Contents

Chapter 3

Series, Operators and Continued Fractions

3.1 Some Basic Facts about Series

3.1.1 Introduction

Series expansions are a very important aid in numerical calculations, especially for quick estimates made in hand calculation—for example, in evaluating functions, integrals, or derivatives. Solutions to differential equations can often be expressed in terms of series expansions. Since the advent of computers it has, however, become more common to treat differential equations directly, using, e.g., finite difference or finite element approximations instead of series expansions. Series have some advantages, especially in problems containing parameters. Automatic methods for formula manipulation and some new numerical methods provide, however, new possibilities for series.

In this section we will discuss general questions concerning the use of infinite series for numerical computations including, e.g., the estimation of remainders, power series and various algorithms for computing their coefficients. Often a series expansion can be derived by simple operations with a known series. We also give an introduction to formal power series. The next section treats perturbation expansions, ill-conditioned and semi-convergent expansions, from the point of view of computing.

Methods and results will sometimes be formulated in terms of *series*, sometimes in terms of *sequences.* These formulations are equivalent, since the sum of an infinite series is defined as the limit of the the sequence s_n of its partial sums

$$S_n = a_1 + a_2 + \ldots + a_n.$$

Conversely, any sequence $S_1, S_2, S_3, \ldots$ can be written as the partial sums of a series,

$$S_1 + (S_2 - S_1) + (S_3 - S_2) + \ldots.$$

We start with some simple examples and some general rules for the approximation of remainders.

Example 3.1.1.

Compute, to five decimals, $y(0.5)$, where $y(x)$ is the solution to the differential equation $y'' = -xy$, with initial conditions $y(0) = 1$, $y'(0) = 0$. The solution cannot be simply expressed in terms of elementary functions. We shall use the method of undetermined coefficients. Thus we try substituting a series of the form:

$$y(x) = \sum_{n=0}^{\infty} c_n x^n = c_0 + c_1 x + c_2 x^2 + \cdots.$$

Differentiating twice we get

$$\begin{aligned} y''(x) &= \sum_{n=0}^{\infty} n(n-1)c_n x^{n-2} \\ &= 2c_2 + 6c_3 x + 12c_4 x^2 + \cdots + (m+2)(m+1)c_{m+2}x^m + \cdots, \\ -xy(x) &= -c_0 x - c_1 x^2 - c_2 x^3 - \cdots - c_{m-1}x^m - \cdots. \end{aligned}$$

Equating coefficients of x^m in these series gives

$$c_2 = 0, \qquad (m+2)(m+1)c_{m+2} = -c_{m-1}, \quad m \geq 1.$$

It follows from the initial conditions that $c_0 = 1$, $c_1 = 0$. Thus $c_n = 0$, if n is not a multiple of 3, and using the recursion we obtain

$$y(x) = 1 - \frac{x^3}{6} + \frac{x^6}{180} - \frac{x^9}{12,960} + \cdots.$$

This gives $y(0.5) = 0.97925$. The x^9 term is ignored, since it is less than $2 \cdot 10^{-7}$. In this example also the first neglected term gives a rigorous bound for the error (i.e. for the remaining terms), since the absolute value of the term decreases, and the terms alternate in sign.

Since the calculation was based on a trial substitution, one should, strictly speaking, prove that the series obtained defines a function which satisfies the given problem. Clearly, the series converges at least for $|x| < 1$, since the coefficients are bounded. (In fact the series converges for all x.) Since a power series can be differentiated term by term in the interior of its interval of convergence, the proof presents no difficulty. Note, in addition, that the finite series obtained for $y(x)$ by breaking off after the x^9-term is the exact solution to the following *modified differential equation*:

$$y'' = -xy - \frac{x^{10}}{12,960}, \qquad y(0) = 1, \qquad y'(0) = 0,$$

where the "perturbation term" $-x^{10}/12,960$ has magnitude less than 10^{-7} for $|x| \leq 0.5$.[1]

[1] We shall see, in Volume III, Chapter 13, how to find a rigorous bound for the difference between the solutions of a differential system and a modified differential system.

In practice, one is seldom seriously concerned about a rigorous error bound when the computed terms decrease rapidly, and it is "obvious" that the terms will continue to decrease equally quickly. One can then break off the series and use either the last included term or a coarse estimate of the **first neglected term** as an estimate of the remainder.

This rule is not very precise. *How rapidly is "rapidly"?* Questions like this occur everywhere in scientific computing. If mathematical rigor costs little effort or little extra computing time, then it should, of course, be used. Often, however, an error bound that is both rigorous and realistic may cost more than what is felt reasonable for (say) a one-off problem.

In problems, where guaranteed error bounds are not asked for, when it is enough to obtain a feeling for the reliability of the results, one can handle these matters in the same spirit as one handles risks in every day life. It is then a matter of experience to formulate a simple and *sufficiently* reliable **termination criterion** based on the automatic inspection of the successive terms.[2]

The unexperienced scientific programmer may, however, find such questions hard, also in simple cases. In the production of general purpose mathematical software, or in a context where an inaccurate numerical result can cause a disaster, such questions are serious and sometimes hard also for the experienced scientific programmer. For this reason, we shall formulate a few theorems, with which one can often transform the feeling that "the remainder is negligible" to a mathematical proof. There are, in addition, actually numerically useful *divergent* series; see Sec. 3.2.6. When one uses such series, estimates of the remainder are clearly essential.

Assume that we want to compute a quantity S, which can be expressed in a series expansion, $S = \sum_{j=0}^{\infty} a_j$, and set

$$S_n = \textstyle\sum_{j=0}^{n} a_j, \qquad R_n = S - S_n.$$

We call $\sum_{j=n+1}^{\infty} a_j$ the **tail** of the series; a_n is the "last included term" and a_{n+1} is the "first neglected term". The *remainder* R_n with reversed sign is called the *truncation error*.[3]

The tail of a convergent series can often be compared to a series with a known sum, for example, a geometric series, or with an integral which can be computed directly.

Theorem 3.1.1. *Comparison with a Geometric Series.*

If $|a_{j+1}| \le k|a_j|$, $\forall j \ge n$, *where* $k < 1$, *then*

$$|R_n| \le \frac{|a_{n+1}|}{1-k} \le \frac{k|a_n|}{1-k}.$$

In particular if $k < 1/2$, *then it is true that the absolute value of the remainder is less than the last included term.*

[2] Termination criteria for iterative methods will be discussed in Sec. 6.1.3.

[3] In this terminology the remainder is the *correction* one has to make in order to eliminate the error.

Proof. By induction, one finds that $|a_j| \le k^{j-1-n}|a_{n+1}|$, $j \ge n+1$, since

$$|a_j| \le k^{j-1-n}|a_{n+1}| \quad \Rightarrow \quad |a_{j+1}| \le k|a_j| \le k^{j-n}|a_{n+1}|.$$

Thus

$$|R_n| \le \sum_{j=n+1}^{\infty} |a_j| \le \sum_{j=n+1}^{\infty} k^{j-1-n}|a_{n+1}| = \frac{|a_{n+1}|}{1-k} \le \frac{k|a_n|}{1-k},$$

according to the formula for the sum of an infinite geometric series. The last statement follows from the inequality $k/(1-k) < 1$, when $k < 1/2$. □

Example 3.1.2. *Power series with slowly varying coefficients.*

Let $a_j = j^{1/2}\pi^{-2j}$. Then $a_6 = 2.4{\cdot}0.0000011 < 3{\cdot}10^{-6}$. Further,

$$\frac{|a_{j+1}|}{|a_j|} \le \frac{(j+1)^{1/2}}{j^{1/2}} \frac{\pi^{2j-2}}{\pi^{-2j}} \le (1+1/6)^{1/2}\pi^{-2} < 0.11,$$

for $j \ge 6$. Thus, by Theorem 3.1.1 $|R_6| < 3{\cdot}10^{-6}\dfrac{0.11}{1-0.11} < 4{\cdot}10^{-7}$.

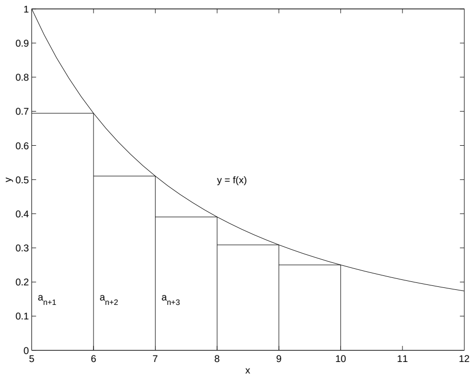

Figure 3.1.1. *Comparison with an integral, (n=5).*

Theorem 3.1.2. *Comparison of a Series with an Integral.*

If $|a_j| \le f(j)$ for all $j \ge n$, where $f(x)$ is a nonincreasing function for $x \ge n$, then

$$|R_n| \le \int_n^{\infty} f(x)dx,$$

which yields an upper bound for $|R_n|$, if the integral is finite.

If $a_j = f(j) > 0$ for all $j \ge n+1$, we also obtain a lower bound for the error, namely $\int_{n+1}^{\infty} f(x)dx$.

Proof. See Figure 3.1.1. □

Example 3.1.3.

When a_j is slowly decreasing, the two error bounds are typically rather close to each other, and are hence rather realistic bounds, much larger than the first neglected term a_{n+1}. Let $a_j = 1/(j^3+1)$, $f(x) = x^{-3}$. It follows that

$$0 < R_n \leq \int_n^\infty x^{-3}dx = n^{-2}/2.$$

In addition this bound gives an asymptotically correct estimate of the remainder, as $n \to \infty$, which shows that R_n is here significantly larger than the first neglected term.

For alternating series, however, the situation is typically quite different.

Definition 3.1.3.

A series is **alternating** *for $j \geq n$ if, for all $j \geq n$, a_j and a_{j+1} have opposite signs, or equivalently* $\text{sign}\, a_j = -\text{sign}\, a_{j+1}$*, where* sign x *(read "signum" of x), is defined by*

$$\text{sign } x = \begin{cases} +1, & \text{if } x > 0; \\ 0, & \text{if } x = 0; \\ -1, & \text{if } x < 0. \end{cases}$$

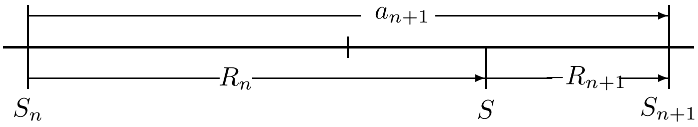

Figure 3.1.2. *Illustration to Theorem 3.1.4*

Theorem 3.1.4.

