinants of 3×3 matrices with the first row of A as coefficients.

42.28 The Volume $V(a_1, \ldots, a_n)$ in $\mathbb{R}^n$

Iterating the row-expansion formula indicated above, we can compute the determinant of an arbitrary $n \times n$ matrix A. As an example we give the expansion formula for a 5×5 matrix $A = (a_{ij})$:

$$\det A = V(a_1, a_2, a_3, a_4, a_5) = a_{11}V(\hat{a}_2, \hat{a}_3, \hat{a}_4, \hat{a}_5) - a_{12}V(\hat{a}_1, \hat{a}_3, \hat{a}_4, \hat{a}_5) + a_{13}V(\hat{a}_1, \hat{a}_2, \hat{a}_4, \hat{a}_5) - a_{14}V(\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{a}_5) + a_{15}V(\hat{a}_1, \hat{a}_2, \hat{a}_3, \hat{a}_4).$$

Evidently, we can formulate the following a rule of sign for the term with the factor a_{ij}: choose the $+$ if $i + j$ is even and the $-$ if $i + j$ is odd. This rule generalizes to expansions with respect to any row of A.

42.29 The Determinant of a Triangular Matrix

Let $A = (a_{ij})$ be a *upper triangular* $n \times n$ matrix, that is $a_{ij} = 0$ for $i > j$. All elements a_{ij} of A below the diagonal are zero. In this case the only non-zero term in the expression for $\det A$, is the product of the diagonal elements of A corresponding to the identity perturbation, so that

$$\det A = a_{11}a_{22} \cdots a_{nn}. \tag{42.54}$$

This formula also applies to a *lower triangular* $n \times n$ matrix $A = (a_{ij})$ with $a_{ij} = 0$ for $i < j$.

42.30 Using the Column Echelon Form to Compute $\det A$

We now present a way to compute $\det A = V(a_1, \ldots, a_n)$, where the a_j are the columns of a $n \times n$ matrix $A = (a_{ij})$, based on reduction to column

echelon form. We then use that the volume does not change if we subtract one column multiplied by a real number from another column, to obtain

$$\det A = V(a_1, a_2, \ldots, a_n) = V(\hat{a}_1, \hat{a}_2, \hat{a}_3, \ldots, \hat{a}_n)$$

where $\hat{a}_{ij} = 0$ if $j > i$, that is the corresponding matrix $\hat{A}$ is a lower triangular matrix. We then compute $V(\hat{a}_1, \hat{a}_2, \hat{a}_3, \ldots, \hat{a}_n)$ by multiplying the diagonal elements. As usual, if we meet a zero diagonal term we interchange columns until we meet a nonzero diagonal term, or if all diagonal terms appearing this way are zero, we proceed to modify the next row. At least one of the diagonal terms in the final triangular matrix will then be zero, and thus the determinant will be zero.

Example 42.9. We show the sequence of matrices in a concrete case:

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 4 & 6 \\ 3 & 4 & 6 \end{pmatrix} \quad \to \begin{pmatrix} 1 & 0 & 0 \\ 2 & 2 & 4 \\ 3 & 1 & 3 \end{pmatrix} \quad \to \begin{pmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 3 & 1 & 1 \end{pmatrix}$$

and conclude that $det\, A = 2$.[TS a]

42.31 The Magic Formula $\det AB = \det A \det B$

Let A and B be two $n \times n$ matrices. We know that AB is the matrix of the composite transformation $f(g(x))$, where $f(y) = Ay$ and $g(x) = Bx$. The volume scale of the mapping $x \to Bx$ is equal to $\det B$ and the volume scale of the mapping $y \to Ay$ is $\det A$, and hence the volume scale of the mapping $x \to ABx$ is equal to $\det A \det B$. This proves that

$$\det AB = \det A \det B,$$

which is one of the corner stones of the calculus of determinants. The proof suggested is a "short proof" avoiding algebraic computations. One can also give a direct algebraic proof using suitable expansion formulas for the determinant.

42.32 Test of Linear Independence

To test the linear independence of a given set of n vectors $\{a_1, a_2, \ldots, a_n\}$ in $\mathbb{R}^n$, we can use the volume $V(a_1, a_2, \ldots, a_n)$. More precisely, we shall prove that $\{a_1, a_2, \ldots, a_n\}$ is linearly independent if and only if $V(a_1, a_2, \ldots, a_n) \neq 0$. First, we note that if $\{a_1, a_2, \ldots, a_n\}$ is linearly dependent, for example if $a_1 = \sum_{j=2}^{n} \lambda_j a_j$ is a linear combination of $\{a_2, \ldots, a_n\}$,

[TS a] Should "det" be made italic.

Editor's or typesetter's annotations (will be removed before the final TEX run)

then $V(a_1, a_2, \ldots, a_n) = \sum_{j=2}^{n} \lambda_j V(a_j, a_2, \ldots, a_n) = 0$, since each factor $V(a_j, a_2, \ldots, a_n)$ has two equal vectors.

Secondly, if $\{a_1, a_2, \ldots, a_n\}$ is linearly independent, i.e., $\{a_1, a_2, \ldots, a_n\}$ is a basis for $\mathbb{R}^n$, then we must have $V(a_1, \ldots, a_n) \neq 0$. We see this as follows. We express each e_j as a linear combination of the set $\{a_1, a_2, \ldots, a_n\}$, for example $e_1 = \sum \lambda_{1j} a_j$. We have, since V is multilinear and vanishes if two arguments are the same, and $V(a_{\pi(1)}, \ldots, a_{\pi(n)}) = \pm V(a_1, \ldots, a_n)$ for any permutation π, that

$$1 = V(e_1, \ldots, e_n) = V\left(\sum_j \lambda_{1j} a_j, e_2, \ldots, e_n\right) = \sum_j \lambda_{1j} V(a_j, e_2, \ldots, e_n)$$

$$= \sum_j \lambda_{1j} V\left(a_j, \sum_k \lambda_{2k} a_k, e_3, \ldots, e_n\right) = \ldots = cV(a_1, \ldots, a_n), \quad (42.55)$$

for some constant c. We conclude that $V(a_1, \ldots, a_n) \neq 0$. We summarize as follows:

Theorem 42.5 *A set $\{a_1, a_2, \ldots, a_n\}$ of n vectors in $\mathbb{R}^n$ is linearly independent if and only if $V(a_1, \ldots, a_n) \neq 0$.*

We may restate this result in matrix language as follows: The columns of an $n \times n$-matrix A are linearly independent if and only if $\det A \neq 0$. We may thus sum up as follows:

Theorem 42.6 *Let A be a $n \times n$ matrix. Then the following statements are equivalent:*

- *The columns of A are linearly independent.*
- *If $Ax = 0$ then $x = 0$.*
- $\det A \neq 0$.

To test linear independence of the columns of a given matrix A we may thus compute $\det A$ and check if $\det A = 0$. We can also use this test in more quantitative form as follows: If $\det A$ is small then the columns are close to being linearly dependent and uniqueness is at risk!

A matrix A with $\det A = 0$ is called *singular*, while matrices with $\det A \neq 0$ are referred to as *non-singular*. Thus an $n \times n$-matrix is non-singular if and only if its columns are linearly independent. Again we can go to quantitative forms and say that a matrix A is close to singular if its determinant is close to zero. The dependence of the solution on the size of the determinant is clearly expressed in the next chapter.

42.33 Cramer's Solution for Non-Singular Systems

Consider again the $n \times n$ linear system of equations

$$Ax = b \tag{42.56}$$

or

$$\sum_{j=1}^{n} a_j x_j = b \tag{42.57}$$

where $A = (a_{ij})$ is an $n \times n$ matrix with columns $a_j = (a_{1j}, \ldots, a_{nj})$, $j = 1, \ldots, n$. Suppose that the columns a_j of A are linearly independent, or equivalently, that $\det A = V(a_1, \ldots, a_n) \neq 0$. We then know that (42.56) has a unique solution $x \in \mathbb{R}^n$ for any given $b \in \mathbb{R}^n$, we shall now seek a formula for the solution x in terms of b and the columns a_j of A.

Using the basic property of the volume function $V(g_1, \ldots, g_n)$ of a set $\{g_1, \ldots, g_n\}$ of n vectors g_i, in particular the property that $V(g_1, \ldots, g_n) = 0$ if any two of the g_i are equal, we obtain the following solution formula (Cramer's formula):

$$\begin{aligned} x_1 &= \frac{V(b, a_2, \ldots, a_n)}{V(a_1, a_2, \ldots, a_n)}, \\ &\ldots \\ x_n &= \frac{V(a_1, \ldots, a_{n-1}, b)}{V(a_1, a_2, \ldots, a_n)}. \end{aligned} \tag{42.58}$$

For example, to obtain the formula for x_1, use that

$$V(b, a_2, \ldots, a_n) = V\left(\sum_j a_j x_j, a_2, \ldots, a_n\right)$$

$$= \sum_{j=1}^{n} x_j V(a_j, a_2, \ldots, a_n) = x_1 V(a_1, a_2, \ldots, a_n).$$

We summarize:

Theorem 42.7 *If A is a $n \times n$ non-singular matrix with $\det A \neq 0$, then the system of equations $Ax = b$ has a unique solution x for any $b \in \mathbb{R}^n$. The solution is given by Cramer's formula* (42.58)*.*

A result like this was first derived by Leibniz and then by Gabriel Cramer (1704-1752) (who got a Ph.D. at the age of 18 with a thesis on the theory of sound) in *Introduction l'analyse des lignes courbes algbraique*. Throughout the book Cramer makes essentially no use of the Calculus in either Leibniz' or Newton's form, although he deals with such topics as tangents, maxima and minima, and curvature, and cites Maclaurin and Taylor in footnotes. One conjectures that he never accepted or mastered Calculus.

Fig. 42.2. Gabriel Cramer: "I am friendly, good-humoured, pleasant in voice and appearance, and possess good memory, judgement and health"

Note that Cramer's solution formula for $Ax = b$ is very computationally demanding, and thus cannot be used for actually computing the solution unless n is small. To solve linear systems of equations other methods are used, like Gaussian elimination and iterative methods, see Chapter *Solving systems of linear equations*.

42.34 The Inverse Matrix

Let A be a non-singular $n \times n$ matrix with $V(a_1, \ldots, a_n) \neq 0$. Then $Ax = b$ can be solved uniquely for all $b \in \mathbb{R}^n$ according to Cramer's solution formula (42.58). Clearly, x depends linearly on b, and the solution x may be expressed as $A^{-1}b$, where A^{-1} is an $n \times n$ matrix which we refer to as the *inverse* of A. The j:th column of A^{-1} is the solution vector corresponding to choosing $b = e_j$. Cramer's formula thus gives the following formula for the inverse A^{-1} of A:

$$A^{-1} = V(a_1, \ldots, a_n)^{-1} \begin{pmatrix} V(e_1, a_2, \ldots, a_n) & .. & V(a_1, \ldots, a_{n-1}, e_1) \\ \cdot & .. & \cdot \\ \cdot & .. & \cdot \\ V(e_n, a_2, \ldots, a_n) & .. & V(a_1, \ldots, a_{n-1}, e_n) \end{pmatrix}. \tag{42.59}$$

The inverse matrix A^{-1} of A satisfies

$$A^{-1}A = AA^{-1} = I,$$

where I is the $n \times n$ identity matrix, with ones on the diagonal and zeros elsewhere.

Evidently, we can express the solution to $Ax = b$ in the form $x = A^{-1}b$ if A is a non-singular $n \times n$ matrix (by multiplying $Ax = b$ by A^{-1} from the left)...

42.35 Projection onto a Subspace

Let V be a subspace of $\mathbb{R}^n$ spanned by the linearly independent set of vectors $\{a_1, \ldots, a_m\}$. In other words, $\{a_1, \ldots, a_m\}$ is a basis for V. The projection Pv of a vector $v \in \mathbb{R}^n$ onto V is defined as the vector $Pv \in V$ satisfying the orthogonality relation

$$(v - Pv, w) = 0 \quad \text{for all vectors } w \in V, \tag{42.60}$$

or equivalently

$$(Pv, a_j) = (v, a_j) \quad \text{for } j = 1, \ldots, m. \tag{42.61}$$

To see the equivalence, we note that (42.60) clearly implies (42.61). Conversely, any $w \in V$ is a linear combination of the form $\sum \mu_j a_j$, and multiplying (42.61) by μ_j and summing over j, we obtain $(Pv, \sum_j \mu_j a_j) = (v, \sum_j \mu_j a_j)$, which is (42.60) with $w = \sum_j \mu_j a_j$ as desired.

Expressing $Pv = \sum_{i=1}^m \lambda_i a_i$ in the basis $\{a_1, \ldots, a_m\}$ for V, the orthogonality relation (42.61) corresponds to the $m \times m$ linear system of equations

$$\sum_{i=1}^m \lambda_i (a_i, a_j) = (v, a_j) \quad \text{for } j = 1, 2, \ldots, m. \tag{42.62}$$

We shall now prove that this system admits a unique solution, which proves that the projection Pv of v onto V exists and is unique. By Theorem 42.6 it is enough to prove uniqueness. We thus assume that

$$\sum_{i=1}^m \lambda_i (a_i, a_j) = 0 \quad \text{for } j = 1, 2, \ldots, m.$$

Multiplying by λ_j and summing we get

$$0 = \left(\sum_{i=1}^m \lambda_i a_i, \sum_{j=1}^m \lambda_j a_j \right) = |\sum_{i=1}^m \lambda_i a_i|^2,$$

which proves that $\sum_i \lambda_i a_i = 0$ and thus $\lambda_i = 0$ for $i = 1, \ldots, m$, since the $\{a_1, \ldots, a_m\}$ is linearly independent.

We have now proved the following fundamental result:

Theorem 42.8 *Let V be a linear subspace of $\mathbb{R}^n$. Then for all $v \in \mathbb{R}^n$ the projection Pv of v onto V, defined by $Pv \in V$ and $(v - Pv, w) = 0$ for all $w \in V$, exists and is unique.*

We note that $P : \mathbb{R}^n \to V$ is a linear mapping. To see this, let v and $\hat{v}$ be two vectors in $\mathbb{R}^n$, and note that since $(v - Pv, w) = 0$ and $(\hat{v} - P\hat{v}, w) = 0$ for all $w \in V$, we have

$$(v + \hat{v} - (Pv + P\hat{v}), w) = (v - Pv, w) + (\hat{v} - P\hat{v}, w) = 0,$$

which shows that $Pv + P\hat{v} = P(v + \hat{v})$. Similarly, $Pw = \lambda Pv$ if $w = \lambda v$, for any $\lambda \in \mathbb{R}$ and $v \in \mathbb{R}^n$. This proves the linearity of $P : \mathbb{R}^n \to V$.

We further note that $PP = P$. We sum up as follows:

Theorem 42.9 *The projection $P : \mathbb{R}^n \to V$ onto a linear subspace V of $\mathbb{R}^n$ is a linear transformation defined by $(v - Pv, w) = 0$ for all $w \in V$, which satisfies $PP = P$.*

42.36 An Equivalent Characterization of the Projection

We shall now prove that the *projection* Pv of a vector $v \in \mathbb{R}^n$ onto V is the vector $Pv \in V$ with minimum distance to v, that is $|v - Pv| \leq |v - w|$ for all $w \in V$.

We state the equivalence of the two definitions of the projection in the following fundamental theorem:

Theorem 42.10 *Let $v \in \mathbb{R}^n$ be given. The vector $Pv \in V$ satisfies the orthogonality relation*

$$(v - Pv, w) = 0 \quad \text{for all vectors } w \in V, \tag{42.63}$$

if and only if Pv minimizes the distance to v in the sense that

$$|v - Pv| \leq |v - w| \quad \text{for all } w \in V. \tag{42.64}$$

Further, the element $Pv \in V$ satisfying (42.63) *and* (42.64) *is uniquely determined.*

To prove the theorem we note that by the orthogonality (42.60), we have for any $w \in V$,

$$\begin{aligned} |v - Pv|^2 &= (v - Pv, v - Pv) \\ &= (v - Pv, v - w) + (v - Pv, w - Pv) = (v - Pv, v - w), \end{aligned}$$

since $w - Pv \in V$. Using Cauchy-Schwarz inequality, we obtain

$$|v - Pv|^2 \leq |v - Pv|\,|v - w|,$$

which shows that $|v - Pv| \leq |v - w|$ for all $w \in V$.

Conversely, if $|v - Pv| \leq |v - w|$ for all $w \in V$, then for all $\epsilon \in \mathbb{R}$ and $w \in V$

$$\begin{aligned} |v - Pv|^2 &\leq |v - Pv + \epsilon w|^2 \\ &= |v - Pv|^2 + \epsilon(v - Pv, w) + \epsilon^2|w|^2, \end{aligned}$$

that is for all $\epsilon > 0$

$$(v - Pv, w) + \epsilon|w|^2 \geq 0,$$

which proves that

$$(v - Pv, w) \geq 0 \quad \text{for all } w \in V.$$

Changing w to $-w$ proves the reverse inequality and we conclude that $(v - Pv, w) = 0$ for all $w \in V$.

Finally, to prove uniqueness, assume that $z \in V$ satisfies

$$(v - z, w) = 0 \quad \text{for all vectors } w \in V.$$

Then $(Pv - z, w) = (Pv - v, w) + (v - z, w) = 0 + 0 = 0$ for all $w \in V$, and $Pv - z$ is a vector in V. Choosing $w = Pv - z$ thus shows that $|Pv - z|^2 = 0$, that is $z = Pv$. The proof of the theorem is now complete.

The argument just given is very fundamental and will be used many times below in various forms, so it is worth taking the time to understand it now.

42.37 Orthogonal Decomposition: Pythagoras Theorem

Let V be a subspace of $\mathbb{R}^n$. Let P be the projection onto V. Any vector x can be written

$$x = Px + (x - Px) \tag{42.65}$$

where $Px \in V$, and further $(x - Px) \perp V$ since by the definition of P we have $(x - Px, w) = 0$ for all $w \in V$. We say that $x = Px + (x - Px)$ is an *orthogonal decomposition* of x since $(Px, x - Px) = 0$.

Define the *orthogonal complement* $V^\perp$ to V by $V^\perp = \{y \in \mathbb{R}^n : y \perp V\} = \{y \in \mathbb{R}^n : y \perp x \quad \text{for all } x \in V\}$. It is clear that $V^\perp$ is a linear subspace of $\mathbb{R}^n$. We have that if $x \in V$ and $y \in V^\perp$, then $(x, y) = 0$. Further, any vector $z \in \mathbb{R}^n$ can be written $z = x + y$, with $x = Pz \in V$ and $y = (x - Px) \in V^\perp$. We can summarize by saying that

$$V \oplus V^\perp = \mathbb{R}^n, \tag{42.66}$$

is an *orthogonal decomposition* of $\mathbb{R}^n$ into the two orthogonal subspaces V and $V^\perp$: $x \in V$ and $y \in V^\perp$ implies $(x, y) = 0$ and any $z \in \mathbb{R}^n$ can be

written uniquely in the form $z = x+y$. The uniqueness of the decomposition $z = Pz + (z - Pz)$ follows from the uniqueness of Pz.

We note the following generalization of Pythagoras theorem: for any $x \in \mathbb{R}^n$, we have

$$|x|^2 = |Px|^2 + |x - Px|^2. \tag{42.67}$$

This follows by writing $x = Px + (x - Px)$ and using that $Px \perp (x - Px)$:

$$|x|^2 = |Px + (x - Px)|^2 = |Px|^2 + 2(Px, x - Px) + |x - Px|^2.$$

More generally, we have if $z = x + y$ with $x \perp y$ (that is $(x, y) = 0$), that

$$|z|^2 = |x|^2 + |y|^2.$$

42.38 Properties of Projections

Let P be the orthogonal projection onto a linear subspace V in $\mathbb{R}^n$. Then $P : \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation that satisfies

$$P^\top = P \quad \text{and} \quad PP = P. \tag{42.68}$$

We have already seen that $PP = P$. To see that $P^\top = P$ we note that

$$(w, P^\top v) = (Pw, v) = (Pw, Pv) = (w, Pv) \quad \text{for all } v, w \in \mathbb{R}^n, \tag{42.69}$$

and thus $P^\top = P$. Conversely, let $P : \mathbb{R}^n \to \mathbb{R}^n$ be a linear transformation which satisfies (42.68). Then P is the orthogonal projection onto a subspace V of $\mathbb{R}^n$. To see this, set $V = R(P)$ and note that since $P^\top = P$ and $PP = P$, we have

$$\begin{aligned}(x - Px, Px) &= (x, Px) - (Px, Px) = (x, Px) - (x, P^\top Px)\\ &= (x, Px) - (x, Px) = 0.\end{aligned}$$

This shows that $x = Px + (x - Px)$ is an orthogonal decomposition, and thus P is the orthogonal projection onto $V = R(P)$.

42.39 Orthogonalization: The Gram-Schmidt Procedure

Let $\{a_1, \ldots, a_m\}$ be a basis for a subspace V of $\mathbb{R}^n$, i.e., $\{a_1, \ldots, a_m\}$ is linearly independent and V is the set of linear combinations of $\{a_1, \ldots, a_m\}$. We try to construct another basis $\{\hat{e}_1, \ldots, \hat{e}_m\}$ for V that is *orthonormal*, i.e. such that the basis vectors $\hat{e}_i$ are mutually orthogonal and have length equal to one or

$$(\hat{e}_i, \hat{e}_j) = 0 \quad \text{for } i \neq j, \quad \text{and } |\hat{e}_i| = 1 \tag{42.70}$$

We choose $\hat{e}_1 = \frac{1}{|a_1|}a_1$ and let V_1 be the subspace spanned by $\hat{e}_1$, or equivalently by a_1. Let P_1 be the projection onto V_1. Define

$$\hat{e}_2 = \frac{1}{|a_2 - P_1 a_2|}(a_2 - P_1 a_2).$$

Then $(\hat{e}_1, \hat{e}_2) = 0$ and $|\hat{e}_2| = 1$. Further, the subspace V_2 spanned by $\{a_1, a_2\}$ is also spanned by $\{\hat{e}_1, \hat{e}_2\}$. We now continue in the same way: Let P_2 be the projection onto V_2 and define

$$\hat{e}_3 = \frac{1}{|a_3 - P_2 a_3|}(a_3 - P_2 a_3)$$

Then the subspace V_3 spanned by $\{a_1, a_2, a_3\}$ is also spanned by the orthonormal set $\{\hat{e}_1, \hat{e}_2, \hat{e}_3\}$.

Continuing, we obtain an orthonormal basis $\{\hat{e}_1, \ldots, \hat{e}_m\}$ for the subspace spanned by $\{a_1, \ldots, a_m\}$ with the property that for $i = 1, \ldots, m$, the subspace spanned by $\{a_1, \ldots, a_i\}$ is spanned by $\{\hat{e}_1, \ldots, \hat{e}_i\}$.

Note that since the basis $\{\hat{e}_1, \ldots, \hat{e}_m\}$ is orthogonal, the system of equations (42.62) corresponding to computing $P_{i-1}a_i$, is diagonal.

42.40 Orthogonal Matrices

Consider the matrix Q with columns $\hat{e}_1, \ldots, \hat{e}_n$, where $\{\hat{e}_1, \ldots, \hat{e}_n\}$ is an orthonormal basis for $\mathbb{R}^n$. Since the vectors $\hat{e}_j$ are pairwise orthogonal and of length one, $Q^\top Q = I$, where I is the $n \times n$ identity matrix. Conversely, if Q is a matrix such that $Q^\top Q = I$, where I is an identity matrix, then the columns of Q must be orthonormal.

An $n \times n$-matrix Q such that $Q^\top Q = I$, is called an *orthogonal matrix*. An orthogonal $n \times n$-matrix can thus be characterized as follows: Its columns form an *orthonormal basis* for $\mathbb{R}^n$, that is a basis consisting of pairwise orthogonal vectors of length, or norm, one.

We summarize:

Theorem 42.11 *An orthogonal matrix Q satisfies $Q^\top Q = QQ^\top = I$, and $Q^{-1} = Q^\top$.*

42.41 Invariance of the Scalar Product Under Orthogonal Transformations

Let Q by an $n \times n$ orthonormal matrix with the columns formed by the coefficients of basis vectors $\hat{e}_j$ of an orthonormal basis $\{\hat{e}_1, \ldots, \hat{e}_n\}$. We then know that the coordinates x of a vector with respect to the standard basis,

and the coordinates $\hat{x}$ with respect to the basis $\{\hat{e}_1, \ldots, \hat{e}_n\}$, are connected by

$$x = Q\hat{x}.$$

We now prove that the scalar product is invariant under the orthonormal change of coordinates $x = Q\hat{x}$. We compute setting $y = Q\hat{y}$,

$$(x, y) = (Q\hat{x}, Q\hat{y}) = (Q^\top Q\hat{x}, \hat{y}) = (\hat{x}, \hat{y}),$$

that is the scalar product is the same in the $\{e_1, \ldots, e_n\}$ coordinates as in the $\{\hat{e}_1, \ldots, \hat{e}_n\}$ coordinates. We summarize:

Theorem 42.12 *If Q is an orthogonal $n \times n$ matrix, then $(x, y) = (Qx, Qy)$ for all $x, y \in \mathbb{R}^n$.*

42.42 The QR-Decomposition

We can give the Gram-Schmidt orthogonalization procedure the following matrix interpretation: Let $\{a_1, \ldots, a_m\}$ be m linearly independent vectors in $\mathbb{R}^n$ and let A be the $n \times m$ matrix with the a_j occurring as columns. Let $\{\hat{e}_1, \ldots, \hat{e}_m\}$ be the corresponding orthonormal set generated by the Gram-Schmidt procedure, and let Q be the $n \times m$ matrix with the $\hat{e}_j$ as columns. Then

$$A = QR, \tag{42.71}$$

where R is a $m \times m$ upper triangular matrix, which expresses each a_j as a linear combination of $\{\hat{e}_1, \ldots, \hat{e}_j\}$.

The matrix Q satisfies $Q^\top Q = I$, where I is the $m \times m$ identity matrix, since the $\hat{e}_j$ are pairwise orthogonal and have length 1. We conclude that a $m \times n$ matrix A with linearly independent columns can be factored into $A = QR$, where Q satisfies $Q^\top Q = I$, and R is upper triangular. The columns of the matrix Q are orthonormal, as in the case of an orthonormal matrix, but if $m < n$, then they do not span all of $\mathbb{R}^n$.

42.43 The Fundamental Theorem of Linear Algebra

We return to our basic question of existence and uniqueness of solutions to the system $Ax = b$ with A a given $m \times n$ matrix and $b \in \mathbb{R}^m$ a given vector. We now allow m to be different from n, remembering that we focussed on the case $m = n$ above. We shall now prove the Fundamental Theorem of Linear Algebra giving an answer of theoretical nature to our basic questions of existence and uniqueness.

We note the following chain of equivalent statements for a $m \times n$-matrix A, where "iff" is shorthand for "if and only if":

$$x \in N(A) \quad \text{iff} \quad Ax = 0 \quad \text{iff} \quad x \perp \text{ rows of } A \quad \text{iff}$$
$$x \perp \text{ columns of } A^\top \quad \text{iff}$$
$$x \perp R(A^\top) \quad \text{iff}$$
$$x \in (R(A^\top)^\perp.$$

Thus $N(A) = (R(A^\top)^\perp$, and since $R(A^\top)^\perp \oplus R(A^\top) = \mathbb{R}^n$, we see that

$$N(A) \oplus R(A^\top) = \mathbb{R}^n. \tag{42.72}$$

As a consequence of this orthogonal splitting, we see that

$$\dim\, N(A) + \dim\, R(A^\top) = n, \tag{42.73}$$

where $\dim V$ is the dimension of the linear space V. We recall that the dimension $\dim V$ of a linear space V is the number of elements in a basis for V. Similarly, replacing A by $A^\top$ and using that $(A^\top)^\top = A$, we have

$$N(A^\top) \oplus R(A) = \mathbb{R}^m, \tag{42.74}$$

and thus in particular,

$$\dim\, N(A^\top) + \dim\, R(A) = m. \tag{42.75}$$

Next we note that, letting $g_1, \ldots, g_k$ be a basis in $N(A)^\perp$ so that $Ag_1, \ldots, Ag_k$ span $R(A)$ and thus $\dim R(A) \leq k$, we have

$$\dim\, N(A) + \dim\, R(A) \leq n, \quad \text{and also} \quad \dim\, N(A^\top) + \dim\, R(A^\top) \leq m. \tag{42.76}$$

Adding (42.73) and (42.75), we conclude that equality holds in (42.76). We summarize in:

Theorem 42.13 (The Fundamental Theorem of Linear Algebra) *Let A be a $m \times n$ matrix. Then*

$$N(A) \oplus R(A^\top) = \mathbb{R}^n \quad N(A^\top) \oplus R(A) = \mathbb{R}^m,$$

$$\textit{dim}\ N(A) + \textit{dim}\ R(A^\top) = n,\ \textit{dim}\ N(A^\top) + \textit{dim}\ R(A) = m,$$

$$\textit{dim}\ N(A) + \textit{dim}\ R(A) = n,\ \textit{dim}\ N(A^\top) + \textit{dim}\ R(A^\top) = m,$$

$$\textit{dim}\ R(A) = \textit{dim}\ R(A^\top).$$

In the special case $m = n$, we have that $R(A) = \mathbb{R}^m$ if and only if $N(A) = 0$ (which we proved above using Cramer's rule), stating that uniqueness implies existence.

We call dim $R(A)$ the *column rank* of the matrix A. The column rank of A is equal to the dimension of the space spanned by the columns of A. Similarly the *row rank* of A is equal to the dimension of the space spanned by the rows of A. The equality dim $R(A) =$ dim $R(A^\top)$ in the Fundamental Theorem expresses that the the column ranks of A and $A^\top$ are equal, that is that the column rank of A is equal to the *row rank* of A. We state this result as:

Theorem 42.14 *The number of linearly independent columns of A is equal to the number of linearly independent rows of A.*

Example 42.10. Returning to Example 41.5, we note that the column echelon form of $A^\top$ is the transpose of the row echelon form of A, that is

$$R(A^\top) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 1 & 3 & 2 & 0 \\ 1 & 6 & 5 & 0 \end{pmatrix}.$$

We check that the columns vectors $(0, 1, -2, 1, 0)$ and $(0, 4, -5, 0, 1)$ spanning $N(A)$ are orthogonal to $R(A^\top)$, that is orthogonal to the columns of the echelon form of $A^\top$. Of course, this is just a restatement of the fact that these vectors are orthogonal to the rows of the row echelon form $\hat{A}$ of A (as is evident from the proof of the Fundamental Theorem). We see that $N(A) \oplus R(A^\top) = \mathbb{R}^5$ as predicted by the Fundamental Theorem.

42.44 Change of Basis: Coordinates and Matrices

Let $\{s_1, \ldots, s_n\}$ be a basis for $\mathbb{R}^n$ where the coordinates of the basis vectors in the standard basis $\{e_1, \ldots, e_n\}$ are given by $s_j = (s_{1j}, \ldots, s_{nj})$. Recalling (42.23), we have the following connection between the coordinates x_i of a vector x with respect to the standard basis and the coordinates $\hat{x}_j$ of x with respect to the basis $\{s_1, \ldots, s_n\}$:

$$x_i = \sum_{j=1}^{n} s_{ij}\hat{x}_j \quad \text{for } i = 1, \ldots, n. \tag{42.77}$$

This follows from taking the scalar product of $\sum_{j=1}^n x_j e_j = \sum_{i=1}^n \hat{x}_j s_j$ with e_i and using that $s_{ij} = (e_i, s_j)$.

Introducing the matrix $S = (s_{ij})$, we thus have the following connection between the coordinates $x = (x_1, \ldots, x_n)$ with respect to $\{e_1, \ldots, e_n\}$, and the coordinates $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_n)$ with respect to $\{s_1, \ldots, s_n\}$:

$$x = S\hat{x}, \quad \text{that is } \hat{x} = S^{-1}x. \tag{42.78}$$

Consider now a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^n$ with matrix $A = (a_{ij})$ with respect to the standard basis $\{e_1, \ldots, e_n\}$, that is with $a_{ij} = f_i(e_j) = (e_i, f(e_j))$, where $f(x) = (f_1(x), \ldots, f_n(x))$ in the standard basis $\{e_1, \ldots, e_n\}$, that is

$$y = f(x) = \sum_i f_i(x)e_i = \sum_{i=1}^n \sum_{j=1}^n a_{ij}x_j e_i = Ax.$$

Writing $y = S\hat{y}$ and $x = S\hat{x}$, we have

$$S\hat{y} = AS\hat{x} \quad \text{that is } \hat{y} = S^{-1}AS\hat{x}$$

This shows that the matrix of the linear transformation $f : \mathbb{R}^n \to \mathbb{R}^n$, with the matrix A with respect to the standard basis, takes the following form in the basis $\{s_1, \ldots, s_n\}$:

$$S^{-1}AS, \tag{42.79}$$

where the the coefficients s_{ij} of the matrix $S = (s_{ij})$ are the coordinates of the basis vectors s_j with respect to the standard basis.

42.45 Least Squares Methods

Consider the $m \times n$ linear system of equations $Ax = b$, or

$$\sum_j^n a_j x_j = b,$$

where $A = (a_{ij})$ is an $m \times n$ matrix with columns $a_j = (a_{1j}, \ldots, a_{mj})$, $j = 1, \ldots, n$. We know that if $b \in R(A)$ then the system can be solved, and if $N(A) = 0$, then the solution is unique. Suppose now that $b \notin R(A)$. Then there is no $x \in \mathbb{R}^n$ such that $Ax = b$, and the system $Ax = b$ has no solution. We can replace the problem by the following *least squares problem*

$$\min_{x \in \mathbb{R}^n} |Ax - b|^2.$$

This problem amounts to seeking the projection Pb of b onto $R(A)$, that is the projection of b onto the space spanned by the columns a_j of A.

By the properties of projections given above, we know that $Pb \in R(A)$ exists and is uniquely determined by the relation

$$(Pb, y) = (b, y) \quad \text{for all } y \in R(A),$$

that is we seek $Pb = A\hat{x}$ for some $\hat{x} \in \mathbb{R}^n$ such that

$$(A\hat{x}, Ax) = (b, Ax) \quad \text{for all } x \in \mathbb{R}^n.$$

This relation can be written

$$(A^\top A\hat{x}, x) = (A^\top b, x) \quad \text{for all } x \in \mathbb{R}^n,$$

which is the same as the matrix equation

$$A^\top A\hat{x} = A^\top b,$$

which we refer to as the *normal equations*.

The matrix $A^\top A$ is an $n \times n$ symmetric matrix. Assume now that the columns a_j of A are linearly independent. Then $A^\top A$ is non-singular, because if $A^\top Ax = 0$, then

$$0 = (A^\top Ax, x) = (Ax, Ax) = |Ax|^2,$$

and thus $Ax = 0$ and therefore $x = 0$, since the columns of A are linearly independent. Thus the equation $A^\top A\hat{x} = A^\top b$ has a unique solution $\hat{x}$ for each right hand side $A^\top b$, given by the formula

$$\hat{x} = (A^\top A)^{-1}A^\top b.$$

In particular, we have the following formula for the projection Pb of b onto $R(A)$,

$$Pb = A(A^\top A)^{-1}A^\top b.$$

We can directly check that $P : \mathbb{R}^m \to \mathbb{R}^m$ defined this way is symmetric and satisfies $P^2 = P$.

If the columns of A are linearly dependent, then $\hat{x}$ is undetermined up to vectors $\hat{x}$ in $N(A)$. It is then natural to single out a unique $\hat{x}$ by requiring that $|\hat{x}|^2$ to be minimal. Using the orthogonal decomposition $\mathbb{R}^n = R(A^\top) \oplus N(A)$, this is equivalent to seeking $\hat{x}$ in $R(A^\top)$, since by Pythagoras theorem this minimizes $|\hat{x}|$. We thus seek $\hat{x}$ so that

- $A\hat{x}$ is equal to the projection Pb of b onto $R(A)$
- $\hat{x} \in R(A^\top)$.

This leads to the following equation for $\hat{x} = A^\top\hat{y}$:

$$(A\hat{x}, AA^\top y) = (b, AA^\top y) \quad \text{for all } y \in \mathbb{R}^m, \tag{42.80}$$

with $\hat{x}$ uniquely determined.

Chapter 42 Problems

42.1. Prove that a plane in $\mathbb{R}^3$ not passing through the origin is not a subspace of $\mathbb{R}^3$.

42.2. (a) What is a vector space? (b) What is a subspace of a vector space?

42.3. Verify (42.17) and (42.18).

42.4. Why must a set of more than n vectors in $\mathbb{R}^n$ be linearly dependent? Why must a set of n linearly independent vectors in $\mathbb{R}^n$ be a basis?

42.5. Verify that $R(A)$ and $N(A)$ are linear subspaces of $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively, and further that the orthogonal complement $V^\top$ of a subspace V of $\mathbb{R}^n$ is also a subspace of $\mathbb{R}^n$.

42.6. (a) Give an example showing that permutations need not commute. (b) Verify the associative law for permutations.

42.7. Compute the determinants of some $n \times n$ matrices with $n = 2, 3, 4, 5$.

42.8. Fill out the details in the proof of Cauchys inequality.

42.9. Write an algorithm for the Gram-Schmidt orthogonalization procedure, and implement it in Matlab, for example.

42.10. Fill in the details in (42.55)

42.11. Verify that for an orthogonal matrix $QQ^\top = I$. Hint: Multiply $Q^\top Q = I$ from the right with C and from the right with Q, where C is the matrix such that $QC = I$.

42.12. Prove for 2×2 matrices A and B that $\det AB = \det A \det B$.

42.13. How many operations are needed to solve an $n \times n$ system of linear equations using Cramer's formula?

42.14. Prove by reduction to column echelon form that a basis for $\mathbb{R}^n$ contains n elements.

42.15. Implement algorithms for reduction to column and row echelon forms.

42.16. Prove that the solution $\hat{x} \in R(A^\top)$ of (42.80) is uniquely determined.

42.17. Construct the row and column echelon forms of different (small) matrices, and check the validity of the Fundamental Theorem.

43
The Spectral Theorem

There seems to be three possibilities (of a Unified Theory of Physics):

1. There really is a complete unified theory, which we will someday discover if we are smart enough.
2. There is no ultimate theory of the Universe, just an infinite sequence of theories that describe the Universe more and more accurately.
3. There is no theory of the Universe; events cannot be predicted beyond a certain extent but occur in a random and arbitrary manner. (Stephen Hawking, in *A Brief History of Time*)

43.1 Eigenvalues and Eigenvectors

Let $A = (a_{ij})$ be a quadratic $n \times n$ matrix. We investigate the situation in which multiplication by A acts like scalar multiplication. To start with, we assume that the elements a_{ij} are real numbers. If $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is a non-zero vector that satisfies

$$Ax = \lambda x, \tag{43.1}$$

where λ is a real number, then we say that $x \in \mathbb{R}^n$ is an *eigenvector* of A and that λ is a corresponding *eigenvalue* of A. An eigenvector x has the property that Ax is parallel to x (if $\lambda \neq 0$), or $Ax = 0$ (if $\lambda = 0$). This is a special property, as easy to verify with almost any example we might make up.