If R_n and R_{n+1} have opposite signs, then S lies between S_n and S_{n+1}. Furthermore

$$S = \frac{1}{2}(S_n + S_{n+1}) \pm \frac{1}{2}|a_{n+1}|.$$

We also have the weaker results:

$$|R_n| \leq |a_{n+1}|, \qquad |R_{n+1}| \leq |a_{n+1}|, \qquad \text{sign}\, R_n = \text{sign}\, a_{n+1}.$$

This theorem has non-trivial applications to practically important divergent sequences; see Sec. 3.2.6.

Proof. The fact that R_{n+1} and R_n have opposite signs means, quite simply, that one of S_{n+1} and S_n is too large and the other is too small, i.e. that S lies between S_{n+1} and S_n. Since $a_{n+1} = S_{n+1} - S_n$, one has for positive values of a_{n+1}, the situation shown in Figure 3.1.2. From this figure, and an analogous one for the case of $a_{n+1} < 0$, the remaining assertions of the theorem clearly follow. □

The actual error of the average $\frac{1}{2}(S_n + S_{n+1})$ is, for slowly convergent alternating series, usually much smaller than the error bound $\frac{1}{2}|a_{n+1}|$. For example, if $S_n = 1 - \frac{1}{2} + \frac{1}{3} - \ldots \pm \frac{1}{n}$, $\lim S_n = \ln 2 \approx 0.6931$, the error bound for $n = 4$ is 0.1, while the actual error is less than 0.01. A systematic exploration of this observation, by means of repeated averaging. is carried out in Sec. 3.4.3.

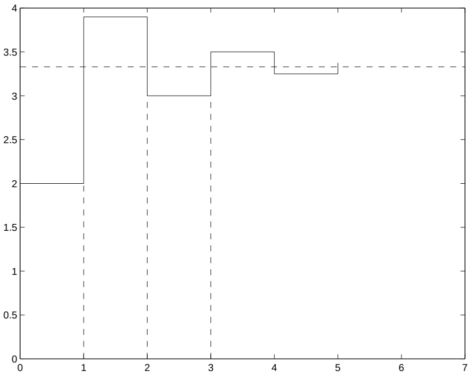

Figure 3.1.3. *The sum of an alternating series.*

In Example 1.2.3 the error function was approximated for $|x| \in [-1, 1]$ by a power series. The series has terms of alternating sign, and the absolute values of the terms decrease monotonically to zero. For such a series the above theorem can be used to prove that the first neglected term gives a rigorous error estimate.

Theorem 3.1.5.

For an alternating series, the absolute values of whose terms approach zero monotonically, the remainder has the same sign as the first neglected term a_{n+1}, and the absolute value of the remainder does not exceed $|a_{n+1}|$. (It is well known that such a series is convergent).

Proof. (Sketch) That the theorem is true is almost clear from Figure 3.1.3. The figure shows how S_j depends on j when the premises of the theorem are fulfilled. A formal proof is left to the reader. □

The use of this theorem was illustrated in Examples 3.1.1 and 3.1.2. An important generalization is given as Problem 3.2.1(g).

In the preceding theorems the ideas of well known convergence criteria are extended to bounds or estimates of the error of a truncated expansion. In Sec. sec3.4, we shall see a further extension of these ideas, namely for *improving the accuracy* obtained from a sequence of truncated expansions. This is known as *convergence acceleration.*

3.1.2 Power Series

Consider an expansion into powers of a complex variable z, and suppose that it is convergent for some $z \neq 0$, and denote its sum by $f(z)$,

$$f(z) = \sum_{j=0}^{\infty} a_j z^j, \qquad z \in \mathbf{C}. \tag{3.1.1}$$

It is then known from complex analysis that the series (3.1.1) either converges for all z, or it has a **circle of convergence** with radius ρ, such that it either converges for all $|z| < \rho$, and diverges for $|z| > \rho$. (For $|z| = \rho$ convergence or divergence is possible). The radius of convergence is determined by the relation

$$\rho = \limsup |a_n|^{-1/n}. \tag{3.1.2}$$

Another formula is $\rho = \lim |a_n|/|a_{n+1}|$, *if this limit exists.*

The function $f(z)$ can be expanded into powers of $z - a$ around any point of analyticity,

$$f(z) = \sum_{j=0}^{\infty} a_j (z-a)^j, \quad z \in \mathbf{C}. \tag{3.1.3}$$

By **Taylor's formula** the coefficients are given by

$$a_0 = f(a), \qquad a_j = f^{(j)}(a)/j!, \quad j \geq 1. \tag{3.1.4}$$

This infinite series is in the general case called a Taylor series, while the special case, $a = 0$, is by tradition called a Maclaurin series.[4]

The function $f(z)$ is analytic inside its circle of convergence, and has at least one singular point on its boundary. The singularity of f, which is closest to the origin, can often be found easily from the expression that defines $f(z)$; so the radius of convergence of a Maclaurin series can often be easily found.

Note that these Taylor coefficients are *uniquely determined* for the function f. This is true also for a non-analytic function, for example if $f \in C^p[a, b]$, although in this case the coefficient a_j exists only for $j \leq p$. Also the remainder formulas (3.1.5), (3.1.7), require only that $f \in C^n$. It is thus not necessary that the infinite expansion converges or even exists.

There are several expressions for the remainder $R_n(z)$, when the expansion for $f(z)$ is truncated after the term that contains z^{n-1}. In order to simplify the notation, *we put $a = 0$, i.e. we consider the Maclaurin series.* The following

[4] Brook Taylor (1685–1731), who announced his theorem in 1712, and Colin Maclaurin (1698–1746) were British mathematicians.

integral form can be obtained by the application of repeated integration by parts to the integral $z\int_0^1 f'(zt)\,dt$:

$$R_n(z) = z^n \int_0^1 \frac{(1-t)^{n-1}}{(n-1)!} f^{(n)}(zt)\,dt; \tag{3.1.5}$$

the details are left for Problem 24 (b). From this follows the upper bound

$$|R_n(z)| \le \frac{1}{n!}|z|^n \max_{0\le t\le 1} |f^{(n)}(zt)|. \tag{3.1.6}$$

This holds also in the complex case; if f is analytic on the segment from 0 to z, one integrates along this segment, i.e. for $0 \le t \le 1$, otherwise another path is to be chosen.

For a real-valued function, **Lagrange's formula**[5] for the remainder

$$R_n(x) = \frac{f^{(n)}(\xi)x^n}{n!}, \qquad \xi \in [0, x], \tag{3.1.7}$$

is obtained by the mean value theorem of integral calculus.

For complex-valued functions and, more generally, for vector-valued functions the mean value theorem and Lagrange's remainder term are not valid with a single ξ. (Sometimes componentwise application with different ξ is possible.) A different form for the remainder, valid in the complex plane is given in Sec. sec3.1.cfft, in terms of the **maximum modulus** $M(r) = \max_{|z|=r} |f(z)|$, which may sometimes be easier to estimate than the nth derivative. A power series is uniformly convergent in any closed bounded region strictly inside its circle of convergence. Roughly speaking, the series can be manipulated like a polynomial, as long as z belongs to such a region;

- it can be integrated or differentiated term by term,
- substitutions can be performed, and terms can be rearranged,
- it can be multiplied by another power series, etc.

Theorem 3.1.6.

If $f(z) = \sum a_j z^j$, $g(z) = \sum b_k z^k$, then

$$f(z)g(z) = \sum c_n z^n, \quad c_n = \sum_{j=0}^{n} a_j b_{n-j}. \tag{3.1.8}$$

The expression on the right side of (3.1.8) *is called the* **convolution** *or the* **Cauchy product** *of the coefficient sequences of f and g.*

[5] Joseph Louis Lagrange (1736–1813) was born in Turin, Italy. In 1766 he succeeded Euler in Berlin but in 1787 went to Paris where he remained until his death. He gave fundamental contributions to most branches of Mathematics and Mechanics.

The use of the Taylor coefficient formula and Lagrange's form of the remainder may be inconvenient, and it is often easier to obtain an expansion by manipulating some known expansions. The geometric series,

$$\frac{1}{1-z} = 1 + z + z^2 + z^3 + \cdots + z^{n-1} + \frac{z^n}{1-z}, \quad z \neq 1, \tag{3.1.9}$$

is of particular importance; note that the remainder $z^n/(1-z)$ is valid even when the expansion is divergent.

Example 3.1.4.

Set $z = -t^2$ in the geometric series, and integrate:

$$\int_0^x (1+t^2)^{-1}\,dt = \sum_{j=0}^{n-1} \int_0^x (-t^2)^j\,dt + \int_0^x (-t^2)^n (1+t^2)^{-1}\,dt.$$

Using the mean-value theorem of integral calculus on the last term we get

$$\arctan x = \sum_{j=0}^{n-1} \frac{(-1)^j x^{2j+1}}{2j+1} + \frac{(1+\xi^2)^{-1}(-1)^n x^{2n+1}}{2n+1}, \tag{3.1.10}$$

for some $\xi \in \operatorname{int}[0,x]$. Both the remainder term and the actual derivation are much simpler than what one would get by using Taylor's formula with Lagrange's remainder term. Note also that Theorem 3.1.4 is applicable to the series obtained above for all x and n, even for $|x| > 1$, when the infinite power series is divergent.

Some useful expansions are collected in Table 3.1.1.These formulas will be used often without a reference; the reader is advised to memorize the expansions. "Remainder ratio" means the *ratio* of the remainder to the first neglected term, if $x \in \mathbf{R}$; ξ means a number between 0 and x. Otherwise these expansions are valid in the unit circle of $\mathbf{C}$ or in the whole of $\mathbf{C}$.

The binomial coefficients are, also for non-integer k, defined by

$$\binom{k}{n} = \frac{k(k-1)\cdots(k-n+1)}{1\cdot 2\cdots n}.$$

Depending on the context, they may be computed by one of the following well known recurrences:

$$\binom{k}{(n+1)} = \binom{k}{n}\frac{(k-n)}{(n+1)}; \quad \text{or} \quad \binom{k+1}{n} = \binom{k}{n} + \binom{k}{n-1}, \tag{3.1.11}$$

with appropriate initial conditions. The latter recurrence follows from the matching of the coefficients of t^n in the equation $(1+t)^{k+1} = (1+t)(1+t)^k$. (Compare the Pascal triangle; see Problem 1.2.3.) The explicit formula $\binom{k}{n} = \frac{k!}{n!(k-n)!}$, for integers k, n, is to be avoided, if k can become large, because $k!$ has overflow for $k \geq 170$ in IEEE double precision.

Table 3.1.1. *Maclaurin expansions for some elementary functions.*

Function	Expansion ($x \in \mathbf{C}$)	Remainder ratio ($x \in \mathbf{R}$)
$(1-x)^{-1}$	$1 + x + x^2 + x^3 + \cdots$ if $\lvert x\rvert < 1$	$(1-x)^{-1}$ if $x \neq 1$
$(1+x)^k$	$1 + kx + \binom{k}{2}x^2 + \cdots$ if $\lvert x\rvert < 1$	$(1+\xi)^{k-n}$ if $x > -1$
$\ln(1+x)$	$x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots$ if $\lvert x\rvert < 1$	$(1+\xi)^{-1}$ if $x > -1$
e^x	$1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$ all x	e^ξ, all x
$\sin x$	$x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$ all x	$\cos\xi$, all x, n odd
$\cos x$	$1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots$ all x	$\cos\xi$, all x, n even
$\frac{1}{2}\ln\left(\frac{1+x}{1-x}\right)$	$x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots$ if $\lvert x\rvert < 1$	$\frac{1}{1-\xi^2}$, $\lvert x\rvert < 1$, n even
$\arctan x$	$x - \frac{x^3}{3} + \frac{x^5}{5} + \cdots$ if $\lvert x\rvert < 1$	$\frac{1}{1+\xi^2}$, all x

The exponent k in $(1+x)^k$ is not necessarily an integer; it can even be an irrational or a complex number. This function may be defined as $(1+x)^k = e^{k\ln(1+x)}$. Since $\ln(1+x)$ is **multi-valued**, $(1+x)^k$ is multi-valued too, unless k is an integer. We can, however, make them single-valued by forbidding the complex variable x to take real values less than -1. In other words, we make a **cut** along the real axis from -1 to $-\infty$ that the complex variable must not cross. (The cut is outside the circle of convergence.) We obtain the **principal branch** by requiring that $\ln(1+x) > 0$ if $x > 0$. Let $1 + x = re^{i\phi}$, $r > 0$, $\phi \to \pm\pi$. Note that

$$1 + x \to -r, \qquad \ln(1+x) \to \ln r + \begin{cases} +i\pi, & \text{if } \phi \to \pi; \\ -i\pi, & \text{if } \phi \to -\pi. \end{cases} \tag{3.1.12}$$

Two important power series, not given in Table 3.1.1, are:

Gauss' hypergeometric function

$$\begin{aligned} F(a,b,c;z) = 1 &+ \frac{ab}{c}\frac{z}{1!} + \frac{a(a+1)b(b+1)}{c(c+1)}\frac{z^2}{2!} \\ &+ \frac{a(a+1)(a+2)b(b+1)(b+2)}{c(c+1)(c+2)}\frac{z^3}{3!} + \dots, \end{aligned} \tag{3.1.13}$$

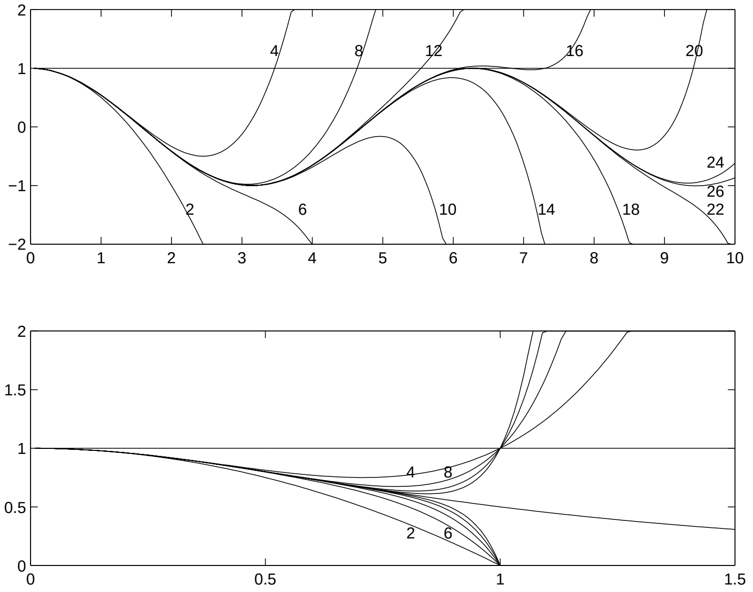

Figure 3.1.4. *The partial sums of the Maclaurin expansions for two functions. The upper curves are for $f(x) = \cos x$, $n = 0 : 2 : 26$, $0 \leq x \leq 10$. This series converges for all x, but the rounding errors cause trouble for large values of x; see Sec. 3.2.5, Ill-conditioned series. The lower curves are for $f(x) = 1/(1+x^2)$, $n = 0 : 2 : 18$, $0 \leq x \leq 1.5$. The convergence radius is 1 in this case.*

where a and b are complex constants and $c \neq -1, -2. \ldots$. The radius of convergence for this series equals unity; see [1, Chap. 15].[6]

Kummer's confluent hypergeometric function[7]

$$M(a,b;z) = 1 + \frac{a}{b}\frac{z}{1!} + \frac{a(a+1)}{b(b+1)}\frac{z^2}{2!} + \frac{a(a+1)(a+2)}{b(b+1)(b+2)}\frac{z^3}{3!} + \ldots, \qquad (3.1.14)$$

converges for all z (see [1, Ch. 13]). It is named "confluent" because

$$M(a,c;z) = \lim_{b\to\infty} F(a,b,c,z/b).$$

The coefficients of these series are easily computed and the functions are easily evaluated by recurrence relations. (You also need some criterion for the truncation of the series, adapted to your demands of accuracy.) In Sec. 3.5, these functions are

[6] This classical Handbook of Mathematical Functions, edited by Milton Abramowitz and Irene A. Stegun, will be used as a reference throughout this book. We will often refer to it just as "the Handbook".

[7] Ernst Eduard Kummer (1810–1893) German mathematician, professor in Berlin from 1855. He extended Gauss work on hypergeometric series. He, together with Weierstrass and Kronecker, made Berlin into one of the leading centers of mathematics.

also expressed in terms of infinite *continued fractions* that typically converge faster and in larger regions than the power series do.

Example 3.1.5.

The following procedure can generally be used in order to find the *expansion of the quotient of two expansions.* We illustrate it in a case, where the result is of interest to us later.

The **Bernoulli**[8] **numbers** B_n are defined by the Maclaurin series

$$\frac{x}{e^x-1} \equiv \sum_{j=0}^{\infty} \frac{B_j x^j}{j!} \tag{3.1.15}$$

For $x = 0$ the left hand side is defined by Hôpital's rule; the value is 1. If we multiply this equation by the denominator, we obtain

$$x \equiv \left(\sum_{i=1}^{\infty} \frac{x^i}{i!}\right)\left(\sum_{j=0}^{\infty} \frac{B_j x^j}{j!}\right).$$

By matching the coefficients of x^n, $n \geq 1$, on both sides, we obtain a recurrence relation for the Bernoulli numbers, which can be written in the form

$$B_0 = 1, \quad \sum_{j=0}^{n-1} \frac{1}{(n-j)!}\frac{B_j}{j!} = 0, \quad n \geq 2, \quad \text{i.e. } \sum_{j=0}^{n-1} \binom{n}{j} B_j = 0. \tag{3.1.16}$$

The last equation is a recurrence that determines B_{n-1} in terms of Bernoulli numbers with smaller subscripts, hence $B_0 = 1$, $B_1 = -\frac{1}{2}$, $B_2 = \frac{1}{6}$, $B_3 = 0$, $B_4 = -\frac{1}{30}$, $B_5 = 0$, $B_6 = \frac{1}{42}, \ldots$.

We see that the Bernoulli numbers are rational. We shall now demonstrate that $B_n = 0$, *when n is odd, except for $n = 1$.*

$$\frac{x}{e^x-1} + \frac{x}{2} = \frac{x}{2}\frac{e^x+1}{e^x-1} = \frac{x}{2}\frac{e^{x/2}+e^{-x/2}}{e^{x/2}-e^{-x/2}} = \sum_{n=0}^{\infty} \frac{B_{2n}x^{2n}}{(2n)!}. \tag{3.1.17}$$

Since the next to last term is an even function, i.e. its value is unchanged when x is replaced by $-x$, its Maclaurin expansion contains only even powers of x, and hence the last expansion is also true.

The recurrence obtained for the Bernoulli numbers by the matching of coefficients in the equation,

$$(e^{x/2} - e^{-x/2})\left(\sum_{n=0}^{\infty} B_{2n}x^{2n}/(2n)!\right) = \tfrac{1}{2}x\left(e^{x/2}+e^{-x/2}\right),$$

[8] Jacob (or James) Bernoulli (1654-1705), Swiss mathematician, one of the earliest to realize how powerful is the infinitesimal calculus. The Bernoulli numbers were published posthumously in 1713, in his fundamental work Ars Conjectandi (on Probability). The notation for Bernoulli numbers varies in the literature. Our notation seems to be the most common in modern texts. Several members of the same family enriched mathematics by their teaching and writing. Their role in the history of mathematics resembles the role of the Bach family in the history of music.

is not the same as the one we found above. It turns out to have better properties of numerical stability. We shall look into this experimentally in Problem 10(g).

The singularities of the function $x/(e^x - 1)$ are poles at $x = 2n\pi i$, $n = \pm 1, \pm 2, \pm 3, \ldots$, hence the radius of convergence is 2π. Further properties of Bernoulli numbers and the related Bernoulli polynomials and periodic functions, are presented in Sec. 3.4.4, where they occur as coefficients in the important Euler–Maclaurin formula.

If r is large the following formula is very efficient; the series on its right hand side then converges rapidly.

$$B_{2r}/(2r)! = (-1)^{r-1}2(2\pi)^{-2r}\Big(1 + \sum_{n=2}^{\infty} n^{-2r}\Big). \tag{3.1.18}$$

This is a particular case ($t = 0$) of a Fourier series for the Bernoulli functions that we shall encounter in Lemma 3.3.1(c). In fact, you obtain IEEE double accuracy for $r > 26$, even if the infinite sum on the right hand side is totally ignored. Thanks to (3.1.18) we do not need to worry much over the instability of the recurrences. When r is *very* large, however, we must be careful about underflow and overflow.

The **Euler numbers** E_n, which will be used later, are similarly defined by the generating function

$$\frac{1}{\cosh z} \equiv \sum_{n=0}^{\infty} \frac{E_n z^n}{n!}, \quad |z| < \frac{\pi}{2}. \tag{3.1.19}$$

Obviously $E_n = 0$ for all odd n. It can be shown that the Euler numbers are integers, $E_0 = 1$, $E_2 = -1$, $E_4 = 5$, $E_6 = -61$; see Problem 7e.

Example 3.1.6.