If x is an eigenvector with corresponding eigenvalue λ then $\bar{x} = \mu x$ for any non-zero real number μ is also an eigenvector corresponding to the eigenvalue λ because

$$\text{if } Ax = \lambda x, \text{ then } A\bar{x} = \mu Ax = \mu\lambda x = \lambda\mu x = \lambda\bar{x}.$$

Thus, we may change the length of an eigenvector without changing the corresponding eigenvalue. For example, we may normalize an eigenvector to have length equal to 1. In essence, the direction of an eigenvector is determined, but not its length.

We shall now study the problem of finding eigenvalues and corresponding eigenvectors of a given a quadratic matrix. We shall see that this is a basic problem of linear algebra arising in many different situations. We shall prove the *Spectral Theorem* stating that if A is a symmetric real $n \times n$ matrix, then there is an orthogonal basis for $\mathbb{R}^n$ consisting of eigenvectors. We shall also briefly discuss the case of non-symmetric matrices.

Rewriting (43.1) as $(A-\lambda I)x = 0$ with $x \in \mathbb{R}^n$ a non-zero eigenvector and I the identity matrix, we see that the matrix $A-\lambda I$ must be singular if λ is an eigenvalue, that is $\det(A - \lambda I) = 0$. Conversely, if $\det(A - \lambda I) = 0$ then $A - \lambda I$ is singular and thus the null-space $N(A - \lambda I)$ is different from the zero vector and thus there is a non-zero vector x such that $(A - \lambda I)x = 0$, that is there is an eigenvector x with corresponding eigenvalue λ. Using the expansion formula for the determinant, we see that $\det(A - \lambda I)$ is a polynomial in λ of degree n with coefficients depending on the coefficients a_{ij} of A. The polynomial equation

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

is called the *characteristic equation*. We summarize:

Theorem 43.1 *The number λ is an eigenvalue of the $n \times n$ matrix A if and only if λ is a root of the characteristic equation* $\det(A - \lambda I) = 0$.

Example 43.1. If $A = (a_{ij})$ is a 2×2 matrix, then the characteristic equation is

$$\det(A - \lambda I) = (a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} = 0,$$

which is a second order polynomial equation in λ. For example, if

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

then the characteristic equation is $\det(A - \lambda I) = \lambda^2 - 1 = 0$ with roots $\lambda_1 = 1$ and $\lambda_2 = -1$. The corresponding normalized eigenvectors are $s_1 = \frac{1}{\sqrt{2}}(1, 1)$ and $s_2 = \frac{1}{\sqrt{2}}(1, -1)$ since

$$(A - \lambda_1 I)\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

and similarly $(A - \lambda_2)s_2 = 0$. We observe that $(s_1, s_2) = s_1 \cdot s_2 = 0$, that is eigenvectors corresponding to different eigenvalues are orthogonal.

43.2 Basis of Eigenvectors

Suppose $\{s_1, \ldots, s_n\}$ is a basis for $\mathbb{R}^n$ consisting of eigenvectors of the $n \times n$ matrix $A = (a_{ij})$ with corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$ so

$$As_i = \lambda_i s_i \quad \text{for } i = 1, \ldots, n. \tag{43.2}$$

Let S be the matrix with the columns equal to the eigenvectors s_j expressed in the standard basis. We can then write (43.2) in matrix form as follows,

$$AS = SD, \tag{43.3}$$

where D is the diagonal matrix with the eigenvalues λ_j on the diagonal. We thus have

$$A = SDS^{-1} \qquad \text{or} \qquad D = S^{-1}AS, \tag{43.4}$$

where D is a diagonal matrix. We say that S transforms A into a diagonal matrix D with the eigenvalues on the diagonal.

Conversely, if we can express a matrix A in the form $A = SDS^{-1}$ with S non-singular and D diagonal then $AS = SD$, which says that the columns of S are eigenvectors with corresponding eigenvalues on the diagonal of D.

Viewing the $n \times n$ matrix A as defining a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^n$ by $f(x) = Ax$, we can express the action of $f(x)$ in a basis of eigenvectors $\{s_1, \ldots, s_n\}$ by the diagonal matrix D since $f(s_i) = \lambda_i s_i$. Thus, the linear transformation $f : \mathbb{R}^n \to \mathbb{R}^n$ is expressed by the matrix A in the standard basis and by the diagonal matrix D matrix in a basis of eigenvectors. The coupling is given by

$$D = S^{-1}AS.$$

Of course, the action of a diagonal matrix is very easy to describe and to understand and this is the motivation for considering eigenvalues and eigenvectors.

We now formulate the following basic question in two equivalent forms:

- Given a $n \times n$ matrix A, is there a basis of eigenvectors of A?
- Given a $n \times n$ matrix A, is there a non-singular matrix S such that $S^{-1}AS$ is diagonal?

As we have seen, the columns of the matrix S are the eigenvectors of A and the diagonal elements are the eigenvalues.

We shall now give the following partial answer: if A is an $n \times n$ symmetric matrix, then there is an orthogonal basis for $\mathbb{R}^n$ consisting of eigenvectors. This is the celebrated *Spectral Theorem for symmetric matrices*. Notice the assumption that A is symmetric and that in this case the basis of eigenvectors may be chosen to be orthogonal.

Example 43.2. Recalling Example 43.1, we see that $s_1 = \frac{1}{\sqrt{2}}(1,1)$ and $s_2 = \frac{1}{\sqrt{2}}(1,-1)$ form an orthogonal basis. By the orthogonality of S, $S^{-1} = S^\top$, and

$$S^{-1}AS = \frac{1}{2}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}$$
$$= \frac{1}{2}\begin{pmatrix}1 & 1\\ -1 & 1\end{pmatrix}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix} = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}.$$

43.3 An Easy Spectral Theorem for Symmetric Matrices

The following version of the Spectral Theorem for symmetric matrices is easy to prove:

Theorem 43.2 *Let A be a symmetric $n \times n$ matrix. Suppose A has n distinct eigenvalues $\lambda_1, \ldots, \lambda_n$ and corresponding normalized eigenvectors $s_1, \ldots, s_n$ with $\|s_j\| = 1$, $j = 1, \ldots, n$. Then, $\{s_1, \ldots, s_n\}$ is an orthonormal basis of eigenvectors. Letting $Q = (q_{ij})$ be the orthogonal matrix with the columns $(q_{1j}, \ldots, q_{nj})$ being the coordinates of the eigenvectors s_j with respect to the standard basis, then $D = Q^{-1}AQ$ is a diagonal matrix with the eigenvalues λ_j on the diagonal and $A = QDQ^{-1}$, where $Q^{-1} = Q^\top$.*

To prove this result, it suffices to prove that eigenvectors corresponding to different eigenvalues are orthogonal. This follows from the assumption that there are n distinct eigenvalues $\lambda_1, \ldots, \lambda_n$ with corresponding normalized eigenvectors $s_1, \ldots, s_n$. If we prove that these eigenvectors are pairwise orthogonal, then they form a basis for $\mathbb{R}^n$ and the proof is complete. Thus, assume that s_i and s_j are eigenvectors corresponding to different eigenvalues λ_i and λ_j. Since A is symmetric and $(Ax, y) = (x, Ay)$ for all $x, y \in \mathbb{R}^n$, we have

$$\lambda_i(s_i, s_j) = (\lambda_i s_i, s_j) = (As_i, s_j) = (s_i, As_j)$$
$$= (s_i, \lambda_j s_j) = \lambda_j(s_i, s_j),$$

which implies that $(s_i, s_j) = 0$ since $\lambda_i \neq \lambda_j$. We state this observation as a theorem because of its basic importance.

Theorem 43.3 *If A is a symmetric $n \times n$ matrix, and s_i and s_j are eigenvectors of A corresponding to the eigenvalues λ_i and λ_j with $\lambda_i \neq \lambda_j$, then $(s_i, s_j) = 0$. In other words, eigenvectors corresponding to different eigenvalues are orthogonal.*

Note that above we prove the Spectral Theorem for a symmetric $n \times n$ matrix A in the case the characteristic equation $\det(A - \lambda I) = 0$

has n different roots. It thus remains to consider the case of multiple roots where there are less than n different roots. We will consider this below. The reader in hurry may skip that proof.

43.4 Applying the Spectral Theorem to an IVP

We show a typical application of the Spectral Theorem. Consider the initial value problem: find $u : [0,1] \to \mathbb{R}^n$ such that

$$\dot{u} = Au, \quad \text{for } 0 < t \leq 1, \quad u(0) = u_0,$$

where $A = (a_{ij})$ is a symmetric $n \times n$ matrix with real coefficients a_{ij} independent of t. Systems of this form arise in many applications and the behavior of such a system may be very complicated.

Suppose now that $\{g_1, \ldots, g_n\}$ is an orthonormal basis of eigenvectors of A and let Q be the matrix with columns comprised of the coordinates of the eigenvectors g_j with respect to the standard basis. Then $A = QDQ^{-1}$, where D is the diagonal matrix with the eigenvalues λ_j on the diagonal. We introduce the new variable $v = Q^{-1}u$, that is we set $u = Qv$, where $v : [0,1] \to \mathbb{R}^n$. Then, the equation $\dot{u} = Au$ takes the form $Q\dot{v} = AQv$, that is $\dot{v} = Q^{-1}AQv = Dv$, where we use the fact that Q is independent of time. Summing up, we get the following diagonal system in the new variable v,

$$\dot{v} = Dv \quad \text{for } 0 < t \leq 1, \quad v(0) = v_0 = Q^{-1}u_0.$$

The solution of this decoupled system is given by

$$v(t) = \begin{pmatrix} \exp(\lambda_1 t) & 0 & 0 & .... & 0 \\ 0 & \exp(\lambda_2 t) & 0 & .... & 0 \\ . & . & . & . & . \\ 0 & 0 & 0 & .... & \exp(\lambda_n) \end{pmatrix} v_0 = \exp(Dt)v_0,$$

where $\exp(Dt)$ is a diagonal matrix with the elements $\exp(\lambda_j t)$ on the diagonal. The dynamics of this system is easy to grasp: each component $v_j(t)$ of $v(t)$ evolves according to $v_j(t) = \exp(\lambda_j t)v_{0j}$.

Transforming back, we get the following solution formula in the original variable $u(t)$,

$$u(t) = Q\exp(Dt)Q^{-1}u_0. \tag{43.5}$$

With A as in Example 43.1,[TS a] we get the solution formula

$$u(t) = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} u_{01} \\ u_{02} \end{pmatrix}$$
$$= \frac{1}{2}\begin{pmatrix} (e^t + e^{-t})u_{01} + (e^t - e^{-t})u_{02} \\ (e^t - e^{-t})u_{01} + (e^t + e^{-t})u_{02} \end{pmatrix}.$$

[TS a] It was (43.1) on hardcopy, please check it.

Editor's or typesetter's annotations (will be removed before the final TEX run)

43.5 The General Spectral Theorem for Symmetric Matrices

Above, we saw that eigenvalues of a matrix A are roots of the characteristic equation $\det(A - \lambda I) = 0$. In principle, we can find the eigenvalues and eigenvectors of given matrix by first solving the characteristic equation to find all the eigenvalues, and then for each eigenvalue λ find corresponding eigenvectors by solving the linear system of equations $(A - \lambda I)x = 0$.

We shall now present an alternative way of constructing/finding the eigenvectors and eigenvalues of a symmetric matrix A that also proves the Spectral Theorem for a symmetric $n \times n$ matrix A in the general case with possibly multiple roots. In the proof, we construct an orthonormal basis of eigenvectors $\{s_1, \ldots, s_n\}$ of A by constructing the eigenvectors one by one starting with s_1.

Constructing the First Eigenvector s_1

To construct the first eigenvector s_1, we consider the minimization problem: find $\bar{x} \in \mathbb{R}^n$ such that

$$F(\bar{x}) = \min_{x \in \mathbb{R}^n} F(x), \tag{43.6}$$

where

$$F(x) = \frac{(Ax, x)}{(x, x)} = \frac{(f(x), x)}{(x, x)} \tag{43.7}$$

is the so-called *Rayleigh quotient*. We note that the function $F(x)$ is *homogenous of degree zero*, that is for any $\lambda \in \mathbb{R}$, $\lambda \neq 0$, we have

$$F(x) = F(\lambda x),$$

because we can simply divide out the factor λ. In particular, for any $x \neq 0$,

$$F(x) = F\left(\frac{x}{\|x\|}\right), \tag{43.8}$$

and thus we may restrict the x in (43.6) to have length one, that is we may consider the equivalent minimization problem: find $\bar{x}$ with $\|\bar{x}\| = 1$ such that

$$F(\bar{x}) = \min_{x \in \mathbb{R}^n \, \|x\|=1} F(x) \tag{43.9}$$

Since $F(x)$ is Lipschitz continuous on the closed and bounded subset $\{x \in \mathbb{R}^n : \|x\| = 1\}$ of $\mathbb{R}^n$, we know by the Chapter Minimization, that the problem (43.9) has a solution $\bar{x}$, and thus also the problem (43.6) has a solution $\bar{x}$. We set $s_1 = \bar{x}$, and check that g_1 is indeed an eigenvector of A, that is an eigenvector of $f : \mathbb{R}^n \to \mathbb{R}^n$.

Since $\bar{x}$ solves the minimization problem (43.6), we have $\nabla F(\bar{x}) = 0$, where ∇F is the gradient of F. Computing $\nabla F(x)$ using the symmetry of $F(x)$ or the matrix A, we find that

$$\nabla F(x) = \frac{(x,x)2Ax - (Ax,x)2x}{(x,x)^2}, \tag{43.10}$$

so that with $x = \bar{x}$ satisfying $(\bar{x},\bar{x}) = 1$,

$$\nabla F(\bar{x}) = 2(A\bar{x} - (A\bar{x},\bar{x})\bar{x}) = 0,$$

that is

$$A\bar{x} = \lambda_1\bar{x}, \tag{43.11}$$

where

$$\lambda_1 = (A\bar{x},\bar{x}) = \frac{(A\bar{x},\bar{x})}{(\bar{x},\bar{x})} = \min_{x\in\mathbb{R}^n} F(x). \tag{43.12}$$

Setting $s_1 = \bar{x}$, we thus have

$$As_1 = \lambda_1 s_1, \quad \lambda_1 = (As_1, s_1), \quad \|s_1\| = 1.$$

We have now constructed the first normalized eigenvector s_1 with corresponding eigenvalue λ_1. We now let V_1 be the orthogonal complement of the space spanned by s_1, consisting of all the vectors $x \in \mathbb{R}^n$ such that $(x, s_1) = 0$. The dimension of V_1 is $n - 1$.

Invariance of A

Note that V_1 is *invariant* with respect to A in the sense that if $x \in V_1$ then $Ax \in V_1$. This follows because if $(x, g_1) = 0$ then $(Ax, s_1) = (x, As_1) = (x, \lambda_1 s_1) = \lambda_1(x, s_1) = 0$. This means that we can now restrict attention to the action of A on V_1, having handled the action of A on the space spanned by the first eigenvector s_1.

Constructing the Second Eigenvector s_2

Consider the minimization problem to find $\bar{x} \in V_1$ such that

$$F(\bar{x}) = \min_{x\in V_1} F(x). \tag{43.13}$$

By the same argument, this problem has a solution which we denote s_2 and which satisfies $As_2 = \lambda_2 s_2$, where $\lambda_2 = \frac{(As_2,s_2)}{(s_2,s_2)}$, and $\|s_2\| = 1$. Because in (43.13) we minimize over a smaller set than in (43.6), $\lambda_2 \geq \lambda_1$. Note that it may happen that $\lambda_2 = \lambda_1$, although V_1 is a subset of $\mathbb{R}^n$. In that case, we say that $\lambda_1 = \lambda_2$ is a *multiple eigenvalue*.

Continuing the Process

Let V_2 be the orthogonal subspace to the space spanned by s_1 and s_2 Again A is invariant on V_2 and the space spanned by $\{s_1, s_2\}$. Continuing this way, we obtain a orthonormal basis $\{s_1, \ldots, s_n\}$ of eigenvectors of A with corresponding real eigenvalues λ_i.

We have now proved the famous

Theorem 43.4 (Spectral Theorem): *If $f : \mathbb{R} \to \mathbb{R}$ is a linear symmetric transformation with corresponding symmetric $n \times n$ matrix A in the standard basis, then there is an orthogonal basis $(g_1, \ldots, g_n)$ of $\mathbb{R}^n$ consisting of eigenvectors g_i of f with corresponding real eigenvalues λ_j satisfying $f(g_j) = Ag_j = \lambda_j g_j$, for $j = 1, \ldots, n$. We have $D = Q^{-1}AQ$ and $A = QDQ^{-1}$, where Q is the orthogonal matrix with the coefficients of the eigenvectors g_j in the standard basis forming the columns, and D is the diagonal matrix with the eigenvalues λ_j on the diagonal.*

43.6 The Norm of a Symmetric Matrix

We recall that we have defined the *Euclidean norm* $\|A\|$ of a $n \times n$ matrix A by

$$\|A\| = \max_{x \in \mathbb{R}^n} \frac{\|Ax\|}{\|x\|}, \tag{43.14}$$

where we maximize over $x \neq 0$. By the definition, we have

$$\|Ax\| \leq \|A\|\,\|x\|, \tag{43.15}$$

and we may thus view $\|A\|$ to be the smallest constant C such that $\|Ax\| \leq C\|x\|$ for all $x \in \mathbb{R}^n$.

We shall now prove that if A is symmetric, then we can directly relate $\|A\|$ to the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A:

$$\|A\| = \max_{i=1,\ldots,n} |\lambda_i|. \tag{43.16}$$

We do this as follows. Using the Spectral theorem, we can write A as $A = Q^\top \Lambda Q$ with Q orthogonal and Λ a diagonal matrix with the eigenvalues λ_i on the diagonal. We recall that (cf. (42.46))

$$\|\Lambda\| = \max_{i=1,\ldots,n} |\lambda_i| = |\lambda_j| \tag{43.17}$$

and thus for all $x \in \mathbb{R}^n$,

$$\|Ax\| = \|Q^\top \Lambda Q x\| = \|\Lambda Q x\| \leq \|\Lambda\|\|Qx\| = \|\Lambda\|\|x\| = \max_{i=1,\ldots,n} |\lambda_i|\|x\|,$$

which proves that $\|A\| \leq \max_{i=1,\ldots,n} |\lambda_i|$. Choosing x to be equal to the eigenvector corresponding to the eigenvalue λ_j of maximal modulus proves

that indeed $\|A\| = \max_{i=1,\ldots,n} |\lambda_i| = |\lambda_j|$. We have proved the following result, which is a corner stone of numerical linear algebra.

Theorem 43.5 *If A is a symmetric $n \times n$ matrix, then $\|A\| = \max |\lambda_i|$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A.*

43.7 Extension to Non-Symmetric Real Matrices

Up until now, we have mainly focussed on the case of *real scalars*, that is we assume that the components of vectors are real numbers. We know that we can also let the components of vectors be *complex numbers*, and we may then allow eigenvalues to be complex numbers. The fundamental theorem of algebra states that a polynomial equation of degree n with complex coefficients, has n complex roots, and thus the characteristic equation $\det(A - \lambda I) = 0$ has n complex roots $\lambda_1, \ldots, \lambda_n$, and thus a $n \times n$ matrix A has n complex eigenvalues $\lambda_1, \ldots, \lambda_n$, if roots are counted with multiplicity. We have in this chapter focussed on *symmetric* matrices A with real coefficients and we have proved that a symmetric matrix with real coefficients has n real eigenvalues, counted with multiplicity. For symmetric matrices we can thus limit ourselves to real roots of the characteristic equation.

Chapter 43 Problems

43.1. Verify (43.10).

43.2. Compute the eigenvalues and eigenvectors of an arbitrary symmetric 2×2 matrix A. Solve the corresponding initial-value problem $\dot{u}(t) = Au(t)$ for $t > 0$, $u(0) = u^0$.

44
Solving Linear Algebraic Systems

All thought is a kind of computation. (Hobbes)

44.1 Introduction

We are interested in solving a system of linear equations

$$Ax = b,$$

where A is a given $n \times n$ matrix and $b \in \mathbb{R}^n$ is a given n-vector and we seek the solution vector $x \in \mathbb{R}^n$. We recall that if A is non-singular with non-zero determinant, then the solution $x \in \mathbb{R}^n$ is theoretically given by Cramer's formula. However if n is large, the computational work in using Cramer's formula is prohibitively large, so we need to find a more efficient means of computing the solution.

We shall consider two types of methods for solving the system $Ax = b$: (i) *direct methods* based on *Gaussian elimination* that theoretically produce a solution after a finite number of arithmetic operations, and (ii) *iterative methods* that produce a generally infinite sequence of increasingly accurate approximations.[TS^b]

44.2 Direct Methods

We begin by noting that some linear systems are easier to solve than others. For example if $A = (a_{ij})$ is *diagonal*, which means that $a_{ij} = 0$ if $i \neq j$, then

the system is solved in n operations: $x_i = b_i/a_{ii}$, $i = 1, \ldots, n$. Further, if the matrix is *upper triangular*, which means that $a_{ij} = 0$ if $i > j$, or *lower triangular*, which means that $a_{ij} = 0$ if $i < j$, then the system can be solved by *backward substitution* or *forward substitution* respectively; see Fig. 44.1 for an illustration of these different types. For example if A is upper triangular, the "pseudo-code" shown in Fig. 44.2 solves the system $Ax = b$ for the vector $x = (x_i)$ given the vector $b = (b_i)$ (assuming that $a_{kk} \neq 0$): In all three cases, the systems have a unique solution as long as the diagonal entries of A are nonzero.

Direct methods are based on Gaussian elimination, which in turn is based on the observation that the solution of a linear system is not changed under the following *elementary row operations*:

- interchanging two equations
- adding a multiple of one equation to another
- multiplying an equation by a nonzero constant.

The idea behind Gaussian elimination is to transform using these operations a given system into an upper triangular system, which is solved by back substitution. For example, to solve the system

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_2 + 2x_3 &= 1 \\ 2x_1 + x_2 + 3x_3 &= 1, \end{aligned}$$

we first subtract 2 times the first equation from the third to get the equivalent system,

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_2 + 2x_3 &= 1 \\ -x_2 + x_3 &= -1. \end{aligned}$$

We define the *multiplier* to be the factor 2. Next, we subtract -1 times the second row from the third to get

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_2 + 2x_3 &= 1 \\ 3x_3 &= 0. \end{aligned}$$

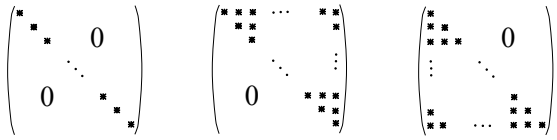

Fig. 44.1. The pattern of entries in diagonal, upper, and lower triangular matrices. A "$*$" denotes a possibly nonzero entry

TS[b] A sentence lacks in comparison with the hardcopy, please check it.

Editor's or typesetter's annotations (will be removed before the final TEX run)

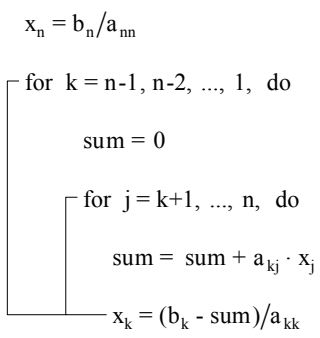

Fig. 44.2. An algorithm for solving an upper triangular system by back substitution

In this case, the multiplier is -1. The system is now upper triangular and using back substitution, we obtain $x_3 = 0$, $x_2 = 1$, and $x_1 = 0$. Gaussian elimination can be coded in a straightforward way using matrix notation.

Matrix Factorization

There is another way to view Gaussian elimination that is useful for the purposes of programming and handling special cases. Namely, Gaussian elimination is equivalent to computing a *factorization* of the coefficient matrix, $A = LU$, where L is a lower triangular and U an upper triangular $n \times n$ matrix. Given such a factorization of A, solving the system $Ax = b$ is straightforward. We first set $y = Ux$, then solve $Ly = b$ by forward substitution and finally solve $Ux = y$ by backward substitution.

To see that Gaussian elimination gives an LU factorization of A, consider the example above. We performed row operations that brought the system into upper triangular form. If we view these operations as row operations on the matrix A, we get the sequence

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 2 & 1 & 3 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & -1 & 1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{pmatrix},$$

which is an upper triangular matrix. This is the "U" in the LU decomposition.

The matrix L is determined by the observation that the row operations can be performed by multiplying A on the left by a sequence of special matrices called *Gauss transformations*. These are lower triangular matrices that have at most one nonzero entry in the off-diagonal positions and 1s down the diagonal. We show a Gauss transformation in Fig. 44.3. Multiplying A on the left by the matrix in Fig. 44.3 has the effect of adding α_{ij}

times row j of A to row i of A. Note that the inverse of this matrix is obtained changing α_{ij} to $-\alpha_{ij}$; we will use this below.

$$\begin{pmatrix}
1 & 0 & & & & \cdots & & & 0 \\
0 & 1 & 0 & & & & & & 0 \\
 & \ddots & 1 & \ddots & & 0 & & & \\
\vdots & & & & & & & & \vdots \\
 & 0 & 0 & 0 & \ddots & 0 & & & \\
 & 0 & \alpha_{ij} & 0 & \ddots & 1 & \ddots & & \\
 & 0 & 0 & 0 & & & \ddots & & \\
 & & & & & & 0 & 1 & 0 \\
0 & & \cdots & & & & & 0 & 1
\end{pmatrix}$$

Fig. 44.3. A Gauss transformation

To perform the first row operation on A above, we multiply A on the left by

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix},$$

to get

$$L_1 A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & -1 & -1 \end{pmatrix}.$$

The effect of pre-multiplication by L_1 is to add $-2\times$ row 1 of A to row 3. Note that L_1 is lower triangular and has ones on the diagonal.

Next we multiply $L_1 A$ on the left by

$$L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix},$$

and get

$$L_2 L_1 A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{pmatrix} = U.$$

L_2 is also lower triangular with ones on the diagonal. It follows that $A = L_1^{-1} L_2^{-1} U$ or $A = LU$, where

$$L = L_1^{-1} L_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & -1 & 1 \end{pmatrix}.$$

It is easy to see that L is also lower triangular with 1's on the diagonal with the multipliers (with sign change) occurring at the corresponding positions. We thus get the factorization

$$A = LU = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{pmatrix}.$$

Note that the entries in L below the diagonal are exactly the multipliers used to perform Gaussian elimination on A.

A general linear system can be solved in exactly the same fashion by Gaussian elimination using a sequence of Gauss transformations to obtain a factorization $A = LU$.

An LU factorization can be performed *in situ* using the storage space allotted to the matrix A. The fragment of code shown in Fig. 44.4 computes the LU factorization of A, storing U in the upper triangular part of A and storing the entries in L below the diagonal in the part of A below the diagonal. We illustrate the storage of L and U in Fig. 44.5.

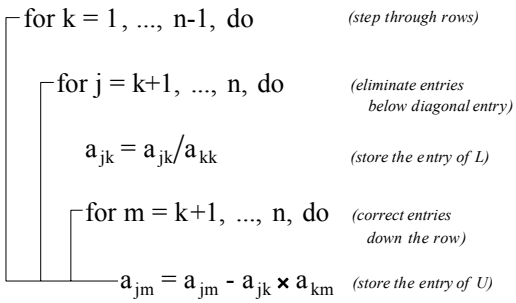

Fig. 44.4. An algorithm to compute the LU factorization of A that stores the entries of L and U in the storage space of A

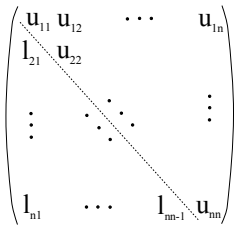

Fig. 44.5. The matrix output from the algorithm in Fig. 44.4. L and U are stored in the space allotted to A

Measuring the Cost

The cost of solving a linear system using a direct method is measured in terms of computer time. In practice, the amount of computer time is proportional to the number of arithmetic and storage operations the computer uses to compute the solution. It is traditional (on a sequential computer) to simplify the cost calculation by equating storing a value, addition, and subtraction and equating multiplication and division when counting operations. Moreover, since multiplication (i.e. multiplications and divisions) generally cost much more than addition on older computers, it is also common to simply count the number of multiplications (=multiplications+divisions).

By this measure, the cost of computing the LU decomposition of an $n \times n$ matrix is $n^3 - n/3 = O(n^3/3)$. We introduce some new notation here, the big "O". The actual count is $n^3/3 - n/3$, however when n is large, the lower order term $-n/3$ becomes less significant. In fact,

$$\lim_{n\to\infty} \frac{n^3/3 - n/3}{n^3/3} = 1, \tag{44.1}$$

and this is the definition of the big "O". (Sometimes the big "O" notation means that the limit of the ratio of the two relevant quantities is any constant). With this notation, the operations count of the LU decomposition is just $O(n^3)$.

The cost of the forward and backward substitutions is much smaller:

Pivoting

During Gaussian elimination, it sometimes happens that the coefficient of a variable in the "diagonal position" becomes zero as a result of previous eliminations. When this happens of course, it is not possible to use that equation to eliminate the corresponding entries in the same column lying below the diagonal position. If the matrix is invertible, it is possible to find a non-zero coefficient in the same column and below the diagonal position, and by switching the two rows, the Gaussian elimination can proceed. This is called *zero pivoting*, or just *pivoting*.

Adding pivoting to the LU decomposition algorithm is straightforward. Before beginning the elimination using the current diagonal entry, we check to see if that entry is non-zero. If it is zero, we search the entries below in the same column for the first non-zero value, then interchange the row corresponding to that non-zero entry with the row corresponding to the current diagonal entry which is zero. Because the row interchanges involve rows in the "un-factored" part of A, the form of L and U are not affected. We illustrate this in Fig. 44.6.

To obtain the correct solution of the linear system $Ax = b$, we have to mirror all pivots performed on A in the data b. This is easy to do with

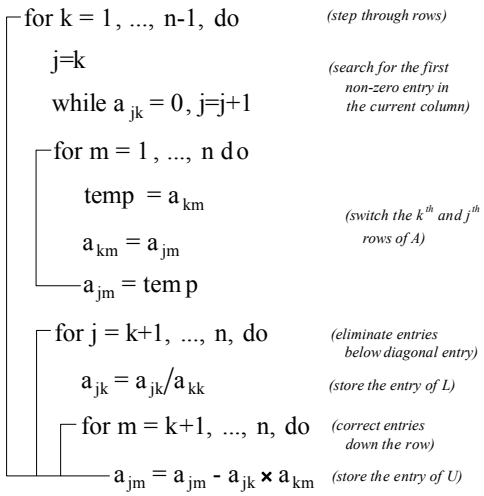

Fig. 44.6. An algorithm to compute the LU factorization of A that used pivoting to avoid zero-valued diagonal entries

the following trick. We define the vector of integers $p = (1,\ 2,\ \ldots,\ n)^\top$. This vector is passed to the LU factorization routine and whenever two rows of A are interchanged, we interchange the corresponding entries in p. After getting the altered p vector back, we pass it to the forward/backward routine. Here, we address the vector b indirectly using the vector p, i.e., we use the vector with entries $(b_{p_i})_{i=1}^n$, which has the effect of interchanging the rows in b in the correct fashion.

There are additional reasons to pivot in practice. As we have noted, the computation of the LU decomposition can be sensitive to errors originating from the finite precision of the computer if the matrix A is close to being non-invertible. We discuss this further below. We mention here however that a special kind of pivoting, called *partial pivoting* can be used to reduce this sensitivity. The strategy behind partial pivoting is to search the entries in the same column and below the current diagonal entry for the largest in absolute value. The row corresponding to the largest entry in magnitude is interchanged with the row corresponding to the current entry at the diagonal. The use of partial pivoting generally gives more accurate results than factorization without partial pivoting. One reason is that partial pivoting insures that the multipliers in the elimination process are kept as small as possible and consequently the errors in each entry are magnified by as little as possible during the course of the Gaussian elimination. We illustrate this with an example. Suppose that we solve

$$\begin{aligned} .000100x_1 + 1.00x_2 &= 1.00 \\ 1.00x_1 + 1.00x_2 &= 2.00 \end{aligned}$$

on a computer that holds three digits. Without pivoting, we get

$$\begin{aligned} .000100x_1 + 1.00x_2 &= 1.00 \\ -10000x_2 &= -10000 \end{aligned}$$

which implies that $x_2 = 1$ and $x_1 = 0$. Note the large multiplier that is required for the elimination. Since the true answer is $x_1 = 1.0001$ and $x_2 = .9999$, the computed result has an error of 100% in x_1. If we switch the two rows before eliminating, which corresponds exactly to the partial pivoting strategy, we get

$$\begin{aligned} 1.00x_1 + 1.00x_2 &= 2.00 \\ 1.00x_2 &= 1.00 \end{aligned}$$

which gives $x_1 = x_2 = 1.00$ as a result.

44.3 Direct Methods for Special Systems

It is often the case that the matrices arising from the Galerkin finite element method applied to a differential equation have special properties that can be useful during the solution of the associated algebraic equations. For example, the stiffness matrix for the Galerkin finite element approximation of the two-point boundary value problem with no convection is symmetric, positive-definite, and tridiagonal. In this section, we examine a couple of different classes of problems that occur frequently.

Symmetric, Positive-Definite Systems

As we mentioned, symmetric, positive-definite matrices are often encountered when discretizing differential equations (especially if the spatial part of the differential equation is of the type called elliptic). If A is symmetric and positive-definite, then it can be factored as $A = BB^\top$ where B is a lower triangular matrix with positive diagonal entries. This factorization can be computed from the LU decomposition of A, but there is a *compact method* of factoring A that requires only $O(n^3/6)$ multiplications called *Cholesky's method*.:

$$\begin{aligned} &b_{11} = \sqrt{a_{11}} \\ &b_{i1} = \frac{a_{i1}}{b_{11}}, \quad 2 \le i \le n, \\ &\begin{cases} b_{jj} = \left(a_{jj} - \sum_{k=1}^{j-1} b_{jk}^2\right)^{1/2}, \\ b_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} b_{ik} b_{jk}\right) / b_{jj}, \end{cases} \qquad 2 \le j \le n,\ j+1 \le i \le n \end{aligned}$$

This is called a compact method because it is derived by assuming that the factorization exists and then computing the coefficients of B directly from the equations obtained by matching coefficients in $BB^\top = A$. For example, if we compute the coefficient in the first row and column of $BB^\top$ we get b_{11}^2, which therefore must equal a_{11}. It is possible to do this because A is positive-definite and symmetric, which implies among other things that the diagonal entries of A remain positive throughout the factorization process and pivoting is not required when computing an LU decomposition.

Alternatively, the square roots in this formula can be avoided by computing a factorization $A = CDC^\top$ where C is a lower triangular matrix with ones on the diagonal and D is a diagonal matrix with positive diagonal coefficients.

Banded Systems

Banded systems are matrices with non-zero coefficients only in some number of diagonals centered around the main diagonal. In other words, $a_{ij} = 0$ for $j \leq i - d_l$ and $j \geq i + d_u$, $1 \leq i, j \leq n$, where d_l is the *lower bandwidth*, d_u is the *upper bandwidth*, and $d = d_u + d_l - 1$ is called the *bandwidth*. We illustrate this in Fig. 44.7. The stiffness matrix computed for the two-point boundary value problem with no convection is an example of a tridiagonal matrix, which is a matrix with lower bandwidth 2, upper bandwidth 2, and bandwidth 3.

When performing the Gaussian elimination used to compute the LU decomposition, we see that the entries of A that are already zero do not have to be reduced further. If there are only relatively few diagonals with non-zero entries, then the potential saving is great. Moreover, there is no

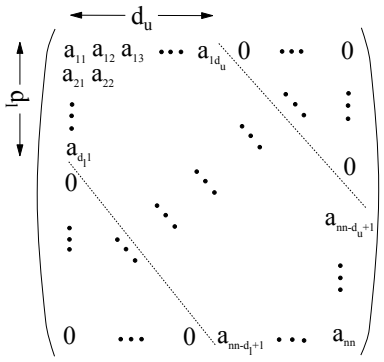

Fig. 44.7. The notation for a banded matrix

need to store the zero-valued entries of A. It is straightforward to adapt the LU factorization and forward/backward substitution routines to a banded pattern, once a storage scheme has been devised. For example, we can store a tridiagonal matrix as a $3 \times n$ matrix:

$$\begin{pmatrix} a_{21} & a_{31} & 0 & & \cdots & & 0 \\ a_{12} & a_{22} & a_{32} & 0 & \cdots & & 0 \\ 0 & a_{13} & a_{23} & a_{33} & 0 & \cdots & \vdots \\ & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & 0 & a_{1n-1} & a_{2n-1} & a_{3n-1} \\ 0 & & \cdots & & 0 & a_{1n} & a_{2n} \end{pmatrix}.$$

The routine displayed in Fig. 44.8 computes the LU factorization, while the routine in Fig. 44.9 performs the forward/backward substitution.

The cost of this routine grows linearly with the dimension, rather than at a cubic rate as in the full case. Moreover, we use only the equivalent of six vectors of dimension n for storage. A more efficient version, derived as a compact method, uses even less.

This algorithm assumes that no pivoting is required to factor A. Pivoting during the factorization of a banded matrix raises the difficulty that the bandwidth becomes larger. This is easy to see in a tridiagonal matrix, in

```
for k = 2, ..., n, do
    a_1k = a_1k / a_2k-1
    a_2k = a_2k - a_1k × a_3k-1
```

Fig. 44.8. A routine for computing the LU factorization of a tridiagonal system

```
y_1 = b_1
for k = 2, ..., n, do
    y_k = b_k - a_1k × y_k-1

x_n = y_n / a_2n
for k = n-1, ..., 1, do
    x_k = (y_k - a_3k × x_k+1) / a_2k
```

Fig. 44.9. Using forward and backward substitution to solve a tridiagonal system given the LU factorization

which case we have to store an extra vector to hold the extra elements above the diagonal that result if two adjacent rows are switched.

As for a tridiagonal matrix, it is straightforward to program special LU factorization and forward/backward substitution routines for a matrix with bandwidth d. The operations count is $O(nd^2/2)$ and the storage requirement is a matrix of dimension $d \times n$ if no pivoting is required. If d is much less than n, the savings in a special approach are considerable.

While it is true that if A is banded, then L and U are also banded, it is also true that in general L and U have non-zero entries in positions where A is zero. This is called *fill-in*. In particular, the stiffness matrix for a boundary value problem in several variables is banded and moreover most of the sub-diagonals in the band have zero coefficients. However, L and U do not have this property and we may as well treat A as if all the diagonals in the band have non-zero entries.

Banded matrices are one example of the class of *sparse* matrices. Recall that a sparse matrix is a matrix with mostly zero entries. As for banded matrices, it is possible to take advantage of sparsity to reduce the cost of factoring A in terms of time and storage. However, it is more difficult to do this than for banded matrices if the sparsity pattern puts non-zero entries at any location in the matrix. One approach to this problem is based on rearranging the equations and variables, or equivalently rearranging the rows and columns to form a banded system.

44.4 Iterative Methods

Instead of solving $Ax = b$ directly, we now consider iterative solution methods based on computing a sequence of approximations $x^{(k)}$, $k = 1, 2, \ldots$, such that

$$\lim_{k\to\infty} x^{(k)} = x \quad \text{or} \quad \lim_{k\to\infty} \|x^{(k)} - x\| = 0,$$

for some norm $\|\cdot\|$.