Let $f(x) = (x^3 + 1)^{-\frac{1}{2}}$. Compute $\int_{10}^{\infty} f(x)dx$ to 9 decimal places, and $f'''(10)$, with at most 1% error. Since x^{-1} is fairly small, we expand in powers of x^{-1}:

$$\begin{aligned} f(x) &= x^{-3/2}(1 + x^{-3})^{-1/2} = x^{-3/2}\Big(1 - \frac{1}{2}x^{-3} + \frac{1\cdot 3}{8}x^{-6} - \ldots\Big) \\ &= x^{-1.5} - \frac{1}{2}x^{-4.5} + \frac{3}{8}x^{-7.5} - \ldots. \end{aligned}$$

By integration,

$$\int_{10}^{\infty} f(x)dx = 2\cdot 10^{-0.5} - \frac{1}{7}10^{-3.5} + \frac{3}{52}10^{-6.5} + \ldots = 0.632410375.$$

Each term is less than 0.001 of the previous term.

By differentiating the series three times, we similarly obtain

$$f'''(x) = -\frac{105}{8}x^{-4.5} + \frac{1,287}{16}x^{-7.5} + \ldots.$$

For $x = 10$ the second term is less than 1% of the first; the terms after the second decrease quickly and are negligible. One can show that the magnitude of each term is less than $8\,x^{-3}$ of the previous term. We get $f'''(10) = -4.12\,10^{-4}$ to the desired accuracy. The reader is advised to carry through the calculation in more detail.

Example 3.1.7. *How to compute* $\sinh x$.

On a computer using IEEE double precision the roundoff unit is $u = 2^{-53} \approx 1.1 \cdot 10^{-16}$. One wishes to compute $\sinh x$ with good *relative* accuracy, both for small and large $|x|$, at least moderately large. Assume that e^x is computed with a relative error less than u in the given interval. The formula $(e^x - e^{-x})/2$ for $\sinh x$ is sufficiently accurate except when $|x|$ is very small and cancellation occurs. Hence for $|x| \ll 1$, e^x and e^{-x} and hence $(e^x - e^{-x})/2$ can have *absolute* errors of order of magnitude (say) u. Then the *relative* error in $(e^x - e^{-x})/2$ can have magnitude $\approx u/|x|$; for example, this is more than 100% for $x \approx 10^{-16}$.

For $|x| \ll 1$ one can instead use (say) two terms in the series expansion for $\sinh x$,

$$\sinh x = x + x^3/3! + x^5/5! + \dots.$$

Then one gets an absolute truncation error which is about $x^5/120$, and a round-off error of the order of $2u|x|$. Thus the formula $x + x^3/6$ is better than $(e^x - e^{-x})/2$ if

$$|x|^5/120 + 2u|x| < u.$$

If $2u|x| \ll u$, we have $|x|^5 < 120u \approx 15{\cdot}2^{-50}$, or $|x| < 15^{1/5} \cdot 2^{-10} \approx 0.00168$, (which shows that $2u|x|$ really could be ignored in this rough calculation). Thus, if one switches from $(e^x - e^{-x})/2$ to $x + x^3/6$ for $|x| < 0.00168$, the relative error will nowhere exceed $u/0.00168 \approx 0.66{\cdot}10^{-14}$. If one needs higher accuracy, one can take more terms in the series, so that the switch can occur at a larger value of $|x|$.

For very large values of $|x|$ one must expect a relative error of order of magnitude $|xu|$ because of round-off error in the argument x. Compare the discussion of range reduction in Sec. 2.2.4 and Problem 2.2.9.

In numerical computation a series should be regarded as a finite expansion together with a remainder. Taylor's formula with the remainder (3.1.5) is valid for any function $f \in C^n[a, a + x]$, but *the infinite series is valid only if the function is analytic in a complex neighborhood of* a.

If a function is not analytic at 0*, it can happen that the Maclaurin expansion converges to a wrong result.* A classical example (see Appendix to Chapter 6 in Courant [15]) is

$$f(x) = e^{-1/x^2}, \quad x \neq 0, \quad f(0) = 0.$$

It can be shown that all its Maclaurin coefficients are zero. This trivial Maclaurin expansion converges for all x, *but the sum is wrong for* $x \neq 0$. There is nothing wrong with the use of Maclaurin's formula as a finite expansion with a remainder. Although the remainder that in this case equals $f(x)$ itself, does not tend to 0 *as* $n \to \infty$ *for a fixed* $x \neq 0$, it tends to 0 faster than any power of x, *as* $x \to 0$*, for any fixed* n. The "expansion" gives, e.g., an *absolute* error less than 10^{-43} for $x = 0.1$, but the *relative* error is 100%. Also note that this function (and there are lots of other examples) can be added to any function without changing its Maclaurin expansion.

From the point of view of complex analysis, however, the origin is a singular point for this function, note, e.g., that $|f(z)| \to \infty$ as $z \to 0$ along the imaginary axis, and this prevents the application of any theorem that would guarantee that the infinite Maclaurin series represents the function. This trouble does not occur for a truncated Maclaurin expansion around a point, where the function under consideration is analytic. The size of the first non-vanishing neglected term then gives a good hint about the truncation error, when $|z|$ is a small fraction of the radius of convergence.

The above example may sound like a purely theoretical matter of curiosity. We emphasize this *distinction between the convergence and the validity of an infinite expansion* in this text, as a background to other expansions of importance in numerical computation, e.g., the Euler–Maclaurin expansion in Sec. 3.4.4, which may converge to the wrong result, also in the application to a well-behaved analytic function. On the other hand, we shall see, e.g., in Sec. 3.1.8, that divergent expansions can sometimes be very useful. The universal recipe *in numerical computation* is to consider an infinite series as a finite expansion plus a remainder term. A more algebraic point of view on a series is, however, often useful *in the design of a numerical method.* See, e.g., Sec. 3.1.5 (Formal Power Series) and Sec. 3.3.2 (The Calculus of Operators). Convergence of an expansion is neither necessary nor sufficient for its success in practical computation.

3.1.3 Analytic Continuation

Analytic functions have many important properties that you may find in any text on complex analysis. A good summary for the purpose of numerical mathematics is found in the first chapter of Stenger [43]. Two important properties are contained in the following lemma.

We remark that the region of analyticity of a function $f(z)$ is an *open* set. If, e.g., we say that $f(z)$ is analytic on a closed real interval, it means that there exists an open set in $\mathbf{C}$ that contains this interval, where $f(z)$ is analytic.

Lemma 3.1.7.

An analytic function can only have a finite number of zeros in a compact subset of the region of analyticity, unless the function is identically zero.

Suppose that two functions f_1 and f_2 are analytic in regions D_1 and D_2, respectively. Suppose that $D_1 \cap D_2$ contains an interval throughout which $f_1(z) = f_2(z)$. Then $f_1(z) = f_2(z)$ in the intersection $D_1 \cap D_2$.

Proof. We refer, for the first part, to any text on Complex Analysis. We here follow Titchmarsh [45] closely. The second part follows by the application of the first part to the function $f_1 - f_2$. □

A consequence of this is known as *the permanence of functional equations*, i.e. in order to prove the validity of a functional equation (or “a formula for a function”) in a region of the complex plane, it may be sufficient to prove its validity in (say) an interval of the real axis, under the conditions specified in the lemma.

Example 3.1.8. *The permanence of functional equations.*

We know from elementary real analysis that the functional equation

$$e^{(p+q)z} = e^{pz}e^{qz}, \quad (p, q \in \mathbf{R}),$$

holds for all $z \in \mathbf{R}$. We also know that all the three functions involved are analytic for all $z \in \mathbf{C}$. Set in the lemma $D_1 = D_2 = \mathbf{C}$, and let "the interval" be any compact interval of $\mathbf{R}$. The lemma then tells us that that the displayed equation holds for all complex z.

The right and the left hand side then have identical power series. Applying the convolution formula and matching the coefficients of z^n, we obtain

$$\frac{(p+q)^n}{n!} = \sum_{j=0}^{n} \frac{p^j}{j!}\frac{q^{n-j}}{(n-j)!}, \quad \text{i.e.,} \quad (p+q)^n = \sum_{j=0}^{n} \frac{n!}{j!(n-j)!} p^j q^{n-j}.$$

This is not a very sensational result. It is more interesting to start from the following functional equation

$$(1+z)^{p+q} = (1+z)^p(1+z)^q.$$

The same argumentation holds, except that—by the discussion around Table 3.1.1—D_1, D_2 should be equal to the complex plane with a cut from -1 to $-\infty$, and that the Maclaurin series is convergent in the unit disk only.

We obtain the equations

$$\binom{p+q}{n} = \sum_{j=0}^{n} \binom{p}{j}\binom{q}{n-j}, \quad n = 0, 1, 2, \ldots. \tag{3.1.20}$$

(They can also be proved by induction, but it is not needed.) This sequence of algebraic identities, where *each identity contains a finite number of terms*, is equivalent to the above functional equation.

We shall see that this observation is useful for motivating certain "*symbolic computations*" with power series, that can provide elegant derivations of useful formulas in numerical mathematics.

Now we may consider the aggregate of values of $f_1(z)$ and $f_2(z)$ at points interior to D_1 or D_2 as a single analytic function f. Thus f is analytic in the union $D_1 \cup D_2$, and $f(z) = f_1(z)$ in D_1, $f(z) = f_2(z)$ in D_2.

The function f_2 may be considered as extending the domain in which f_1 is defined, and it is called a (single-valued) **analytic continuation** of f_1. In the same way f_1 is an analytic continuation of f_2. Analytic continuation denotes both this process of extending the definition of a given function, and the result of the process. We shall see examples of this, e.g., in Sec. 3.1.3. Under certain conditions the analytic continuation is unique.

Theorem 3.1.8.

Suppose that a region D is overlapped by regions D_1, D_2, and that $(D_1 \cap D_2) \cap D$ contains an interval. Let f be analytic in D, and let f_1 be an analytic continuation of f to D_1, and let f_2 an analytic continuation of f to D_2, so that

$$f(z) = f_1(z) = f_2(z) \quad \text{in} \quad (D_1 \cap D_2) \cap D.$$

Then either of these functions provides a single-valued analytic continuation of f to $D_1 \cap D_2$. The results of the two processes are the same.

Proof. Since $f_1 - f_2$ is analytic in $D_1 \cap D_2$, and $f_1 - f_2 = 0$ in the set $(D_1 \cap D_2) \cap D$, which contains an interval, it follows from the lemma that $f_1(z) = f_2(z)$ in $D_1 \cap D_2$, which proves the theorem. □

If the set $(D_1 \cap D_2) \cap D$ is *void*, the conclusion in the theorem *may not be valid*. We may still consider the aggregate of values as a single analytic function, but *this function can be multi-valued in* $D_1 \cap D_2$.

Example 3.1.9.

For $|x| < 1$ the important formula

$$\arctan x = \frac{1}{2i} \ln \left(\frac{1 + ix}{1 - ix} \right)$$

easily follows from the expansions in the Table 3.1.1. The function $\arctan x$ has an analytic continuation as single-valued functions in the complex plane with cuts along the imaginary axis from i to ∞ and from $-i$ to $-\infty$. It follows from the theorem that "the important formula" is valid in this set.

3.1.4 Manipulating Power Series

In some contexts, algebraic recurrence relations can be used for the computation of the coefficients in Maclaurin expansions, in particular if only a moderate number of coefficients are wanted. We shall study a few examples.

Example 3.1.10. *Expansion of a composite function.*

Let $g(x) = b_0 + b_1 x + b_2 x^2 + \dots$, $f(z) = a_0 + a_1 z + a_2 z^2 + \dots$, be given functions, analytic at the origin. Find the power series

$$h(x) = f(g(x)) \equiv c_0 + c_1 x + c_2 x^2 + \dots.$$

In particular, we shall study the case $f(z) = e^z$.

The first idea we may think of is to substitute the expansion $b_0 + b_1 x + b_2 x^2 + \dots$ for z into the power series for $f(z)$. This is, however, *no good unless* $g(0) = b_0 = 0$, because

$$(g(x))^k = b_0^k + k b_0^{k-1} b_1 x + \dots$$

gives a contribution to, e.g., c_0, c_1, for every k, so we cannot successively compute the c_j by *finite* computation.

Now suppose that $b_0 = 0$, $b_1 = 1$, i.e. $g(x) = x + b_2 x^2 + b_3 x^3 + \dots$. (The assumption that $b_1 = 1$ is not important, but it simplifies the writing.) Then c_j depends only on b_k, a_k, $k \le j$, since $(g(x))^k = x^k + k b_2 x^{k+1} + \dots$. We obtain

$$h(x) = a_0 + a_1 x + (a_1 b_2 + a_2) x^2 + (a_1 b_3 + 2 a_2 b_2 + a_3) x^3 + \dots,$$

and the coefficients of $h(x)$ come out recursively,

$$c_0 = a_0; \quad c_1 = a_1, \quad c_2 = a_1 b_2 + a_2, \quad c_3 = a_1 b_3 + 2a_2 b_2 + a_3, \ldots.$$

Now consider the case $f(z) = e^z$, i.e. $a_n = 1/n!$. We first see that it is then also easy to handle the case that $b_0 \neq 0$, since

$$e^{g(x)} = e^{b_0} e^{b_1 x + b_2 x^2 + b_3 x^3 + \ldots}.$$

But there exists a more important simplification if $f(z) = e^z$. Note that h satisfies the differential equation $h'(x) = g'(x)h(x)$, $h(0) = e^{b_0}$. Hence

$$\sum_{n=0}^{\infty} (n+1) c_{n+1} x^n \equiv \sum_{j=0}^{\infty} (j+1) b_{j+1} x^j \sum_{k=0}^{\infty} c_k x^k.$$

Set $c_0 = e^{b_0}$, apply the convolution formula (3.1.8), and match the coefficients of x^n on the two sides:

$$(n+1)c_{n+1} = b_1 c_n + 2b_2 c_{n-1} + \ldots + (n+1) b_{n+1} c_0, \quad (n = 0, 1, 2, \ldots).$$

This recurrence relation is more easily programmed than the general procedure indicated above. Other functions that satisfy appropriate differential equations can be treated similarly; see Problem 8. More information is found in Knuth [29, Sec. 4.7].

Formulas like these are often used in packages for **symbolic differentiation** and for **automatic** or **algorithmic differentiation**. Expanding a function into a Taylor series is equivalent to finding the sequence of derivatives of the function at a given point. The goal of *symbolic* differentiation is to obtain analytic *expressions* for derivatives of functions given in analytic form. This is handled by computer algebra systems, e.g., Maple or Mathematica.

In contrast, the goal of **automatic** or **algorithmic differentiation** is to extend an algorithm (a program) for the computation of the *numerical values* of a few functions to an algorithm that also computes *the numerical values* of a few derivatives of these functions, without truncation errors. A simple example, Horner's scheme for computing values and derivatives for a polynomial was given in Sec. 1.3.1. At the time of writing, there is a lively activity about automatic differentiation—theory, software development and applications. Typical applications are in the solution of ordinary differential equations by Taylor expansion; see Example 3.1.1. Such techniques are also used in optimization for partial derivatives of low order, e.g., for the computation of Jacobian and Hessian matrices.

Sometimes power series are needed with many terms, although rarely more than 30 (say). (The ill-conditioned series are exceptions; see Sec. 3.2.5.) The determination of the coefficients can be achieved by the **Toeplitz matrix method** using floating point computation and an interactive matrix language. Computational details will be given in Problems 9–12 of this section for MATLAB . (Systems like Maple and Mathematica that include exact arithmetic and other features, are evidently also useful here.) An alternative method, the **Cauchy–FFT method**, will be described in Sec. 3.2.2.

Both methods will be applied later in the book. See in particular Sec. 3.3.4, where they are used for deriving approximation formulas in the form of *expansions in powers of elementary difference or differential operators.* In such applications, the coefficient vector, v (say), is obtained in floating point (usually in a very short time). Very accurate rational approximations to v, often even the exact values, can be obtained (again in a very short time) by applying MATLAB function $[N, D] =$ `rat`(z, Tol) to the results, with two different values of the tolerance. This function is based on a continued fraction algorithm, given in Example 3.5.2 for finding the best rational approximation to a real number. This can be used for the *"cleaning"* of numerical results which have, for practical reasons, been computed by floating point arithmetic, although the exact results are known to be (or strongly believed to be) rather simple rational numbers. The algorithm attempts to remove the *"dirt"* caused by computational errors. In Sec. 3.5.1 you also find some comments of importance for the interpretation of the results, e.g., for judging whether the rational numbers are exact results or good approximations only.

Let

$$f(z) = \sum_{j=0}^{\infty} a_j z^j$$

be the power series of a function analytic at $z = 0$. With this power series we can associate an infinite upper triangular **semicirculant matrix**

$$C_f = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & \dots \\ & a_0 & a_1 & a_2 & \dots \\ & & a_0 & a_1 & \dots \\ & & & a_0 & \dots \\ & & & & \ddots \end{pmatrix}, \tag{3.1.21}$$

Similarly, a truncated power series $f_N(z) = \sum_{j=0}^{N-1} a_j z^j$ is represented by the finite leading principal $N \times N$ submatrix of C_f (see Definition A.3.1), which can be written as

$$f_N(S_N) = \sum_{j=0}^{N-1} a_j S_N^j, \tag{3.1.22}$$

where S_N is a **shift matrix**. For example, with $N = 4$,

$$f_N(S_N) = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 \\ 0 & a_0 & a_1 & a_2 \\ 0 & 0 & a_0 & a_1 \\ 0 & 0 & 0 & a_0 \end{pmatrix}, \qquad S_N = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The following properties of S_N explains the term "shift matrix":

$$S_N \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} x_2 \\ x_3 \\ x_4 \\ 0 \end{pmatrix}, \qquad (x_1, x_2, x_3, x_4) S_N = (0, x_1, x_2, x_3).$$

An $N \times N$ matrix, with a structure analogous to $f_4(S_4)$, is known as an upper triangular **Toeplitz matrix.**[9]. Matrices (not necessarily triangular), whose entries are constant along each diagonal, are called *Toeplitz matrices.*

What do the powers of S_N look like? Note that $S_N^N = 0$, i.e. S_N is a **nilpotent** matrix. This is one of the reasons why the Toeplitz matrix representation is convenient for work with truncated power series, since it follows that

$$f(S_N) = \sum_{j=0}^{\infty} a_j S_N^j = \sum_{j=0}^{N-1} a_j S_N^j = f_N(S_N).$$

It is easily verified that a **product** of upper triangular Toeplitz matrices is of the same type. Also note that the multiplication of such matrices is *commutative.* It is also evident that a **linear combination** of such matrices is of the same type. Further it holds that

$$\begin{aligned}(f \cdot g)(S_N) &= f(S_N)g(S_N) = f_N(S_N)g_N(S_N);\\ (\alpha f + \beta g)(S_N) &= \alpha f_N(S_N) + \beta g_N(S_N).\end{aligned}$$

(In general, Toeplitz matrices are not nilpotent, and the product of two *non-triangular* Toeplitz matrices is not a Toeplitz matrix. Similarly for the inverse. In this section we shall only deal with upper triangular Toeplitz matrices.)

Denote by e_1^T the first row of the unit matrix of a size appropriate in the context. An upper triangular Toeplitz matrix of order N is uniquely *determined by its first row* r by means of a simple and fast algorithm that we call $toep\,(r, N)$. For example, the unit matrix of order N is $I_N = toep\,(e_1^T, N)$, and the shift matrix is $S_N = toep\,([0\ e_1^T], N)$. A MATLAB implementation is given in Problem 9.

Now it will be indicated how one can save CPU time and memory space by working on the row vector level, with the first rows instead of with the full triangular matrices.[10] We shall *denote by $f1$, $g1$, the row vectors with the first N coefficients of the Maclaurin expansions of $f(z)$, $g(z)$.* They are equal to the first rows of the matrices $f(S_N)$, $g(S_N)$, respectively. Suppose that $f1$, $g1$ are given and we shall compute $f \cdot g1$, i.e. the first row of $f(S_N) \cdot g(S_N)$ in a similar notation. Then

$$f \cdot g1 = e_1^T(f(S_N) \cdot g(S_N)) = (e_1^T f(S_N)) \cdot g(S_N) = f1 \cdot \mathrm{toep}(g_1, N). \tag{3.1.23}$$

Note that you never have to multiply two triangular matrices, if you work with the first rows only. So, only about $N^2/2$ flops and (typically) an application of the toep(r, N) algorithm, are needed instead of about $N^3/6$ if two upper triangular matrices are multiplied; see Sec. 1.4.1, where the operation count for matrix multiplication is discussed.

Similarly the **quotient** *of two upper triangular Toeplitz matrices*, (say)

$$Q(S_N) = f(S_N) \cdot g(S_N)^{-1},$$

[9] Otto Toeplitz (1881-1940), German mathematician

[10] In interactive computations with rather short series the gain of time may sometimes be neutralized by an increased number of manual operations. See the computer exercises.

is also a matrix of the same type. (A hint to a proof is given in Problem 9.[11]) Note that $Q(S_N) \cdot g(S_N) = f(S_N)$. With similar notations as above, we obtain for the first row of this matrix equation the following triangular linear system where the *row* vector $q1$ is the unknown.

$$q1 \cdot \mathrm{toep}(g1, N) = f1. \tag{3.1.24}$$

Although the discussion in Sec. 1.3.4 is concerned with a linear system with a *column* as the unknown (instead of a row), we draw from it the conclusion that only about $N^2/2$ scalar flops (including N scalar divisions) and one application of the toep algorithm, are needed, instead of the $N^3/6$ needed in the solution of the matrix equation $Q \cdot g(S_N) = f(S_N)$.[12]

A library called toeplib consists of short Matlab scripts mainly based on Table 3.1.2. It is given in Problem 10 (a). In Problem 10 (b), etc., the series of the library are combined by elementary operations to become interesting examples of the Toeplitz matrix method. The convenience, the accuracy and the execution time are probably much better than you expect; even the authors were surprised.