Note that the finite precision of a computer has a different effect on an iterative method than it has on a direct method. A theoretically convergent sequence can not reach its limit in general on a computer using a finite number of digits. In fact, at the point at which the change from one iterate to the next occurs outside the range of digits held by the computer, the sequence simply stops changing. Practically speaking, there is no point computing iterations past this point, even if the limit has not been reached. On the other hand, it is often sufficient to have less accuracy than the limit of machine precision, and thus it is important to be able to estimate the accuracy of the current iterate.

Minimization Algorithms

We first construct iterative methods for a linear system $Ax = b$ where A is symmetric and positive-definite. In this case, the solution x can be characterized equivalently as the solution of the quadratic minimization problem: find $x \in \mathbb{R}^n$ such that

$$F(x) \leq F(y) \quad \text{for all } y \in R^n, \tag{44.2}$$

where

$$F(y) = \frac{1}{2}(Ay, y) - (b, y),$$

with $(\cdot, \cdot)$ denoting the usual Euclidean scalar product.

We construct an iterative method for the solution of the minimization problem (44.2) based on the following simple idea: given an approximation $x^{(k)}$, compute a new approximation $x^{(k+1)}$ such that $F(x^{(k+1)}) < F(x^{(k)})$. On one hand, since F is a quadratic function, there must be a "downhill" direction from the current position, unless we are at the minimum. On the other hand, we hope that computing the iterates so that their function values are strictly decreasing, will force the sequence to converge to the minimum point x. Such an iterative method is called a *minimization method*.

Writing $x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$, where $d^{(k)}$ is a *search direction* and α_k is a *step length*, by direct computation we get

$$F(x^{(k+1)}) = F(x^{(k)}) + \alpha_k \big(Ax^{(k)} - b, d^{(k)}\big) + \frac{\alpha_k^2}{2}\big(Ad^{(k)}, d^{(k)}\big),$$

where we used the symmetry of A to write $(Ax^{(k)}, d^{(k)}) = (x^{(k)}, Ad^{(k)})$. If the step length is so small that the second order term in α_k can be neglected, then the direction $d^{(k)}$ in which F decreases most rapidly, or the direction of *steepest descent*, is

$$d^{(k)} = -(Ax^{(k)} - b) = -r^{(k)},$$

which is the opposite direction to the residual error $r^{(k)} = Ax^{(k)} - b$. This suggests using an iterative method of the form

$$x^{(k+1)} = x^{(k)} - \alpha_k r^{(k)}. \tag{44.3}$$

A minimization method with this choice of search direction is called a *steepest descent method*. The direction of steepest descent is perpendicular to the *level curve* of F through $x^{(k)}$, which is the curve in the graph of F generated by the points where F has the same value as at $x^{(k)}$. We illustrate this in Fig. 44.10.

It remains to choose the step lengths α_k. Staying with the underlying principle, we choose α_k to give the minimum value of F in the direction

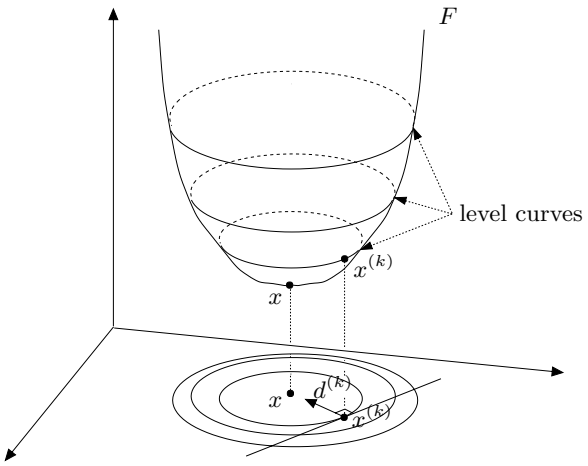

Fig. 44.10. The direction of steepest descent of F at a point is perpendicular to the level curve of F through the point

of $d^{(k)}$ starting from $x^{(k)}$. Differentiating $F(x^{(k)} + \alpha_k r^{(k)})$ with respect to α_k and setting the derivative zero gives

$$\alpha_k = -\frac{\big(r^{(k)}, d^{(k)}\big)}{\big(d^{(k)}, Ad^{(k)}\big)}. \tag{44.4}$$

As a simple illustration, we consider the case

$$A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \quad 0 < \lambda_1 < \lambda_2, \tag{44.5}$$

and $b = 0$, corresponding to the minimization problem

$$\min_{y \in \mathbb{R}^n} \frac{1}{2}\big(\lambda_1 y_1^2 + \lambda_2 y_2^2\big),$$

with solution $x = 0$.

Applying (44.3) to this problem, we iterate according to

$$x^{(k+1)} = x^{(k)} - \alpha_k A x^{(k)},$$

using for simplicity a constant step length with $\alpha_k = \alpha$ instead of (44.4). In Fig. 44.11, we plot the iterations computed with $\lambda_1 = 1$, $\lambda_2 = 9$, and $x^{(0)} = (9, 1)^\top$. The convergence in this case is quite slow. The reason is that if $\lambda_2 \gg \lambda_1$, then the search direction $-(\lambda_1 x_1^{(k)}, \lambda_2 x_2^{(k)})^\top$ and the direction $-(x_1^{(k)}, x_2^{(k)})^\top$ to the solution at the origin, are very different. As a result the iterates swing back and forth across the long, narrow "valley".

It turns out that the rate at which the steepest descent method converges in general depends on the *condition number* $\kappa(A) = \lambda_n/\lambda_1$ of A, where

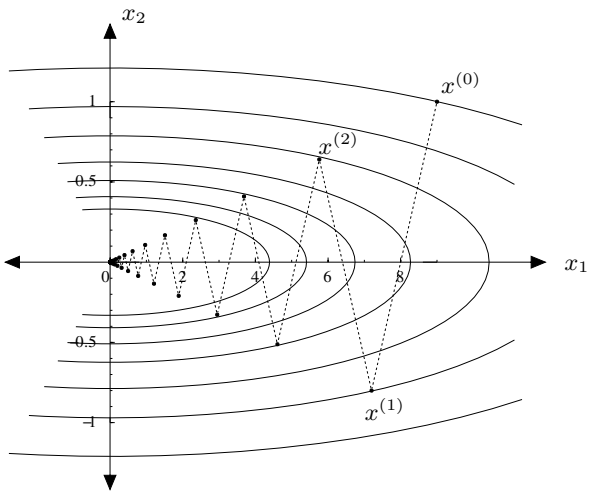

Fig. 44.11. A sequence generated by the steepest descent method for (44.5) plotted together with some level curves of F

$\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ are the eigenvalues of A (counted with multiplicity). In other words, the condition number of a symmetric positive definite matrix is the ratio of the largest eigenvalue to the smallest eigenvalue.

The general definition of the condition number of a matrix A in terms of a norm $\|\cdot\|$ is $\kappa(A) = \|A\|\|A^{-1}\|$. In the $\|\cdot\|_2$ norm, the two definitions are equivalent for symmetric matrices. Using any definition, a matrix is said to be *ill-conditioned* if the $\log(\kappa(A))$ is of the order of the number of digits used in the computer. As we said, we can expect to have difficulty solving an ill-conditioned system; which in terms of direct methods means large errors due to rounding errors and in terms of iterative methods means slow convergence.

We now analyze the steepest descent method for $Ax = b$ in the case of a onstant step length α, where we iterate according to

$$x^{(k+1)} = x^{(k+1)} - \alpha(Ax^{(k)} - b).$$

Since the exact solution x satisfies $x = x - \alpha(Ax - b)$, we get the following equation for the error $e^{(k)} = x - x^{(k)}$:

$$e^{(k+1)} = (I - \alpha A)e^{(k)}.$$

The iterative method converges if the error tend to zero. Taking norms, we get

$$\|e^{(k+1)}\| \leq \mu \|e^{(k)}\| \tag{44.6}$$

where we use the spectral estimate (43.16) to write

$$\mu = \|I - \alpha A\| = \max_j |1 - \alpha\lambda_j|,$$

since the eigenvalues of the matrix $I-\alpha A$ are $1-\alpha\lambda_j$, $j = 1, \ldots, n$. Iterating this estimate we get

$$\|e^{(k+1)}\| \leq \mu^k \, \|e^{(0)}\|, \tag{44.7}$$

where $e^{(0)}$ is the initial error.

To understand when (44.6), or (44.7), guarantees convergence, consider the scalar sequence $\{\lambda^k\}$ for $k \geq 0$. If $|\lambda| < 1$, then $\lambda^k \to 0$; if $\lambda = 1$, then the sequence is always 1; if $\lambda = -1$, the sequence alternates between 1 and -1 and does not converge; and if $|\lambda| > 1$, then the sequence diverges. Therefore if we want the iteration to converge for any initial value, then we must choose α so that $\mu < 1$. Since the λ_j are positive by assumption, $1 - \alpha\lambda_j < 1$ automatically, and we can guarantee that $1 - \alpha\lambda_j > -1$ if α satisfies $\alpha < 2/\lambda_n$. Choosing $\alpha = 1/\lambda_n$, which is not so far from optimal, we get

$$\mu = 1 - 1/\kappa(A).$$

If $\kappa(A)$ is large, then the convergence can be slow because then the reduction factor $1 - 1/\kappa(A)$ is close to one. More precisely, the number of steps required to lower the error by a given amount is proportional to the condition number.

When an iteration converges in this fashion, i.e. the error decreases (more or less) by a given factor in each iteration, then we say that the iteration converges *linearly*. We define the *rate of convergence* to be $-\log(\mu)$. The motivation is that the number of iterations are required to reduce the error by a factor of 10^{-m} is approximately $-m\log(\mu)$. Note that a faster rate of convergence means a smaller value of μ.

This is an *a priori* estimate of the error reduction per iteration, since we estimate the error before the computation. Such an analysis must account for the slowest possible rate of convergence because it holds for all initial vectors.

Consider the system $Ax = 0$ with

$$A = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}, \tag{44.8}$$

where $0 < \lambda_1 < \lambda_2 < \lambda_3$. For an initial guess $x^{(0)}=(x_1^0, x_2^0, x_3^0)^\top$, the steepest descent method with $\alpha = 1/\lambda_3$ gives the sequence

$$x^{(k)} = \left(\left(1 - \frac{\lambda_1}{\lambda_3}\right)^k x_1^0, \left(1 - \frac{\lambda_2}{\lambda_3}\right)^k x_2^0, 0 \right), \quad k = 1, 2, \ldots,$$

and,

$$\|e^{(k)}\| = \sqrt{\left(1 - \frac{\lambda_1}{\lambda_3}\right)^{2k} \left(x_1^0\right)^2 + \left(1 - \frac{\lambda_2}{\lambda_3}\right)^{2k} \left(x_2^0\right)^2}, \quad k = 1, 2, \ldots$$

Thus for a general initial guess, the size of the error is given by the root mean square average of the corresponding iterate and the rate that the errors decrease is the root mean square average of the rates of decrease of the components. Therefore, depending on the initial vector, initially the iterates will generally converge more quickly than the rate of decrease of the first, i.e. slowest, component. In other words, more quickly than the rate predicted by (44.6), which bounds the rate of decrease of the errors by the rate of decrease in the slowest component. However, as the iteration proceeds, the second component eventually becomes much smaller than the first component (as long as $x_1^0 \neq 0$) and we can neglect that term in the expression for the error, i.e.

$$\|e^{(k)}\| \approx \left(1 - \frac{\lambda_1}{\lambda_3}\right)^k |x_1^0| \quad \text{for } k \text{ sufficiently large.} \tag{44.9}$$

In other words, the rate of convergence of the error for almost all initial vectors eventually becomes dominated by the rate of convergence of the slowest component. It is straightforward to show that the number of iterations that we have to wait for this approximation to be valid is determined by the relative sizes of the first and second components of $x^{(0)}$.

This simple error analysis does not apply to the unmodified steepest descent method with varying α_k. However, it is generally true that the rate of convergence depends on the condition number of A, with a larger condition number meaning slower convergence. If we again consider the 2×2 example (44.5) with $\lambda_1 = 1$ and $\lambda_2 = 9$, then the estimate (44.6) for the simplified method suggests that the error should decrease by a factor of $1 - \lambda_1/\lambda_2 \approx .89$ in each iteration. The sequence generated by $x^{(0)} = (9, 1)^\top$ decreases by exactly .8 in each iteration. The simplified analysis overpredicts the rate of convergence for this particular sequence, though not by a lot. By way of comparison, if we choose $x^{(0)} = (1, 1)^\top$, we find that the ratio of successive iterations alternates between $\approx .126$ and $\approx .628$, because α_k oscillates in value, and the sequence converges much more quickly than predicted. On the other hand, there are initial guesses leading to sequences that converge at the predicted rate.

The stiffness matrix A of a linear second order two-point boundary value problem with no convection is symmetric and positive-definite, and its condition number $\kappa(A) \propto h^{-2}$. Therefore the convergence of the steepest descent method is very slow if the number of mesh points is large.

A General Framework for Iterative Methods

We now briefly discuss iterative methods for a general, linear system $Ax = b$, following the classical presentation of iterative methods in Isaacson and Keller ([13]). Recall that some matrices, like diagonal and triangular matrices, are relatively easy and cheap to invert, and Gaussian elimination

can be viewed as a method of factoring A into such matrices. One way to view an iterative method is an attempt to approximate A^{-1} by the inverse of a part of A that is easier to invert. This is called an approximate inverse of A, and we use this to produce an approximate solution to the linear system. Since we don't invert the matrix A, we try to improve the approximate solution by repeating the partial inversion over and over. With this viewpoint, we start by *splitting* A into two parts:

$$A = N - P,$$

where the part N is chosen so that the system $Ny = c$ for some given c is relatively inexpensive to solve. Noting that the true solution x satisfies $Nx = Px + b$, we compute $x^{(k+1)}$ from $x^{(k)}$ by solving

$$Nx^{(k+1)} = Px^{(k)} + b \quad \text{for } k = 1, 2, \ldots, \tag{44.10}$$

where $x^{(0)}$ is an initial guess. For example, we may choose N to be the diagonal of A:

$$N_{ij} = \begin{cases} a_{ij}, & i = j, \\ 0, & i \neq j, \end{cases}$$

or triangular:

$$N_{ij} = \begin{cases} a_{ij}, & i \geq j, \\ 0, & i < j. \end{cases}$$

In both cases, solving the system $Nx^{(k+1)} = Px^{(k)} + b$ is cheap compared to doing a complete Gaussian elimination on A. so we could afford to do it many times.

As an example, suppose that

$$A = \begin{pmatrix} 4 & 1 & 0 \\ 2 & 5 & 1 \\ -1 & 2 & 4 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}, \tag{44.11}$$

and we choose

$$N = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 4 \end{pmatrix} \quad \text{and} \quad P = \begin{pmatrix} 0 & -1 & 0 \\ -2 & 0 & -1 \\ 1 & -2 & 0 \end{pmatrix},$$

in which case the equation $Nx^{(k+1)} = Px^{(k)} + b$ reads

$$\begin{aligned} 4x_1^{k+1} &= -x_2^k + 1 \\ 5x_2^{k+1} &= -2x_1^k - x_3^k \\ 4x_3^{k+1} &= x_1^k - 2x_2^k + 3. \end{aligned}$$

Being a diagonal system it is easily solved, and choosing an initial guess and computing, we get

$$x^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},\ x^{(1)} = \begin{pmatrix} 0 \\ -.6 \\ .5 \end{pmatrix},\ x^{(2)} = \begin{pmatrix} .4 \\ -.1 \\ 1.05 \end{pmatrix},\ x^{(3)} = \begin{pmatrix} .275 \\ -.37 \\ .9 \end{pmatrix},$$

$$x^{(4)} = \begin{pmatrix} .3425 \\ -.29 \\ 1.00375 \end{pmatrix},\ \cdots\ x^{(15)} = \begin{pmatrix} .333330098 \\ -.333330695 \\ .999992952 \end{pmatrix},\ \cdots$$

The iteration appears to converge to the true solution $(1/3, -1/3,1)^\top$.

In general, we could choose $N = N_k$ and $P = P_k$ to vary with each iteration.

To analyze the convergence of (44.10), we subtract (44.10) from the equation $Nx = Px + b$ satisfied by the true solution to get an equation for the error $e^{(k)} = x - x^{(k)}$:

$$e^{(k+1)} = Me^{(k)},$$

where $M = N^{-1}P$ is the *iteration matrix*. Iterating on k gives

$$e^{(k+1)} = M^{k+1}e^{(0)}. \tag{44.12}$$

Rephrasing the question of convergence, we are interested in whether $e^{(k)} \to 0$ as $k \to \infty$. By analogy to the scalar case discussed above, if M is "small", then the errors $e^{(k)}$ should tend to zero. Note that the issue of convergence is independent of the data b.

If $e^{(0)}$ happens to be an eigenvector of M, then it follows from (44.12)

$$\|e^{(k+1)}\| = |\lambda|^{k+1}\|e^{(0)}\|,$$

and we conclude that if the method converges then we must have $|\lambda| < 1$ (or $\lambda = 1$). Conversely, one can show that if $|\lambda| < 1$ for all eigenvalues of M, then the method (44.10) indeed does converge:

Theorem 44.1 *An iterative method converges for all initial vectors if and only if every eigenvalue of the associated iteration matrix is less than one in magnitude.*

This theorem is often expressed using the *spectral radius* $\rho(M)$ of M, which is the maximum of the magnitudes of the eigenvalues of A. An iterative method converges for all initial vectors if and only if $\rho(M) < 1$. In general, the asymptotic limit of the ratio of successive errors computed in $\|\ \|_\infty$ is close to $\rho(M)$ as the number of iterations goes to infinity. We define the *rate of convergence* to be $R_M = -\log(\rho(M))$. The number of iterations required to reduce the error by a factor of 10^m is approximately m/R_M.

Practically speaking, "asymptotic" means that the ratio can vary as the iteration proceeds, especially in the beginning. In previous examples, we

saw that this kind of a priori error result can underestimate the rate of convergence even in the special case when the matrix is symmetric and positive-definite (and therefore has an orthonormal basis of eigenvectors) and the iterative method uses the steepest descent direction. The general case now considered is more complicated, because interactions may occur in direction as well as magnitude, and a spectral radius estimate may overestimate the rate of convergence initially. As an example, consider the non-symmetric (even non-normal) matrix

$$A = \begin{pmatrix} 2 & -100 \\ 0 & 4 \end{pmatrix} \tag{44.13}$$

choosing

$$N = \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix} \text{ and } P = \begin{pmatrix} 8 & 100 \\ 0 & 6 \end{pmatrix} \text{ gives } M = \begin{pmatrix} .9 & 10 \\ 0 & .8 \end{pmatrix}.$$

In this case, $\rho(M) = .9$ and we expect the iteration to converge. Indeed it does converge, but the errors become quite large before they start to approach zero. We plot the iterations starting from $x^{(0)} = (1, 1)^\top$ in Fig. 44.12.

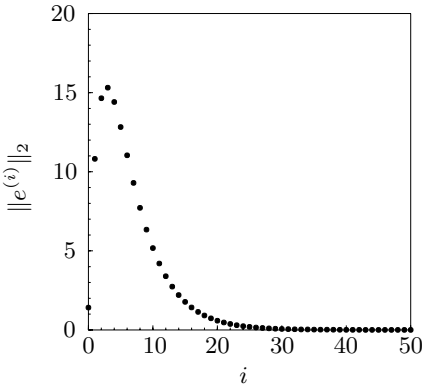

Fig. 44.12. The results of an iterative method computed using a non-normal matrix

The goal is obviously to choose an iterative method so that the spectral radius of the iteration matrix is small. Unfortunately, computing $\rho(M)$ in the general case is much more expensive than solving the original linear system and is impractical in general. We recall that $|\lambda| \leq \|A\|$ holds for any norm and any eigenvalue λ of A. The following theorem indicates a practical way to check for convergence.

Theorem 44.2 *Assume that* $\|N^{-1}P\| \leq \mu$ *for some constant* $\mu < 1$ *and matrix norm* $\|\cdot\|$. *Then the iteration converges and* $\|e^{(k)}\| \leq \mu^k \|e^{(0)}\|$ *for* $k \geq 0$.

This theorem is also an a priori convergence result and suffers from the same deficiency as the analysis of the simplified steepest descent method presented above. In fact, choosing an easily computable matrix norm, like $\| \ \|_\infty$, generally leads to an even more inaccurate estimate of the convergence rate than would be obtained by using the spectral radius. In the worst case, it is entirely possible that $\rho(M) < 1 < \|M\|$ for the chosen norm, and hence the iterative method converges even though the theorem does not apply. The amount of "slack" in the bound in Theorem 44.2 depends on how much larger $\|A\|_\infty$ is than $\rho(A)$.

For the 3×3 example (44.11), we compute $\|N^{-1}P\|_\infty = 3/4 = \lambda$ and therefore we know the sequence converges. The theorem predicts that the error will get reduced by a factor of $3/4$ every iteration. If we examine the error of each iterate along with the ratios of successive errors after the first iteration:

i	$\|e^{(i)}\|_\infty$	$\|e^{(i)}\|_\infty / \|e^{(i-1)}\|_\infty$
0	1.333	
1	.5	.375
2	.233	.467
3	.1	.429
4	.0433	.433
5	.0194	.447
6	.00821	.424
7	.00383	.466
8	.00159	.414
9	.000772	.487

we find that after the first few iterations, the errors get reduced by a factor in the range of .4–.5 each iteration and not the factor $3/4$ predicted above. The ratio of $e^{(40)}/e^{(39)}$ is approximately .469. If we compute the eigenvalues of M, we find that $\rho(M) \approx .476$ which is close to the ratio of successive errors. To decrease the initial error by a factor of 10^{-4} using the predicted decrease of .75 per iteration, we would compute 33 iterations, while only 13 iterations are actually needed.

We get different methods, and different rates of convergence, by choosing different N and P. The method used in the example above is called the *Jacobi* method. In general, this consists of choosing N to be the "diagonal part" of A and P to be the negative of the "off-diagonal" part of A. This gives the set of equations

$$x_i^{k+1} = -\frac{1}{a_{ii}} \left(\sum_{j \neq i} a_{ij} x_j^k - b_i \right), \quad i = 1, \ldots, n.$$

To derive a more sophisticated method, we write out these equations in Fig. 44.13. The idea behind the *Gauss-Seidel* method is to use the new values of the approximation in these equations as they become known. The substitutions are drawn in Fig. 44.13. Presumably, the new values are more

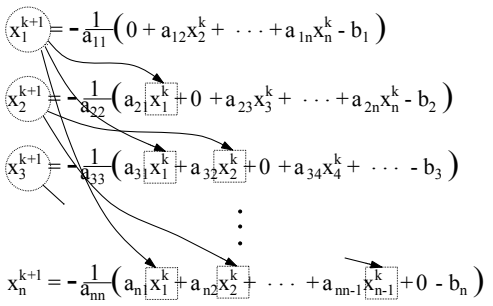

Fig. 44.13. The Gauss-Seidel method substitutes new values of the iteration as they become available

accurate than the old values, hence we might guess that this iteration will converge more quickly. The equations can be written

$$x_i^{k+1} = \frac{1}{a_{ii}}\left(-\sum_{j=1}^{i-1} a_{ij}x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij}x_j^k + b_i\right).$$

If we decompose A into the sum of its lower triangular L, diagonal D, and upper triangular U parts, $A = L+D+U$, then the equations can be written $Dx^{(k+1)} = -Lx^{(k+1)} - Ux^{(k)} + b$ or

$$(D + L)x^{(k+1)} = -Ux^{(k)} + b.$$

Therefore, $N = D + L$ and $P = -U$. The iteration matrix is $M_{\mathrm{GS}} = N^{-1}P = -(D + L)^{-1}U$.

A diagonally dominant matrix often occurs when a parabolic problem is discretized. We have already seen the other case, if A is symmetric and positive-definite then the Gauss-Seidel method converges. This is quite hard to prove, see Isaacson and Keller ([13]) for details.

44.5 Estimating the Error of the Solution

The issue of estimating the error of the numerical solution of a linear system $Ax = b$ arises both in Gaussian elimination, because of the cumulative effects of round-off errors, and when using iterative methods, where we need a stopping criterion. Therefore it is important to be able to estimate the error in some norm with a fair degree of accuracy.

We discussed this problem in the context of iterative methods in the last section when we analyzed the convergence of iterative methods and Theorem 44.2 gives an *a priori* estimate for the convergence rate. It is an a priori estimate because the error is bounded before the computation

begins. Unfortunately, as we saw, the estimate may not be very accurate on a particular computation, and it also requires the size of the initial error. In this section, we describe a technique of *a posteriori* error estimation that uses the approximation after it is computed to give an estimate of the error of that particular approximation.

We assume that x_c is a numerical solution of the system $Ax = b$ with exact solution x, and we want to estimate the error $\|x - x_c\|$ in some norm $\|\cdot\|$. We should point out that we are actually comparing the approximate solution $\tilde{x}_c$ of $\tilde{A}\tilde{x} = \tilde{b}$ to the true solution $\tilde{x}$, where $\tilde{A}$ and $\tilde{b}$ are the finite precision computer representations of the true A and b respectively. The best we can hope to do is compute $\tilde{x}$ accurately. To construct a complete picture, it would be necessary to examine the effects of small errors in A and b on the solution x. To simplify things, we ignore this part of the analysis and drop the $\tilde{}$. In a typical use of an iterative method, this turns out to be reasonable. It is apparently less reasonable in the analysis of a direct method, since the errors arising in direct methods are due to the finite precision. However, the initial error caused by storing A and b on a computer with a finite number of digits occurs only once, while the errors in the arithmetic operations involved in Gaussian elimination occur many times, so even in that case it is not an unreasonable simplification.

We start by considering the *residual error*

$$r = Ax_c - b,$$

which measures how well x_c solves the exact equation. Of course, the residual error of the exact solution x is zero but the residual error of x_c is not zero unless $x_c = x$ by some miracle. We now seek to estimate the unknown error $e = x - x_c$ in terms of the computable residual error r.

By subtracting $Ax - b = 0$ from $Ax_c - b = r$, we get an equation relating the error to the residual error:

$$Ae = -r. \tag{44.14}$$

This is an equation of the same from as the original equation and by solving it numerically by the same method used to compute x_c, we get an approximation of the error e. This simple idea will be used in a more sophisticated form below in the context of a posteriori error estimates for Galerkin methods.

We now illustrate this technique on the linear system arising in the Galerkin finite element discretization of a two-point boundary value problem with no convection. We generate a problem with a known solution so that we can compute the error and test the accuracy of the error estimate. We choose the true solution vector x with components $x_i = \sin(\pi i h)$, where $h = 1/(M+1)$, corresponding to the function $\sin(\pi x)$ and then compute the data by $b = Ax$, where A is the stiffness matrix. We use the Jacobi method, suitably modified to take advantage of the fact that A is tridiagonal, to solve the linear system. We use $\|\ \| = \|\ \|_2$ to measure the error.

We compute the Jacobi iteration until the residual error becomes smaller than a given *residual tolerance* RESTOL. In other words, we compute the residual $r^{(k)} = Ax^{(k)} - b$ after each iteration and stop the process when $\|r^{(k)}\| \leq$ RESTOL. We present computations using the stiffness matrix generated by a uniform discretization with $M = 50$ elements yielding a finite element approximation with an error of .0056 in the l_2 norm. We choose the value of RESTOL so that the error in the computation of the coefficients of the finite element approximation is about 1% of the error of the approximation itself. This is reasonable since computing the coefficients of the approximation more accurately would not significantly increase the overall accuracy of the approximation. After the computation of $x^{(k)}$ is complete, we use the Jacobi method to approximate the solution of (44.14) and compute the estimate of the error.

Using the initial vector $x^{(0)}$ with all entries equal to one, we compute 6063 Jacobi iterations to achieve $\|r\| <$ RESTOL $= .0005$. The actual error of $x^{(6063)}$, computed using the exact solution, is approximately .0000506233. We solve (44.14) using the Jacobi method for 6063 iterations, reporting the value of the error estimate every 400 iterations:

Iter.	Error Est.	Iter.	Error Est.	Iter.	Error Est.
1	0.00049862	2001	0.000060676	4001	0.000050849
401	0.00026027	2401	0.000055328	4401	0.000050729
801	0.00014873	2801	0.000052825	4801	0.000050673
1201	0.000096531	3201	0.000051653	5201	0.000050646
1601	0.000072106	3601	0.000051105	5601	0.000050634

We see that the error estimate is quite accurate after 6001 iterations and sufficiently accurate for most purposes after 2000 iterations. In general, we do not require as much accuracy in the error estimate as we do in the solution of the system, so the estimation of the accuracy of the approximate solution is cheaper than the computation of the solution.

Since we estimate the error of the computed solution of the linear system, we can stop the Jacobi iteration once the error in the coefficients of the finite element approximation is sufficiently small so that we are sure the accuracy of the approximation will not be affected. This is a reasonable strategy given an estimate of the error. If we do not estimate the error, then the best strategy to guarantee that the approximation accuracy is not affected by the solution error is to compute the Jacobi iteration until the residual error is on the order of roughly 10^{-p}, where p is the number of digits that the computer uses. Certainly, there is not much point to computing further Jacobi iterations after this. If we assume that the computations are made in single precision, then $p \approx 8$. It takes a total of 11672 Jacobi iterations to achieve this level of residual error using the same initial guess as above. In fact, estimating the error and computing the coefficients of the approximation to a reasonable level of accuracy costs significantly less than this crude approach.

This approach can also be used to estimate the error of a solution computed by a direct method, provided the effects of finite precision are included. The added difficulty is that in general the residual error of a solution of a linear system computed with a direct method is small, even if the solution is inaccurate. Therefore, care has to be taken when computing the residual error because the possibility that subtractive cancellation makes the calculation of the residual error itself inaccurate. *Subtractive cancellation* is the name for the fact that the difference of two numbers that agree to the first i places has i leading zeroes. If only the first p digits of the numbers are accurate then their difference can have at most $p-i$ accurate significant digits. This can have severe consequences on the accuracy of the residual error if Ax_c and b agree to most of the digits used by the computer. One way to avoid this trouble is to compute the approximation in single precision and the residual in double precision (which means compute the product Ax_c in double precision, then subtract b). The actual solution of (44.14) is relatively cheap since the factorization of A has already been performed and only forward/backward substitution needs to be done.

44.6 The Conjugate Gradient Method

We learned above that solving an $n \times n$ linear system of equations $Ax = b$ with A symmetric positive definite using the gradient method, requires a number of iterations, which is proportional to the condition number $\kappa(A) = \lambda_n/\lambda_1$, where $\lambda_1 \leq \ldots \leq \lambda_n$ are the eigenvalues of A. Thus the number of iteration will be large, maybe prohibitively so, if the condition number $\kappa(A)$ is large.

We shall now present a variant of the gradient method, referred as the *conjugate gradient method*, where the number of iterations scales instead like $\sqrt{\kappa(A)}$, which may be much smaller than $\kappa(A)$ if $\kappa(A)$ is large.

In the conjugate gradient method each new search direction is chosen to be orthogonal, with respect to the scalar product induced by the positive definite symmetric matrix A, which prevents choosing inefficient search directions as in the usual gradient method.

The conjugate gradient method may be formulated as follows: for $k = 1, 2, \ldots$ compute an approximate solution $x^k \in \mathbb{R}^n$ as the solution of the minimization problem

$$\min_{y \in K_k(A)} F(y) = \min_{y \in K_k(A)} \frac{1}{2}(Ay, y) - (b, y)$$

where $K_k(A)$ is the *Krylov space* spanned by the vectors $\{b, Ab, \ldots, A^{k-1}b\}$.

This is the same as defining x^k to be the projection of x onto $K_k(A)$ with respect to the scalar product $\langle y, z\rangle$ on $\mathbb{R}^n \times \mathbb{R}^n$ defined by $\langle y, z\rangle = (Ay, z)$,

because we have using the symmetry of A and that $Ax = b$:

$$\frac{1}{2}(Ay, y) - (b, y) = \frac{1}{2}\langle y - x, y - x\rangle - \frac{1}{2}\langle x, x\rangle.$$

In particular, the conjugate gradient method has the following minimization property

$$\|x - x^k\|_A = \min_{y \in K_k(A)} \|x - y\|_A \leq \|p_k(A)x\|_A$$

where $p_k(x)$ is a polynomial of degree k with $p(0) = 1$, and $\|\cdot\|_A$ is the norm associated with the scalar product $\langle\cdot,\cdot\rangle$, that is, $\|y\|_A^2 = \langle y, y\rangle$. This follows by using that since $b = Ax$, we have that $K_k(A)$ is spanned by the vectors $\{Ax, A^2x, \ldots, A^kx\}$. In particular, we conclude that for all polynomials $p_k(x)$ of degree k such that $p_k(0) = 1$, we have

$$\|x - x^k\|_A \leq \max_{\lambda \in \Lambda} |p_k(\lambda)| \|x\|_A \tag{44.15}$$

where Λ is the set of eigenvalues of A. By choosing the polynomial $p_k(x)$ properly, e.g as a so-called *Chebyshev polynomial* $q_k(x)$ with the property that $q_k(x)$ is small on the interval $[\lambda_1, \lambda_n]$ containing the eigenvalues of A, one can prove that the number of iterations scales like $\sqrt{\kappa(A)}$ if n is large.

If n is not large, we have in particular from (44.15) that we get the exact solution after at most n iterations, since we may choose the polynomial $p_k(x)$ to be zero at the n eigenvalues of A.

We have now defined the conjugate gradient method through it structural properties: projection onto a Krylov space with respect to a certain scalar product, and we now address the problem of actually computing the sequence x^k step by step. This is done as follows: For $k = 0, 1, 2, \ldots,$

$$x^{k+1} = x^k + \alpha_k d^k, \quad \alpha_k = -\frac{(r^k, d^k)}{\langle d^k, d^k\rangle}, \tag{44.16}$$

$$d^{k+1} = -r^{k+1} + \beta_k d^k, \quad \beta_k = \frac{\langle r^{k+1}, d^k\rangle}{\langle d^k, d^k\rangle}, \tag{44.17}$$

where $r^k = Ax^k - b$ is the residual of the approximation x^k, and we choose $x^0 = 0$ and $d_0 = b$. Here, (44.17) signifies that the new search direction d^{k+1} gets new directional information from the new residual r^{k+1} and is chosen to be orthogonal (with respect to the scalar product $\langle\cdot,\cdot\rangle$) to the old search direction d^k. Further, (44.16), expresses that x^{k+1} is chosen so as to to minimize $F(x^{(k)} + \alpha d^k)$ in α, corresponding to projection onto $K_{k+1}(A)$. We prove these properties in a sequence of problems below.

Note that if we choose the initial approximation x^0 different from zero, then we may reduce to the above case by considering instead the problem $Ay = b - Ax^0$ in y, where $y = x - x^0$.

44.7 GMRES

The conjugate gradient method for solving an $n \times n$ system $Ax = b$ builds on the matrix A being symmetric and positive definite. If A is non-symmetric or non-positive definite, but yet non-singular, then we may apply the conjugate gradient method to the least squares problem $A^\top Ax = A^\top b$, but the since the condition number of $A^\top A$ typically is the square of the condition number of A, the required number of iterations may be too large for efficiency.

Instead we may try the *Generalized Minimum Residual* method referred to as *GMRES*, which generates a sequence of approximations x^k of the solution x of $Ax = b$, satisfying for any polynomial $p_k(x)$ of degree at most k with $p_k(0) = 1$

$$\|Ax^k - b\| = \min_{y \in K_k(A)} \|Ay - b\| \leq \|p_k(A)b\|, \tag{44.18}$$

that is x^k is the element in the Krylov space $K_k(A)$ which minimizes the Euclidean norm of the residual $Ay - b$ with $y \in K_k(A)$. Assuming that the matrix A is *diagonalizable*, there exist a nonsingular matrix V so that $A = VDV^{-1}$, where D is a diagonal matrix with the eigenvalues of A on the diagonal. We then have that

$$\|Ax^k - b\| \leq \kappa(V) \max_{\lambda \in \Lambda} |p_k(\lambda)| \|b\|, \tag{44.19}$$

where Λ is the set of eigenvalues of A.

In the actual implementation of GMRES we use the *Arnoldi iteration*, a variant of the Gram-Schmidt orthogonalization, that constructs a sequence of matrices Q_k whose orthogonal column vectors span the successive Krylov spaces $K_k(A)$, and we write $x^k = Q_k c$ to get the following least squares problem:

$$\min_{c \in \mathbb{R}^k} \|AQ_n c - b\|. \tag{44.20}$$

The Arnoldi iteration is based on the identity $AQ_k = Q_{k+1}H_k$, where H_k is an *upper Hessenberg matrix* so that $h_{ij} = 0$ for all $i > j + 1$. Using this identity and multiplying from the left by Q_{k+1}^T gives us another equivalent least squares problem:

$$\min_{c \in \mathbb{R}^k} \|H_k c - Q_{k+1}^T b\|. \tag{44.21}$$

Recalling the construction of the Krylov spaces $K_k(A)$, in particular that $K_1(A)$ is spanned by b, we find that $Q_{k+1}^T b = \|b\| e_1$, where $e_1 = (1, 0, 0, \ldots)$, and we obtain the final form of the least squares problem to be solved in the GMRES iteration:

$$\min_{c \in \mathbb{R}^k} \|H_k c - \|b\| e_1\|. \tag{44.22}$$

This problem is now easy to solve due to the simpel structure of the Hessenberg matrix H_k.

In Fig. 44.14 we compare the performance of the conjugate gradient method and GMRES for system with a tridiagonal 200×200 matrix with 1 on the diagonal, and random off-diagonal entries that take values in $(-0.5, 0.5)$ and the right hand side a random vector with values in $[-1, 1]$. The system matrix in this case is not symmetric, but it is strictly diagonally dominant and thus may be viewed asa perturbation of the identity matrix and should be easy to solve iteratively. We see that both the conjugate gradient method and GMRES converge quite rapidly, with GMRES winning in number of iterations.

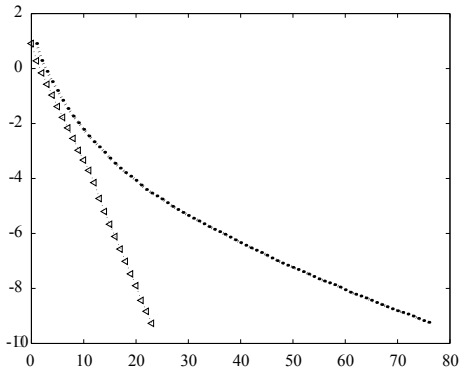

Fig. 44.14. Log-plot of the residual versus the number of iterations for diagonal dominant random matrix, using the conjugate gradient method ($'\cdot'$) and GMRES ('triangles')

In GMRES we need to store the basis vectors for the increasing Krylov space, which may be prohibitive for large systems requiring many iterations. To avoid this problem, we may restart GMRES when we have reached a maximal number of stored basis vector, by using as initial approximation x^0 the last approximation before restart. The trade-off is of course that a retarted GMRES may require more iterations for the same accuracy than GMRES without restart.

We now consider the more challenging problem of solving a 200×200 *stiffness matrix* system, that is a system with a tridiagonal matrix with 2 on the diagonal, and -1 on the off-diagonal (which is not strictly diagonally dominant). We will meet this type of system matrix in Chapter FEM for Two-Point Boundary Value Problems below, and we will see that it has a condition number proportional to the square of the number of unknowns. We thus expect the conjugate gradient method to require about the same number of iterations as the number of unknowns. In Fig. 44.15 we compare again the performance of the conjugate gradient method with the GMRES

method, now restarted after 100 iterations. We find that the conjugate gradient method as expected converges quite slowly (and non monotonically), until immediate convergence at iteration 200 as predicted by theory. The GMRES iteration on the other hand has a monotone but still quite slow convergence in particular after each restart when the Krylov subspace is small.