Next we shall study how a **composite function** $h(z) = f(g(z))$ can be expanded in powers of z. *Suppose that* $f(z)$ and $g(z)$ are analytic at $z = 0$, $f(z) = \sum_{j=1}^{\infty} f1(j)z^{j-1}$. An **important assumption** is that $g(0) = 0$. Then we can set $g(z) = z\bar{g}(z)$, hence $(g(z))^n = z^n(\bar{g}(z))^n$ and, because $S_N^n = 0$, $n \geq N$, we obtain

$$(g(S_N))^n = S_N^n \cdot (\bar{g}(S_N))^n = 0, \quad \text{if } n \geq N \text{ and } g(0) = 0,$$

$$h(S_N) \equiv f(g(S_N)) = \sum_{j=1}^{N} f1(j)(g(S_N))^{j-1}, \quad \text{if } g(0) = 0. \tag{3.1.25}$$

This matrix polynomial can be computed by a matrix version of Horner's scheme. The row vector version of this equation is written

$$h1 = \mathrm{comp}(f1, g1, N). \tag{3.1.26}$$

A Matlab implementation of the function comp is listed and applied in Problem 11.

If $g(0) \neq 0$, Equation (3.1.25) still provides an "expansion", but it is **wrong**; see Problem 11 (c). Suppose that $|g(0)|$ is less than the radius of convergence of the Maclaurin expansion of $f(x)$. Then a correct expansion is obtained by a different decomposition. Set $\tilde{g}(z) = g(z) - g(0)$, $\tilde{f}(x) = f(x + g(0))$. Then $\tilde{f}$, $\tilde{g}$ are analytic

[11] In the terminology of algebra, the set of upper triangular $N \times N$ Toeplitz matrices, i.e. $\{\sum_{j=0}^{N-1} \alpha_j S_N^j\}$, $\alpha_j \in \mathbf{C}$, is a **commutative integral domain** that is isomorphic with the set of polynomials $\sum_{j=0}^{N-1} \alpha_j x^j$ modulo x^N, where x is an indeterminate.

[12] The equations (3.1.23) and (3.1.24) are mathematically equivalent to the convolution product in (3.1.8) and the procedure demonstrated in Example 3.1.6, respectively. Sometimes both procedures suffer from the growth of the effects of rounding errors when n is very large, in particular when the power series are ill- conditioned; see Sec. 3.1.11. An advantage of the Toeplitz matrix method is that the coding, in a language with convenient matrix handling, becomes easier.

at $z = 0$. $\tilde{g}(0) = 0$ and $\tilde{f}(\tilde{g}(z)) = f(g(z)) = h(z)$. So, (3.1.25) and its row vector implementations can be used if $\tilde{f}, \tilde{g}$ are substituted for f, g.

Analytic functions of matrices are defined, in terms of their Taylor series; see Sec. 9.2.5 in Vol. II. For example, the series

$$e^A = I + A + \frac{A^2}{2!} + \frac{A^2}{2!} + \cdots,$$

converges elementwise for any matrix A. There exist several algorithms for computing e^A, $\sqrt{A}$, $\log A$, where A is a square matrix. One can make linear combinations, products, quotients and composite functions of them. For example, a "principal matrix value" of $Y = (I + A)^\alpha$ is obtained by

$$B = \log(I + A), \quad Y = e^{\alpha B}.$$

For a composite function $f(g(A))$, it is here *not* necessary that $g(0) = 0$, but it is *important* that $g(z)$ and $f(g(z))$ are analytic when z is an eigenvalue of A. We obtain *truncated power series if* $A = S_N$; note that S_N has a multiple eigenvalue at 0. The coding, and the manual handling in interactive computing, are convenient with matrix functions, but the computer has to perform more operations on full triangular matrices than with the row vector level algorithms described above. So, for *very* long expansions the earlier algorithms are notably faster.

If the given power series, $f(x)$, $g(x), \ldots$ have *rational coefficients*, then the exact results of a sequence of additions, multiplications, divisions, compositions, differentiations, integrations will have rational coefficients, because the algorithms are all formed by a finite number of scalar additions, multiplications and divisions. As mentioned above, very accurate rational approximations, often even the exact values, can be quickly obtained by applying a continued fraction algorithm that is presented in Sec. 3.5.2 to the results of a floating point computation. See also Problems 9–12.

If $f(x)$ is an even function, its power series contains only even powers of x. You gain space and time, by letting the shift matrix S_N correspond to x^2 (instead of x). Similarly, if $f(x)$ is an odd function, you can instead work with the even function $f(x)/x$, and let S_N correspond to x^2. See Problems 9–12.

Finally we consider a classical problem of mathematics, known as **power series reversion**. The task is to find the power series for the **inverse function** $x = g(y)$ of the function $y = f(x) = \sum_{j=0}^{\infty} a_j x^j$, in the particular case where $a_0 = 0$, $a_1 = 1$. Note that even if the series for $f(x)$ is finite, the series for $g(y)$ is in general infinite!

The following simple cases of power series reversion are often sufficient and useful in low order computations with paper and pencil.

$$\begin{aligned} & y = x + ax^k + \ldots, \quad (k > 1), \\ \Rightarrow\ & x = y - ax^k - \ldots = y - ay^k - \ldots; \qquad (3.1.27) \\ & y = f(x) \equiv x + a_2x^2 + a_3x^3 + a_4x^4 + \ldots, \\ \Rightarrow\ & x = g(y) \equiv y - a_2y^2 + (2a_2^2 - a_3)y^3 - (5a_2^2 - 5a_2a_3 + a_4)y^4 + \ldots. \qquad (3.1.28) \end{aligned}$$

An application of power series reversion occurs in the derivation of a family of iterative methods of arbitrary high order for solving scalar non-linear equations; see Sec. 6.3.5.

Knuth [29] devotes Sec. 4.7 to the manipulation of power series. He presents several algorithms for power series reversion, e.g., a classical algorithm due to Lagrange 1768 that requires $O(N^3)$ operations to compute the first N terms. Knuth also includes a recent algorithm due to Brent and Kung [7]. It is based on an adaptation, to formal power series, of Newton's method (1.2.3) for solving a numerical algebraic equation. For power series reversion, the equation to be solved reads

$$f(g(y)) = y, \tag{3.1.29}$$

where the coefficients of g are the unknowns. The number of correct terms is roughly doubled in each iteration, as long as N is not exceeded. In the usual numerical application of Newton's method to a scalar non-linear equation (see Secs. 1.2 and 6.3) it is the number of significant digits that is (approximately) doubled, so-called quadratic convergence. Brent–Kung's algorithm can be implemented in about $150\,(N \log N)^{3/2}$ scalar flops.

In Problem 12, a convenient Toeplitz matrix implementation of the idea of Brent and Kung is presented. It requires about $cN^3 \log N$ scalar flops with a moderate value of c. It is thus much inferior to the original algorithm if N is very large. In some interesting interactive applications, however, N rarely exceeds 30. In such cases our implementation is satisfactory, unless (say) hundreds of series are to be reversed.

3.1.5 Formal Power Series

A power series is not only a means for numerical computation; it is also an aid for deriving formulas in numerical mathematics and in other branches of applied mathematics. Then one has another, more algebraic, aspect of power series that we shall briefly introduce. A more rigorous and detailed treatment is found in Henrici [25, Chapter 1], and in the literature quoted there.

In a **formal power series**, $\mathbf{P} = a_0 + a_1\mathbf{x} + a_2\mathbf{x}^2 + \cdots$, the coefficients a_j may be real or complex numbers (or elements in some other field), while $\mathbf{x}$ is an algebraic **indeterminate**; $\mathbf{x}$ and its powers can be viewed as place keepers. No real or complex values are assigned to $\mathbf{x}$ and $\mathbf{P}$. Convergence, divergence and remainder term have no relevance for formal power series. The coefficients of a formal power series may even be such that the series diverges for any non-zero complex value that you substitute for the indeterminate, e.g. the series

$$\mathbf{P} = 0! - 1!\mathbf{x} + 2!\mathbf{x}^2 - 3!\mathbf{x}^3 + \cdots. \tag{3.1.30}$$

In algebraic terminology, the set of formal power series is an integral domain. The sum of $\mathbf{P}$ and another formal power series, $\mathbf{Q} = b_0 + b_1\mathbf{x} + b_2\mathbf{x}^2 + \cdots$, is *defined* as

$$\mathbf{P} + \mathbf{Q} = (a_0 + b_0) + (a_1 + b_1)\mathbf{x} + (a_2 + b_2)\mathbf{x}^2 + \cdots.$$

Similarly, the *Cauchy product* is *defined* as

$$\mathbf{PQ} = c_0 + c_1\mathbf{x} + c_2\mathbf{x}^2 + \cdots,$$

where the coefficients are given by the convolution formula (3.1.8). The division of two formal power series is defined by a recurrence, as indicated in Example 3.1.5, iff the first coefficient of the the denominator is not zero.

Other operations are defined without surprises, e.g., the derivative of $\mathbf{P}$ is *defined* as $\mathbf{P}' = 1a_1 + 2a_2\mathbf{x} + 3a_3\mathbf{x}^2 + \ldots$. The limit process, by which the derivative is defined in Calculus, does not exist for formal power series. The usual rules for differentiation are still valid, and as an exercise you may verify that the formal power series defined by (3.1.30) satisfies the formal differential equation $\mathbf{x}^2\mathbf{P}' = \mathbf{x} - \mathbf{P}$.

Formal power series can be used for deriving identities. In most applications in this book difference operators or differential operators are substituted for the indeterminates, and the identities are then used in the deriving of approximation formulas, e.g. for interpolation, numerical differentiation and integration etc.

The formal definitions of the Cauchy product, (i.e. convolution) and division are rarely used in practical calculation. It is easier to work with upper triangular $N \times N$ Toeplitz matrices, as in Sec. 3.1.5, where N is any natural number. Algebraic calculations with these matrices are isomorphic with calculations with formal power series modulo $\mathbf{x}^N$.

If you perform operations on matrices $f_M(S)$, $g_M(S), \ldots$, where $M < N$, the results are equal to the principal $M \times M$ submatrices of the results obtained with the matrices $f_N(S)$, $g_N(S), \ldots$. This fact follows directly from the equivalence with power series manipulations. It is also related to the fact that in the multiplication etc. of block upper triangular matrices, the diagonal blocks of the product equals the products of the diagonal blocks, and no new off-diagonal blocks enter.