In Fig. 44.16 we compare different restart conditions for GMRES, and we find that there is a trade-off between the convergence rate and the memory consumption: few restarts give a faster convergence, but require more memory to store more basis vectors for the Krylov space. On the other hand we save memory by using more restarts, but then the convergence rate deteriorates.

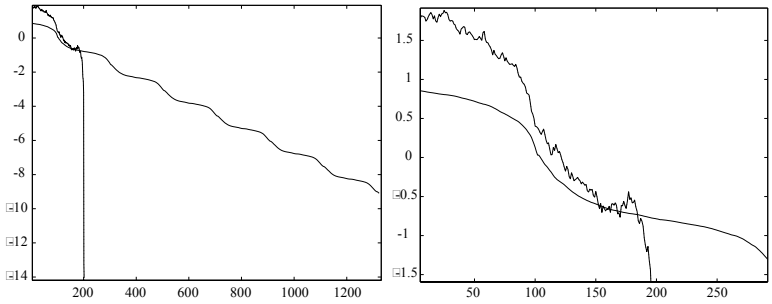

Fig. 44.15. Log-plot of the residual versus the number of iterations for stiffness matrix, using the conjugate gradient method and GMRES, restarted after 100 iterations

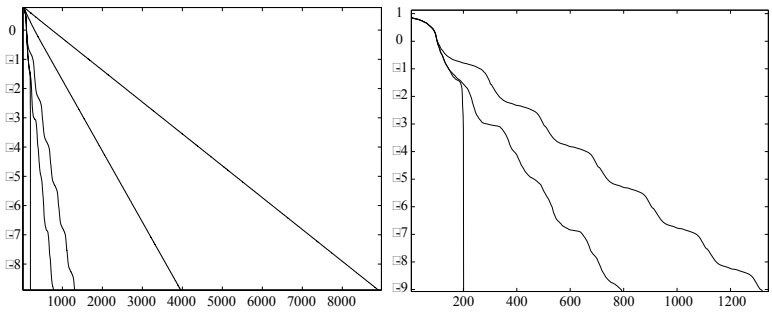

Fig. 44.16. Log-plot of the residual versus the number of iterations for stiffness matrix using GMRES and restarted GMRES, restarted after 20,50,100,150 iterations (left), and a close-up on the cases of no restart and restart after 100 and 150 iterations (right)

Chapter 44 Problems

44.1. Using a similar format, write down algorithms to solve a diagonal system and then a lower triangular system using forward substitution. Determine the number of arithmetic operations needed to compute the solution.

44.2. Prove that multiplying a square matrix A on the left by the matrix in Fig. 44.3 has the effect of adding α_{ij} times row j of A to row i of A. Prove that the inverse of the matrix in Fig. 44.3 is obtained changing α_{ij} to $-\alpha_{ij}$

44.3. Show that the product of two Gauss transformations is a lower triangular matrix with ones on the diagonal and the inverse of a Gauss transformation is a Gauss transformation.

44.4. Solve the system

$$\begin{aligned} x_1 - x_2 - 3x_3 &= 3 \\ -x_1 + 2x_2 + 4x_3 &= -5 \\ x_1 + x_2 &= -2 \end{aligned}$$

by computing an LU factorization of the coefficient matrix and using forward/backward substitution.

44.5. On some computers, dividing two numbers is up to ten times more expensive than computing the reciprocal of the denominator and multiplying the result with the numerator. Alter this code to avoid divisions. Note: the reciprocal of the diagonal element a_{kk} has to be computed just once.

44.6. Write some pseudo-code that uses the matrix generated by the code in Fig. 44.4 to solve the linear system $Ax = b$ using forward/backward substitution. Hint: the only missing entries of L are the 1s on the diagonal.

44.7. Show that the cost of a backward substitution using an upper triangular matrix of dimension $n \times n$ is $O(n^2/2)$.

44.8. Determine the cost of multiplying a $n \times n$ matrix with another.

44.9. One way to compute the inverse of a matrix is based on viewing the equation $AA^{-1} = I$ as a set of linear equations for the columns of A^{-1}. If $a^{(j)}$ denotes the j^{th} column of A^{-1}, then it satisfies the linear system

$$Aa^{(j)} = e_j$$

where e_j is the standard basis vector of $\mathbb{R}^n$ with a one in the j^{th} position. Use this idea to write a pseudo-code for computing the inverse of a matrix using LU factorization and forward/backward substitution. Note that it suffices to compute the LU factorization only once. Show that the cost of computing the inverse in this fashion is $O(4n^3/3)$.

44.10. Solve the system

$$\begin{aligned} x_1 + x_2 + x_3 &= 2 \\ x_1 + x_2 + 3x_3 &= 5 \\ -x_1 - 2x_3 &= -1. \end{aligned}$$

This requires pivoting.

44.11. Alter the LU decomposition and forward/backward routines to solve a linear system with pivoting.

44.12. Modify the code in Problem 44.11 to use partial pivoting.

44.13. Count the cost of Cholesky's method.

44.14. Compute the Cholesky factorization of

$$\begin{pmatrix} 4 & 2 & 1 \\ 2 & 3 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

44.15. Show that the operations count for solving a tridiagonal system using the solver described in Fig. 44.9 is $O(5n)$.

44.16. Find an algorithm to solve a tridiagonal system that stores only four vectors of dimension n.

44.17. A factorization of a tridiagonal solver can be derived as a compact method. Assume that A can be factored as

$$A = \begin{pmatrix} \alpha_1 & 0 & & \cdots & & 0 \\ \beta_2 & \alpha_2 & 0 & & & \vdots \\ 0 & \beta_3 & \alpha_3 & & & \\ \vdots & & & \ddots & 0 & \\ 0 & \cdots & 0 & \beta_n & \alpha_n & \end{pmatrix} \begin{pmatrix} 1 & \gamma_1 & 0 & \cdots & & 0 \\ 0 & 1 & \gamma_2 & 0 & & \\ \vdots & & \ddots & \ddots & & \\ & & & 1 & \gamma_{n-1} & \\ 0 & \cdots & & 0 & 1 & \end{pmatrix}$$

Multiply out the factors and equate the coefficients to get equations for α, β, and γ. Derive some code based on these formulas.

44.18. Write some code to solve the tridiagonal system resulting from the Galerkin finite element discretization of a two-point boundary value problem. Using 50 elements, compare the time it takes to solve the system with this tridiagonal solver to the time using a full LU decomposition routine.

44.19. Show that the operations count of a banded solver for a $n \times n$ matrix with bandwidth d is $O(nd^2/2)$.

44.20. Write code to solve a linear system with bandwidth five centered around the main diagonal. What is the operations count for your code?

44.21. Prove that the solution of (44.2) is also the solution of $Ax = b$.

44.22. Prove that the direction of steepest descent for a function F at a point is perpendicular to the level curve of F through the same point.

44.23. Prove (44.4).

44.24. Prove that the level curves of F in the case of (44.5) are ellipses with major and minor axes proportional to $1/\sqrt{\lambda_1}$ and $1/\sqrt{\lambda_2}$, respectively.

44.25. Compute the iteration corresponding to $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = 3$, and $x^{(0)} = (1, 1, 1)^\top$ for the system $Ax = 0$ with A defined in (44.8). Make a plot of the ratios of successive errors versus the iteration number. Do the ratios converge to the ratio predicted by the error analysis?

44.26. Prove that the estimate (44.9) generalizes to any symmetric positive-definite matrix A, diagonal or not. Hint: use the fact that there is a set of eigenvectors of A that form an orthonormal basis for $\mathbb{R}^n$ and write the initial vector in terms of this basis. Compute a formula for the iterates and then the error.

44.27. (a) Compute the steepest descent iterations for (44.5) corresponding to $x^{(0)} = (9, 1)^\top$ and $x^{(0)} = (1, 1)^\top$, and compare the rates of convergence. Try to make a plot like Fig. 44.11 for each. Try to explain the different rates of convergence.

(b) Find an initial guess which produces a sequence that decreases at the rate predicted by the simplified error analysis.

44.28. Prove that the method of steepest descent corresponds to choosing

$$N = N_k = \frac{1}{\alpha_k} I, \text{ and } P = P_k = \frac{1}{\alpha_k} I - A,$$

with suitable α_k in the general iterative solution algorithm.

44.29. Compute the eigenvalues and eigenvectors of the matrix A in (44.13) and show that A is not normal.

44.30. Prove that the matrix $\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$ is normal.

44.31. Prove Theorem 44.2.

44.32. Compute 10 Jacobi iterations using the A and b in (44.11) and the initial guess $x^{(0)} = (-1,\ 1,\ -1)^\top$. Compute the errors and the ratios of successive errors and compare to the results above.

44.33. Repeat Problem 44.32 using

$$A = \begin{pmatrix} 4 & 1 & 100 \\ 2 & 5 & 1 \\ -1 & 2 & 4 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}.$$

Does Theorem 44.2 apply to this matrix?

44.34. Show that for the Jacobi iteration, $N = D$ and $P = -(L + U)$ and the iteration matrix is $M_{\mathrm{J}} = -D^{-1}(L + U)$

44.35. (a) Solve (44.11) using the Gauss-Seidel method and compare the convergence with that of the Jacobi method. Also compare $\rho(M)$ for the two methods. (b) Do the same for the system in Problem 44.33.

44.36. (Isaacson and Keller ([13])) Analyze the convergence of the Jacobi and Gauss-Seidel methods for the matrix

$$A = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

in terms of the parameter ρ.

In general it is difficult to compare the convergence of the Jacobi method with that of the Gauss-Seidel method. There are matrices for which the Jacobi method converges and the Gauss-Seidel method fails and vice versa. There are two special classes of matrices for which convergence can be established without further computation. A matrix A is *diagonally dominant* if

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad i = 1, \ldots, n.$$

If A is diagonally dominant then the Jacobi method converges.

44.37. Prove this claim.

44.38. Derive an algorithm that uses the Jacobi method to solve a tridiagonal system. Use as few operations and as little storage as possible.

44.39. Devise an algorithm to estimate the error of the solution of a linear system using single and double precision as suggested. Repeat the example using a tridiagonal solver and your algorithm to estimate the error.

44.40. Show that the sequences $\{x^k\}$ and $\{d^k\}$ generated by the conjugate gradient method (44.16)-(44.17), with $x^1 = 0$ and $d^1 = b$, satisfies for $k = 1, 2, \ldots$, (a) $x^k \in K_k(A) = \{b, \ldots, A^{k-1}b\}$, (b) d^{k+1} is orthogonal to $K_k(A)$, (c) x^k is the projection of x onto $K_k(A)$ with respect to the scalar product $\langle y, z \rangle = (Ay, z)$.

44.41. The Cbebyshev polynomial $q_k(x)$ if degree k is defined for $-1 \leq x \leq 1$ by the formula $q_k(x) = \cos(k \arccos(x))$. Show that $q_k'(0) \approx k^2$. Deduce from this result that the number of iterations in the conjugate gradient method scales like $\sqrt{\kappa A)}$ TS^C.

44.42. Compare the GMRES-algorithm for $Ax = b$ with the conjugate gradient method fro the normal equations $A^\top A = A^\top b$.

TS^C Please check this right parenthesis.

Editor's or typesetter's annotations (will be removed before the final TEX run)

44.43. The formula $AQ_k = Q_{k+1}H_k$, with H_k an upper Hessenberg matrix ($h_{ij} = 0$ for all $i > j + 1$), defines a recurrence relation for the column vector q_{k+1} of Q_{k+1} in terms of itself and the previous Krylov vectors. (a) Derive this recurrence relation. (b) Implement an algorithm that computes Q_{k+1} and H_k, given a matrix A (this is the *Arnoldi iteration*).

44.44. Prove that $Q_{k+1}^T b = \|b\|e_1$.

44.45. Implement the GMRES-method.

45
Linear Algebra Tool Bag

45.1 Linear Algebra in $\mathbb{R}^2$

Scalar product of two vectors $a = (a_1, a_2)$ and $b = (b_1, b_2)$ in $\mathbb{R}^2$:

$$a \cdot b = (a, b) = a_1 b_1 + a_2 b_2.$$

Norm: $|a| = (a_1^2 + a_2^2)^{1/2}$.

Angle between two vectors a and b in $\mathbb{R}^2$: $\cos(\theta) = \frac{a \cdot b}{|a|\,|b|}$.
The vectors a and b are orthogonal if and only if $a \cdot b = 0$.

Vector product of two vectors $a = (a_1, a_2)$ and $b = (b_1, b_2)$ in $\mathbb{R}^3$:

$$a \times b = a_1 b_2 - a_2 b_1.$$

Properties of vector product: $|a \times b| = |a||b||\sin(\theta)|$, where θ is the angle between a and b. In particular, a and b are parallel if and only if $a \times b = 0$.

Volume of parallelogram spanned by two vectors $a, b \in \mathbb{R}^2$:

$$V(a, b) = |a \times b| = |a_1 b_2 - a_2 b_1|.$$

45.2 Linear Algebra in $\mathbb{R}^3$

Scalar product of two vectors $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$ in $\mathbb{R}^3$:

$$a \cdot b = \sum_{i=1}^{3} a_i b_i = a_1 b_1 + a_2 b_2 + a_3 b_3.$$

Norm: $|a| = (a_1^2 + a_2^2 + a_3^2)^{1/2}$.

Angle between two vectors a and b in $\mathbb{R}^3$: $\cos(\theta) = \frac{a \cdot b}{|a|\,|b|}$.
The vectors a and b are orthogonal if and only if $a \cdot b = 0$.
Vector product of two vectors $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$ in $\mathbb{R}^3$:

$$a \times b = (a_2 b_3 - a_3 b_2, a_3 b_1 - a_1 b_3, a_1 b_2 - a_2 b_1).$$

Properties of vector product: The vector product $a \times b$ of two non-zero vectors a and b in $\mathbb{R}^3$, is orthogonal to both a and b, and $|a \times b| = |a||b||\sin(\theta)|$, where θ is the angle between a and b. In particular, a and b are parallel if and only if $a \times b = 0$.

Volume of parallelepiped spanned by three vectors $a, b, c \in \mathbb{R}^3$:

$$V(a, b, c) = |c \cdot (a \times b)|.$$

45.3 Linear Algebra in $\mathbb{R}^n$

Definition of $\mathbb{R}^n$: The set of ordered n-tuples, $x = (x_1, \ldots, x_n)$ with components $x_i \in \mathbb{R}$, $i = 1, \ldots, n$.

Vector addition and scalar multiplication: For $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ in $\mathbb{R}^n$ and $\lambda \in \mathbb{R}$, we define

$$x + y = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n), \quad \lambda x = (\lambda x_1, \ldots, \lambda x_n).$$

Scalar product: $x \cdot y = (x, y) = \sum_{i=1}^{n} x_i y_i$. **Norm**: $|x| = (\sum_{i=1}^{n} x_i^2)^{1/2}$.

Cauchy's inequality: $|(x, y)| \leq |x|\,|y|$.

Angle of two vectors x and y in $\mathbb{R}^n$: $\cos(\theta) = \frac{(x,y)}{|x||y|}$.

Standard basis: $\{e_1, \ldots, e_n\}$, where $e_i = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$ with a single coefficient 1 at position i.

Linear independence: A set $\{a_1, \ldots, a_n\}$ of vectors in $\mathbb{R}^m$ is said to be *linearly independent* if none of the vectors a_i can be expressed as a linear combination of the others, that is, if $\sum_{i=1}^{n} \lambda_i a_i = 0$ with $\lambda_i \in \mathbb{R}$ implies that $\lambda_i = 0$ for $i = 1, \ldots, n$.

A basis for $\mathbb{R}^n$ is a linearly independent set of vectors whose linear combinations span $\mathbb{R}^n$. Any basis of $\mathbb{R}^n$ has n elements. Further, a set of n vectors in $\mathbb{R}^n$ span $\mathbb{R}^n$ if and only if it is linearly independent, that is, a set of n vectors in $\mathbb{R}^n$ that spans $\mathbb{R}^n$ or is independent, must be a basis. Also, a set of fewer than n vectors in $\mathbb{R}^n$ cannot span $\mathbb{R}^n$, and a set of more than n vectors in $\mathbb{R}^n$ must be linearly dependent.

45.4 Linear Transformations and Matrices

An $m \times n$ real (or complex) *matrix* $A = (a_{ij})$ is rectangular array with rows $(a_{i1}, \ldots, a_{in})$, $i = 1, \ldots, m$, and columns $(a_{1j}, \ldots, a_{mj})$, $j = 1, \ldots, n$, where $a_{ij} \in \mathbb{R}$ (or $a_{ij} \in \mathbb{C}$).

Matrix addition: Given two $m \times n$ matrices $A = (a_{ij})$ and $B = (b_{ij})$, we define $C = A+B$ as the $m \times n$ matrix $C = (c_{ij})$ with elements $c_{ij} = a_{ij}+b_{ij}$, corresponding to elementwise addition.

Multiplication by scalar Given a $m \times n$ matrix $A = (a_{ij})$ and a real number λ, we define the $m \times n$ matrix λA with elements (λa_{ij}), corresponding to multiplying all elements of A by the real number λ.

Matrix multiplication: Given a $m \times p$ matrix A and a $p \times n$ matrix B we define a $m \times n$ matrix AB with elements $(AB)_{ij} = \sum_{k=1}^{p} a_{ik}b_{kj}$. Matrix multiplication is not commutative, that is, $AB \neq BA$ in general. In particular, BA is defined only if $n = m$.

A linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ can be expressed as $f(x) = Ax$, where $A = (a_{ij})$ is an $m \times n$ matrix with elements $a_{ij} = f_i(e_j) = (e_i, f(e_j))$, where $f(x) = (f_1(x), \ldots, f_m(x))$. If $g : \mathbb{R}^n \to \mathbb{R}^p$ and $f : \mathbb{R}^p \to \mathbb{R}^m$ are two linear transformations with corresponding matrices A and B, then the matrix of $f \circ g : \mathbb{R}^n \to \mathbb{R}^m$ is given by AB.

Transpose: If $A = (a_{ij})$ is a real $m \times n$ matrix, then the transpose $A^\top$ is an $n \times m$ matrix with elements $a_{ji}^\top = a_{ij}$, and $(Ax, y) = (x, A^\top y)$ for all $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$.

Matrix norms:

$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}|, \quad \|A\|_\infty = \max_{i+1,\ldots,m} \sum_{j=1}^{n} |a_{ij}|, \quad \|A\| = \max_{x \in \mathbb{R}^n} \frac{\|Ax\|}{\|x\|}.$$

If $A = (\lambda_i)$ is a diagonal $n \times n$ matrix with diagonal elements $a_{ii} = \lambda_i$, then

$$\|A\| = \max_{i=1,\ldots,n} |\lambda_i|.$$

Lipschitz constant of a linear transformation: The Lipschitz constant of a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ given by a $m \times n$ matrix $A = (a_{ij})$ is equal to $\|A\|$.

45.5 The Determinant and Volume

The **determinant** $\det A$ of an $n \times n$ matrix $A = (a_{ij})$, or the volume $V(a_1, \ldots, a_n)$ spanned by the column vectors of A, is defined by

$$\det A = V(a_1, \ldots, a_n) = \sum_{\pi} \pm a_{\pi(1)\,1} a_{\pi(2)\,2} \cdots a_{\pi(n)\,n},$$

where we sum over all permutations π of the set $\{1, \ldots, n\}$, and the sign indicates if the permutation is even (+) or odd (-). We have $\det A = \det A^\top$.

Volume $V(a_1, a_2)$ in $\mathbb{R}^2$:

$$\det A = V(a_1, a_2) = a_{11}a_{22} - a_{21}a_{12}.$$

Volume $V(a_1, a_2, a_3)$ in $\mathbb{R}^3$:

$$\begin{aligned}\det A &= V(a_1, a_2, a_3) = a_1 \cdot a_2 \times a_3 \\ &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}).\end{aligned}$$

Volume $V(a_1, a_2, a_3, a_4)$ in $\mathbb{R}^4$:

$$\begin{aligned}\det A = V(a_1, a_2, a_3, a_4) &= a_{11}V(\hat{a}_2, \hat{a}_3, \hat{a}_4) - a_{12}V(\hat{a}_1, \hat{a}_3, \hat{a}_4) \\ &\quad + a_{13}V(\hat{a}_1, \hat{a}_2, \hat{a}_4) - a_{14}V(\hat{a}_1, \hat{a}_2, \hat{a}_3),\end{aligned}$$

where the $\hat{a}_j$, $j = 1, 2, 3, 4$ are the 3-column vectors corresponding to cutting out the first coefficient of the a_j.

Determinant of a triangular matrix: If $A = (a_{ij})$ is a *upper triangular* $n \times n$ matrix, that is $a_{ij} = 0$ for $i > j$, then

$$\det A = a_{11}a_{22} \cdots a_{nn}.$$

This formula also applies to a *lower triangular* $n \times n$ matrix $A = (a_{ij})$ with $a_{ij} = 0$ for $i < j$.

The magic formula: $\det AB = \det A \det B$.

Test of linear independence: A set $\{a_1, a_2, \ldots, a_n\}$ of n vectors in $\mathbb{R}^n$ is linearly independent if and only if $V(a_1, \ldots, a_n) \neq 0$. The following statements are equivalent for an $n \times n$ matrix A: (a) The columns of A are linearly independent, (b) If $Ax = 0$, then $x = 0$, (c) $\det A \neq 0$.

45.6 Cramer's Formula

If A is a $n \times n$ non-singular matrix with $\det A \neq 0$, then the system of equations $Ax = b$ has a unique solution $x = (x_1, \ldots, x_n)$ for any $b \in \mathbb{R}^n$. given by

$$x_i = \frac{V(a1, \ldots, a_{i-1}, b, a_{i+1}, \ldots, a_n)}{V(a_1, a_2, \ldots, a_n)}, \quad i = 1, \ldots, n.$$

45.7 Inverse

A nonsingular $n \times n$ matrix A has a inverse matrix A^{-1} satisfying:

$$A^{-1}A = AA^{-1} = I,$$

where I is the $n \times n$ identity matrix.

45.8 Projections

The projection $Pv \in V$ of $v \in \mathbb{R}^n$, where V is a linear subspace of $\mathbb{R}^n$, is uniquely defined defined by $(v - Pv, w) = 0$ for all $w \in V$ and satisfies $|v - Pv| \leq |v - w|$ for all $w \in V$. Further, $PP = P$ and $P^\top = P$.

45.9 The Fundamental Theorem of Linear Algebra

If A is a $m \times n$ matrix with null space $N(A) = \{x \in \mathbb{R}^n : Ax = 0\}$ and range $R(A) = \{y = Ax : x \in \mathbb{R}^n\}$, then

$$N(A) \oplus R(A^\top) = \mathbb{R}^n \quad N(A^\top) \oplus R(A) = \mathbb{R}^m,$$

$$\text{dim } N(A) + \text{dim } R(A^\top) = n, \text{ dim } N(A^\top) + \text{dim } R(A) = m,$$

$$\text{dim } N(A) + \text{dim } R(A) = n, \text{ dim } N(A^\top) + \text{dim } R(A^\top) = m,$$

$$\text{dim } R(A) = \text{dim } R(A^\top),$$

The number of linearly independent columns of A is equal to the number of linearly independent rows of A.

45.10 The QR-Decomposition

An $n \times m$ matrix A can be expressed in the form

$$A = QR,$$

where Q is a $n \times m$ matrix with orthogonal columns and R is a $m \times m$ upper triangular matrix.

45.11 Change of Basis

A linear transformation $f : \mathbb{R}^n \to \mathbb{R}^n$, with matrix A with respect to the standard basis, has the following matrix in a basis $\{s_1, \ldots s_n\}$:

$$S^{-1}AS,$$

where the the coefficients s_{ij} of the matrix $S = (s_{ij})$ are the coordinates of the basis vectors s_j with respect to the standard basis.

45.12 The Least Squares Method

The least squares solution of the linear system $Ax = b$ with A an $m \times n$ matrix minimizing $|Ax - b|^2$ satisfies $A^\top Ax = A^\top b$, and is unique if the columns of A are linearly independent.

45.13 Eigenvalues and Eigenvectors

If A is an $n \times n$ matrix and $x =\in \mathbb{R}^n$ is a non-zero vector which satisfies $Ax = \lambda x$, where λ is a real number, then we say that $x \in \mathbb{R}^n$ is an *eigenvector* of A and that λ is a corresponding *eigenvalue* of A. The number λ is an eigenvalue of the $n \times n$ matrix A if and only if λ is a root of the characteristic equation $\det(A - \lambda I) = 0$.

45.14 The Spectral Theorem

If A is a symmetric $n \times n$ matrix A, then there is an orthonormal basis $\{q_1, \ldots, q_n\}$ of $\mathbb{R}^n$ consisting of eigenvectors q_j of A with corresponding real eigenvalues λ_j, satisfying $Aq_j = \lambda_j q_j$, for $j = 1, \ldots, n$. We have $D = Q^{-1}AQ$ and $A = QDQ^{-1}$, where Q is the orthogonal matrix with the eigenvectors q_j in the standard basis forming the columns, and D is the diagonal matrix with the eigenvalues λ_j on the diagonal. Further, $\|A\| = \max_{i=1,\ldots,n} |\lambda_i|$.

45.15 The Conjugate Gradient Method for $Ax = b$

For $k = 0, 1, 2, \ldots$, with $r^k = Ax^k - b$, $x^0 = 0$ and $d^0 = b$, do

$$x^{k+1} = x^k + \alpha_k d^k, \quad \alpha_k = -\frac{(r^k, d^k)}{(d^k, Ad^k)},$$

$$d^{k+1} = -r^{k+1} + \beta_k d^k, \quad \beta_k = \frac{(r^{k+1}, Ad^k)}{(d^k, Ad^k)}.$$

46
The Matrix Exponential $\exp(xA)$

> I tell them that if they will occupy themselves with the study of mathematics, they will find in it the best remedy against the lusts of the flesh. (Thomas Mann (1875-1955))

An important special case of the general initial value problem (40.1) is the linear system

$$u'(x) = Au(x) \quad \text{for } 0 < x \leq T,\ u(0) = u^0, \tag{46.1}$$

where A is a *constant* $d \times d$ matrix, $u^0 \in \mathbb{R}^d$, $T > 0$ and the solution $u(x) \in \mathbb{R}^d$ is a column vector. By the general existence result of the previous chapter we know that that a unique solution exists. Recalling that the solution of the scalar problem $u' = au$, a constant, is $u = e^{xa}u^0$, we denote the solution of (46.1) by

$$u(x) = \exp(xA)u_0 = e^{xA}u^0. \tag{46.2}$$

This definition can be extended to $x < 0$ in the obvious way.

To make sense of $\exp(xA)$, we may view $\exp(xA) = e^{xA}$ as the $d \times d$ *matrix* with column i denoted by $\exp(xA)_i$ being the solution vector $u(x)$ with initial data $u^0 = e_i$, where the e_i are the standard basis vectors. This means that $\exp(xA)_i = \exp(xA)e_i$. By linearity we can write the solution $u(x)$ with general initial data $u^0 = \sum_{i=1}^d u_i^0 e_i$ in the form

$$u(x) = \exp(xA)u^0 = \sum_{i=1}^d \exp(xA)u_i^0 e_i = \sum_{i=1}^d u_i^0 \exp(xA)_i.$$

Example 46.1. If A is diagonal with diagonal elements d_i, then $exp(xA)$[TS d] is a diagonal matrix with diagonal elements $\exp(xd_i)$.

We may express the basic property of the matrix exponential $\exp(xA)$ as follows:

$$\frac{d}{dx}\exp(xA) = A\exp(xA) = Ae^{xA}, \quad \text{for } x \in \mathbb{R} \tag{46.3}$$

We also note the following related basic property, which generalizes a familiar property of the usual exponential (for a proof, see Problem 46.1):

$$\exp(xA)\exp(yA) = \exp((x+y)A) \quad \text{for } x, y \in \mathbb{R}. \tag{46.4}$$

46.1 Computation of $\exp(xA)$ when A Is Diagonalizable

We have defined $\exp(xA)$ through the solution of the initial value problem (46.1), but we do not yet have an analytical formula for $\exp(xA)$, except in the "trivial" case with A diagonal. It turns out that we can find a formula in the case with A diagonalizable. This formula helps to give an idea of the structure of the matrix function $\exp(xA)$ in terms of the eigenvalues and eigenvectors of A, and in particular gives a chance of identifying cases when $\exp(xA)$ is exponentially decaying as x increases.

We consider the system (46.1) with the matrix A diagonalizable so that there is a nonsingular matrix S such that $S^{-1}AS = D$ or $A = SDS^{-1}$, where D is a diagonal matrix with diagonal elements d_i (the eigenvalues of A), and the columns of S corresponding eigenvectors, see the Chapter The Spectral Theorem. If A is symmetric then S may be chosen to be orthogonal with $S^{-1} = S^\top$. Introducing the new dependent variable $v = S^{-1}u$ so that $u = Sv$, we can rewrite $\dot{u} = Au$ in the form $\dot{v} = S^{-1}ASv = Dv$, and letting $\exp(xD)$ be the diagonal matrix with diagonal elements equal to $\exp(xd_i)$, we have $v(x) = \exp(xD)v(0)$, which we can write in the form $S^{-1}u(x) = \exp(xD)S^{-1}u^0$, that is

$$u(x) = S\exp(xD)S^{-1}u^0.$$

Since according to the previous section we have also decided to write $u(x) = \exp(xA)u^0$, we conclude that

$$\exp(xA) = S\exp(xD)S^{-1}.$$

This shows that indeed $\exp(xA)$ may be viewed as a matrix, and we also get an analytic formula to compute $\exp(xA)$ without having to directly solve $\dot{u} = Ax$. We note that each element of $\exp(xA)$ is a certain linear combination of terms of the form $\exp(xd_i)$ with d_i and eigenvalue of A. We give some basic examples.

[TS d] Should "exp" be italic here.

Editor's or typesetter's annotations (will be removed before the final TEX run)

Example 46.2. (A symmetric with real eigenvalues) Suppose

$$A = \begin{pmatrix} a & 1 \\ 1 & a \end{pmatrix}.$$

The eigenvalues of A are $d_1 = a - 1$ and $d_2 = a + 1$ and the corresponding matrix S of normalized eigenvectors (which is orthogonal since A is symmetric), is given by

$$S = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad S^{-1} = S^T = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

We compute and find that

$$\exp(xA) = S \exp(xD) S^T = \exp(ax) \begin{pmatrix} \cosh(x) & \sinh(x) \\ \sinh(x) & \cosh(x) \end{pmatrix},$$

and we see that each element of $\exp(xA)$ is a linear combination of $\exp(d_j x)$ with d_j an eigenvalue of A. If $a < -1$, then all elements of $\exp(xA)$ are exponentially decaying.

Example 46.3. (A anti-symmetric with purely imaginary eigenvalues)

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

The eigenvalues of A are purely imaginary, $d_1 = -i$ and $d_2 = i$, and the corresponding matrix S of eigenvectors is given by

$$S = \begin{pmatrix} -i & i \\ 1 & 1 \end{pmatrix}, \quad S^{-1} = \frac{1}{2i} \begin{pmatrix} -1 & i \\ 1 & i \end{pmatrix}$$

We compute and find that

$$\exp(xA) = S \exp(xD) S^{-1} = \begin{pmatrix} \cos(x) & -\sin(x) \\ \sin(x) & \cos(x) \end{pmatrix},$$

and we see again that each element of $\exp(xA)$ is a linear combination of $\exp(d_j x)$ with d_j an eigenvalue of A. In this case the problem $\dot{u} = Au$ takes the form of the scalar linear oscillator $\ddot{v} + v = 0$ with $u_1 = v$ and $u_2 = \dot{v}$.

Example 46.4. (A non-normal) The matrix

$$A = \begin{pmatrix} a & 1 \\ \epsilon^2 & a \end{pmatrix}$$

with $a \in \mathbb{R}$, and ϵ small positive has the eigenvalues $d_1 = a - \epsilon$ and $d_2 = a + \epsilon$, and the corresponding matrix of eigenvectors S is given by

$$S = \begin{pmatrix} 1 & 1 \\ -\epsilon & \epsilon \end{pmatrix}, \quad S^{-1} = \frac{1}{2} \begin{pmatrix} 1 & \epsilon^{-1} \\ 1 & -\epsilon^{-1} \end{pmatrix}.$$

Computing we find that

$$\exp(xA) = S\exp(xD)S^{-1} = \exp(ax)\begin{pmatrix}\cosh(\epsilon x) & \epsilon^{-1}\sinh(\epsilon x)\\ \epsilon\sinh(\epsilon x & \cosh(\epsilon x)\end{pmatrix}.$$

Again we note that each element of exp(xA) is a linear combination of exp($d_j x$) with d_j and eigenvalue. We further that if ϵ is small, then S is nearly singular (the two eigenvectors are almost parallel), and the inverse S^{-1} contains large numbers (ϵ^{-1}).

46.2 Properties of exp(Ax)

If $A = D = (d_j)$ is diagonal with diagonal elements d_i, then the nature of exp(xA) is easy to describe. In this case, exp(xA) is diagonal and if $d_j > 0$, then the corresponding diagonal element exp($d_j x$) is exponentially increasing and if $d_j < 0$, then exp($d_j x$) is exponentially decreasing. If all the d_j are positive (negative), then we may describe exp(xA) as exponentially increasing (decaying). If $d_i = a + ib$ is complex with $b \neq 0$, then exp($d_j x$) = exp(ax) exp(ibx) oscillates with an exponentially increasing or decreasing amplitude depending on the sign of $a > 0$.

If A is diagonalizable with $S^{-1}AS = D$ with $D = (d_j)$ diagonal, then exp(xA) = $S\exp(xD)S^{-1}$, and it follows that the elements of exp(AX) are a certain linear combinations of the exponentials exp($d_j x$). If all the d_j are negative, then all elements of exp(xA) will be exponentially decaying. We will pay particular attention to this case below.

If A is not diagonalizable, then the structure of exp(xA) is more complex: the elements of exp(xA) will then be of the form $p(x)\exp(d_j x)$ with d_j an eigenvalue of A and $p(x)$ a polynomial with degree less than the multiplicity of d_j.

46.3 Duhamel's Formula

We can generalize the previous discussion to the non-homogeneous problem

$$u'(x) = Au(x) + f(x) \quad \text{for } 0 < x \leq 1,\ u(0) = u^0, \tag{46.5}$$

where $f(x)$ is a given function. The solution $u(x)$ can be expressed in the form of a *Duhamel formula* generalizing the formula that holds for a scalar problem,

$$u(x) = \exp(xA)u^0 + \int_0^x \exp((x-y)A)f(y)\,dy. \tag{46.6}$$

This is readily verified by differentiation,

$$\begin{aligned}\dot{u}(x) &= \frac{d}{dx}\exp(xA)u^0 + \frac{d}{dx}\int_0^x \exp((x-y)A)f(y)\,dy \\ &= A\exp(xA)u^0 + \exp((x-x)A)f(x) + \int_0^x A\exp((x-y)A)f(y)\,dy \\ &= Au(x) + f(x).\end{aligned} \tag{46.7}$$

Chapter 46 Problems

46.1. Prove (46.4). Hint: Assuming that $u' = Au$ we may write $u(x+y) = \exp(xA)u(y) = \exp(xA)\exp(yA)u(0)$ and also $u(x+y) = \exp((x+y)A)u(0)$.

46.2. Rewrite the second order scalar constant coefficient problem $\ddot{v} + a_1\dot{v} + a_0 v = 0$ in first order system form $\dot{u} = Au$ by setting $u_1 = v$ and $u_2 = \dot{v}$, and connect the analysis of this chapter to the analysis of the linear oscillator in the Chapter N-body systems. Generalize to higher order scalar problems.

47

Lagrange and the Principle of Least Action*

> Dans les modifications des mouvements, l'action devient ordinairement un Maximum ou un Minimum. (Leibniz)
>
> Whenever any action occurs in nature, the quantity of action employed by this change is the least possible. (Maupertuis 1746)
>
> From my earliest recollection I have had an irresistible liking for mechanics and the physical laws on which mechanics as a science is based. (Reynolds)

47.1 Introduction

Lagrange (1736-1813), see Fig. 47.1, found a way to formulate certain dynamical problems in mechanics using a *Principle of Least Action*. This principle states that the *state* $u(t)$ of a system changes with time t over a given time interval $[t_1, t_2]$, so that the *action integral*

$$I(u) = \int_{t_1}^{t_2} (T(\dot{u}(t)) - V(u(t))\, dt \tag{47.1}$$

is *stationary*, where $T(\dot{u}(t))$ with $\dot{u} = \frac{du}{dt}$ is the *kinetic energy*, and $V(u(t))$ is the *potential energy* of the *state* $u(t)$. We here assume that the state $u(t)$ is a function $u : [t_1, t_2] \to \mathbb{R}$ satisfying $u(t_1) = u_1$ and $u(t_2) = u_2$, where u_1 and u_2 are given initial and final values. For example, $u(t)$ may be the position of a moving mass at time t. The action integral of a state is thus

Fig. 47.1. Lagrange, Inventor of the Principle of Least Action: "I regard as quite useless the reading of large treatises of pure analysis: too large a number of methods pass at once before the eyes. It is in the works of applications that one must study them; one judges their ability there and one apprises the manner of making use of them"

the difference between the kinetic and potential energies integrated in time along the state.

We shall now get acquainted with Lagrange's famous Principle of Least Action and we shall see that it may be interpreted as a reformulation of Newton's law stating that mass times acceleration equals force. To this end, we first need to explain what is meant by the statement that the *action integral is stationary* for the actual solution $u(t)$. Our tool is Calculus, at its best!

Following in the foot-steps of Lagrange, consider a *perturbation* $v(t) = u(t) + \epsilon w(t) = (u + \epsilon w)(t)$ of the state $u(t)$, where $w(t)$ is a function on $[t_1, t_2]$ satisfying $w(t_1) = w(t_2) = 0$ and ϵ is a small parameter. The function $v(t)$ corresponds to changing $u(t)$ with the function $\epsilon w(t)$ inside (t_1, t_2) while keeping the values $v(t_1) = u_1$ and $v(t_2) = u_2$. The Principle of Least Action states that the actual path $u(t)$ has the property that for all such functions $w(t)$, we have

$$\frac{d}{d\epsilon} I(u + \epsilon w) = 0 \quad \text{for } \epsilon = 0. \tag{47.2}$$

The derivative $\frac{d}{d\epsilon} I(u + \epsilon w)$ at $\epsilon = 0$, measures the rate of change with respect to ϵ at $\epsilon = 0$ of the value of the action integral with $u(t)$ replaced by $v(t) = u(t) + \epsilon w(t)$. The Principle of Least Action says this rate of change is zero if u is the actual solution, which expresses the stationarity of the action integral.

We now present a couple of basic applications illustrating the use of the Principle of Least Action.