So, we can easily *define the product of two infinite upper triangular matrices*, $C = AB$, by stating that if $i \le j \le n$ then c_{ij} has the same value that it has in the $N \times N$ submatrix $C_N = A_N B_N$ for every $N \ge n$. In particular C is upper triangular, and note that there are no conditions on the behaviour of the elements a_{ij}, b_{ij} as $i, j \to \infty$. One can show that this product is associative and distributive. For the infinite triangular Toeplitz matrices it is commutative too.[13]

Henrici [25, Sec. 1.3], calls this a representation of formal power series by *infinite* upper triangular Toeplitz matrices, (which he names *semicirculants*), and proves that *the mapping of the set of formal power series onto the set of semicirculants is an isomorphism.* If the formal power series $a_0 + a_1\mathbf{x} + a_2\mathbf{x}^2 + \cdots$, and its reciprocal series, which exists iff $a_0 \ne 0$, are represented by the semicirculants A and B, respectively, Henrici proves that $AB = BA = I$, where I is the unit matrix of infinite order. This indicates how to define the inverse of any infinite upper triangular matrix if all diagonal elements $a_{ii} \ne 0$.

If a function f of a complex variable z is analytic at the origin, then *we define*[14] $f(\mathbf{x})$ as the formal power series with the same coefficients as the Maclaurin series for $f(z)$. In the case of a multivalued function we take the principal branch.

[13] For infinite *non-triangular* matrices the definition of a product generally contains conditions on the behaviour of the elements as $i, j \to \infty$, but we shall not discuss this here.

[14] Henrici, loc. cit., does not use this concept—it may not be established.

We do *not* consider formal power series with several indeterminates. There may occur expressions with several bold-type symbols. Only one of them is the indeterminate, and the others are shorthand notations for analytic functions of this indeterminate.

There is a kind of "permanence of functional equations" also for the generalization from a function $g(z)$ of a complex variable that is analytic at the origin, to the formal power series $g(\mathbf{x})$. We illustrate a *general principle* on an important special example that we formulate as a lemma, since we shall need it in the next section.

Lemma 3.1.9.

$$(e^{\mathbf{x}})^{\theta} = e^{\theta\mathbf{x}}, \quad (\theta \in \mathbf{R}). \tag{3.1.31}$$

Proof. Let the coefficient of $\mathbf{x}^j$ in the expansion of the left hand side be $\phi_j(\theta)$. The corresponding coefficient for the right hand side is $\theta^j/j!$. If we replace $\mathbf{x}$ by a complex variable z, the power series coefficients are the same, and we know that $(e^z)^{\theta} = e^{\theta z}$, hence $\phi_j(\theta) = \theta^j/j!$, $j = 1, 2, 3 \ldots$, hence $\sum_0^{\infty} \phi_j(\theta)\mathbf{x}^j = \sum_0^{\infty}(\theta^j/j!)\mathbf{x}^j$, and the lemma follows. □

Example 3.1.11.

Find (if possible) a formal power series $\mathbf{Q} = 0 + b_1\mathbf{x} + b_2\mathbf{x}^2 + b_3\mathbf{x}^3 + \ldots$, that satisfies the equation

$$e^{-\mathbf{Q}} = 1 - \mathbf{x}, \tag{3.1.32}$$

where $e^{-\mathbf{Q}} = 1 - \mathbf{Q} + \mathbf{Q}^2/2! - \ldots$.

We can, in principle, determine an arbitrarily long sequence $b_1, b_2, b_3, \ldots b_k$ by matching the coefficients of $\mathbf{x}, \mathbf{x}^2, \mathbf{x}^3, \ldots \mathbf{x}^k$, in the two sides of the equation. We display the first three equations.

$$1 - (b_1\mathbf{x} + b_2\mathbf{x}^2 + b_3\mathbf{x}^3 + \ldots) + (b_1\mathbf{x} + b_2\mathbf{x}^2 + \ldots)^2/2 - (b_1\mathbf{x} + \ldots)^3/6 + \ldots$$
$$= 1 - 1\mathbf{x} + 0\mathbf{x}^2 + 0\mathbf{x}^3 + \ldots.$$

For any natural number k, the matching condition is of the form

$$-b_k + \phi_k(b_{k-1}, b_{k-2}, \ldots, b_1) = 0.$$

This shows that the coefficients are *uniquely* determined.

$$\begin{aligned} -b_1 = -1 \Rightarrow&\ b_1 = 1; \\ -b_2 + b_1^2/2 = 0 \Rightarrow&\ b_2 = 1/2; \\ -b_3 + b_1b_2 - b_1/6 = 0 \Rightarrow&\ b_3 = 1/3; \end{aligned}$$

There exists, however, *a much easier way* to determine the coefficients. For the analogous problem with a complex variable z, we know that the solution is unique:

$q(z) = -\ln(1-z) = \sum_1^\infty z^j/j$ (the principal branch, where $b_0 = 0$), and hence $\sum_1^\infty \mathbf{x}^j/j$ is the unique formal power series that solves the problem, and we can use the notation $\mathbf{Q} = -\ln(1-\mathbf{x})$ for it.[15]

This example will be applied in Example 3.2.18 to the derivation of formulas for numerical differentiation.

The theory of formal power series can in a similar way justify many elegant "symbolic" applications of power series for deriving mathematical formulas.

Review Questions

1. (a) Formulate three general theorems that can be used for estimating the remainder term in numerical series.

(b) What can you say about the remainder term, if the nth term is $O(n^{-k})$, $k > 1$? Suppose in addition that the series is alternating. What further condition should you add, in order to guarantee that the remainder term will be $O(n^{-k})$?

2. Give, with convergence conditions, the Maclaurin series for $\ln(1+x)$, e^x, $\sin x$, $\cos x$, $(1+x)^k$, $(1-x)^{-1}$, $\ln\dfrac{1+x}{1-x}$, $\arctan x$.

3. Describe the main features of a few methods to compute the Maclaurin coefficients of, e.g., $\sqrt{2e^x - 1}$.

4. Give generating functions of the Bernoulli and the Euler numbers. Describe generally how to derive the coefficients in a quotient of two Maclaurin series.

5. If a functional equation, e.g. $4(\cos x)^3 = \cos 3x + 3\cos x$, is known to be valid for real x, how do you know that it holds also for all complex x? Explain what is meant by the statement that it holds also for formal power series, and why is this true?

6. (a) Show that multiplying two arbitrary upper triangular matrices of order N uses $\sum_{k=1}^N k(N-k) \approx N^3/6$ flops, compared to $\sum_{k=1}^N k \approx N^2/2$ for the product of a row vector and an upper triangular matrix.

(b) Show that if $g(x)$ is a power series and $g(0) = 0$, then $g(S_N)^n = 0$, $n \geq N$. Make an operation count for the evaluation of the matrix polynomial $f(g(S_N))$ by the matrix version of Horner's scheme.

(c) Consider the product $f(S_N)g(S_N)$, where $f(x)$ and $g(x)$ are two power series. Show, using rules for matrix multiplication, that for any $M < N$ the leading $M \times M$ block of the product matrix equals $f(S_M)g(S_M)$.

7. Consider a power series $y = f(x) = \sum_{j=0}^\infty a_j x^j$, where $a_0 = 0$, $a_1 = 1$. What is meant by reversion of this power series? In the Brent–Kung method the problem of reversion of a power series is formulated as a nonlinear equation. Write this equation for the Toeplitz matrix representation of the series.

[15] The three coefficients b_j computed above agree, of course, with $1/j$, $j = 1:3$.

8. Let $\mathbf{P} = a_0 + a_1\mathbf{x} + a_2\mathbf{x}^2 + \cdots$ and $\mathbf{Q} = b_0 + b_1\mathbf{x} + b_2\mathbf{x}^2 + \cdots$ be two formal power series. Define the sum $\mathbf{P} + \mathbf{Q}$ and the Cauchy product $\mathbf{PQ}$.

Problems and Computer Exercises

1. In how large a neighborhood of $x = 0$ does one get, respectively, four and six correct decimals using the following approximations?

(a) $\sin x \approx x$; (b) $(1+x^2)^{-1/2} \approx 1-x^2/2$; (c) $(1+x^2)^{-1/2}e^{\sqrt{\cos x}} \approx e(1-\frac{3}{4}x^2)$.

Comment: The truncation error is asymptotically qx^p where you know (?) p. An alternative to an exact algebraic calculation of q, is a numerical estimation of q, by means of the actual error for a suitable value of x—neither too big nor too small (!). (Check the estimate of q for another value of x.)

2. (a) Let a, b, be the lengths of the two smaller sides of a right angle triangle, $b \ll a$. Show that the hypotenuse is approximately $a + b^2/(2a)$ and estimate the error of this approximation. If $a = 100$, how large is b allowed to be, in order that the absolute error should be less than 0.01?

(b) How large a relative error do you commit, when you approximate the length of a small circular arc by the length of the chord? How big is the error if the arc is 100 km on a great circle of the earth? (Approximate the earth by a ball of radius $40000/(2\pi)$ km.)

(c) $P(x) = 1 - \frac{1}{2}x^2 + \frac{1}{24}x^4$ is a polynomial approximation to $\cos x$ for small values of $|x|$. Estimate the errors of $P(x)$, $P'(x)$, $\frac{1}{x}\int_0^x P(t)\,dt$, and compare them, e.g., for $x = 0.1$.

(d) How accurate is the formula $\arctan x \approx \pi/2 - 1/x$ for $x \gg 1$?

3. (a) Compute $10-(999.999)^{1/3}$ to 9 significant digits, by the use of the binomial expansion. Compare with the result obtained by a computer in IEEE double precision, directly from the first expression.

(b) How many terms of the Maclaurin series for $\ln(1 + x)$ would you need in order to compute $\ln 2$ with an error less than 10^{-6} ? How many terms do you need, if you use instead the the series for $\ln{(1 + x)/(1 - x)}$, with an appropriate choice of x?

4. Give an approximate expression of the form $ah^bf^{(c)}(0)$ for the error of the estimate of the integral $\int_{-h}^{h} f(x)dx$ obtained by Richardson extrapolation (according to Sec. 1.2.2) from the trapezoidal values $T(h)$ and $T(2h)$.

5. Compute, by means of appropriate expansions, not necessarily in powers of t, the following integrals to (say) five correct decimals.
(This is for paper, pencil and a pocket calculator.)

$$\text{(a)} \int_0^{0.1} (1 - 0.1\sin t)^{1/2}\,dt; \qquad \text{(b)} \int_{10}^{\infty} (t^3 - t)^{-1/2}\,dt.$$

6. (a) Expand $\arcsin x$ into powers of x by the integration of the expansion of $(1-x^2)^{-1/2}$.