47.2 A Mass-Spring System

We consider a system of a mass m sliding on a on a horizontal frictionless x-axis and being connected to the origin with a weight-less Hookean spring with spring constant k, see the Chapter Galileo, Newton et al. We know that this system may be described by the equation $m\ddot{u} + ku = 0$, where $u(t)$ is the length of the spring at time t. We derive this model by using the Principle of Least Action. In this case,

$$T(\dot{u}(t)) = \frac{m}{2}\dot{u}^2(t) \quad \text{and} \quad V(u(t)) = \frac{k}{2}u^2(t),$$

and thus

$$I(u) = \int_{t_1}^{t_2} \left(\frac{m}{2}\dot{u}^2(t) - \frac{k}{2}u^2(t) \right) dt.$$

To motivate the expression $V(u(t)) = \frac{k}{2}u^2(t)$ for the potential energy, we use the definition of the potential energy as the total work required to move the mass from position $u = 0$ to position $u(t)$. The work to move the mass from position v to $v + \Delta v$ is equal to $kv\Delta v$ following the principle that work = force $\times$ displacement. The total work is thus

$$V(u(t)) = \int_0^{u(t)} kv\, dv = \frac{k}{2}u^2(t),$$

as announced.

To see how the equation $m\ddot{u}+ku = 0$ arises, we compute the derivative of $I(u+\epsilon w)$ with respect to ϵ and then set $\epsilon = 0$, where $w(x)$ is a perturbation satisfying $w(t_1) = w(t_2) = 0$. Direct computation based on moving $\frac{d}{d\epsilon}$ inside the integral, which is allowed since the limits of integration are fixed,

$$\frac{d}{d\epsilon}I(u + \epsilon w) = \int_{t_1}^{t_2} \frac{d}{d\epsilon}\left(\frac{m}{2}\dot{u}\dot{u} + \epsilon m\dot{u}\dot{w} + \frac{m}{2}\epsilon^2\dot{w}\dot{w} - \frac{k}{2}u^2 - k\epsilon uw - \frac{k}{2}\epsilon^2 w^2 \right) dt$$
$$= \int_{t_1}^{t_2} \big(m\dot{u}\dot{w} - kuw\big)\, dt \quad \text{for } \epsilon = 0.$$

Integrating by parts in the term $m\dot{u}\dot{w}$, we get

$$\int_{t_1}^{t_2} \big(m\ddot{u} + ku\big)w\, dt = 0,$$

for all $w(t)$ with $w(t_1) = w(t_2) = 0$. This implies that $m\ddot{u}+ku = 0$ in $[t_1, t_2]$, since $w(t)$ can vary arbitrarily in the interval (t_1, t_2), and we obtain the desired equation.

47.3 A Pendulum with Fixed Support

We consider a pendulum in the form of a body of mass one attached to a weightless string of unit length fixed to the ceiling under the action of a vertical gravity force normalized to one. The action integral of the difference between kinetic and potential energy is given by

$$I(u) = \int_{t_1}^{t_2} \left(\frac{1}{2}\dot{u}^2(t) - (1 - \cos(u(t)) \right) dt,$$

where $u(t)$ represents the angle of the pendulum in radians at time t, measured from the vertical position, see Fig. 47.2.

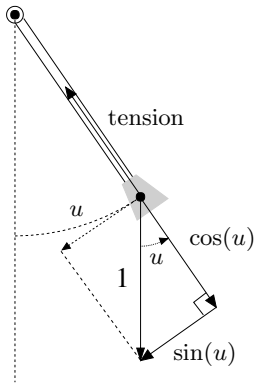

Fig. 47.2. A pendulum

The potential energy in this case is equal to the work of lifting the mass from the bottom position to the level $(1-\cos(v))$, which is exactly equal to $(1-\cos(v))$ if the gravitational constant is normalized to one. Stationarity of the action integral requires that for all perturbations $w(t)$ satisfying $w(t_1) = w(t_2) = 0$, we have

$$0 = \frac{d}{d\epsilon} \int_{t_1}^{t_2} \left(\frac{1}{2}(\dot{u} + \epsilon\dot{w})^2(t) - (1 - \cos(u(t) + \epsilon w(t)) \right) dt \quad \text{for } \epsilon = 0,$$

which gives as above

$$\int_{t_1}^{t_2} \big(\ddot{u} + \sin(u(t))\big) w \, dt = 0.$$

This yields the initial value problem

$$\begin{cases} \ddot{u} + \sin(u) = 0 & \text{for } t > 0 \\ u(0) = u_0,\ \dot{u}(0) = u_1, \end{cases} \tag{47.3}$$

TS[e] Please check the parentheses.
TS[f] Please check the parentheses.

Editor's or typesetter's annotations (will be removed before the final TEX run)

where we added initial conditions for position and velocity.

The resulting differential equation $\ddot{u} = -\sin(u)$ is an expression of Newton's Law, since $\ddot{u}$ is the angular acceleration and $-\sin(u)$ is the restoring force in the angular direction. We conclude that the Principle of Least Action in the present case is a reformulation of Newton's Law.

If the angle of the pendulum stays small during the motion, then we can approximate $\sin(u)$ by u and obtain the linear equation $\ddot{u} + u = 0$, with solutions being linear combinations of $\sin(t)$ and $\cos(t)$.

47.4 A Pendulum with Moving Support

We now generalize to a pendulum with a support that is subject to a prescribed motion. Consider thus a body of mass m swinging in a weightless string of length l that is attached to a support moving according to a given function $r(t) = (r_1(t), r_2(t))$ in a coordinate system with the x_1-axis horizontal and the x_2-axis vertical upward. Let $u(t)$ be the angle of the string at time t measured from the vertical.

The potential energy is again equal to the height of the body, measured from some reference position, times mg with g the gravitational constant. Thus, we may choose

$$V(u(t)) = mg(r_2(t) - l\cos(u)).$$

To express the kinetic energy, we need to take into account the motion of the support. The velocity of the body relative to the support is $(l\dot{u}\cos u, l\dot{u}\sin u)$, and the total velocity is thus $(\dot{r}_1(t) + l\dot{u}\cos u, \dot{r}_2(t) + l\dot{u}\sin u)$. The kinetic energy is $m/2$ times the square of the *modulus of the velocity*, and thus

$$T = \frac{m}{2}\left[(\dot{r}_1 + l\dot{u}\cos u)^2 + (\dot{r}_1 + l\dot{u}\sin u)^2\right].$$

Using the Principle of Least Action, we obtain the following equation:

$$\ddot{u} + \frac{g}{l}\sin u + \frac{\ddot{r}_1}{l}\cos u + \frac{\ddot{r}_2}{l}\sin u = 0, \tag{47.4}$$

together with initial conditions for $u(0)$ and $\dot{u}(0)$.

If the support is fixed with $\ddot{r}_1 = \ddot{r}_2 = 0$, then we recover the equation (47.3) setting $l = m = g = 1$.

47.5 The Principle of Least Action

We now consider a mechanical system that is described by a vector function $u(t) = (u_1(t), u_2(t))$. We may think of a system consisting of two bodies

with positions given by the functions $u_1(t)$ and $u_2(t)$. The action integral is

$$I(u_1, u_2) = I(u) = \int_{t_1}^{t_2} L(u(t)\, dt,$$

where

$$L(u_1(t), u_2(t)) = L(u(t)) = T(\dot{u}(t)) - V(u(t))$$

is the difference of the kinetic energy $T(\dot{u}(t)) = T(\dot{u}_1(t), \dot{u}_2(t))$ and the potential energy $V(u(t)) = V(u_1(t), u_2(t))$. We refer to $L(u(t))$ as the *Lagrangean* of the state $u(t)$.

The Principle of Least Action states that the action integral is stationary at the true state $u(t)$ in the sense that for all perturbations $w_1(t)$ and $w_2(t)$ with $w_1(t_1) = w_2(t_1) = w_1(t_2) = w_2(t_2)$, we have for $\epsilon = 0$,

$$\frac{d}{d\epsilon} I(u_1 + \epsilon w_1, u_2) = 0$$
$$\frac{d}{d\epsilon} I(u_1, u_2 + \epsilon w_2) = 0.$$

Assuming that

$$T(\dot{u}_1(t), \dot{u}_2(t)) = \frac{m_1}{2}\dot{u}_1^2(t) + \frac{m_2}{2}\dot{u}_2^2(t),$$

we obtain performing the differentiation with respect to ϵ as above and setting $\epsilon = 0$,

$$\int_{t_1}^{t_2} \big(m\dot{u}_1(t)\dot{w}_1(t) - \frac{\partial V}{\partial u_1}(u_1(t), u_2(t))w_1(t)\big)\, dt = 0,$$
$$\int_{t_1}^{t_2} \big(m\dot{u}_2(t)\dot{w}_2(t) - \frac{\partial V}{\partial u_2}(u_1(t), u_2(t))w_2(t)\big)\, dt = 0.$$

Integrating by parts as above and letting w_1 and w_2 vary freely over (t_1, t_2), we obtain

$$\begin{aligned} m\ddot{u}_1(t) &= -\frac{\partial V}{\partial u_1}(u_1(t), u_2(t)), \\ m\ddot{u}_2(t) &= -\frac{\partial V}{\partial u_2}(u_1(t), u_2(t)). \end{aligned} \tag{47.5}$$

If we set

$$F_1 = -\frac{\partial V}{\partial u_1}, \quad F_2 = -\frac{\partial V}{\partial u_2},$$

then we can write the equations derived from the Principle of Least Action as

$$\begin{aligned} m\ddot{u}_1(t) &= F_1(u_1(t), u_2(t)), \\ m\ddot{u}_2(t) &= F_1(u_1(t), u_2(t)), \end{aligned} \tag{47.6}$$

which can be viewed as Newton's Law if F_1 and F_2 are interpreted as forces.

47.6 Conservation of the Total Energy

Defining the *total energy*

$$E(u(t)) = T(\dot{u}(t)) + V(u(t))$$

as the sum of the kinetic and potential energies and using (47.5), we get

$$\begin{aligned}\frac{d}{dt}E(u(t)) &= m_1\dot{u}_1\ddot{u}_1 + m_2\dot{u}_2\ddot{u}_2 + \frac{\partial V}{\partial u_1}\dot{u}_1 + \frac{\partial V}{\partial u_2}\dot{u}_2 \\ &= \dot{u}_1\left(m_1\ddot{u}_1 + \frac{\partial V}{\partial u_1}\right) + \dot{u}_2\left(m_2\ddot{u}_2 + \frac{\partial V}{\partial u_2}\right) = 0.\end{aligned}$$

We conclude that the total energy $E(u(t))$ is constant in time, that is the energy is *conserved*. Obviously, energy conservation is not a property of all systems, and thus the Principle of Least Action only applies to so called *conservative systems*, where the total energy is conserved. In particular, effects of *friction* are not present.

47.7 The Double Pendulum

We now consider a *double pendulum* consisting of two bodies of masses m_1 and m_2, where the first body of mass m_1 hangs on a weightless string of length l_1 attached to a fixed support and the second body of mass m_2 hangs on a weightless string of length l_2 attached to the first body. We shall now apply the Principle of Least Action to derive the equations of motion for this system.

To describe the state of the system, we use the angles $u_1(t)$ and $u_2(t)$ of the two bodies measured from vertical position.

We now seek expressions for the kinetic and potential energies of the system of the two bodies. The contributions from the second body is obtained from the expressions for a pendulum with moving support derived above if we set $(r_1(t), r_2(t)) = (l_1 \sin u_1, -l_1 \cos u_1)$.

The potential energy of the first pendulum is $-mgl_1 \cos u_1$ and the total potential energy is

$$V(u_1(t), u_2(t)) = -m_1 g l_1 \cos u_1(t) - m_2 g \left(l_1 \cos u_1(t) + l_2 \cos u_2(t)\right).$$

The total kinetic energy is obtained similarly adding the kinetic energies of the two bodies:

$$\begin{aligned}T(\dot{u}_1(t), \dot{u}_2(t)) = \frac{m_1}{2} l_1^2 \dot{u}_1^2 + \frac{m_2}{2}\big[&(l_1\dot{u}_1 \cos u_1 + l_2\dot{u}_2 \cos u_2)^2 \\ &+ (l_1\dot{u}_1 \sin u_1 + l_2\dot{u}_2 \sin u_2)^2\big].\end{aligned}$$

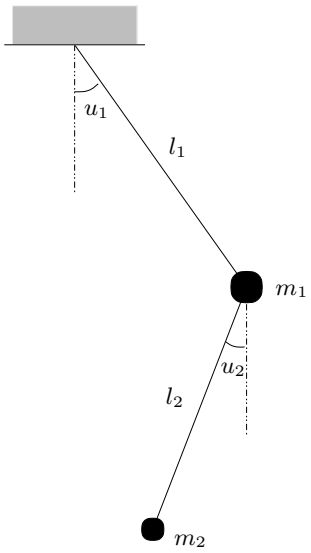

Fig. 47.3. Double pendulum

Using the identities $\sin^2 u + \cos^2 u = 1$ and $\cos(u_1 - u_2) = \cos u_1 \cos u_2 + \sin u_1 \sin u_2$, we can rewrite this expression as

$$T = \frac{m_1}{2} l_1^2 \dot{u}_1^2 + \frac{m_2}{2} \left[l_1^2 \dot{u}_1^2 + l_2^2 \dot{u}_2^2 + 2 l_1 l_2 \dot{u}_1 \dot{u}_2 \cos(u_1 - u_2) \right].$$

Applying the Principle of Least Action, we obtain the following system of equations for a double pendulum:

$$\begin{aligned} \ddot{u}_1 + \frac{m_2}{m_1 + m_2} \frac{l_2}{l_1} \left[\ddot{u}_2 \cos(u_2 - u_1) - \dot{u}_2^2 \sin(u_2 - u_1) \right] + \frac{g}{l_1} \sin u_1 = 0, \\ \ddot{u}_2 + \frac{l_1}{l_2} \left[\ddot{u}_1 \cos(u_2 - u_1) + \dot{u}_1^2 \sin(u_2 - u_1) \right] + \frac{g}{l_2} \sin u_2 = 0. \end{aligned} \tag{47.7}$$

We note that if $m_2 = 0$, then the first equation is just the equation for a simple pendulum, and that if $\ddot{u}_1 = \dot{u}_1 = 0$, then the second equation is again the equation for a simple pendulum.

47.8 The Two-Body Problem

We consider the *two-body* problem for a small mass orbiting around a heavy mass, such as the Earth moving around the Sun neglecting the influence of the other planets. We assume that the motion takes place in a plane and use polar coordinates (r, θ) with the origin at the center of the heavy mass

to describe the position of the light body. Assuming that the heavy body is fixed, the action integral representing the difference between the kinetic and potential energy of the small mass is given by

$$\int_{t_1}^{t_2} \left(\frac{1}{2}\dot{r}^2 + \frac{1}{2}(\dot{\theta}r)^2 + \frac{1}{r} \right) dt \tag{47.8}$$

because the velocity is $(\dot{r}, r\dot{\theta})$ in the radial and angular directions respectively, and the gravity potential is $-r^{-1} = -\int_r^\infty s^{-2}\,ds$ corresponding to the work needed to move a particle of unit mass a distance r from the orbit center to infinity. The corresponding Euler-Lagrange equations are

$$\begin{cases} \ddot{r} - r\dot{\theta}^2 = -\frac{1}{r^2}, & t > 0, \\ \frac{d}{dt}(r^2\dot{\theta}) = 0, & t > 0, \end{cases} \tag{47.9}$$

which is a second order system to be complemented with initial values for position and velocity.

We construct the analytical solution of this system in a set of problems below, which may be viewed as a short course on Newton's *Principia Matematica*. We invite the reader to take this opportunity of getting on speaking terms with Newton himself.

47.9 Stability of the Motion of a Pendulum

The linearization of the equation for a pendulum at $\bar{u} \in \mathbb{R}$, $\ddot{u} + \sin(u) = 0$, is obtained by setting $u = \bar{u} + \varphi$ and noting that $\sin(u) \approx \sin(\bar{u}) + \cos(\bar{u})\varphi$. This leads to

$$0 = \ddot{u} + \sin(u) \approx \ddot{\varphi} + \sin(\bar{u}) + \cos(\bar{u})\varphi.$$

Assuming first that $\bar{u} = 0$, we obtain the following linearized equation for the perturbation φ,

$$\ddot{\varphi} + \varphi = 0, \tag{47.10}$$

with solution being a linear combination of $\sin(t)$ and $\cos(t)$. For example, if $\varphi(0) = \delta$ and $\dot{\varphi}(0) = 0$, then $\varphi(t) = \delta\cos(t)$, and we see that an initially small perturbation is kept small for all time: the pendulum stays close to the bottom position under small perturbations.

Setting next $\bar{u} = \pi$, we obtain

$$\ddot{\varphi} - \varphi = 0 \tag{47.11}$$

with the solution being a linear combination of $\exp(\pm t)$. Since $\exp(t)$ grows very quickly, the state $\bar{u} = \pi$ corresponding to the pendulum in the top position is *unstable*. A small perturbation will quickly develop into a large perturbation and the pendulum will move way from the top position.

We will return to the topic of this section in Chapter *Linearization and stability of initial value problems*

Chapter 47 Problems

47.1. Supply the missing details in the derivation of the equation for the pendulum. If the angle u stays small during the motion, then the simpler *linearized* model $\ddot{u} + u = 0$ may be used. Solve this equation analytically and compare with numerical results for the nonlinear pendulum equation to determine limits of validity of the linear model.

47.2. Carry out the details in the derivation of the equations for the pendulum with moving support and the double pendulum.

47.3. Study what happens for the double pendulum in the extreme cases, i.e. at zero and infinity, for the parameters m_1, m_2, l_1 and l_2.

47.4. Derive the second equation of motion for the double pendulum from the result for the pendulum with moving support by setting $(r_1(t), r_2(t)) = (l_1 \sin u_1, -l_1 \cos u_1)$.

47.5. Derive the equation of motion for a bead sliding on a frictionless plane vertical curve under the action of gravity.

47.6. In the foot-steps of Newton give an analysis and analytical solution of the two-body problem modeled by (47.9) through the following sequence of problems: (i) Prove that a stationary point of the action integral (47.8) satisfies (47.9). (ii) Prove that the total energy is constant in time. (iii) Introducing the change of variables $u = r^{-1}$, show that $\dot{\theta} = cu^2$ for c constant. Use this relation together with the fact that the chain rule implies that

$$\frac{dr}{dt} = \frac{dr}{du}\frac{du}{d\theta}\frac{d\theta}{dt} = -c\frac{du}{d\theta} \quad \text{and} \quad \ddot{r} = -c^2u^2\frac{d^2u}{d\theta^2}$$

to rewrite the system (47.9) as

$$\frac{d^2u}{d\theta^2} + u = c^{-2}. \tag{47.12}$$

Show that the general solution of (47.12) is

$$u = \frac{1}{r} = \gamma\cos(\theta - \alpha) + c^{-2},$$

where γ and α are constants. (iii) Prove that the solution is either an ellipse, parabola, or hyperbola. Hint: Use the fact that these curves can be described as the loci of points for which the ratio of the distance to a fixed point and to a fixed straight line, is constant. Polar coordinates are suitable for expressing this relation. (iv) Prove Kepler's three laws for planetary motion using the experience from the previous problem.

47.7. Study the linearizations of the double pendulum at $(u_1, u_2) = (0, 0)$ and $(u_1, u_2) = (\pi, \pi)$ and draw conclusions about stablity.

47.8. Attach an elastic string to a simple pendulum in some way and model the resulting system.

47.9. Compute solutions of the presented models numerically.

48
N-Body Systems*

> The reader will find no figures in this work. The methods which I set forth do not require either geometrical or mechanical reasonings, but only algebraic operations, subject to a regular and uniform rule of procedure. (Lagrange in *Méchanique Analytique*)

48.1 Introduction

We shall now model systems of N bodies interacting through mechanical forces that result from springs and dashpots, see Fig. 48.1, or from gravitational or electrostatic forces. We shall use two different modes of description. In the first formulation, we describe the system through the coordinates of (the centers of gravity of) the bodies. In the second, we use the *displacements* of the bodies measured from an initial reference configuration. In the latter case, we also *linearize* under an assumption of small displacements to obtain a linear system of equations. In the first formulation, the initial configuration is only used to initialize the system and is "forgotten" at a later time in the sense that the description of the system only contains the present position of the masses. In the second formulation, the reference configuration is retrievable through the evolution since the unknown is the displacement from the reference position. The different formulations have different advantages and ranges of application.

48.2 Masses and Springs

We consider the motion in $\mathbb{R}^3$ of a system of N bodies connected by a set of Hookean springs. For $i = 1, \ldots, N$, let the position at time t of body i be given by the vector function $u_i(t) = (u_{i1}(t), u_{i2}(t), u_{i3}(t))$, with $u_{ik}(t)$ denoting the x_k coordinate, $k = 1, 2, 3$, and suppose the mass of body i is m_i. Let body i be connected to body j with a Hookean spring of spring constant $k_{ij} \geq 0$ for $i, j = 1, \ldots N$. Some of the k_{ij} may be zero, which effectively means that there is no spring connection between body i and body j. In particular $k_{ii} = 0$. We assume to start with that the reference length of the spring corresponding to zero spring tension is equal to zero. This means that the spring forces are always attractive.

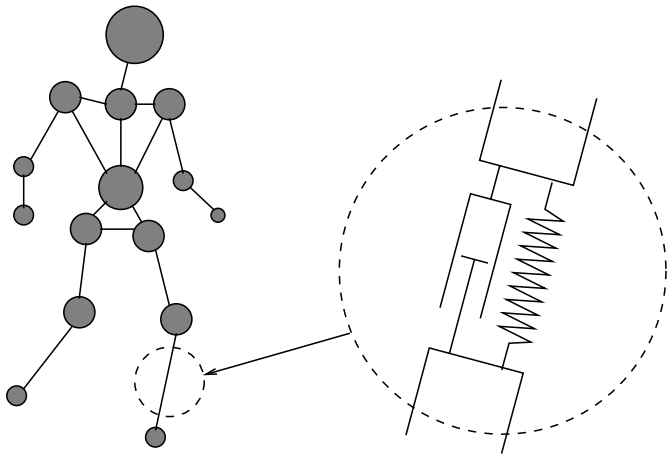

Fig. 48.1. A typical system of masses, springs and dashpots in motion

We now derive the equations of motion for the mass-spring system using the Principle of Least Action. We assume first that the gravitational force is set to zero. The potential energy of the configuration $u(t)$ is given by

$$\begin{aligned} V(u(t)) &= \sum_{i,j=1}^{N} \frac{1}{2} k_{ij} |u_i - u_j|^2 \\ &= \sum_{i,j=1}^{N} \frac{1}{2} k_{ij} \big((u_{i1} - u_{j1})^2 + (u_{i2} - u_{j2})^2 + (u_{i3} - u_{j3})^2 \big), \end{aligned} \tag{48.1}$$

with the time dependence of the coordinates u_{ik} suppressed for readability. This is because the length of the spring connecting the body i and body j is equal to $|u_i - u_j|$, and the work to stretch the spring from zero length to length l is equal to $\frac{1}{2} k_{ij} l^2$.

The action integral is

$$I(u) = \int_{t_1}^{t_2} \left(\sum_{i=1}^{N} \frac{1}{2} m_i \big(\dot{u}_{i1}^2 + \dot{u}_{i2}^2 + \dot{u}_{i3}^2 \big) - V(u(t)) \right) dt,$$

and using the Principle of Least Action and the fact that

$$\frac{\partial V(u)}{\partial u_{ik}} = \sum_{j=1}^{N} k_{ij}(u_{ik} - u_{jk}),$$

we obtain the following system of equations of motion:

$$m_i \ddot{u}_{ik} = -\sum_{j=1}^{N} k_{ij}(u_{ik} - u_{jk}), \quad k = 1, 2, 3,\ i = 1, \ldots, N, \tag{48.2}$$

or in vector form

$$m_i \ddot{u}_i = -\sum_{j=1}^{N} k_{ij}(u_i - u_j),\ i = 1, \ldots, N, \tag{48.3}$$

together with initial conditions for $u_i(0)$ and $\dot{u}_i(0)$. We can view these equations as expressing Newton's Law

$$m_i \ddot{u}_i = F_i^s, \tag{48.4}$$

with the total spring force $F_i^s = (F_{i1}^s, F_{i2}^s, F_{i3}^s)$ acting on body i being equal to

$$F_i^s = -\sum_{j=1}^{N} k_{ij}(u_i - u_j). \tag{48.5}$$

Inclusion of gravity forces in the direction of the negative x_3 axis, adds a component $-m_i g$ to F_{i3}^s, where g is the gravitational constant.

The system (48.3) is linear in the unknowns $u_{ij}(t)$. If we assume that the reference length with zero spring force of the spring connecting body i and j is equal to $l_{ij} > 0$, then the potential changes to

$$V(u(t)) = \sum_{i,j=1}^{N} \frac{1}{2} k_{ij}(|u_i - u_j| - l_{ij})^2, \tag{48.6}$$

and the resulting equations of motion are no longer linear. Below, we shall consider a linearized form assuming $|u_i - u_j| - l_{ij}$ is small compared to l_{ij}.

48.3 The N-Body Problem

By tradition, a "N-body" problem refers to a system of N bodies in motion in $\mathbb{R}^3$ under the influence of mutual gravitational forces. An example is given by our solar system with 9 planets orbiting around the Sun, where we typically disregard moons, asteroids, and comets.

Let the position at time t of (the center of gravity of) body i be given by the vector function $u_i(t) = (u_{i1}(t), u_{i2}(t), u_{i3}(t))$, with $u_{ik}(t)$ denoting the x_k coordinate in $\mathbb{R}^3$, $k = 1, 2, 3$, and suppose the mass of body i is m_i. Newton's inverse square law of gravitation states that the gravitational force from the body j on the body i is given by

$$-\frac{\gamma m_i m_j}{|u_i(t) - u_j(t)|^2}\frac{u_i(t) - u_j(t)}{|u_i(t) - u_j(t)|} = -\gamma m_i m_j \frac{u_i(t) - u_j(t)}{|u_i(t) - u_j(t)|^3},$$

where γ is a gravitational constant. We thus obtain the following system of equations modeling the N-body problem:

$$m_i \ddot{u}_i = -\gamma m_i m_j \sum_{j \neq i} \frac{u_i - u_j}{|u_i(t) - u_j(t)|^3}, \tag{48.7}$$

together with initial conditions for $u_i(0)$ and $\dot{u}_i(0)$.

Alternatively, we may derive these equations using the Principle of Least Action using the gravity potential

$$V(u) = -\sum_{i,j=1,\, i \neq j}^{N} \frac{\gamma m_i m_j}{|u_i - u_j|},$$

and the fact that

$$\frac{\partial V}{\partial u_{ik}} = \sum_{j \neq i} \frac{\gamma m_i m_j}{|u_i - u_j|^3}(u_{ik} - u_{jk}). \tag{48.8}$$

The expression for the gravity potential is obtained by noticing that the work to bring body i from a distance r of body j to infinity is equal to

$$\int_r^\infty \frac{\gamma m_i m_j}{s^2}\, ds = \gamma m_i m_j [-\frac{1}{s}]_{s=r}^{s=\infty} = \frac{\gamma m_i m_j}{r}.$$

Notice the minus sign of the potential, arising from the fact that the body i loses potential energy as it approaches body j.

Analytical solutions are available only in the case of the 2-body problem. The numerical solution of for example the 10-body problem of our solar system is very computationally demanding in the case of long time simulation. As a result, the long time stability properties of our Solar system are unknown. For example, it does not seem to be known if eventually

the Earth will change orbit with Mercury, Pluto will spin away to another galaxy, or some other dramatic event will take place.

The general result of existence guarantees a solution, but the presence of the stability factor $\exp(tL_f)$ brings the accuracy in long-time simulation seriously in doubt.

48.4 Masses, Springs and Dashpots: Small Displacements

We now give a different description of the mass-spring system above. Let the initial position of body i, which is now chosen as reference position, be $a_i = (a_{i1}, a_{i2}, a_{i3})$, and let the actual position at time $t > 0$ be given by $a_i + u_i(t)$ where now $u_i(t) = (u_{i1}(t), u_{i2}(t), u_{i3}(t))$ is the *displacement* of body i from its reference position a_i.

The potential energy of the configuration $u(t)$ is given by

$$\begin{aligned} V(u(t)) &= \sum_{i,j=1}^{N} \frac{1}{2} k_{ij} \big(|a_i + u_i - (a_j + u_j)| - |a_i - a_j| \big)^2 \\ &= \frac{1}{2} k_{ij} \big(|a_i - a_j + (u_i - u_j)| - |a_i - a_j| \big)^2, \end{aligned}$$

assuming zero spring forces if the springs have the reference lengths $a_i - a_j$.

We now specialize to small displacements, assuming that $|u_i - u_j|$ is small compared to $|a_i - a_j|$. We then use that if $|b|$ is small compared to $|a|$, where $a, b \in \mathbb{R}^3$, then

$$\begin{aligned} |a+b| - |a| &= \frac{(|a+b| - |a|)(|a+b| + |a|)}{|a+b| + |a|} \\ &= \frac{|a+b|^2 - |a|^2}{|a+b| + |a|} = \frac{(a+b)\cdot(a+b) - a \cdot a}{|a+b| + |a|} \approx \frac{a \cdot b}{|a|}. \end{aligned}$$

Thus, if $|u_i - u_j|$ is small compared to $|a_i - a_j|$, then

$$|a_i - a_j + (u_i - u_j)| - |a_i - a_j| \approx \frac{(a_i - a_j) \cdot (u_i - u_j)}{|a_i - a_j|},$$

and we obtain the following approximation of the potential energy

$$\hat{V}(\text{TS}^{\text{g}} u(t) = \sum_{i,j=1}^{N} \frac{1}{2} k_{ij} \frac{\big((a_i - a_j) \cdot (u_i - u_j) \big)^2}{|a_i - a_j|^2}.$$

Using the Principle of Least Action we thus obtain the following linearized system of equations

$$m_i \ddot{u}_{ik} = -\sum_{j=1}^{N} \frac{k_{ij}(a_i - a_j) \cdot (u_i - u_j)}{|a_i - a_j|^2} (a_{ik} - a_{jk}), \quad k = 1, 2, 3,\ i = 1, \ldots, N,$$

TS^g Please check the parentheses.

Editor's or typesetter's annotations (will be removed before the final TEX run)

or in vector form

$$m_i\ddot{u}_i = -\sum_{j=1}^{N} \frac{k_{ij}(a_i - a_j)\cdot(u_i - u_j)}{|a_i - a_j|^2}(a_i - a_j), \quad i = 1, \ldots, N. \tag{48.9}$$

together with initial conditions for $u_i(0)$ and $\dot{u}_i(0)$. We can view these equations as expressing Newton's Law

$$m_i\ddot{u}_i = F_i^s, \quad i+1, \ldots, N, \tag{48.10}$$

with the spring force F_i^s acting on body i given by

$$F_i^s = -\sum_{j=1}^{N} b_{ij}e_{ij},$$

where

$$e_{ij} = \frac{a_i - a_j}{|a_i - a_j|}$$

is the normalized vector connecting a_j and a_i, and

$$b_{ij} = k_{ij}e_{ij} \cdot (u_i - u_j). \tag{48.11}$$

48.5 Adding Dashpots

A *dashpot* is a kind of shock absorber which may be thought of as consisting of a piston that moves inside a cylinder filled with oil or some other viscous fluid, see Fig. 48.2. As the piston moves, the flow of the fluid past the piston

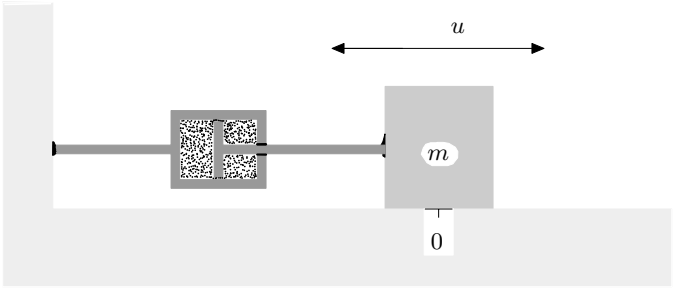

Fig. 48.2. Cross section of a dashpot connected to a mass

creates a force opposite to the motion, which we assume is proportional to the velocity with the constant of proportionality representing the coefficient of *viscosity* of the dashpot.

We now expand the above mass-spring model to include springs and dashpots coupled in parallel. For each pair of nodes i and j, we let k_{ij} and μ_{ij} be the coefficients of a spring and a dashpot coupled in parallel, with $k_{ij} = 0$ and $\mu_{ij} = 0$ if the spring or dashpot is absent, and in particular $k_{ii} = \mu_{ii} = 0$. The dashpot force F_i^d acting on body i will then be given by

$$F_i^d = -\sum_{j=1}^{N} d_{ij} e_{ij},$$

where

$$d_{ij} = \mu_{ij} e_{ij} \cdot (\dot{u}_i - \dot{u}_j). \tag{48.12}$$

To get this result, we use the fact that

$$e_{ij} \cdot (\dot{u}_i - \dot{u}_j) e_{ij}$$

is the projection of $\dot{u}_i - \dot{u}_j$ onto e_{ij}. We thus assume that the dashpot reacts with a force that is proportional to the projection of $\dot{u}_i - \dot{u}_j$ onto the direction $a_i - a_j$.

This leads to the linearized mass-spring-dashpot model:

$$m_i \ddot{u}_i = F_i^s + F_i^d, \quad i = 1, \ldots, N, \tag{48.13}$$

together with initial conditions for $u_i(0)$ and $\dot{u}_i(0)$. We can write these equations as a system in the form

$$M\ddot{u} + D\dot{u} + Ku = 0, \tag{48.14}$$

with constant coefficient matrices matrices M, D and K, where u is a $3N$-vector listing all the components u_{ik}. The matrix M is diagonal with the masses m_i as entries, and D and K are symmetric positive semi-definite (see the problem section).

A system with dashpots is not conservative, since the dashpots consume energy, and therefor cannot be modeled using the Principle of Least Action.

The linear system (48.14) models a wide range of phenomena and can be solved numerically with appropriate solvers. We return to this issue below. We now consider the simplest example of one mass connected to the origin with a spring and a dashpot in parallel.

48.6 A Cow Falling Down Stairs

In Fig. 48.3 and Fig. 48.4 we show the result of computational simulation of a cow falling down a staircase. The computational model consists of a skeleton in the form of a mass-spring-dashpot-system together with a surface model built upon the skeleton. The skeleton deforms under the action of gravity forces and contact forces from the staircase and the surface model conforms to the deformation.

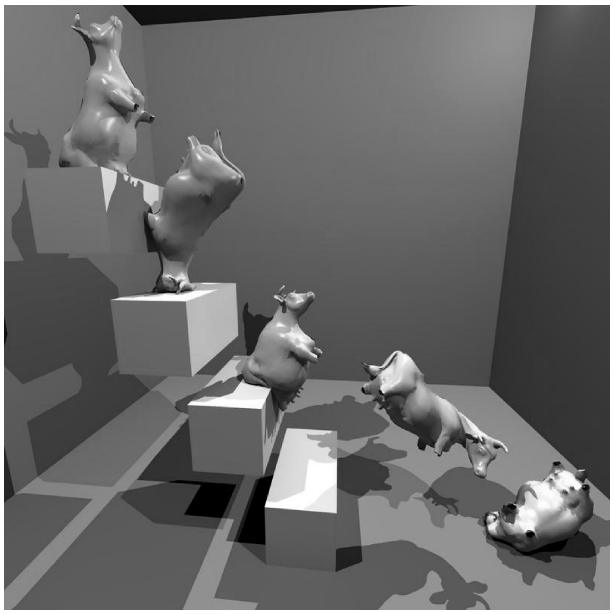

Fig. 48.3. Cow falling down a staircase (simulation created by Mr. Johan Jansson)

48.7 The Linear Oscillator

We now consider the simplest example consisting of one body of mass 1 connected to one end of a Hookean spring connected to the origin with the motion taking place along the x_1-axis. Assuming the spring has zero length at zero tension, the system is described by

$$\begin{cases} \ddot{u} + ku = 0 & \text{for } t > 0, \\ u(0) = u_0,\ \dot{u}(0) = \dot{u}_0. \end{cases} \tag{48.15}$$

with $u(t)$ denoting the x_1 coordinated of the body at time t, and u_0 and $\dot{u}_0$ given initial conditions. The solution is given by

$$u(t) = a\cos(\sqrt{k}t) + b\sin(\sqrt{k}t) = \alpha\cos(\sqrt{k}(t-\beta)), \tag{48.16}$$

where the constants a and b, or α and β, are determined by the initial conditions. We conclude that the motion of the mass is periodic with *frequency* $\sqrt{k}$ and *phase shift* β and *amplitude* α, depending on the initial data. This model is referred to as the *linear oscillator*. The solution is periodic with period $\frac{2\pi}{\sqrt{k}}$, and the *time scale* is similar.

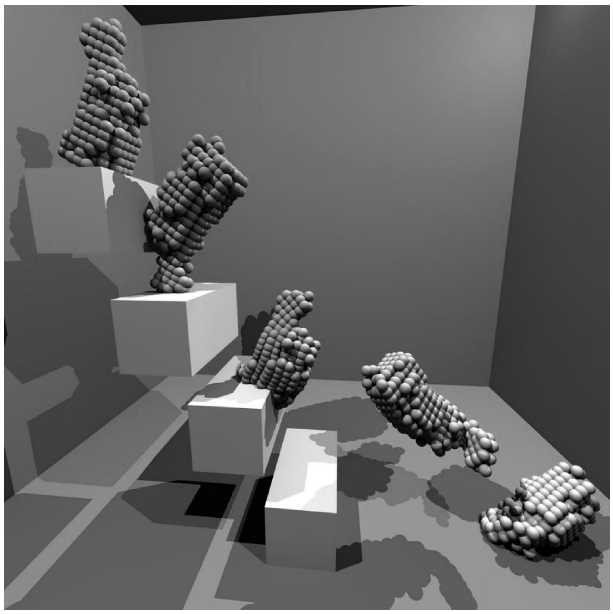

Fig. 48.4. N-body cow-skeleton falling down a staircase (simulation created by Mr. Johan Jansson)

48.8 The Damped Linear Oscillator

Adding a dashpot in parallel with the spring in the model above gives the model of a damped linear oscillator

$$\begin{cases} \ddot{u} + \mu\dot{u} + ku = 0, & \text{for } t > 0, \\ u(0) = u_0, \quad \dot{u}(0) = \dot{u}_0. \end{cases} \tag{48.17}$$

In the case $k = 0$, we obtain the model

$$\begin{cases} \ddot{u} + \mu\dot{u} = 0 & \text{for } t > 0, \\ u(0) = u_0, \quad \dot{u}(0) = \dot{u}_0, \end{cases} \tag{48.18}$$

with the solution

$$u(t) = -\frac{\dot{u}_0}{\mu}\exp(-\mu t) + u_0 + \frac{\dot{u}_0}{\mu}.$$

We see that the mass approaches the fixed position $u = u_0 + \frac{\dot{u}_0}{\mu}$ determined by the initial data as t increases to infinity. The time scale is of size $\frac{1}{\mu}$.

The characteristic polynomial equation for the full model $\ddot{u}+\mu\dot{u}+ku = 0$, is

$$r^2 + \mu r + kr = 0.$$

Completing the square we can write the characteristic equation in the form

$$\left(r + \frac{\mu}{2}\right)^2 = \frac{\mu^2}{4} - k = \frac{1}{4}(\mu^2 - 4k). \tag{48.19}$$

If $\mu^2 - 4k > 0$, then there are two real roots $-\frac{1}{2}(\mu \pm \sqrt{\mu^2 - 4k})$, and the solution $u(t)$ has the form (see the Chapter The exponential function),

$$u(t) = ae^{-\frac{1}{2}(\mu+\sqrt{\mu^2-4k})t} + be^{-\frac{1}{2}(\mu-\sqrt{\mu^2-4k})t},$$

with the constants a and b determined by the initial conditions. In this case, the viscous damping of the dashpot dominates over the spring force, and the solution converges exponentially to a rest position, which is equal to $u = 0$ if $k > 0$. The fastest time scale is again of size $\frac{1}{\mu}$.

If $\mu^2 - 4k < 0$, then we introduce the new variable $v(t) = e^{\frac{\mu t}{2}}u(t)$, with the objective of transforming the characteristic equation (48.19) into an equation of the form $s^2 + (k - \frac{\mu^2}{4}) = 0$. Since $u(t) = e^{-\frac{\mu t}{2}}v(t)$, we have

$$\dot{u}(t) = \frac{d}{dt}\big(e^{-\frac{\mu t}{2}}v(t)\big) = \left(\dot{v} - \frac{\mu}{2}v\right)e^{-\frac{\mu t}{2}},$$
$$\ddot{u}(t) = \left(\ddot{v} - \mu\dot{v} + \frac{\mu^2}{4}\right)e^{-\frac{\mu t}{2}},$$

and thus the differential equation $\ddot{u} + \mu\dot{u} + ku = 0$ is transformed into

$$\ddot{v} + \left(k - \frac{\mu^2}{4}\right)v = 0,$$

with the solution $v(t)$ being a linear combination of $\cos(\frac{t}{2}\sqrt{4k - \mu^2})$ and $\sin(\frac{t}{2}\sqrt{4k - \mu^2})$. Transforming back to the variable $u(t)$ we get the solution formula

$$u(t) = ae^{-\frac{1}{2}\mu t}\cos\left(\frac{t}{2}\sqrt{4k - \mu^2}\right) + be^{-\frac{1}{2}\mu t}\sin\left(\frac{t}{2}\sqrt{4k - \mu^2}\right).$$

The solution again converges to the zero rest position as time passes if $\mu > 0$, but now it does so in an oscillatory fashion. Now two time scales appear: a time scale of size $\frac{1}{\mu}$ for the exponential decay and a time scale $1/\sqrt{k - \mu^2/4}$ of the oscillations.

Finally, in the limit case $\mu^2 - 4k = 0$ the solution $v(t)$ of the corresponding equation $\ddot{v} = 0$ is given by $v(t) = a + bt$, and thus

$$u(t) = (a + bt)e^{-\frac{1}{2}\mu t}.$$

This solution exhibits initial linear growth and eventually converges to a zero rest position as time tends to infinity. We illustrate the three possible behaviors in Fig. 48.5.

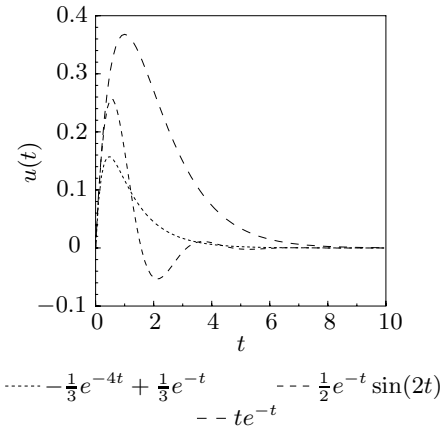

Fig. 48.5. Three solutions of the mass-spring-dashpot model (48.17) satisfying the initial conditions $u(0) = 0$ and $\dot{u}(0) = 1$. The first solution corresponds to $\mu = 5$ and $k = 4$, the second to $\mu = 2$ and $k = 5$, and the third to $\mu = 2$ and $k = 1$

48.9 Extensions

We have above studied systems of bodies interacting through Hookean springs, linear dashpots and gravitational forces. We can generalize to systems of non-linear springs, dashpots, and other mechanical devices like springs reacting to changes of angles between the bodies, or other forces like electrostatic forces. In this way, we can model very complex systems from macroscopic scales of galaxies to microscopic molecular scales. For example, electrostatic forces are related to potentials of the form

$$V^e(u) = \pm c \sum_{i,j=1}^{N} \frac{q_i q_j}{|u_i - u_j|}$$

where q_i is the charge of body i and c is a constant, and thus have a form similar to that of gravitational forces.

In particular, models for molecular dynamics take the form of N-body systems interacting through electrostatic forces and forces modeled by various springs reacting to bond lengths and bond angles between the atoms. In these applications, N may be of the order 10^4 and the smallest time scale of the dynamics may be of size 10^{-14} related to very stiff bond length springs. Needless to say, simulations with such models may be very computationally demanding and is often out of reach with present day computers. For more precise information, we refer to the survey article *Molecular modeling of proteins and mathematical prediction of protein structure*, SIAM REV. (39), No 3, 407-460, 1997, by A. Neumair.

Chapter 48 Problems

48.1. Verify the solution formulas for the three solutions shown in Fig. 48.5.

48.2. Write down the model (48.2) in a simple case of a system with a few bodies.

48.3. Derive the equations of motion with the potential (48.6).

48.4. Generalize the mass-spring-dashpot model to arbitrary displacements.

48.5. Generalize the mass-spring model to different non-linear springs.

48.6. Model the vertical motion of a floating buoy. Hint: use that by Archimedes' Principle, the upward force on a cylindrical vertical buoy from the water is proportional to the immersed depth of the buoy.

48.7. Prove that the matrices D and K in (48.14) are symmetric positive semi-definite.

49
The Crash Model*

> On October 24, 1929, people began selling their stocks as fast as they could. Sell orders flooded market exchanges. On a normal day, only 750-800 members of the New York Stock Exchange started the Exchange. However, there were 1100 members on the floor for the morning opening. Furthermore, the Exchange directed all employees to be on the floor since there were numerous margin calls and sell orders placed overnight and extra telephone staff was arranged at the members' boxes around the floor. The Dow Jones Industrial Index closed at 299 that day. October 29 was the beginning of the Crash. Within the first few hours the stock market was open, prices fell so far as to wipe out all the gains that had been made in the previous year. The Dow Jones Industrial Index closed at 230. Since the stock market was viewed as the chief indicator of the American economy, public confidence was shattered. Between October 29 and November 13 (when stock prices hit their lowest point) over $30 billion disappeared from the American economy. It took nearly twenty-five years for many stocks to recover. (www.arts.unimelb.edu.au/amu/ucr/student/1997/Yee/1929.htm)

49.1 Introduction

Why did the Wall fall on November 9 1989? Why did the Soviet Union dissolve in January 1992? Why did the Stock market collapse in October 1929 and 1987? Why did Peter and Mary break up last Fall after 35 years of marriage? What caused the September 11 attack? Why does the flow in

the river go from orderly laminar to chaotic turbulent at a certain specific point? All the situations behind these questions share a common feature: Nothing particularly dramatic preceded the sudden transition from stable to unstable, and in each case the rapid and dramatic change away from normality came as big surprise to almost everyone.

We now describe a simple mathematical model that shows the same behavior: the solution stays almost constant for a long time and then quite suddenly the solution explodes.

We consider the following initial value problem for a system of two ordinary differential equations: find $u(t) = (u_1(t), u_2(t))$ such that

$$\begin{cases} \dot{u}_1 + \epsilon u_1 - \lambda u_2 u_1 = \epsilon & t > 0, \\ \dot{u}_2 + 2\epsilon u_2 - \epsilon u_1 u_2 = 0 & t > 0, \\ u_1(0) = 1, u_2(0) = \kappa\epsilon, \end{cases} \tag{49.1}$$

where ϵ is a small positive constant of size say 10^{-2} or smaller and λ and κ are positive parameters of moderate size ≈ 1. If $\kappa = 0$, then the solution $u(t) = (1, 0)$ is constant in time, which we view as the *base solution*. In general, for $\kappa > 0$, we think of $u_1(t)$ as a primary part of solution with initial value $u_1(0) = 1$, and $u_2(t)$ as a small secondary part with an initial value $u_2(0) = \kappa\epsilon$ that is small because ϵ is small. Both components $u_1(t)$ and $u_2(t)$ will correspond to physical quantities that are non-negative and $u_1(0) = 1$ and $u_2(0) = \kappa\epsilon \geq 0$.

49.2 The Simplified Growth Model

The system (49.1) models an interaction between a primary quantity $u_1(t)$ and a secondary quantity $u_2(t)$ through the terms $-\lambda u_1 u_2$ and $-\epsilon u_2 u_1$. If we keep just these terms, we get a simplified system of the form

$$\begin{cases} \dot{w}_1(t) = \lambda w_1(t) w_2(t) & t > 0, \\ \dot{w}_2(t) = \epsilon w_2(t) w_1(t) & t > 0, \\ w_1(0) = 1, \quad w_2(0) = \kappa\epsilon. \end{cases} \tag{49.2}$$

We see that the coupling terms are *growth terms* in the sense that both the equation $\dot{w}_1(t) = \lambda w_1(t) w_2(t)$ and $\dot{w}_2(t) = \epsilon w_2(t) w_1(t)$ say that $\dot{w}_1(t)$ and $\dot{w}_2(t)$ are positive if $w_1(t) w_2(t) > 0$. In fact, the system (49.1) always blow up for $\kappa > 0$ because the two components propel each other to infinity as t increases in the sense that the right hand sides get bigger with $w_1(t) w_2(t)$ and this increases the growth rates $\dot{w}_1(t)$ and $\dot{w}_2(t)$, which in turn makes $w_1 w_2(t)$ even bigger, and so on towards blow up, see Fig. 49.1.

We can study the blow up in (49.2) analytically assuming for simplicity that $\lambda = \kappa = 1$. In this case, it turns out that the two components $w_1(t)$

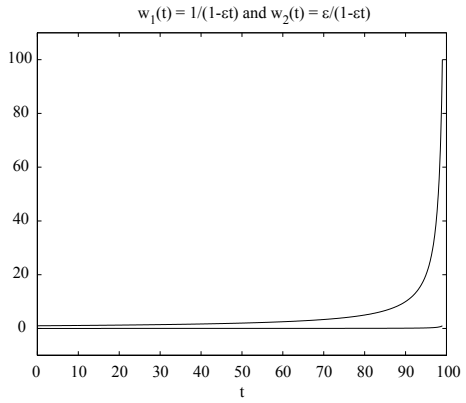

Fig. 49.1. Solution of simplified growth model

and $w_2(t)$ for all t are coupled by the relation $w_2(t) = \epsilon w_1(t)$, that is $w_2(t)$ is always the same multiple of $w_1(t)$. We check this statement by first verifying that $w_2(0) = \epsilon w_1(0)$ and then by dividing the two equations to see that $\dot{w}_2(t)/\dot{w}_1(t) = \epsilon$. So, $\dot{w}_2(t) = \epsilon \dot{w}_1(t)$, that is $w_2(t) - w_2(0) = \epsilon w_1(t) - \epsilon w_2(0)$, and we get the desired conclusion $w_2(t) = \epsilon w_1(t)$ for $t > 0$. Inserting this relation into the first equation of (49.2), we get

$$\dot{w}_1(t) = \epsilon w_1^2(t) \qquad \text{for } t > 0,$$

which can be written as

$$-\frac{d}{dt}\frac{1}{w_1(t)} = \epsilon \qquad \text{for } t > 0.$$

Recalling the initial condition $w_1(0) = 1$, we get

$$-\frac{1}{w_1(t)} = \epsilon t - 1 \qquad \text{for } t \geq 0,$$

which gives the following solution formula in the case $\lambda = \kappa = 1$:

$$w_1(t) = \frac{1}{1-\epsilon t}, \quad w_2(t) = \frac{\epsilon}{1-\epsilon t} \qquad \text{for } t \geq 0. \tag{49.3}$$

This formula shows that the solution tends to infinity as t increases towards $1/\epsilon$, that is, the solution explodes at $t = 1/\epsilon$. We notice that the time of blow up is $1/\epsilon$, and that the *time scale* before the solution starts to increase noticeably, is of size $\frac{1}{2\epsilon}$, which is a long time since ϵ is small. Thus, the solution changes very slowly for a long time and then eventually blows up quite a bit more rapidly, see Fig. 49.1.

49.3 The Simplified Decay Model

On the other hand, if we forget about the growth terms, we get another simplified system:

$$\begin{cases} \dot{v}_1 + \epsilon v_1 = \epsilon & t > 0, \quad t > 0, \\ \dot{v}_2 + 2\epsilon v_2 = 0 & t > 0, \\ v_1(0) = 1 + \delta, \quad v_2(0) = \kappa\epsilon, \end{cases} \tag{49.4}$$

where we have also introduced a small perturbation δ in $v_1(0)$. Here the two terms ϵv_1 and $2\epsilon v_2$ are so called *dissipative* terms that cause the solution $v(t)$ to return to the base solution $(1, 0)$ regardless of the perturbation, see Fig. 49.2. This is clear in the equation $\dot{v}_2 + 2\epsilon v_2 = 0$ with solution $v_2(t) = v_2(0)\exp(-2\epsilon t)$, which decays to zero as t increases. Rewriting the equation $\dot{v}_1 + \epsilon v_1 = \epsilon$ as $\dot{V}_1 + \epsilon V_1 = 0$, setting $V_1 = v_1 - 1 = \exp(-\epsilon t)$, we find that $v_1(t) = \delta\exp(-\epsilon t) + 1$, and thus $v_1(t)$ approaches 1 as t increases. We summarize: the solution $(v_1(t), v_2(t))$ of (49.4) satisfies

$$v_1(t) = \delta\exp(-\epsilon t) + 1 \to 1, \quad v_2(t) = \kappa\epsilon\exp(-2\epsilon t) \to 0, \quad \text{as } t \to \infty.$$

We say that (49.4) is a *stable* system because the solution always returns from $(1+\delta, \kappa\epsilon)$ to the base solution $(1, 0)$ independently of the perturbation $(\delta, \kappa\epsilon)$ of $(v_1(0), v_2(0)$.

We note that the time scale is again of size $1/\epsilon$, because of the presence of the factors $\exp(-\epsilon t)$ and $\exp(-2\epsilon t)$.

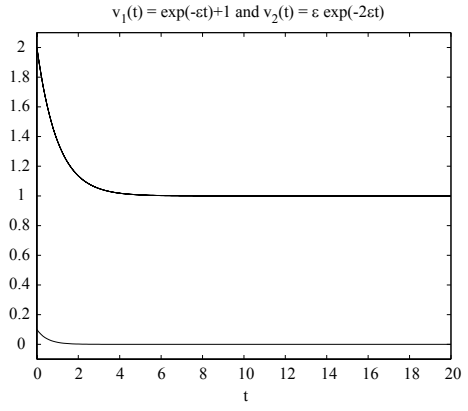

Fig. 49.2. Solution of simplified decay model

49.4 The Full Model

We can now sum up: The real system (49.1) is a combination of the unstable system (49.2) that includes only the growth terms only and whose the solution always blows up, and the stable system (49.4) that excludes the growth terms. We shall see that depending on the size of $\lambda\kappa$ the unstable or stable feature will take over. In Fig. 49.3 and Fig. 49.4, we show different solutions for different values of the parameters λ and κ with different initial values $u(0) = (u_1(0), u_2(0)) = (1, \kappa\epsilon)$. We see that if $\lambda\kappa$ is sufficiently large, then the solution $u(t)$ eventually blows up after a time of size $1/\epsilon$, while if $\lambda\kappa$ is sufficiently small, then the solution $u(t)$ returns to the base solution $(1, 0)$ as t tends to infinity.

Thus, there seems to be a *threshold value* for $\lambda\kappa$ above which the initially disturbed solution eventually blows up and below which the initially disturbed solution returns to the base solution. We can view κ as a measure of the size if the initial disturbance, because $u_2(0) = \kappa\epsilon$. Further, we can view the factor λ as a quantitative measure of the *coupling* between the growth components $u_2(t)$ and $u_1(t)$ through the growth term $\lambda u_1 u_2$ in the evolution equation for u_1.

Our main conclusion is that if the initial disturbance times the coupling is sufficiently large, then the system will blow up. Blow up thus requires both the initial disturbance and the coupling to be sufficiently large. A large initial disturbance will not cause blow up unless there some coupling. A strong coupling will not cause blow up unless there is an initial disturbance.

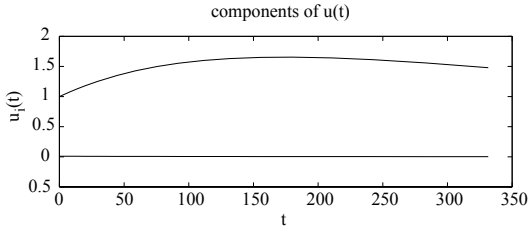

Fig. 49.3. Return to the base solution if $\lambda\kappa$ is small enough

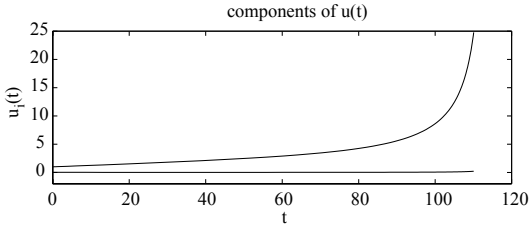

Fig. 49.4. Blow up if $\lambda\kappa$ is large enough

We now investigate the qualitative behavior of (49.1) in a little more detail. We see that $\dot{u_1}(0)/u_1(0) = \lambda\kappa\epsilon$, while $\dot{u_2}(0)/u_2(0) = -\epsilon$, which shows that initially $u_1(t)$ grows and $u_2(t)$ decays at relative rates of size ϵ. Now, $u_1(t)$ will continue to grow as long as $\lambda u_2(t) > \epsilon$, and further $u_2(t)$ will start to grow as soon as $u_1(t) > 2$. Thus, if $u_1(t)$ manages to become larger than 2, before $u_2(t)$ has decayed below ϵ/λ, then both components will propel each other to cause a blow up to infinity. This happens if $\lambda\kappa$ is above a certain threshold.

We notice that the time scale for significant changes in both u_1 and u_2 is of size ϵ^{-1}, because the growth rates are of size ϵ. This conforms with the experience from the simplified models. The scenario is thus that the primary part $u_1(t)$ grows slowly starting from 1 at a rate of size ϵ and the secondary part $u_2(t)$ decays slowly at a rate of size ϵ^2, over a time of size $1/\epsilon$. If $\lambda\kappa$ is above a certain threshold, then $u_1(t)$ reaches the value 2, at which point $u_2(t)$ starts to grow and eventually blow up follows on a somewhat shorter time scale. If $u_1(t)$ does not reach the value 2 in time, then $(u_1(t), u_2(t))$ returns to the base solution $(1, 0)$ as t increases.

We hope the presented scenario is quite easy to grasp intuitively, and conforms with every-day experiences of quit sudden blow-up, as a result of an accumulation of small events over a long period.

We can give the Crash model very many interpretations in real life, such as

- stock market (u_1 stock prize of big company, u_2 stock prize of small innovative company),
- chemical reaction (u_1 main reactant, u_2 catalyst),
- marriage crisis (u_1 main discontent, u_2 small irritation factor),
- spread of infection (u_1 infected people, u_2 amount of germs),
- symbiosis (u_1 main organism, u_2 small parasite),
- population model (u_1 rabbits, u_2 vitalizing carrots),

and many others.

In particular, the model describes an essential aspect of the process of transition from laminar to turbulent flow in for example a pipe. In this case u_1 represents a flow component in the direction of the pipe and u_2 represents a small perturbation of the flow in the transversal direction. The time to explosion corresponds to the time it takes for the flow to go turbulent starting as laminar flow at the inlet. In the famous experiment of Reynolds from 1888, ink is injected at the inlet of a transparent pipe and the observer can follow the streamline traced by the ink, which forms a straight line in the laminar part and then successively becomes more and more wavy until it breaks down to completely turbulent flow at some

distance from the inlet. The distance to breakdown varies with the flow speed and viscosity and perturbations resulting from e.g. roughness of the surface of the pipe or a heavy-weight truck passing by at some distance from the experimental set-up.

Chapter 49 Problems

49.1. Develop the indicated applications of the Crash model.

49.2. Solve the full system (49.1) numerically for various values of λ and κ and try to pin down the threshold value of $\lambda\kappa$.

49.3. Develop a *Theory of Capitalism* based on (49.1) as a simple model of the economy in a society, with u_1 representing the value of a basic resource like land, and u_2 some venture capital related to the exploitation of new technology, with $(1, 0)$ a base solution without the new technology, and with the coefficient λ of the u_1u_2 term in the first equation representing the positive interplay between base and new technology, and the terms ϵu_i representing stabilizing effects of taxes for example. Show that the possible pay-off $u_1(t) - u_1(0)$ of a small investment $u_2(0) = \kappa\epsilon$ may be large, and that an exploding economy may result if $\lambda\kappa$ is large enough. Show that no growth is possible if $\lambda = 0$. Draw some conclusions from the model coupled to for example the role of the interest rate for controlling the economy.

49.4. Interpret (49.1) as a simple model of a stock market with two stocks, and discuss scenarios of overheating. Extend to a model for the world stock market, and predict the next crash.

49.5. Consider the linear model

$$\begin{aligned} \dot\varphi_1 + \epsilon\varphi_1 - \lambda\varphi_2 &= 0 \qquad t > 0, \\ \dot\varphi_2 + \epsilon\varphi_2 &= 0 \qquad t > 0, \\ \varphi_1(0) = 0, \quad \varphi_2(0) &= \kappa\epsilon, \end{aligned} \tag{49.5}$$

which is obtained from (49.1) by setting $\varphi_1 = u_1 - 1$ and $\varphi_2 = u_2$ and replacing $u_1\varphi_2$ by φ_2 assuming u_1 is close to 1. Show that the solution of (49.5) is given by

$$\varphi_2(t) = \kappa\epsilon\exp(-\epsilon t), \quad \varphi_1(t) = \lambda\kappa\epsilon t\exp(-\epsilon t).$$

Conclude that

$$\frac{\varphi_1(\frac{1}{\epsilon})}{\varphi_2(0)} = \lambda\frac{\exp(-1)}{\epsilon},$$

and make an interpretation of this result.

49.6. Expand the Crash model (49.1) to[TS:h]

$$\begin{aligned} \dot u_1 + \epsilon u_1 - \lambda u_1u_2 + \mu_1u_2^2 &= \epsilon \qquad t > 0, \\ \dot u_2 + 2\epsilon u_2 - \epsilon u_2u_1 + \mu_2u_1^2 &= 0 \qquad t > 0, \\ u_1(0) = 1, \quad u_2(0) &= \kappa\epsilon, \end{aligned}$$

[TS:h] The following equation was numbered with (49.6) in the hardcopy, please check it.

Editor's or typesetter's annotations (will be removed before the final TEX run)

with decay terms $\mu_1 u_2^2$ and $\mu_2 u_1^2$, where μ_1 and μ_2 are positive coefficients. (a) Study the stabilizing effect of such terms numerically. (b) Seek to find values of μ_1 and μ_2, so that the corresponding solution starting close to $(1, 0)$ shows an intermittent behavior with repeated periods of blow up followed by a decay back to a neighborhood of $(1, 0)$. (c) Try to find values of μ_1 and μ_2 so that multiplication of the first equation with a positive multiple of u_1 and the second by u_2, leads to bounds on $|\epsilon u_1(t)|^2$ and $|u_2(t)|^2$ in terms of initial data. Hint: Try for example $\mu_1 \approx 1/\epsilon$, and $\mu_2 \approx \epsilon^2$.

49.7. Study the initial value problem $\dot{u} = f(u)$ for $t > 0$, $u(0) = 0$, where $f(u) = \lambda u - u^3$, with different values of $\lambda \in \mathbb{R}$. Relate the time-behavior of $u(t)$ to the set of solutions $\bar{u}$ of $f(u) = 0$, that is, $\bar{u} = 0$ if $\lambda \leq 0$, and $\bar{u} = 0$ or $\bar{u} = \pm\sqrt{\lambda}$ if $\lambda > 0$. Study the linearized models $\dot{\varphi} - \lambda\varphi + 3\bar{u}^2\varphi = 0$ for the different $\bar{u}$. Study the behavior of the solution assuming $\lambda(t) = t - 1$.

49.8. Study the model

$$\begin{aligned} \dot{w}_1 + w_1 w_2 + \epsilon w_1 &= 0, \qquad t > 0, \\ \dot{w}_2 - \epsilon w_1^2 + \epsilon w_2 &= -\gamma\epsilon, \qquad t > 0, \end{aligned} \tag{49.6}$$

with given initial data $w(0)$, where γ is a parameter and $\epsilon > 0$. This problem admits the stationary "trivial branch" solution $\bar{w} = (0, -\gamma)$ for all γ. If $\gamma > \epsilon$, then also $\bar{w} = (\pm\sqrt{\gamma - \epsilon}, -\epsilon)$ is a stationary solution. Study the evolution of the solution for different values of γ. Study the corresponding linearized problem, linearized at $\bar{w}$.

50
Electrical Circuits*

> We can scarcely avoid the conclusion that light consists in the transverse undulations of the same medium which is the cause of electric and magnetic phenomena. (Maxwell 1862)

50.1 Introduction

There is an analogy between models of masses, dashpots, and springs in mechanics and models of *electrical circuits* involving *inductors*, *resistors*, and *capacitors* respectively. The basic model for an electrical circuit, see Fig. 50.1, with these three components has the form

$$L\ddot{q}(t) + R\dot{q}(t) + \frac{q(t)}{C} = f(t), \quad \text{for} \quad t > 0, \tag{50.1}$$

together with initial conditions for q and $\dot{q}$. Here $f(t)$ represents an applied *voltage* and $q(t)$ is a primitive function of the *current* $i(t)$. The equation (50.1) says that the applied voltage $f(t)$ is equal to the sum of the *voltage drops* $L\frac{di}{dt}$, Ri and u/C across the inductor, resistor, and capacitor respectively, where L, R and C are the coefficients of *inductance*, *resistance* and *capacitance*. Note that the integral $q(t)$ of the current represents the *charge*.

The system (50.1), which is referred to as an LCR-circuit, takes the same form as the mass-dashpot-spring system (48.17), and the discussion above concerning the case $f(t) = 0$ applies to the LCR-circuit. In the absence of a resistor, a non-zero solution oscillates between extreme states with the

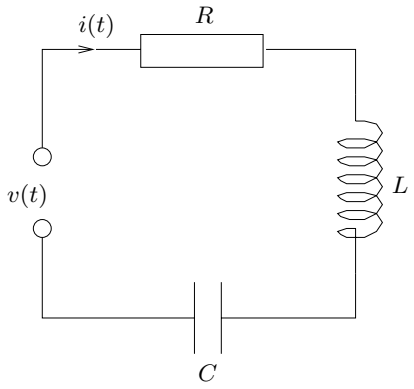

Fig. 50.1. A circuit with inductor, resistor and capacitor

charge $|u(t)|$ in the capacitor maximal and $\dot{u} = 0$ over an intermediate state with zero capacitor charge and $|\dot{u}|$ maximal. This is analogous to the mass-spring system with the potential energy corresponding to the capacitor and the velocity corresponding to $\dot{u}$. Further, the influence of a resistor is analogous to that of a dash pot, causing a damping of the oscillations.

We now proceed to describe the components of the electrical circuit in some more detail and show how complex circuits may be constructed combining the components in series or parallel in different configurations.

50.2 Inductors, Resistors and Capacitors

The voltage drop $v(t)$ across a capacitor satisfies

$$v(t) = \frac{q(t)}{C},$$

where $q(t)$ is the *charge* defined by

$$q(t) = \int_0^t i(t)\,dt,$$

assuming $q(0) = 0$, where $i(t)$ is the current and the constant C is the capacitance. Differentiating, we have

$$i(t) = \dot{q}(t) = C\dot{v}(t).$$

The voltage drop across a resistor is according to Ohm's Law $v(t) = Ri(t)$, where the constant R is the resistance. Finally, the voltage drop across an inductor is

$$v(t) = L\frac{di}{dt}(t),$$

where the constant L is the inductance.

50.3 Building Circuits: Kirchhoff's Laws

By joining the components of inductors, resistors and capacitors by electrical wires, we can build *electrical circuits*, with the wires joined at *nodes*. A closed loop in the circuit is a series of wires connecting components and nodes that leads back to the node of departure. To model the circuit, we use the two Kirchhoff laws:

- the sum of all currents entering a node is zero (first law),
- the sum of the voltage drops around any closed loop of the circuit is zero (second law).

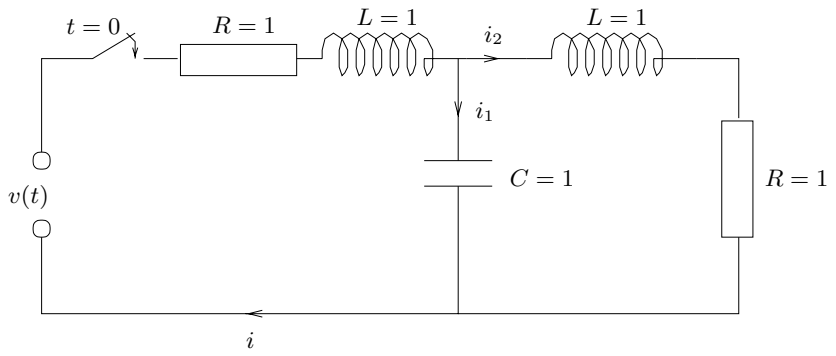

Fig. 50.2. A circuit with two loops

Example 50.1. We consider the following circuit consisting of two loops, see Fig. 50.3, and assume that $v(t) = 10$. Suppose that the switch is turned on at $t = 0$, with the charge in the capacitor being zero at $t = 0$ and $i_1(0) = i_2(0) = 0$. By Kirchhoffs second law applied to the two closed loops, we have

$$i + \frac{di}{dt} + \int_0^t i_1(s)\,ds = 10,$$
$$i_2 + \frac{i_2}{dt} - \int_0^t i_1(s)\,ds = 0.$$

Inserting Kirchhoffs first law stating that $i = i_1 + i_2$ into the first equation, and eliminating $i_2 + \frac{i_2}{dt}$ using the second equation, we get

$$i_1 + \frac{di_1}{dt} + 2\int_0^t i_1(s)\,ds = 10, \tag{50.2}$$

which upon differentiation gives the following second order equation,

$$\frac{d^2 i_1}{dt^2} + \frac{di_1}{dt} + 2i_1 = 0,$$

with characteristic equation $r^2 + r + 2 = 0$. Completing the square, we get the equation $(r+\frac{1}{2})^2 + \frac{7}{4} = 0$, and thus using the initial condition $i_1(0) = 0$, we have

$$i_1(t) = c \exp\left(-\frac{t}{2}\right) \sin\left(\frac{t\sqrt{7}}{2}\right),$$

with c a constant. Inserting into (50.2) gives $c = \frac{20}{\sqrt{7}}$, and we have determined the current $i_1(t)$ as a function of time. We can now solve for i_2 using the second loop equation.

50.4 Mutual Induction

Two inductors in different loops of the circuit may be coupled through *mutual inductance*, for instance by letting the inductors share an iron core. We give an example in Fig. 50.3.

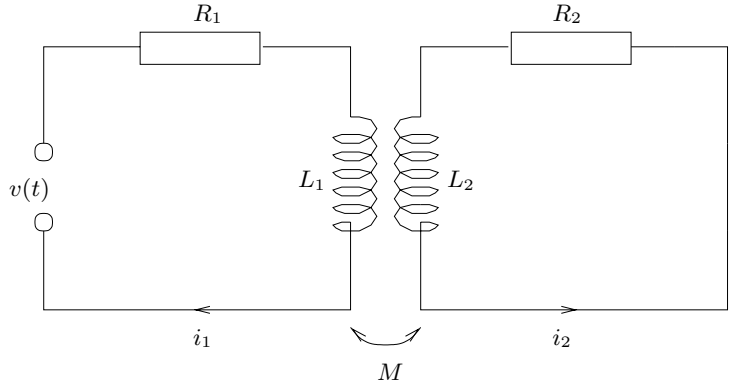

Fig. 50.3. A circuit with mutual inductance

Kirchhoffs second law applied to each of the circuits now takes the form

$$R_1 i_1 + L_1 \frac{di_1}{dt} + M \frac{di_2}{dt} = v(t),$$
$$R_2 i_2 + L_1 \frac{di_2}{dt} + M \frac{di_1}{dt} = 0,$$

where L_1 and L_2 are the inductances in the two circuits and M is the coefficient of mutual inductance.

Chapter 50 Problems

50.1. Design circuits with the indicated components. Study resonance phenomena, and amplifiers.

50.2. Study the charging of a capacitor through a resistor.

50.3. Derive the effective resistance of n resistors coupled (a) in parallel, (b) in series. Do the same for inductors and capacitors.

51
String Theory*

> It's because physicists dream of a unified theory: a single mathematical framework in which all fundamental forces and units of matter can be described together in a manner that is internally consistent and consistent with current and future observation. And it turns out that having extra dimensions of space makes it possible to build candidates for such a theory. Superstring theory is a possible unified theory of all fundamental forces, but superstring theory requires a 10 dimensional spacetime, or else bad quantum states called ghosts with unphysical negative probabilities become part of the spectrum (http://superstringtheory.com/experm/exper5.html).

51.1 Introduction

We now study a couple of basic mass-spring models with different response characteristics depending on the geometry. These simple models lead to separable equations in the phase plane. The models can generalized and coupled into systems of high complexity.

Consider a horizontal elastic string attached at $(-1, 0$ and $(1, 0)$ in a $x - y$-plane with the y-axis vertical downward, and with a body of mass 1 attached to the string at its midpoint, see Fig. 51.1. To describe the dynamics of this system, we seek the vertical force required to displace the midpoint of the string the vertical distance y. Pythagoras theorem implies that the length of half the string after displacement is equal to $\sqrt{1+y^2}$, see Fig. 51.1. The elongation is thus $\sqrt{1+y^2}-1$ and assuming that the

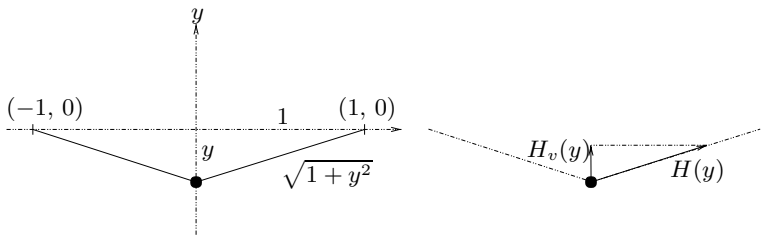

Fig. 51.1. A horizontal elastic string **fix fig: y-axis downward and x-axis**

tension in the string before displacement is H and that the string is linear elastic with constant 1, we find that the tension of the string after displacement is equal to $H + (\sqrt{1+y^2} - 1)$. By similarity, the vertical component $H_v(y)$ of the string force is

$$H_v(y) = (H + \sqrt{1+y^2} - 1)\frac{y}{\sqrt{1+y^2}},$$

and the total downward force $f(y)$ required to pull the spring downward the distance y is thus

$$f(y) = 2H_v(y), \tag{51.1}$$

where the factor 2 comes from the fact that the upward force has contributions from both sides of the string, see Fig. 51.1. Using Newton's Law, we thus obtain the following model for the mass-spring system

$$\ddot{y} + f(y) = 0,$$

if we neglect gravity forces.

51.2 A Linear System

We now assume that y so small that we can replace $\sqrt{1+y^2}$ by 1, and assuming that $H = \frac{1}{2}$, we then obtain the linear harmonic oscillator model:

$$\ddot{y} + y = 0, \tag{51.2}$$

with the solution $y(t)$ being a linear combination of $\sin(t)$ and $\cos(t)$. For example, if $y(0) = \delta$ with δ small and $\dot{y}(0) = 0$, then the solution is $y(t) = \delta\cos(t)$ corresponding to small vibrations around the rest position $y = 0$.

51.3 A Soft System

We next assume that $H = 0$ so that the string is without tension with $y = 0$. By Taylor's theorem for y small (with an error proportional to y^4),

$$\sqrt{1+y^2} \approx 1 + \frac{y^2}{2},$$

and we obtain assuming y to be small

$$f(y) \approx \frac{y^3}{\sqrt{1+y^2}} \approx y^3,$$

and are thus led to the model

$$\ddot{y} + y^3 = 0.$$

In this model the restoring force y^3 is much smaller than the y in the linear model for y small, and thus corresponds to a "soft" system.

51.4 A Stiff System

We consider now a system of two nearly vertical bars each of length one connected by a frictionless joint with the lower bar fixed to the ground with a frictionless joint and with a body of mass 1 at the top of the upper bar, see Fig. 51.2. Let y be the vertical displacement downward of the weight from the top position with the bars fully vertical, and let z be the corresponding elongation of the horizontal spring. The corresponding horizontal spring force is $H = z$ assuming the spring is Hookean with zero natural length and spring constant equal to 1. With V denoting the vertical component of the bar force, momentum balance gives, see Fig. 51.2,

$$Vz = H\left(1 - \frac{y}{2}\right) = z\left(1 - \frac{y}{2}\right),$$

and thus the vertical force reaction on the body from the spring-bar system is given by

$$f(y) = -V = -\left(1 - \frac{y}{2}\right).$$

We see that the vertical force from the spring-bar system in this system is almost constant and equal to 1 for small y. Thus the system reacts with an almost constant response, which contrasts to the linear response y and the cubic response y^3 met above. In the present context, we may refer to this response as being "stiff" in contrast to the more or less soft response of the above systems. We are thus led to the following model in the present case

$$\ddot{y} + \left(1 - \frac{y}{2}\right) = 0, \tag{51.3}$$

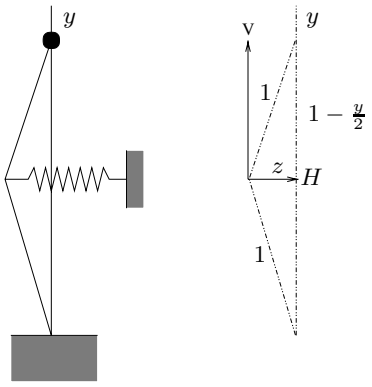

Fig. 51.2. A "stiff" bar-spring system

where we assume that $0 \leq y \leq 1$. In particular, we may consider the initial conditions $y(0) = 0$, and $\dot{y}(0) = y_0 > 0$ and follow the evolution until $y(t)$ reaches the value 1 or 0.

51.5 Phase Plane Analysis

We rewrite the second order equation $\ddot{y} + f(y) = 0$ as the first order system

$$\dot{v} + f(u) = 0, \quad \dot{u} = v$$

by setting $u = y$ and $v = \dot{y}$. The corresponding phase plane equation reads

$$\frac{du}{dv} = -\frac{v}{f(u)},$$

which is a separable equation

$$f(u)\,du = -v\,dv$$

with solution curves in the $u - v$-plane satisfying

$$F(u) + \frac{v^2}{2} = C,$$

where $F(u)$ is a primitive function of $f(u)$. With $f(u) = u$ as in the linear case, the phase plane curves are circles,

$$u^2 + v^2 = 2C.$$

In the soft case with $f(u) = u^3$, the phase plane curves are given by

$$\frac{u^4}{4} + \frac{v^2}{2} = C,$$

which represent a sort of ellipses. In the stiff case with $f(u) = (1 - \frac{u}{2})$, the phase-plane curves are given by

$$-\left(1 - \frac{u}{2}\right)^2 + \frac{v^2}{2} = C,$$

which represent hyperbolas.

Chapter 51 Problems

51.1. Compare the three systems from the point of view of efficiency as catapults, assuming the system can be "loaded" by applying a force of a certain maximal value and distance of action. Hint: Compute the work to load the system.

51.2. Design other mass-spring systems, for instance by coupling the elementary systems considered in series and parallel, and find their corresponding mathematical models. Solve the systems numerically.

51.3. Add gravity force to the above systems.

51.4. Develop an analog of the stiff system above of the form $\ddot{y} + (1 - y)$ if $0 \leq y \leq 1$ and $\ddot{y} - (1 + y) = 0$ if $-1 \leq y \leq 0$ allowing y to take arbitrary values in $-1 \leq y \leq 1$. Hint: mirror the given system.

51.5. Visualize the indicated phase-plane plots.

52
Piecewise Linear Approximation

> The beginners mind is empty, free of the habits of the expert, ready to accept, or doubt, and open to all the possibilities. It is a kind of mind which can see things as they are. (Shunryu Suzuki)

52.1 Introduction

Approximating a complicated function to arbitrary accuracy by "simpler" functions is a basic tool of applied mathematics. We have seen that piecewise polynomials are very useful for this purpose, and that is why approximation by piecewise polynomials plays a very important role in several areas of applied mathematics. For example, the *Finite Element Method* FEM is an extensively used tool for solving differential equations that is based on piecewise polynomial approximation, see the Chapters FEM for two-point boundary value problems and FEM for Poisson's equation.

In this chapter, we consider the problem of approximating a given real-valued function $f(x)$ on an interval $[a, b]$ by piecewise linear polynomials on a subdivision of $[a, b]$. We derive basic error estimates for interpolation with piecewise linear polynomials and we consider an application to least squares approximation.

52.2 Linear Interpolation on $[0, 1]$

Let $f : [0, 1] \to \mathbb{R}$ be a given Lipschitz continuous function. Consider the function $\pi f : [0, 1] \to \mathbb{R}$ defined by

$$\pi f(x) = f(0)(1 - x) + f(1)x = f(0) + (f(1) - f(0))x.$$

Clearly, $\pi f(x)$ is a *linear* function in x,

$$\pi f(x) = c_0 + c_1 x,$$

where $c_0 = f(0)$, $c_1 = f(1) - f(0)$, and $\pi f(x)$ *interpolates* $f(x)$ at the end-points 0 and 1 of the interval $[0, 1]$, by which we mean that πf takes the same values as f at the end-points, i.e.

$$\pi f(0) = f(0), \quad \pi f(1) = f(1).$$

We refer to $\pi f(x)$ as a linear *interpolant* of $f(x)$ that interpolates $f(x)$ at the end-points of the interval $[0, 1]$.

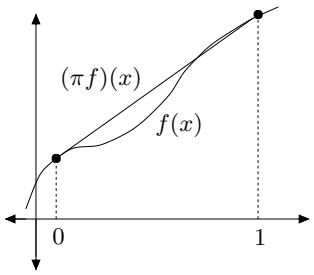

Fig. 52.1. The linear interpolant πf of a function f

We now study the *interpolation error* $f(x) - \pi f(x)$ for $x \in [0, 1]$. Before doing so we get some perspective on the space of linear functions on $[0, 1]$ to which the interpolant πf belongs.

The Space of Linear Functions

We let $\mathcal{P} = \mathcal{P}(0, 1)$ denote the set of first order (linear) polynomials

$$p(x) = c_0 + c_1 x,$$

defined for $x \in [0, 1]$, where the real numbers c_0 and c_1 are the coefficients of p. We recall that two polynomials $p(x)$ and $q(x)$ in $\mathcal{P}$ may be added to give a new polynomial $p + q$ in $\mathcal{P}$ defined by $(p + q)(x) = p(x) + q(x)$, and that a polynomial $p(x)$ in $\mathcal{P}$ may be multiplied by a scalar α to give a polynomial αp in $\mathcal{P}$ defined by $(\alpha p)(x) = \alpha p(x)$. Adding two polynomials

is carried out by adding their coefficients, and multiplying a polynomial by a real number is carried out by multiplying the coefficients by the real number.

We conclude that $\mathcal{P}$ is a vector space where each vector is a particular first order polynomial $p(x) = c_0 + c_1 x$ determined by the two real numbers c_0 and c_1. As a basis for $\mathcal{P}$ we may choose $\{1, x\}$. To see this, we note that each $p \in \mathcal{P}$ can be uniquely expressed as a linear combination of 1 and x: $p(x) = c_0 + c_1 x$, and we may refer to the pair (c_0, c_1) as the coordinates of the polynomial $p(x) = c_0 + c_1 x$ with respect to the basis $\{1, x\}$. For example, the coordinates of the polynomial $p(x) = x$ with respect to the basis $\{1, x\}$, are $(0, 1)$, right? Since there are two basis functions, we say that the dimension of the vector space $\mathcal{P}$ is equal to two.

We now consider an alternative basis $\{\lambda_0, \lambda_1\}$ for $\mathcal{P}$ consisting of the two functions λ_0 and λ_1 defined

$$\lambda_0(x) = 1 - x, \quad \lambda_1(x) = x.$$

Each of these functions takes the value 0 at one end-point and the value 1 at the other end-point, namely

$$\lambda_0(0) = 1,\ \lambda_0(1) = 0, \qquad \text{and } \lambda_1(0) = 0, \lambda_1(1) = 1.$$

See Fig. 52.2.

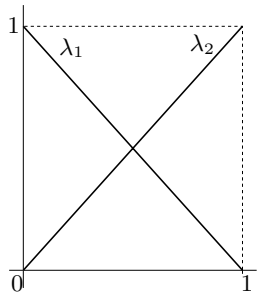

Fig. 52.2. The basis functions λ_0 and λ_1

Any polynomial $p(x) = c_0 + c_1 x$ in $\mathcal{P}$ can be expressed as a linear combination of the functions $\lambda_0(x)$ and $\lambda_1(x)$, i.e.

$$\begin{aligned} p(x) &= c_0 + c_1 x = c_0(1 - x) + (c_1 + c_0)x = c_0\lambda_0(x) + (c_1 + c_0)\lambda_1(x) \\ &= p(0)\lambda_0(x) + p(1)\lambda_1(x). \end{aligned}$$

A very nice feature of these functions is that the coefficients $p(0)$ and $p(1)$ are the values of $p(x)$ at $x = 0$ and $x = 1$. Moreover, λ_0 and λ_1 are linearly independent, since if

$$a_0\lambda_0(x) + a_1\lambda_1(x) = 0 \quad \text{for } x \in [0, 1],$$

then setting $x = 0$ and $x = 1$ shows that $a_1 = a_0 = 0$. We conclude that $\{\lambda_0, \lambda_1\}$ is a basis for $\mathcal{P}$.

In particular, we can express the interpolant $\pi f \in \mathcal{P}$ in the basis $\{\lambda_0, \lambda_1\}$ as follows:

$$\pi f(x) = f(0)\lambda_0(x) + f(1)\lambda_1(x), \tag{52.1}$$

where the end-point values $f(0)$ and $f(1)$ appear as coefficients.

The Interpolation Error

We want to estimate the interpolation error $f(x) - \pi f(x)$ for $x \in [0, 1]$. We prove that

$$|f(x) - \pi f(x)| \leq \frac{1}{2}x(1-x) \max_{y\in[0,1]} |f''(y)|, \quad x \in [0, 1]. \tag{52.2}$$

Since (convince yourself!)

$$0 \leq x(1-x) \leq \frac{1}{4} \quad \text{for } x \in [0, 1],$$

we can state the interpolation error estimate in the form

$$\max_{x\in[0,1]} |f(x) - \pi f(x)| \leq \frac{1}{8} \max_{y\in[0,1]} |f''(y)|. \tag{52.3}$$

This estimate states that the maximal value of the interpolation error $|f(x)-\pi f(x)|$ over $[0, 1]$ is bounded by a constant times the maximum value of the second derivative $|f''(y)|$ over $[0, 1]$, i.e. to the degree of concavity or convexity of f, or the amount that f curves away from being linear, see Fig. 52.3.

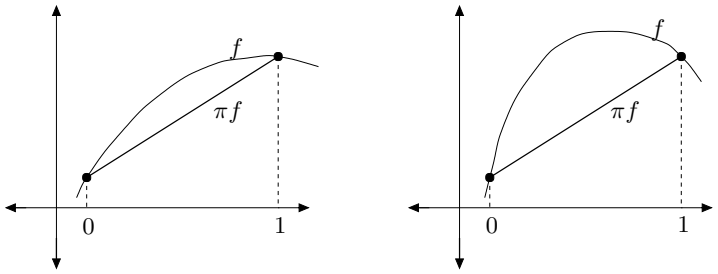

Fig. 52.3. The error of a linear interpolant depends on the size of $|f''|$, which measures the degree that f curves away from being linear. Notice that the error of the linear interpolant of the function on the *right* is much larger than of the linear interpolant of the function on the *left* and the function on the *right* has a larger second derivative in magnitude

To prove (52.2), we fix x in $(0,1)$ and use Taylor's theorem to express the values $f(0)$ and $f(1)$ in terms of $f(x)$, $f'(x)$, $f''(y_0)$ and $f''(y_1)$ where $y_0 \in (0,x)$ and $y_1 \in (x,1)$. This gives

$$\begin{aligned} f(0) &= f(x) + f'(x)(-x) + \frac{1}{2}f''(y_0)(-x)^2, \\ f(1) &= f(x) + f'(x)(1-x) + \frac{1}{2}f''(y_1)(1-x)^2. \end{aligned} \tag{52.4}$$

Substituting the Taylor expansions (52.4) into (52.1) and using the identities

$$\begin{aligned} \lambda_0(x) + \lambda_1(x) &= (1-x) + x \equiv 1, \\ (-x)\lambda_0(x) + (1-x)\lambda_1(x) &= (-x)(1-x) + (1-x)x \equiv 0, \end{aligned} \tag{52.5}$$

we obtain the *error representation*

$$f(x) - \pi f(x) = -\frac{1}{2}\big(f''(y_0)(-x)^2(1-x) + f''(y_1)(1-x)^2 x\big).$$

Using the identity $(-x)^2(1-x)+(1-x)^2x = x(1-x)(x+1-x) = x(1-x)$ gives (52.2),

$$|f(x) - \pi f(x)| \le \frac{1}{2}x(1-x)\max_{y\in[0,1]}|f''(y)| \le \frac{1}{8}\max_{y\in[0,1]}|f''(y)|. \tag{52.6}$$

Next, we prove the following estimate for the error in the first derivative,

$$|f'(x) - (\pi f)'(x)| \le \frac{x^2 + (1-x)^2}{2}\max_{y\in[0,1]}|f''(y)|, \quad x \in [0,1]. \tag{52.7}$$

Since $0 \le x^2 + (1-x)^2 \le 1$ for $x \in [0,1]$,

$$\max_{x\in[0,1]}|f'(x) - (\pi f)'(x)| \le \frac{1}{2}\max_{y\in[0,1]}|f''(y)|.$$

We illustrate in Fig. 52.4.

To prove (52.7), we differentiate (52.1) with respect to x (note that the x-dependence is carried by $\lambda_0(x)$ and $\lambda_1(x)$) and use (52.4) together with the obvious identities

$$\begin{aligned} \lambda_0'(x) + \lambda_1'(x) &= -1 + 1 \equiv 0, \\ (-x)\lambda_0'(x) + (1-x)\lambda_1'(x) &= (-x)(-1) + (1-x) \equiv 1. \end{aligned}$$

This gives the error representation:

$$f'(x) - (\pi f)'(x) = -\frac{1}{2}\big(f''(y_0)(-x)^2(-1) + f''(y_1)(1-x)^2\big),$$

where again $y_0 \in (0,x)$ and $y_1 \in (x,1)$. This proves the desired result.

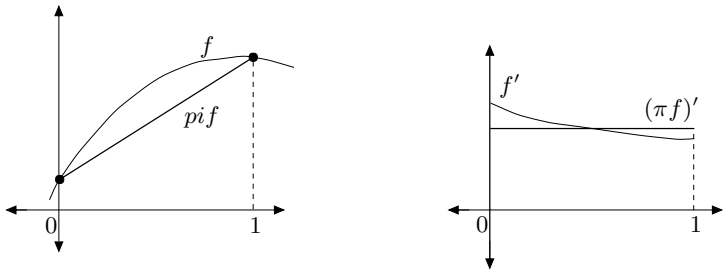

Fig. 52.4. The derivative of a linear interpolant of f approximates the derivative of f. We show f and the linear interpolant πf on the *left* and their derivatives on the *right*

Finally, we prove an estimate for $|f(x) - \pi f(x)|$ using only the first derivative f'. This is useful when the second derivative f'' does not exist. The Mean Value theorem implies

$$f(0) = f(x) + f'(y_0)(-x), \quad f(1) = f(x) + f'(y_1)(1-x), \tag{52.8}$$

where $y_0 \in [0, x]$ and $y_1 \in [x, 1]$. Substituting into (52.1), we get

$$|f(x) - \pi f(x)| = |f'(y_0)x(1-x) - f'(y_1)(1-x)x| \leq 2x(1-x) \max_{y\in[0,1]} |f'(y)|.$$

Since $2x(1-x) \leq \frac{1}{2}$ for $0 \leq x \leq 1$, we thus find that

$$\max_{x\in[0,1]} |f(x) - \pi f(x)| \leq \frac{1}{2} \max_{y\in[0,1]} |f'(y)|.$$

We summarize in the following theorem.

Theorem 52.1 *The linear polynomial $\pi f \in \mathcal{P}(0,1)$, which interpolates the given function $f(x)$ at $x = 0$ and $x = 1$, satisfies the following error bounds:*

$$\begin{aligned}
\max_{x\in[0,1]} |f(x) - \pi f(x)| &\leq \frac{1}{8} \max_{y\in[0,1]} |f''(y)|, \\
\max_{x\in[0,1]} |f(x) - \pi f(x)| &\leq \frac{1}{2} \max_{y\in[0,1]} |f'(y)|, \\
\max_{x\in[0,1]} |f'(x) - (\pi f)'(x)| &\leq \frac{1}{2} \max_{y\in[0,1]} |f''(y)|.
\end{aligned} \tag{52.9}$$

The corresponding estimates for an arbitrary interval $I = [a, b]$ of length $h = b - a$ takes the following form, where of course $\mathcal{P}(a, b)$ denotes the set of linear functions on $[a, b]$. Observe how the length $h = b-a$ of the interval enters, with the factor h^2 in the estimate for $f(x) - \pi f(x)$ with f'', and h in the estimate for $f'(x) - (\pi f)'(x)$.

Theorem 52.2 *The linear polynomial $\pi f \in \mathcal{P}(a,b)$, which interpolates the given function $f(x)$ at $x = a$ and $x = b$, satisfies the following error bounds:*

$$\begin{aligned}
\max_{x\in[a,b]} |f(x) - \pi f(x)| &\leq \frac{1}{8} \max_{y\in[a,b]} |h^2 f''(y)|,\\
\max_{x\in[a,b]} |f(x) - \pi f(x)| &\leq \frac{1}{2} \max_{y\in[a,b]} |hf'(y)|,\\
\max_{x\in[a,b]} |f'(x) - (\pi f)'(x)| &\leq \frac{1}{2} \max_{y\in[a,b]} |hf''(y)|,
\end{aligned} \tag{52.10}$$

where $h = b - a$.

If we define the *maximum norm* over $I = [a,b]$ by

$$\|v\|_{L_\infty(I)} = \max_{x\in[a,b]} |v(x)|,$$

then we can state (52.9) as follows

$$\begin{aligned}
\|f - \pi f\|_{L_\infty(I)} &\leq \frac{1}{8} \|h^2 f''\|_{L_\infty(I)},\\
\|f - \pi f\|_{L_\infty(I)} &\leq \frac{1}{2} \|hf'\|_{L_\infty(I)},\\
\|f' - (\pi f)'\|_{L_\infty(I)} &\leq \frac{1}{2}\|hf''\|_{L_\infty(I)}.
\end{aligned} \tag{52.11}$$

Below we shall use an analog of this estimate with the $L_\infty(I)$-norm replaced by the $L_2(I)$-norm.

52.3 The Space of Piecewise Linear Continuous Functions

For a given interval $I = [a,b]$, we let $a = x_0 < x_1 < x_2 < \cdots < x_N = b$ be a *partition* of I into N sub-intervals $I_i = (x_{i-1}, x_i)$ of length $h_i = x_i - x_{i-1}$, $i = 1,\ldots,N$. We denote by $h(x)$ the *mesh function* defined by $h(x) = h_i$ for $x \in I_i$ and we use $\mathcal{T}_h = \{I_i\}_{i=1}^N$ to denote the set of intervals or *mesh* or partition.

We introduce the vector space V_h of continuous piecewise linear functions on the mesh $\mathcal{T}_h$. A function $v \in V_h$ is linear on each subinterval I_i and is continuous on $[a,b]$. Adding two functions in V_h or multiplying a function in V_h by a real number gives a new function in V_h, and thus V_h is indeed a vector space. We show an example of such a function in Fig. 53.2.

We now present a particularly important basis for V_h that consists of the *hat functions* or *nodal basis functions* $\{\varphi_i\}_{i=0}^N$ illustrated in Fig. 52.5.

The hat-function $\varphi_i(x)$ is a function in V_h satisfying

$$\varphi_i(x_j) = 1 \quad \text{if } j = i, \quad \varphi_i(x_j) = 0 \quad \text{if } j \neq i.$$

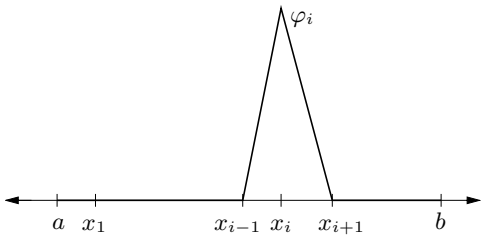

Fig. 52.5. The hat function φ_i associated to node x_i

and is given by the formula:

$$\varphi_i(x) = \begin{cases} 0, & x \notin [x_{i-1}, x_{i+1}], \\ \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, & x \in [x_{i-1}, x_i], \\ \dfrac{x - x_{i+1}}{x_i - x_{i+1}}, & x \in [x_i, x_{i+1}]. \end{cases}$$

The basis functions φ_0 and φ_N associated to the boundary nodes x_0 and x_N look like "half hats". Observe that each hat function $\varphi_i(x)$ is defined on the whole interval $[a, b]$ and takes the value zero outside the interval $[x_{i-1}, x_{i+1}]$ (or $[a, x_1]$ if $i = 0$ and $[x_{N-1}, b]$ if $i = N$).

The set of hat-functions $\{\varphi_i\}_{i=0}^N$ is a basis for V_h because each $v \in V_h$ has the unique representation

$$v(x) = \sum_{i=0}^{N} v(x_i)\varphi_i(x),$$

where the nodal values $v(x_i)$ appear as coefficients. To see this, it is sufficient to realize that the functions on the left and right hand side are both continuous and piecewise linear and take the same values at the nodes, and thus coincide. Since the number of basis functions φ_i is equal to $N + 1$, the dimension of V_h is equal to $N + 1$.

The continuous piecewise linear interpolant $\pi_h f \in V_h$ of a given Lipschitz continuous function $f(x)$ on $[0, 1]$ is defined by

$$\pi_h f(x_i) = f(x_i) \quad \text{for } i = 0, 1, \ldots, N,$$

that is, $\pi_h f(x)$ interpolates $f(x)$ at the nodes x_i, see Fig. 52.6. We can express $\pi_h f$ in terms of the basis of hat functions $\{\varphi_i\}_{i=0}^N$ as follows:

$$\pi_h f = \sum_{i=0}^{N} f(x_i)\varphi_i \quad \text{or} \quad \pi_h f(x) = \sum_{i=0}^{N} f(x_i)\varphi_i(x) \quad \text{for } x \in [0, 1], \tag{52.12}$$

with the x-dependence indicated.

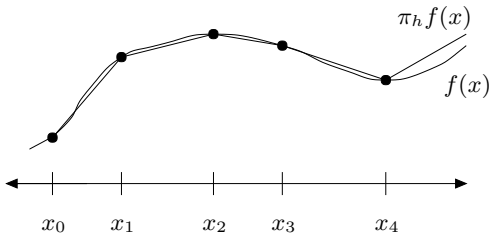

Fig. 52.6. An example of a continuous piecewise linear interpolant

Since $\pi_h f(x)$ is linear on each subinterval I_i and interpolates $f(x)$ at the end-points of I_i, we can express $\pi f(x)$ analytically on I_i as follows:

$$\pi_h f(x) = f(x_{i-1})\frac{x - x_i}{x_{i-1} - x_i} + f(x_i)\frac{x - x_{i-1}}{x_i - x_{i-1}} \quad \text{for } x_{i-1} \le x \le x_i,$$

for $i = 1, \ldots, N$.

Using Theorem 52.2, we obtain the following error estimate for piecewise linear interpolation:

Theorem 52.3 *The piecewise linear interpolant $\pi_h f(x)$ of a twice differentiable function $f(x)$ on a partition of $[a, b]$ with mesh function $h(x)$ satisfies*

$$\begin{aligned} \|f - \pi_h f\|_{L_\infty(a,b)} &\le \frac{1}{8}\|h^2 f''\|_{L_\infty(a,b)}, \\ \|f' - (\pi_h f)'\|_{L_\infty(a,b)} &\le \frac{1}{2}\|h f''\|_{L_\infty(a,b)}. \end{aligned} \tag{52.13}$$

If $f(x)$ is only once differentiable, then

$$\|f - \pi_h f\|_{L_\infty(a,b)} \le \frac{1}{2}\|h f'\|_{L_\infty(a,b)}. \tag{52.14}$$

Note that since the mesh function $h(x)$ may have jumps at the nodes, we interpret $\|h^2 f''\|_{L_\infty(a,b)}$ as

$$\max_{i=1,\ldots,N} \max_{y \in [x_{i-1}, x_i]} |h^2(y) f''(y)|,$$

where $h(y) = x_i - x_{i-1}$ for $y \in [x_{i-1}, x_i]$.

52.4 The L_2 Projection into V_h

Let $f(x)$ be a given function on an interval $I = [a, b]$ and V_h denote the space of continuous piecewise linear functions V_h on a partition $a = x_0 < \ldots < x_N = b$ of I with mesh function $h(x)$.

The *orthogonal projection* $P_h f$ of the function f into V_h is the function $P_h f \in V_h$ such that

$$\int_I (f - P_h f)v \, dx = 0 \quad \text{for } v \in V_h. \tag{52.15}$$

Recalling the definition of the $L_2(I)$-scalar product

$$(v, w)_{L_2(I)} = \int_I v(x)w(x) \, dx,$$

with the corresponding $L_2(I)$-norm

$$\|v\|_{L_2(I)} = \left(\int_I v^2(x) \, dx \right)^{1/2},$$

we can write (52.15) in the form

$$(f - P_h f, v)_{L_2(I)} = 0 \quad \text{for } v \in V_h.$$

This says that $f - P_h f$ is orthogonal to V_h with respect to the $L_2(I)$ scalar product. We also call $P_h f$ the $L_2(I)$-*projection* of f onto V_h.

We first show that $P_h f$ is uniquely defined and then prove that $P_h f$ is the best V_h-approximation of f in the $L_2(I)$-norm.

To prove uniqueness and existence, we express $P_h f$ in the nodal basis $\{\varphi_i\}_{i=0}^N$:

$$P_h f(x) = \sum_{j=0}^N c_j \varphi_j(x),$$

where the $c_j = (P_h f)(x_j)$ are the nodal values of $P_h f$ that have to be determined. We insert this representation into (52.15) and choose $v = \varphi_i$ with $i = 0, \ldots, N$, to get for $i = 0, \ldots, N$,

$$\int_I \sum_{j=0}^N c_j \varphi_j(x)\varphi_i(x) \, dx = \sum_{j=0}^N c_j \int_I \varphi_j(x)\varphi_i(x) \, dx$$
$$= \int_I f\varphi_i \, dx \equiv b_i \,, \tag{52.16}$$

where we changed the order of integration and summation. This gives the following system of equations

$$\sum_{j=0}^N m_{ij} c_j = \int_I f\varphi_i \, dx \equiv b_i \quad i = 0, 1, \ldots, N, \tag{52.17}$$

where

$$m_{ij} = \int_I \varphi_j(x)\varphi_i(x) \, dx, \quad i, j = 0, \ldots, N.$$

We can write (52.17) in matrix form as

$$Mc = b$$

where $c = (c_0, \ldots, c_N)$ is a $N+1$-vector of the unknown coefficients c_j, and $b = (b_0, \ldots, b_N)$ is computable from $f(x)$, and $M = (m_{ij})$ is a $(N+1)\times(N+1)$-matrix that depends on the basis functions φ_i, but not on the function $f(x)$. We refer to the matrix M as the *mass matrix*.

We can now easily prove the uniqueness of $P_h f$. Since the difference $P_h f - \bar{P}_h f$ of two functions $P_h f \in V_h$ and $\bar{P}_h f \in V_h$ satisfying the relation (52.15), also satisfy

$$\int_I (P_h f - \bar{P}_h f) v \, dx = 0 \quad \text{for } v \in V_h,$$

by choosing $v = P_h f - \bar{P}_h f$, we get

$$\int_I (P_h f - \bar{P}_h f)^2 \, dx = 0,$$

and thus $P_h f(x) = \bar{P}_h f(x)$ for $x \in I$. Solutions of the system $Mc = b$ are therefore unique, and since M is a square matrix, existence follows from the Fundamental Theorem of Linear Algebra. We sum up:

Theorem 52.4 *The $L_2(I)$-projection $P_h f$ of a given function f onto the set of piecewise linear functions V_h on I is uniquely defined by* (52.15) *or the equivalent system of equations $Mc = b$, where $c_j = P_h f(x_j)$ are the nodal values of $P_h f$, M is the mass matrix with coefficients $m_{ij} = (\varphi_j, \varphi_i)_{L_2(I)} = (\varphi_i, \varphi_j)_{L_2(I)}$ and the coefficients of the right hand side b are given by $b_i = (f, \varphi_i)$.*

Example 52.1. We compute the mass matrix M in the case of a uniform subdivision with $h(x) = h = (b-a)/N$ for $x \in I$. We get by a direct computation

$$m_{ii} = \int_{x_{i-1}}^{x_{i+1}} \varphi_i^2(x) \, dx = \frac{2h}{3} \quad i = 1, \ldots N-1, \qquad m_{00} = m_{NN} = \frac{h}{3},$$

$$m_{i,i+1} = \int_{x_{i-1}}^{x_{i+1}} \varphi_i(x)\varphi_{i+1}(x) \, dx = \frac{h}{6} \quad i = 1, \ldots N-1.$$

The corresponding "lumped" mass matrix $\hat{M} = (\hat{m}_{ij})$, which is a diagonal matrix with the diagonal element in each row being the sum of the elements in the corresponding row of M, takes the form

$$\hat{m}_{ii} = h \quad i = 1, \ldots, N-1, \quad \hat{m}_{00} = \hat{m}_{NN} = h/2.$$

We see that $\hat{M}$ may be viewed as a h-scaled variant of the identity matrix and M can be viewed as an h-scaled approximation of the identity matrix.

We now prove that the $L_2(I)$-projection $P_h f$ of a function f satisfies

$$\|f - P_h f\|_{L_2(I)} \leq \|f - v\|_{L_2(I)}, \quad \text{for all } v \in V_h. \tag{52.18}$$

This implies that $P_h f$ is the element in V_h with smallest deviation from f in the $L_2(I)$-norm. Applying Cauchy's inequality to (52.15) with $v \in V_h$ gives

$$\begin{aligned}
&\int_I (f - P_h f)^2 \, dx \\
&= \int_I (f - P_h f)(f - P_h f) \, dx + \int_I (f - P_h f)(P_h f - v) \, dx \\
&= \int_I (f - P_h f)(f - v) \, dx \leq \left(\int_I (f - P_h f)^2 \, dx \right)^{1/2} \left(\int_I (f - v)^2 \, dx \right)^{1/2},
\end{aligned}$$

which proves the desired result. We summarize:

Theorem 52.5 *The $L_2(I)$-projection P_h into V_h defined by* (52.15)*, is the unique element in V_h which minimizes $\|f - v\|_{L_2(I)}$ with v varying over V_h.*

In particular, choosing $v = \pi_h f$ in (52.18), we obtain

$$\|f - P_h f\|_{L_2(I)} \leq \|f - \pi_h f\|_{L_2(I)},$$

where $\pi_h f$ is the nodal interpolant of f introduced above. One can prove the following analog of (52.13)

$$\|f - \pi_h f\|_{L_2(I)} \leq \frac{1}{\pi^2} \|h^2 f''\|_{L_2(I)},$$

where the interpolation constant happens to be π^{-2}. We thus conclude the following basic result:

Theorem 52.6 *The $L_2(I)$-projection P_h into the space of piecewise linear functions V_h on I with mesh function $h(x)$, satisfies the following error estimate:*

$$\|f - P_h f\|_{L_2(I)} \leq \frac{1}{\pi^2} \|h^2 f''\|_{L_2(I)}. \tag{52.19}$$

Chapter 52 Problems

52.1. Give a different proof of the first estimate of Theorem[TS a] Theorem 52.1 by considering for a given $x \in (0, 1)$, the function

$$g(y) = f(y) - \pi f(y) - \gamma(x) y(1 - y), \quad y \in [0, 1],$$

where $\gamma(x)$ is chosen so that $g(x) = 0$. Hint: the function $g(y)$ vanishes at 0, x and 1. Show by repeated use of the Mean Value theorem that g'' vanishes at some point ξ, from which it follows that $\gamma(x) = -f''(\xi)/2$.

[TS a] Please check it.

Editor's or typesetter's annotations (will be removed before the final TEX run)

52.2. Prove Theorem 52.2 from Theorem 52.1 by using the change of variables $x = a + (b-a)z$ transforming the interval $[0,1]$ onto $[a,b]$, setting $F(z) = f(a+(b-a)z)$ and using that by the Chain Rule, $F' = \frac{dF}{dz} = (b-a)f' = (b-a)\frac{df}{dx}$.

52.3. Develop approximation/interpolation with piecewise constant (discontinuous) functions on a partition of an interval. Consider interpolation at left-hand endpoint, right-hand endpoint, midpoint and mean value for each subinterval. Prove error estimates of the form $\|u - \pi_h u\|_{L_\infty(I)} \leq C\|hu'\|_{L_\infty(I)}$, with $C = 1$ or $C = \frac{1}{2}$.

53

FEM for Two-Point Boundary Value Problems

> The results, however, of the labour and invention of this century are not to be found in a network of railways, in superb bridges, in enormous guns, or in instantaneous communication. We must compare the social state of the inhabitants of the country with what it was. The change is apparent enough. The population is double what it was a century back; the people are better fed and better housed, and comforts and even luxuries that were only within the reach of the wealthy can now be obtained by all classes alike. . . . But with these advantages there are some drawbacks. These have in many cases assumed national importance, and it has become the province of the engineer to provide a remedy. (Reynolds, 1868)

53.1 Introduction

We begin by deriving a model that is based on a *conservation principle* which states:

> The rate at which a specified quantity changes in a region is equal to the rate that the quantity leaves and enters the region plus the rate at which the quantity is created and destroyed inside the region.

Such a conservation principle holds for a wide variety of quantities, including animals, automobiles, bacteria, chemicals, fluids, heat and energy, etc. So the model we derive has a wide application.

In this chapter, we assume that the quantity to be modeled exists in a very small diameter "tube" with constant cross section and that the quantity varies in the direction along the tube but not at all within a fixed cross section, see Fig. 53.1. We use x to denote the position along the length of the tube and let t denote time. We assume that the quantity in the tube is sufficiently abundant that it makes sense to talk about a *density* $u(x,t)$, measured in amount of the quantity per unit volume, that varies continuously with the position x and time t. This is certainly valid for quantities such as heat and energy, and may be more or less valid for quantities such as bacteria and chemicals provided there is a sufficient number of creatures or molecules respectively.

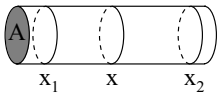

Fig. 53.1. Variation in a very narrow "tube"

We next express the conservation principle mathematically. We consider a small region of the tube of width dx and cross-sectional area A. The amount of quantity in this region is $u(x,t)A\,dx$. We let $q(x,t)$ denote the *flux* at position x and time t, or the amount of the quantity crossing the section at x at time t measured in amount per unit area per unit time. We choose the orientation so that q is positive when the flow is to the right. The amount of quantity crossing the section at position x at time t is therefore $Aq(x,t)$. Lastly, we let $f(x,t)$ denote the rate that the quantity is created or destroyed within the section at x at time t measured in amount per unit volume per unit time. So, $f(x,t)Adx$ is the amount of the quantity created or destroyed in the small region of width dx per unit time.

The conservation principle for a fixed length of pipe between $x = x_1$ and $x = x_2$ implies that the rate of change of the quantity in this section must equal the rate at which it flows in at $x = x_1$ minus the rate at which it flows out at $x = x_2$ plus the rate at which it is created in $x_1 \le x \le x_2$. In mathematical terms,

$$\frac{\partial}{\partial t}\int_{x_1}^{x_2} u(x,t)A\,dx = Aq(x_1,t) - Aq(x_2,t) + \int_{x_1}^{x_2} f(x,t)A\,dx.$$

or

$$\frac{\partial}{\partial t}\int_{x_1}^{x_2} u(x,t)\,dx = q(x_1,t) - q(x_2,t) + \int_{x_1}^{x_2} f(x,t)\,dx. \tag{53.1}$$

Equation (53.1) is called the *integral formulation* of the conservation principle.

We can reformulate (53.1) as a partial differential equation provided $u(x,t)$ and $q(x,t)$ are sufficiently smooth. For we can write,

$$\frac{\partial}{\partial t}\int_{x_1}^{x_2} u(x,t)\,dx = \int_{x_1}^{x_2} \frac{\partial}{\partial t}u(x,t)\,dx,$$
$$q(x_1,t) - q(x_2,t) = \int_{x_1}^{x_2} \frac{\partial}{\partial x}q(x,t)\,dx,$$

and therefore collecting terms,

$$\int_{x_1}^{x_2} \left(\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}q(x,t) - f(x,t)\right)\,dx = 0.$$

Since x_1 and x_2 are arbitrary, the integrand must be zero at each point, or

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}q(x,t) = f(x,t). \tag{53.2}$$

Equation (53.2) is the *pointwise* or *differential formulation* of the conservation principle.

So far we have one equation for two unknowns. To complete the model, we use a *constitutive relation* that describes the relation between the flux and the density. This relation is specific to the physical properties of the quantity being modeled, yet it is often unclear exactly how to model these properties. A constitutive relation used in practice is often only an approximation to the true unknown relation.

Many quantities have the property that the quantity flows from regions of high concentration to regions of low concentration, and the rate of flow increases as the differences in concentration increases. As a first approximation, we assume a simple linear relation

$$q(x,t) = -a(x,t)\frac{\partial}{\partial x}u(x,t), \tag{53.3}$$

where $a(x,t) > 0$ is the *diffusion coefficient*. In case u represents heat, (53.3) is known as *Newton's Heat Law*. In general, equation (53.3) is known as *Fick's Law*. Note that the choice of sign of a guarantees for example that flow is to the right if $u_x < 0$, i.e. if u decreases across the section at x. Substituting (53.3) into (53.2), we obtain the general time-dependent reaction-diffusion equation,

$$\frac{\partial}{\partial t}u(x,t) - \frac{\partial}{\partial x}\left(a(x,t)\frac{\partial}{\partial x}u(x,t)\right) = f(x,t).$$

To simplify the notation, we use $\dot{u}$ to denote $\partial u/\partial t$ and u' to denote $\partial u/\partial x$. This yields

$$\dot{u}(x,t) - (a(x,t)u'(x,t))' = f(x,t). \tag{53.4}$$

Convection or transport is another important process to take into account in this model.

Example 53.1. When modeling populations of animals, diffusion reflects the natural tendency of most creatures to spread out over a region due to randomly occurring interactions between pairs of creatures, while convection models phenomena such as migration.

Convection is modeled by assuming a constitutive relation in which the flux is proportional to the density, i.e.

$$\varphi(x,t) = b(x,t)u(x,t),$$

which results in a convection term in the differential equation of the form $(bu)'$. The convection coefficient $b(x,t)$ determines the rate and direction of transport of the quantity being modeled.

In general, many quantities are modeled by a constitutive relation of the form

$$\varphi(x,t) = -a(x,t)u'(x,t) + b(x,t)u(x,t)$$

which combines diffusion and convection. Arguing as above, we obtain the general reaction-diffusion-convection equation

$$\dot{u}(x,t) - (a(x,t)u'(x,t))' + (b(x,t)u(x,t))' = f(x,t). \tag{53.5}$$

53.2 Initial Boundary-Value Problems

We have to add suitable data to (53.4) or (53.5) in order to specify a unique solution. We model the amount of substance in a fixed length of tube located between $x = 0$ and $x = 1$, as in Fig. 53.1, and specify some information about u called *boundary conditions* at $x = 0$ and $x = 1$. We also need to give some initial data at some initial time, which we take to be $t = 0$. The *evolutionary* or *time-dependent initial two point boundary value problem* reads: find $u(x,t)$ such that

$$\begin{cases} \dot{u} - (au')' + (bu)' = f & \text{in } (0,1)\times(0,T), \\ u(0,t) = u(1,t) = 0 & \text{for } t \in (0,T) \\ u(x,0) = u_0(x) & \text{for } x \in (0,1), \end{cases} \tag{53.6}$$

where a, b, c are given coefficients and f and g are given data. The boundary values $u(0,t) = u(1,t) = 0$ are known as *homogeneous Dirichlet boundary conditions.*

Example 53.2. In the case that we use (53.4) to model the heat u in a long thin wire, the coefficient a represents the heat conductivity of the metal in the wire, f is a given heat source, and the homogeneous Dirichlet boundary conditions at the end-points means that the temperature of the wire is held fixed at 0 there. Such conditions are realistic for example if the wire is attached to very large masses at the ends.

Other boundary conditions found in practice include: *nonhomogeneous Dirichlet boundary conditions* $u(0) = u_0$, $u(1) = u_1$ with constants u_0, u_1; one homogeneous Dirichlet $u(0) = 0$ and one *nonhomogeneous Neumann boundary condition* $a(1)u'(1) = g_1$ with constant g_1; and more general *Robin boundary conditions*

$$-a(0)u'(0) = \gamma(0)(u_0 - u(0)), \quad a(1)u'(1) = \gamma(1)(u_1 - u(1))$$

with constants $\gamma(0)$, u_0, $\gamma(1)$, u_1.

53.3 Stationary Boundary Value Problems

In many situations, u is independent of time and the model reduces to the *stationary* reaction-diffusion equation

$$-(a(x)u'(x))' = f(x) \tag{53.7}$$

in the case of pure diffusion and

$$-(a(x)u'(x))' + (b(x)u(x))' = f(x) \tag{53.8}$$

in case there is convection as well. For these problems, we only need to specify boundary conditions. For example, we consider the *two-point boundary value problem*: find the function $u(x)$ satisfying

$$\begin{cases} -(au')' = f & \text{in } (0,1), \\ u(0) = u(1) = 0 \end{cases} \tag{53.9}$$

and when there is convection: find $u(x)$ such that

$$\begin{cases} -(au')' + (bu)' = f & \text{in } (0,1), \\ u(0,t) = u(1,t) = 0. \end{cases} \tag{53.10}$$

53.4 The Finite Element Method

We begin the discussion of discretization by studying the simplest model above, namely the two-point boundary value problem for the stationary reaction-diffusion model (53.9).

We can express the solution $u(x)$ of (53.9) analytically in terms of data by integrating twice (setting $w = au'$)

$$u(x) = \int_0^x \frac{w(y)}{a(y)}\,dy + \alpha_1, \quad w(y) = -\int_0^y f(z)\,dz + \alpha_2,$$

where the constants α_1 and α_2 are chosen so that $u(0) = u(1) = 0$. We can use this solution formula to compute the value of the solution $u(x)$ for any given $x \in (0,1)$ by evaluating the integrals analytically or numerically using quadrature. However, this is very time consuming if we want the solution at many points in $[0,1]$. This motivates consideration of an alternative way of computing the solution $u(x)$ using the *Finite Element Method* (FEM), which is a general method for solving differential equations numerically. FEM is based on rewriting the differential equation in *variational form* and seeking an approximate solution as a piecewise polynomial.

Note that we do not use the solution by integration outlined above, one important consequence of that procedure is that u is "twice as differentiable" as the data f, since we integrate twice to get from f to u.

We present FEM for (53.9) based on continuous piecewise linear approximation. We let $\mathcal{T}_h : 0 = x_0 < x_1 < \ldots < x_{M+1} = 1$, be a *partition* (or *triangulation*) of $I = (0,1)$ into sub-intervals $I_j = (x_{j-1}, x_j)$ of length $h_j = x_j - x_{j-1}$. We look for an approximate solution in the set V_h of continuous piecewise linear functions $v(x)$ on $\mathcal{T}_h$ such that $v(0) = 0$ and $v(1) = 0$. We show an example of such a function in Fig. 53.2. In Chapter 52, we saw that V_h is a finite dimensional vector space of dimension M with a basis consisting of the hat functions $\{\varphi_j\}_{j=1}^M$ illustrated in Fig. 52.5, associated with the *interior nodes* $x_1, \cdots, x_M$. The coordinates of a function v in V_h in this basis are the values $v(x_j)$ at the interior nodes since a function $v \in V_h$ can be written

$$v(x) = \sum_{j=1}^{M} v(x_j)\varphi_j(x).$$

Note that because $v \in V_h$ is zero at 0 and 1, we do not include φ_0 and φ_{M+1} in the set of basis functions for V_h.

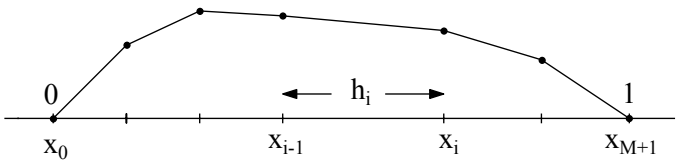

Fig. 53.2. A continuous piecewise linear function in V_h

The finite element method is based on restating the differential equation $-(au')' = f$ in an average or *variational* form

$$-\int_0^1 (au')'v\,dx = \int_0^1 fv\,dx, \tag{53.11}$$

where the function v varies over an appropriate set of *test functions*. The variational form results from multiplying the differential equation

$-(au')' = f$ by the test function $v(x)$ and integrating over the interval $(0,1)$. The variational formulation says that the residual $-(au')' - f$ of the true solution is orthogonal to all test functions v with respect to the $L_2(0,1)$ scalar product.

The basic idea of FEM is to compute an approximate solution $U \in V_h$ that satisfies (53.11) for a restricted set of test functions. This approach to computing an approximate solution is known as the *Galerkin method* in memory of the Russian engineer and scientist Galerkin (1871-1945), see Fig. 53.3. He invented his method while imprisoned for anti-Tsarist activities during 1906-7. We call the set V_h, where we seek the FEM-solution U, the *trial space* and we call the space of test functions the *test space*. In the present case of homogeneous Dirichlet boundary conditions, we usually choose the test space to be equal to V_h. Consequently, the dimensions of the trial and test spaces are equal, which is necessary for the existence and uniqueness of the approximate solution U.

Fig. 53.3. Boris Galerkin, inventor of the Finite Element Method: "It is really quite simple; just multiply by $v(x)$ and then integrate"

However since the functions in V_h do not have second derivatives, we can not simply plug a potential approximate solution U in V_h directly into (53.11). To get around this difficulty, we use integration by parts to move one derivative from $(au')'$ onto v, noting that functions in V_h are piecewise differentiable. Assuming v is differentiable and $v(0) = v(1) = 0$:

$$-\int_0^1 (au')'v\,dx = -a(1)u'(1)v(1) + a(0)u'(0)v(0) + \int_0^1 au'v'\,dx$$
$$= \int_0^1 au'v'\,dx.$$

This leads to the *continuous Galerkin finite element method of order* 1 (*cG(1)-method*) for (53.9): compute $U \in V_h$ such that

$$\int_0^1 aU'v'\,dx = \int_0^1 fv\,dx \quad \text{for all } v \in V_h. \tag{53.12}$$

We note that the derivatives U' and v' of the functions U and $v \in V_h$ are piecewise constant functions of the form depicted in Fig. 53.4 and are not

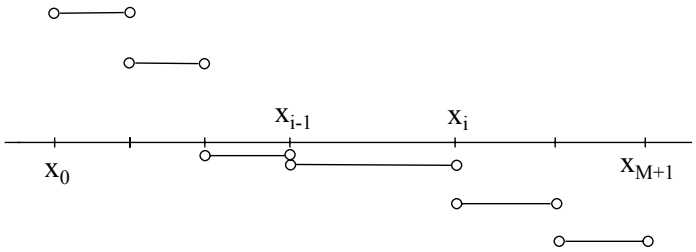

Fig. 53.4. The derivative of the continuous piecewise linear function in Fig. 53.2

defined at the nodes x_i. However, the value of an integral is independent of the value of the integrand at isolated points. Therefore, the integral (53.12) with integrand $aU'v'$ is uniquely defined as the sum of the integrals over the sub-intervals I_j.

Discretization of the Stationary Reaction-Diffusion-Convection Problem

To solve (53.10) numerically let $0 = x_0 < x_1 < \ldots < x_{L+1} = 1$ be a partition of $(0,1)$, and let V_h be the corresponding space of continuous piecewise linear functions $v(x)$ such that $v(0) = v(1) = 0$. The cG(1) FEM for (53.10) takes the form: compute $U \in V_h$ such that

$$\int_0^1 (aU')v' + (bU)'v)\,dx = \int_0^1 fv\,dx \quad \text{for all } v \in V_h.$$

53.5 The Discrete System of Equations

We have not yet proved that the set of equations (53.12) has a unique solution nor discussed what is involved in computing the solution U. This is an important issue considering we constructed the FEM precisely because the original problem is likely impossible to solve analytically.

We prove that the cG(1)-method (53.12) corresponds to a square linear system of equations for the unknown nodal values $\xi_j = U(x_j)$, $j =$

$1, \ldots, M$. We write U using the basis of hat functions as

$$U(x) = \sum_{j=1}^{M} \xi_j \varphi_j(x) = \sum_{j=1}^{M} U(x_j)\varphi_j(x).$$

Substituting into (53.12), we change the order of summation and integration to obtain

$$\sum_{j=1}^{M} \xi_j \int_0^1 a\varphi_j' v' \, dx = \int_0^1 f v \, dx, \tag{53.13}$$

for all $v \in V_h$. Now, it suffices to check (53.13) with v varying over the set of basis functions $\{\varphi_i\}_{i=1}^M$, since any function in V_h can be expressed as a linear combination of the basis functions. We are thus led to the $M \times M$ linear system of equations

$$\sum_{j=1}^{M} \xi_j \int_0^1 a\varphi_j' \varphi_i' \, dx = \int_0^1 f\varphi_i \, dx, \quad i = 1, \ldots, M, \tag{53.14}$$

for the unknown coefficients $\xi_1, \ldots, \xi_M$. We let $\xi = (\xi_1, \ldots, \xi_M)^\top$ denote the M-vector of unknown coefficients and define the $M \times M$ *stiffness matrix* $A = (a_{ij})$ with elements

$$a_{ij} = \int_0^1 a\varphi_j' \varphi_i' \, dx, \qquad i, j = 1, \ldots, M,$$

and the *load vector* $b = (b_i)$ with

$$b_i = \int_0^1 f\varphi_i \, dx, \qquad i = 1, \ldots, M.$$

These names originate from early applications of the finite element method in *structural mechanics* describing deformable structures like the body and wing of an aircraft or buildings. Using this notation, (53.14) is equivalent to the system of linear equations

$$A\xi = b. \tag{53.15}$$

In order to solve for the unknown vector ξ of nodal values of U, we first have to compute the stiffness matrix A and the load vector b. In the first instance, we assume that $a(x) = 1$ for $x \in [0, 1]$. We note that a_{ij} is zero unless $i = j-1$, $i = j$, or $i = j+1$ because otherwise either $\varphi_i(x)$ or $\varphi_j(x)$ is zero on each sub-interval occurring in the integration. We illustrate this in Fig. 53.5. We compute a_{ii} first. Using the definition of the hat function φ_i,

$$\varphi_i(x) = \begin{cases} (x - x_{i-1})/h_i, & x_{i-1} \le x \le x_i, \\ (x_{i+1} - x)/h_{i+1}, & x_i \le x \le x_{i+1}, \\ 0, & \text{elsewhere}, \end{cases}$$

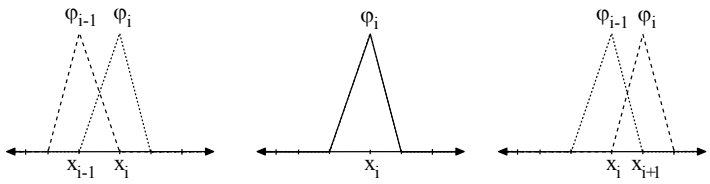

Fig. 53.5. Three possibilities to obtain a non-zero element in the stiffness matrix

the integration breaks down into two integrals:

$$a_{ii} = \int_{x_{i-1}}^{x_i} \left(\frac{1}{h_i}\right)^2 dx + \int_{x_i}^{x_{i+1}} \left(\frac{-1}{h_{i+1}}\right)^2 dx = \frac{1}{h_i} + \frac{1}{h_{i+1}} \text{ for } i = 1, 2, \ldots, M,$$

since $\varphi_i' = 1/h_i$ on (x_{i-1}, x_i) and $\varphi_i' = -1/h_{i+1}$ on (x_i, x_{i+1}) and φ_i is zero on the other sub-intervals. Similarly,

$$a_{i\,i+1} = \int_{x_i}^{x_{i+1}} \frac{-1}{(h_{i+1})^2}\, dx = -\frac{1}{h_{i+1}} \text{ for } i = 1, 2, \ldots, M,$$

while $a_{i\,i-1} = -1/h_i$ for $i = 2, 3, \ldots, M$.

We compute the elements of the load vector of b in the same way to get

$$b_i = \int_{x_{i-1}}^{x_i} f(x)\frac{x - x_{i-1}}{h_i} dx + \int_{x_i}^{x_{i+1}} f(x)\frac{x_{i+1} - x}{h_{i+1}} dx, \quad i = 1, \ldots, M.$$

The matrix A is a *sparse* matrix in the sense that most of its entries are zero. In particular, A is a *banded* matrix with non-zero entries occurring only in the diagonal, super-diagonal and sub-diagonal positions. A is also called a tri-diagonal matrix. Moreover, A is a *symmetric* matrix since $\int_0^1 \varphi_i'\varphi_j'\, dx = \int_0^1 \varphi_j'\varphi_i'\, dx$. Finally, A is *positive-definite* in the sense that

$$\eta^\top A\eta = \sum_{i,j=1}^{M} \eta_i a_{ij} \eta_j > 0,$$

unless $\eta_i = 0$ for $i = 1, \ldots, M$. This follows by noting that if $v(x) = \sum_{j=1}^{M} \eta_j \varphi_j(x)$ then by reordering the summation (check!)

$$\begin{aligned}\sum_{i,j=1}^{M} \eta_i a_{ij} \eta_j &= \sum_{i,j=1}^{M} \eta_i \int_0^1 a\varphi_j'\varphi_i'\, dx\, \eta_j \\ &= \int_0^1 a \sum_{j=1}^{M} \eta_j \varphi_j' \sum_{i=1}^{M} \eta_i \varphi_i'\, dx = \int_0^1 a v'(x) v'(x)\, dx > 0\end{aligned}$$

unless $v'(x) = 0$ for all $x \in [0, 1]$, that is $v(x) = 0$ for $x \in [0, 1]$, since $v(0) = 0$, that is $\eta_i = 0$ for $i = 1, \ldots, M$. This implies that A is invertible, so that (53.15) has a unique solution for all data b.

We sum up: the stiffness matrix A is sparse, symmetric and positive definite, and thus in particular the system $A\xi = b$ has a unique solution for all b.

We expect the accuracy of the approximate solution to increase as M increases since the work involved in solving for U increases. Systems of dimension $10^2 - 10^3$ in one space dimension and up to 10^6 in two or three space dimensions are common. An important issue is the efficient numerical solution of the system $A\xi = b$.

53.6 Handling Different Boundary Conditions

We consider briefly the discretization of the two-point boundary value problem $-(au')' = f$ in $(0,1)$ with the different boundary conditions.

Non-Homogeneous Dirichlet Boundary Conditions

We begin with the boundary conditions $u(0) = u_0$ and $u(1) = u_1$, where u_0 and u_1 are given boundary values, where the conditions are non-homogeneous if $u_0 u_1 \neq 0$. In this situation, we compute an approximate solution in the trial space V_h of continuous piecewise linear functions $v(x)$ on a partition $\mathcal{T}_h : 0 = x_0 < x_1 < \ldots < x_{M+1} = 1$, satisfying the boundary conditions $v(0) = u_0$, $v(1) = u_1$, and we let the test functions vary over the space V_h^0 of continuous piecewise linear functions $v(x)$ satisfying the homogeneous boundary conditions $v(0) = v(1) = 0$. The trial and test spaces are different in this case, but we note that they have equal dimension (equal to the number M of internal nodes). Multiplying by a test function and integrating by parts, we are led to the following method: compute $U \in V_h$ such that

$$\int_0^1 aU'v'\,dx = \int_0^1 fv\,dx \quad \text{for all } v \in V_h^0. \tag{53.16}$$

As above this leads to a symmetric positive definite system of equations in the internal unknown nodal values $U(x_1), \ldots, U(x_M)$.

Neumann Boundary Conditions

We now consider the problem

$$\begin{cases} -(au')' = f, & \text{in } (0,1), \\ u(0) = 0,\ a(1)u'(1) = g_1, \end{cases} \tag{53.17}$$

with a non-homogeneous Neumann boundary condition at $x = 1$, which in the case of modeling heat in a wire, corresponds to prescribing the heat flux $a(1)u'(1)$ at $x = 1$ to be g_1.

To derive a variational formulation of this problem, we multiply the differential equation $-(au')' = f$ by a test function v and integrate by parts to get

$$\int_0^1 fv\,dx = -\int_0^1 (au')'v\,dx = \int_0^1 au'v'\,dx - a(1)u'(1)v(1) + a(0)u'(0)v(0).$$

Now $a(1)u'(1) = g_1$ is specified but $a(0)u'(0)$ is unknown. So it is convenient to assume that v satisfies the homogeneous Dirichlet condition $v(0) = 0$. Correspondingly, we define V_h to be the space of continuous functions v that are piecewise linear on a partition $\mathcal{T}_h$ of $(0,1)$ satisfying $v(0) = 0$. Replacing $a(1)u'(1)$ by g_1, we are led to the following FEM for (53.17): compute $U \in V_h$ such that

$$\int_0^1 aU'v'\,dx = \int_0^1 fv\,dx + g_1v(1) \quad \text{for all } v \in V_h. \tag{53.18}$$

We substitute $U(x) = \sum_{i=1}^{M+1} \xi_i\varphi_i(x)$, noting that the value $\xi_{M+1} = U(x_{M+1})$ at the node x_{M+1} is now undetermined, into (53.18) and choose $v = \varphi_1, \cdots, \varphi_{M+1}$ to get a $(M+1)\times(M+1)$ system of equations for ξ. We show the form of the resulting stiffness matrix with $a = 1$ and load vector in Fig. 53.6. Note that the last equation

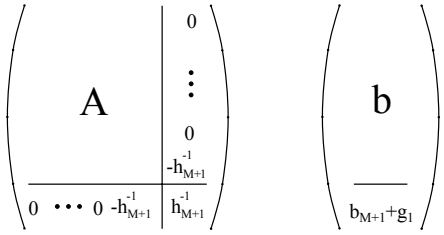

Fig. 53.6. The stiffness matrix and load vector computed from (53.18) in the case that $a \equiv 1$. A and b are the stiffness matrix and load vector previously obtained in the problem with homogeneous Dirichlet boundary conditions and $b_{M+1} = \int_0^1 f\varphi_{M+1}\,dx$

$$\frac{U(x_{M+1}) - U(x_M)}{h_{M+1}} = b_{M+1} + g_1$$

is a discrete analog of the boundary condition $u'(1) = g_1$ since $b_{M+1} \approx \frac{h_{M+1}}{2}f(1)$.

To conclude, a Neumann boundary condition, unlike a Dirichlet condition, is not explicitly enforced in the trial space. Instead, the Neumann condition is automatically satisfied as a consequence of the variational formulation by letting the test functions vary freely at the corresponding boundary

point. In the case of Neumann boundary conditions, we thus simply can "forget" the boundary conditions in the definition of the trial space V_h and let the test space coincide with V_h. A Dirichlet boundary condition is called an *essential* boundary condition and a Neumann condition is called a *natural* boundary condition. An essential boundary condition is imposed explicitly in the definition of the trial space, i.e. it is a *strongly imposed* boundary condition, and the test space satisfy the corresponding homogeneous boundary condition. A natural boundary condition is not imposed in the trial space and becomes automatically satisfied through the variational formulation by letting the test functions vary freely at the corresponding boundary point.

Robin Boundary Conditions

A natural generalization of Neumann conditions for the problem $-(au')' = f$ in $(0,1)$ are called *Robin* boundary conditions. These take the form

$$-a(0)u'(0) = \gamma(0)(u_0 - u(0)), \quad a(1)u'(1) = \gamma(1)(u_1 - u(1)). \tag{53.19}$$

In the case of modeling heat in a wire, $\gamma(0)$ and $\gamma(1)$ are given (non-negative) boundary heat conductivities and u_0 and u_1 are given "outside temperatures". The Robin boundary condition at $x = 0$ states that the heat flux $-a(0)u'(0)$ is proportional to the temperature difference $u_0 - u(0)$ between the outside and inside temperature. If $u_0 > u(0)$ then heat will flow from outside to inside and if $u_0 < u(0)$ then heat will flow from inside out.

Example 53.3. We may experience this kind of boundary condition with $\gamma(0)$ quite large in a poorly insulated house on a cold winter day. The size of the boundary heat conductivity γ is an important issue in the real estate business in the north of Sweden.

When $\gamma = 0$, (53.19) reduces to a homogeneous Neumann boundary condition. Conversely, letting γ tend to infinity, the Robin boundary condition $-a(0)u'(0) = \gamma(0)(u_0 - u(0))$ approaches the Dirichlet boundary condition $u(0) = u_0$.

Robin boundary conditions are natural boundary conditions like Neumann conditions, Therefore, we let V_h be the space of continuous piecewise linear functions on a partition of $(0,1)$ without any boundary conditions imposed. Multiplying the equation $-(au')' = f$ by a function $v \in V_h$ and integrating by parts, we get

$$\int_0^1 fv\,dx = -\int_0^1 (au')'v\,dx = \int_0^1 au'v'\,dx - a(1)u'(1)v(1) + a(0)u'(0)v(0).$$

Replacing $a(0)u'(0)$ and $a(1)u'(1)$ using the Robin boundary conditions, we get

$$\int_0^1 fv\,dx = \int_0^1 au'v'\,dx + \gamma(1)(u(1)-u_1)v(1) + \gamma(0)(u(0)-u_0)v(0).$$

Collecting data on the right hand side, we are led to the following cG(1) method: compute $U \in V_h$ such that

$$\int_0^1 aU'v'\,dx + \gamma(0)u(0)v(0) + \gamma(1)u(1)v(1) = \int_0^1 fv\,dx + \gamma(0)u_0v(0) + \gamma(1)u_1v(1)$$

for all $v \in V_h$.

An even more general Robin boundary condition has the form $-a(0)u'(0) = \gamma(0)(u_0 - u(0)) + g_0$, where g_0 is a given heat flux. This Robin boundary condition thus includes Neumann boundary conditions ($\gamma = 0$) and Dirichlet boundary conditions (letting $\gamma \to \infty$). The implementation of a Robin boundary conditions is facilitated by the fact that the trial and test space are the same.

53.7 Error Estimates and Adaptive Error Control

When conducting scientific experiments in a laboratory or building a suspension bridge, for example, there is always a lot of worry about the errors in the process. In fact, if we were to summarize the philosophy behind the scientific revolution, a main component would be the modern emphasis on the quantitative analysis of error in measurements during experiments and the reporting of the errors along with the results. The same issue comes up in computational mathematical modeling: whenever we make a computation on a practical problem, we must be concerned with the accuracy of the results and the related issue of how to compute efficiently. These issues naturally fit into a wider framework which also addresses how well the differential equation models the underlying physical situation and what effect errors in data and the model have on the conclusions we can draw from the results.

We address these issues by deriving two kinds of error estimates for the error $u - U$ of the finite element approximation. First we prove an *a priori* error estimate which shows that the Galerkin finite element method for (53.9) produces the best possible approximation in V_h of the solution u in a certain sense. If u has continuous second derivatives, then we know that V_h contains good approximations of u, for example the piecewise linear interpolant. So the a priori estimate implies that the error of the finite

element approximation can be made arbitrarily small by refining the mesh provided that the solution u is sufficiently smooth to allow the interpolation error to go to zero as the mesh is refined. This kind of result is called an *a priori* error estimate because the error bound does not depend on the approximate solution to be computed. One the other hand, it does requires knowledge about the derivatives of the (unknown) exact solution.

After that, we prove an *a posteriori* error bound that bounds the error of the finite element approximation in terms of its residual error. This error bound can be evaluated once the finite element solution has been computed and used to estimate the error. Through the a posteriori error estimate, it is possible to estimate and adaptively control the finite element error to a desired tolerance level by suitably refining the mesh.

To measure the size of the error $e = u - U$, we shall use the *weighted* L_2 *norm*

$$\|w\|_a = \left(\int_0^1 a\, w^2\, dx \right)^{1/2},$$

with *weight* a. More precisely we shall estimate the quantity

$$\|(u - U)'\|_a$$

which we refer to as the *energy norm* of the error $u - U$.

We will use the following variations of Cauchy's inequality with the weight a present:

$$\left| \int_0^1 av'w'\, dx \right| \leq \|v'\|_a \|w'\|_a \quad \text{and} \quad \left| \int_0^1 vw\, dx \right| \leq \|v\|_a \|w\|_{a^{-1}}\,. \quad (53.20)$$

An A Priori Error Estimate

We shall prove that the finite element approximation $U \in V_h$ is the best approximation of u in V_h with respect to the energy norm. This is a consequence of the *Galerkin orthogonality* built into the finite element method expressed by

$$\int_0^1 a(u - U)'v'\, dx = 0 \quad \text{for all } v \in V_h \quad (53.21)$$

which results from subtracting (53.12) from (53.11) (integrated by parts) with $v \in V_h$. This is analogous to the best approximation property of the L_2 projection studied in the Chapter Piecewise linear approximation.

We have for any $v \in V_h$,

$$\begin{aligned}\|(u-U)'\|_a^2 &= \int_0^1 a(u-U)'(u-U)'\,dx \\ &= \int_0^1 a(u-U)'(u-v)'\,dx + \int_0^1 a(u-U)'(v-U)'\,dx \\ &= \int_0^1 a(u-U)'(u-v)'\,dx,\end{aligned}$$

where the last line follows because $v - U \in V_h$. Estimating using Cauchy's inequality, we get

$$\|(u-U)'\|_a^2 \le \|(u-U)'\|_a \|(u-v)'\|_a,$$

so that

$$\|(u-U)'\|_a \le \|(u-v)'\|_a \quad \text{for all } v \in V_h.$$

This is the best approximation property of U. We now choose in particular $v = \pi_h u$, where $\pi_h u \in V_h$ is the nodal interpolant of u, and use the following weighted analog of (52.11)

$$\|(u - \pi_h u)'\|_a \le C_i \|hu''\|_a,$$

where C_i is an interpolation constant that depends only on (the variation of) a. We then obtain the following error estimate.

Theorem 53.1 *The finite element approximation U satisfies $\|(u-U)'\|_a \le \|(u-v)'\|_a$ for all $v \in V_h$. In particular, there is a constant C_i depending only on a such that*

$$\|u' - U'\|_a \le C_i \|hu''\|_a.$$

This energy norm estimate says that the derivative of the error of the finite element approximation converges to zero at a first order rate in the mesh size h. By integration it follows that the error itself, say pointwise or in the L_2 norm, also tends to zero. One can also prove a more precise bound for the error $u - U$ itself that is second order in the mesh size h.

An A Posteriori Error Estimate

We shall now estimate the energy norm error $\|u' - U'\|_a$ in terms of the residual $R(U) = (aU')' + f$ of the finite element solution U on each subinterval. The residual measures how well U solves the differential equation and it is completely computable once U has been computed.

We start by using the variational form of (53.11) with $v = e = u - U$ to find an expression for $\|u - U\|_a^2$:

$$\begin{aligned}\|e'\|_a^2 &= \int_0^1 ae'e'\,dx = \int_0^1 au'e'\,dx - \int_0^1 aU'e'\,dx \\ &= \int_0^1 fe\,dx - \int_0^1 aU'e'\,dx.\end{aligned}$$

We then use (53.12), with $v = \pi_h e$ denoting the nodal interpolant of e in V_h, to obtain

$$\begin{aligned}\|e'\|_a^2 &= \int_0^1 f\,(e - \pi_h e)\,dx - \int_0^1 aU'(e - \pi_h e)'\,dx \\ &= \int_0^1 f\,(e - \pi_h e)\,dx - \sum_{j=1}^{M+1} \int_{I_j} aU'(e - \pi_h e)'\,dx.\end{aligned}$$

Now, we integrate by parts over each sub-interval I_j in the last term and use the fact that all the boundary terms disappear because $(e - \pi_h e)(x_j) = 0$ to get the *error representation formula*

$$\|e'\|_a^2 = \int_0^1 R(U)(e - \pi_h e)\,dx, \tag{53.22}$$

where the residual error $R(U)$ is the discontinuous function defined on $(0, 1)$ by

$$R(U) = f + (aU')' \quad \text{on each sub-interval } I_j.$$

From the weighted Cauchy inequality (53.20) (inserting factors h and h^{-1}), we get

$$\|e'\|_a^2 \leq \|hR(U)\|_{a^{-1}} \|h^{-1}(e - \pi_h e)\|_a.$$

One can prove the following analog of the second estimate of (52.11)

$$\|h^{-1}(e - \pi_h e)\|_a \leq C_i \|e'\|_a,$$

where C_i is an interpolation constant depending on a, and we notice the appearance of the factor h^{-1} on the left hand side. This proves the basic a posteriori error estimate:

Theorem 53.2 *There is an interpolation constant C_i depending only on a such that the finite element approximation U satisfies*

$$\|u' - U'\|_a \leq C_i \|hR(U)\|_{a^{-1}}. \tag{53.23}$$

Adaptive Error Control

Since the a posteriori error estimate (53.23) indicates the size of the error of an approximation on a given mesh in terms of computable information, it is natural to try to use this information to compute an accurate approximation. This is the basis of *adaptive error control*.

The computational problem that arises once a two-point boundary value problem is specified is to find a mesh such that the finite element approximation achieves a given level of accuracy, or in other words, such that the error of the approximation is bounded by an *error tolerance* TOL. In practice, we are also concerned with efficiency, which means that we want to determine a mesh with the fewest number of elements that yields an approximation with the desired accuracy. We try to reach this optimal mesh by starting with a coarse mesh and successively refining based on the size of the a posteriori error estimate. By starting with a coarse mesh, we try to keep the number of elements as small as possible.

More precisely, we choose an initial mesh $\mathcal{T}_h$, compute the corresponding cG(1) approximation U, and then check whether or not

$$C_i\|hR(U)\|_{a^{-1}} \leq \text{TOL}.$$

This is the *stopping criterion*, which guarantees that $\|u' - U'\|_a \leq \text{TOL}$ by (53.23). Therefore when the stopping criterion is satisfied, U is sufficiently accurate. If the stopping criterion is not satisfied, we try to construct a new mesh $\mathcal{T}_{\tilde{h}}$ of mesh size $\tilde{h}$ with as few elements as possible such that

$$C_i\|\tilde{h}R(U)\|_{a^{-1}} = \text{TOL}.$$

This is the *mesh modification criterion* from which the new mesh size $\tilde{h}$ is computed based on the size of the residual error $R(U)$ of the approximation on the old mesh. In order to minimize the number of mesh points, it turns out that the mesh size should be chosen to *equidistribute* the residual error in the sense that the contribution from each element to the integral giving the total residual error is roughly the same. In practice, this means that elements with large residual errors are refined, while elements in intervals where the residual error is small are combined together to form bigger elements.

We repeat the adaptive cycle of mesh modification followed by solution on the new mesh until the stopping criterion is satisfied. By the a priori error estimate, we know that if u'' is bounded then the error tends to zero as the mesh is refined. Hence, the stopping criterion will be satisfied eventually. In practice, the adaptive error control rarely requires more than a few iterations.

TS[b] Please supply a shorter running title. Thank you.

Editor's or typesetter's annotations (will be removed before the final TEX run)

53.8 Discretization of Time-Dependent Reaction-Diffusion-Convection Problems

We now return to original time dependent problem (53.6).

To solve (53.6) numerically, we apply the cG(1) method for time discretization and the cG(1) FEM for discretization in space. More precisely, let $0 = x_0 < x_1 < \ldots < x_{L+1} = 1$ be a partition of $(0,1)$, and let V_h be the corresponding space of continuous piecewise linear functions $v(x)$ such that $v(0) = v(1) = 0$. Let $0 = t_0 < t_1 < t_2 < \ldots < t_N = T$ be a sequence of discrete time levels with corresponding time intervals $I_n = (t_{n-1}, t_n)$ and time steps $k_n = t_n - t_{n-1}$, for $n = 1, \ldots, N$. We look for a numerical solution $U(x,t)$ that is linear in t on each time interval I_n. For $n = 1, \ldots, N$, we compute $U^n \in V_h$ such that for all $v \in V_h$,

$$\begin{aligned} &\int_{I_n} \int_0^1 \dot{U} v \, dx \, dt + \int_{I_n} \int_0^1 (aU')v' + (bU)'v) \, dx \, dt \\ &= \int_{I_n} \int_0^1 fv \, dx \, dt + \int_{I_n} (g(0,t)v(0) + g(1,t)v(1)) \, dt, \end{aligned} \tag{53.24}$$

where $U(t_n, x) = U^n(x)$ denotes the *time nodal value* for $n = 1, 2, \ldots, N$ and $U^0 = u_0$, assuming that $u_0 \in V_h$. Since U is linear on each time interval, it is determined completely once we have computed its nodal values.

Arguing as above using the expansion in terms of the basis functions for V_h leads to a sequence of systems of equations for $n = 1, \ldots, N$,

$$MU^n + k_n A_n U^n = MU^{n-1} + k_n b^n, \tag{53.25}$$

where M is the mass matrix corresponding to V_h and A_n is a stiffness matrix related to time interval I_n. Solving this system successively for $n = 1, 2, \ldots, N$, we obtain an approximate solution U of (53.10).

53.9 Non-Linear Reaction-Diffusion-Convection Problems

In many situations, the coefficients or data depend on the solution u, which leads to a nonlinear problem. For example if f depends on u, we get a problem of the form

$$\begin{cases} \dot{u} - (au')' + (bu)' = f(u) & \text{in } (0,1) \times (0,T), \\ u(0,t) = u(1,t) = 0, & \text{for } t \in (0,T), \\ u(x,0) = u_0(x) & \text{for } x \in (0,1). \end{cases} \tag{53.26}$$

Discretization as above eventually yields a discrete system of the form

$$MU^n + k_n A_n U^n = MU^{n-1} + k_n b^n(U^n), \tag{53.27}$$

where b^n depends on U^n. This nonlinear system may be solved by fixed point iteration or Newton's method.

We conclude this section by presenting some examples of systems of nonlinear reaction-diffusion-convection problems arising in physics, chemistry and biology. These systems may be solved numerically by a direct extension of the cG(1) method in space and time presented above. In all examples, a and the α_i are a positive constants.

Example 53.4. The bistable equation for ferro-magnetism

$$\dot{u} - au'' = u - u^3, \tag{53.28}$$

with a small.

Example 53.5. Model of a superconductivity of a fluid

$$\begin{aligned} \dot{u}_1 - au_1'' &= (1 - |u|^2)u_1, \\ \dot{u}_2 - au_2'' &= (1 - |u|^2)u_2. \end{aligned} \tag{53.29}$$

Example 53.6. Model of flame propagation

$$\begin{aligned} \dot{u}_1 - au_1'' &= -u_1 e^{-\alpha_1/u_2}, \\ \dot{u}_2 - au_2'' &= \alpha_2 u_1 e^{-\alpha_1/u_2}. \end{aligned} \tag{53.30}$$

Example 53.7. Field-Noyes equations for chemical reactions

$$\begin{aligned} \dot{u}_1 - au_1'' &= \alpha_1(u_2 - u_1u_3 + u_1 - \alpha_2 u_1^2), \\ \dot{u}_2 - au_2'' &= \alpha^{-1}(\alpha_3 u_3 - u_2 - u_1u_2), \\ \dot{u}_2 - au_2'' &= \alpha_4(u_1 - u_3). \end{aligned} \tag{53.31}$$

Example 53.8. Spread of rabies in foxes

$$\begin{aligned} \dot{u}_1 - au_1'' &= \alpha_1(1 - u_1 - u_2 - u_3) - u_3u_1, \\ \dot{u}_2 - au_2'' &= u_3u_1 - (\alpha_2 + \alpha_3 + \alpha_1u_1 + \alpha_1u_1 + \alpha_1u_3)u_2, \\ \dot{u}_2 - au_2'' &= \alpha_2u_2 - (\alpha_4 + \alpha_1u_1 + \alpha_1u_1 + \alpha_1u_3)u_3, \end{aligned} \tag{53.32}$$

where $\alpha_4 < (1 + (\alpha_3 + \alpha_1)/\alpha_2)^{-1} - \alpha_1$.

Example 53.9. Interaction of two species

$$\begin{aligned} \dot{u}_1 - au_1'' &= u_1M(u_1, u_2), \\ \dot{u}_2 - au_2'' &= u_2N(u_1, u_2), \end{aligned} \tag{53.33}$$

where $M(u_1, u_2)$ and $N(u_1, u_2)$ are given functions describing various situations such as (i) predator-prey ($M_{u_2} < 0$, $N_{u_1} > 0$) (ii) competing species ($M_{u_2} < 0$, $N_{u_1} < 0$) and (iii) symbiosis ($M_{u_2} > 0$, $N_{u_1} > 0$).

Example 53.10. Morphogenesis of patterns (zebra or tiger)

$$\begin{aligned} \dot{u}_1 - au_1'' &= -u_1u_2^2 + \alpha_1(1-u_1) \\ \dot{u}_2 - au_2'' &= u_1u_2^2 - (\alpha_1+\alpha_2)u_2. \end{aligned} \tag{53.34}$$

Example 53.11. Fitz-Hugh-Nagumo model for transmission of axons

$$\begin{aligned} \dot{u}_1 - au_1'' &= -u_1(u_1-\alpha_1)(u_1-1) - u_2 \\ \dot{u}_2 - au_2'' &= \alpha_2 u_1 - \alpha_3 u_2, \end{aligned} \tag{53.35}$$

$0 < \alpha_1 < 1$.

Chapter 53 Problems

53.1. Compute the stiffness matrix and load vector for the cG(1) method on a uniform partition for (53.9) with $a(x) = 1+x$ and $f(x) = \sin(x)$. Use quadrature if exact integration is inconvenient.

53.2. Formulate the cG(1) method for the problem $-(au')' + cu = f$ in $(0,1)$, $u(0) = u(1) = 0$, where $a(x)$ and $c(x)$ are positive coefficients. Compute the corresponding stiffness matrix when $a = c = 1$, assuming a uniform partition. Is the stiffness matrix still symmetric, positive-definite, and tridiagonal?

53.3. Determine the resulting system of equations corresponding to the cG(1) method (53.16) with non-homogeneous Dirichlet boundary conditions.

53.4. Prove a priori and a posteriori error estimates for cG(1) for $-(au')' = f$ in $(0,1)$ with Robin boundary conditions (a positive).

53.5. Prove a priori and a posteriori error estimates for cG(1) for $-(au')'+cu = f$ in $(0,1)$ with Robin boundary conditions (a and c positive).

> The "classical" phase of my career was summed up in the book *The Large Scale Structure of Spacetime* which Ellis and I wrote in 1973. I would not advise readers of this book to consult that work for further information: it is highly technical, and quite unreadable. I hope that since then I have learned how to write in a manner that is easier to understand. (Stephen Hawking in *A Brief History of Time*)

References

[1] L. Ahlfors, *Complex Analysis*, McGraw-Hill Book Company, New York, 1979.

[2] K. Atkinson, *An Introduction to Numerical Analysis*, John Wiley and Sons, New York, 1989.

[3] L. Bers, *Calculus*, Holt, Rinehart, and Winston, New York, 1976.

[4] M. Braun, *Differential Equations and their Applications*, Springer-Verlag, New York, 1984.

[5] R. Cooke, *The History of Mathematics. A Brief Course*, John Wiley and Sons, New York, 1997.

[6] R. Courant and F. John, *Introduction to Calculus and Analysis*, vol. 1, Springer-Verlag, New York, 1989.

[7] R. Courant and H. Robbins, *What is Mathematics?*, Oxford University Press, New York, 1969.

[8] P. Davis and R. Hersh, *The Mathematical Experience*, Houghton Mifflin, New York, 1998.

[9] J. Dennis and R. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, New Jersey, 1983.

[10] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Computational Differential Equations*, Cambridge University Press, New York, 1996.

[11] I. Grattan-Guiness, *The Norton History of the Mathematical Sciences*, W.W. Norton and Company, New York, 1997.

[12] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley and Sons, New York, 1962.

[13] E. Isaacson and H. Keller, *Analysis of Numerical Methods*, John Wiley and Sons, New York, 1966.

[14] M. Kline, *Mathematical Thought from Ancient to Modern Times*, vol. I, II, III, Oxford University Press, New York, 1972.

[15] J. O'Connor and E. Robertson, *The MacTutor History of Mathematics Archive*, School of Mathematics and Statistics, University of Saint Andrews, Scotland, 2001. `http://www-groups.dcs.st-and.ac.uk/~history/`.

[16] W. Rudin, *Principles of Mathematical Analysis*, McGraw–Hill Book Company, New York, 1976.

[17] T. Ypma, *Historical development of the Newton-Raphson method*, SIAM Review, 37 (1995), pp. 531–551.

Index