(b) Use the result in (a) to prove the expansion

$$x = \sinh x - \frac{1}{2}\frac{\sinh^3 x}{3} + \frac{1\cdot 3}{2\cdot 4}\frac{\sinh^5 x}{5} - \frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}\frac{\sinh^7 x}{7} + \dots$$

7. (a) Consider the power series for

$$(1+x)^{-\alpha}, \quad x > 0, \quad 0 < \alpha < 1.$$

Show that it is equal to the hypergeometric function $F(\alpha, 1, 1, -x)$. Is it true that the expansion is alternating, and that the remainder has the same sign as the first neglected term, also if $x > 1$, where the series is divergent? What do the Theorems 3.1.3 and 3.1.4 tell you in the cases $x < 1$ and $x > 1$?

Comment: An application of the divergent case for $\alpha = \frac{1}{2}$ is found in Problem 3.2.9 (c).

(b) Express the coefficients of the power series expansions of $y \cot y$ and $\ln(\sin y/y)$ in terms of the Bernoulli numbers.

Hint: Set $x = 2iy$ into (3.1.17). Differentiate the second function.

(c) Find a recurrence relation for the Euler numbers E_n, see Example 3.1.5, and use it for showing that these numbers are integers.

(d) Show that

$$\ln\Big(\frac{z+1}{z-1}\Big) = 2\Big(\frac{1}{z} + \frac{1}{3z^3} + \frac{1}{5z^5} + \dots\Big), \quad |z| > 1.$$

Find a recurrence relation for the coefficients of the expansion

$$\Big(\ln\Big(\frac{z+1}{z-1}\Big)\Big)^{-1} = \frac{1}{2}z - \mu_1 z^{-1} - \mu_3 z^{-3} - \mu_5 z^{-5} - \dots, \quad |z| > 1.$$

Compute μ_1, μ_3, μ_5 and determine $\sum_0^\infty \mu_{2j+1}$ by letting $z \downarrow 1$. (Full rigor is not required.)

Hint: Look at Example 3.1.5.

8. The power series expansion $g(x) = b_1 x + b_2 x^2 + \dots$ is given. Find recurrence relations for the coefficients of the expansion for $h(x) \equiv f(g(x)) = c_0 + c_1 x + c_2 x^2 + \dots$ in the following cases:

(a) $h(x) = \ln(1 + g(x))$, $f(x) = \ln(1 + x)$.

Hint: Show that $h'(x) = g'(x) - h'(x)g(x)$. Then proceed analogously to Example 3.1.10.

Answer:

$$c_0 = 0, \quad c_n = b_n - \frac{1}{n}\sum_{j=1}^{n-1}(n-j)c_{n-j}b_j.$$

(b) $h(x) = (1 + g(x))^k$, $f(x) = (1 + x)^k$, $k \in \mathbf{R}$, $k \neq 1$.

Hint: Show that $g(x)h'(x) = kh(x)g'(x) - h'(x)$. Then proceed analogously to Example 3.1.10.

Answer:

$$c_0 = 1, \qquad c_n = \frac{1}{n}\sum_{j=1}^{n}\big((k+1)j - n\big)c_{n-j}b_j,$$

$n = 1, 2, \ldots$. The recurrence relation is known as the J. C. P. Miller formula.

(c) $h_1(y) = \cos g(x)$, $h_2(y) = \sin g(x)$, simultaneously.

Hint: Consider instead $h(y) = e^{ig(x)}$, and separate real and imaginary parts afterwards.

9. In Problems 9–12 we use notations and results from the Toeplitz matrix representation in Sec. 3.1.4. For example, S_N denotes the Nth order shift matrix and $f1$ is the first row vector of the matrix $f(S_N)$. We also assume that you are familiar with simple MATLAB . We now present some more advanced notions from MATLAB that are important in our context.

The solution to the linear system (3.1.24), i.e. $q1 \cdot \text{toep}(g1, N) = f1$, can in MATLAB be neatly written as `q1 = f1/toep(g1,N)`. Note that this is the *vector by matrix* division of MATLAB .

If x is a vector, `cumprod(x)` is the *cumulative product* of the elements of x, e.g., `cumprod(2:5) = [2 6 24 120];` cumsum is analogously defined.

If some of the arguments of a function, in the sense of MATLAB , are optional, `nargin` is the number of input arguments actually used; `nargout` is defined analogously.

The MATLAB function `[Nu,De] = rat(v,Tol)` returns two integer vectors so that `abs(Nu./De - v) <= Tol*abs(v)`. There are several variants of the function `rat`; see the help file. This function is based on a version of the continued fraction algorithm presented in Sec. 3.5.2. Take at least two different values for TOL, and compare the results. Use the rational form of a result, only if it seems reliable and shorter than the floating form.

Choose $N = 6$ while you test that a code is correct. When you apply it or examine the properties of the algorithm, choose N in the range $[12, 24]$. (Even then the computing time may be too short to be measured by the "the stopwatch timer" tic ... toc; `tic` starts the timer; `toc`, by itself, displays the elapsed time since tic was used. (You can also *save* the elapsed time by a statement like `t = toc`.) If you choose N very large you may risk exponent underflow or overflow, or some other nuisance.

In most of the following examples, the algorithms are reasonably stable. Numerical instability can occur, however, depending on the functions $f, g, \ldots$ that they are applied to, and it is a good habit to try to compare the results of two "independent" ways to derive an expansion. In applications to ill-conditioned power series; see Sec. 3.2.5, high values of N are needed, and the results may sometimes be ruined by numerical instability, unless multiple precision is used.

(a) Convince yourself that the following function expands the row r to a triangular Toeplitz matrix. What is done if `length(r) < N` and why? What is the default value of the optional input argument `N`?

```
function T = toep(r,N);
% toep expands the row vector r into an upper
% triangular Toeplitz matrix T.
%N is an optional argument.
%
lr= length(r);
if (nargin==1 | lr > N), N = lr; end;
if lr<N,  r=[r, zeros(1,N-lr)]; end;
gs = zeros(N,N);
for i=1:N,
   gs(i,i:N) = r(1:N-i+1);
end
T = gs;
```

(b) If you want $N > 3$, terms in your results, although the number of terms in the given expression for $f(x)$, e.g., $(1+x+x^2)/(1-x+x^2)$ is smaller, you must augment this by sequences of zeros, so that the order of Toeplitz matrix becomes N. Show experimentally and theoretically that the first row of

$$(I_N + S_N + S_N^2)/(I_N - S_N + S_N^2)$$

is, e.g., obtained by the statement

```
[1, 1, 1, zeros(1,N-3)]/toep([1, -1, 1, zeros(1,N-3)]).
```

(c) Let $f(z) = -z^{-1}\ln(1-z)$. Compute the first six coefficients of the Maclaurin series for the functions $f(z)^k$, $k = 1:5$ in floating point, and convert them to rational form. (The answer and an application to numerical differentiation are given in Example 3.3.6.)

Comment: If you choose an appropriate tolerance in the Matlab function rat you will obtain an accurate rational approximation, but it is not necessarily exact. Try to judge which of the coefficients are exact.

(d) Compute in floating point the coefficients μ_{2j-1}, $j = 1:11$, defined in Problem 7 (d), and convert them to rational form.

Hint: First seek an equivalent problem for an expansion in ascending powers.

(e) Prove that $Q = f(S_N)g(S_N)^{-1}$ is an upper triangular Toeplitz matrix.

Hint: Define $Q = \text{toep}(q1, N)$, where $q1$ is defined by (3.1.24), and show that each row of the equation $Q \cdot g(S_N) = f(S_N)$ is satisfied.

10. (a) Study the following "library" of Matlab lines for common applications of the Toeplitz matrix method for arbitrary given values of N; the shift matrix S_N corresponds to the variable x. You are welcome to add new "cases", e.g., for some of the exercises below.

```
function y = toeplib(cas,N,par)
% cas is a string parameter;
% par is an optional real or complex scalar;
% the default value is 1.
% All series are truncated to N terms.
if  nargin == 2, par = 1; end
if  cas == ’bin’,
      y=[1  cumprod(par:-1:par-N+2)./cumprod(1:N-1)];
%     y = 1st row of binomial series (1+x)^par, par in R;
elseif   cas == ’con’,
  y = cumprod([1  par*ones(1,N-1)]);
%  The array multiplication y.*f1 returns the first
%  row of f(par*S_N);
%  sum(y.*f1) evaluates f(par). See also Problem~(b).
elseif cas == ’exp’,
   y = [1  cumprod(par./[1:(N-1)]];
%  y = 1st row of exponential \exp(par*x).
%  Since par can be complex, circular
%  (or trigonometric) functions can also be expanded.
elseif cas == ’log’,
    y=[0  1./[1:(N-1)]].*cumprod ([-1 -par*ones(1:N-1)]);
%   y= 1st row of logarithm \ln(1+par*x).
elseif cas == ’e1t’,
    y=[1  zeros(1,N-1)];
%   y=e_1^T , i.e. 1st row of (eye)N.
elseif cas == ’SN1’, y = [0  1  zeros(1,N-2)];
%   y=1st row of S_N .
elseif cas == ’dif’,  y = [0  1:(N-1)];
% The array multiplication y.*f1 returns xf’(x).
else  cas == ’int’,  y =1./[1:N].
% The array multiplication y.*f1 returns
% {1\over x}\int_0^x f(t) dt.
end
```

(b) *Evaluation of $f(x)$.* Given `N` and `f1` of your own choice, set `fterms = toeplib(’con’,N,x).*f1`. What is `sum(fterms)` and `cumsum(fterms)`? When can `sum(fliplr(fterms))` be useful?

(c) Write a code that, for arbitrary given N, returns the 1st rows of the Toeplitz matrices for $\cos x$ and $\sin x$, with S_N corresponding to x, and then transforms them to to 1st rows for Toeplitz matrices with S_N corresponding to x^2. Apply this, for (say) $N = 36$, to determine the errors of the coefficients of $4(\cos x)^3 - 3\cos x - \cos 3x$.

(d) Find out how a library "`toeplib2`" designed for Toeplitz matrices for *even* functions, where S_N corresponds to x^2, must be different from `toeplib`. For example how are `cas == ’dif’` and `cas == ’int’` to be changed?

(e) Unfortunately, a `toeplib` "case" has at most one parameter, namely `par`. Write a code that calls toeplib twice for finding the Maclaurin coefficients of the three parameter function

$$y = (a + bx)^{\alpha}, \quad a > 0,$$

b, α real. Compute the coefficients in two different ways for $N = 24$; $a = 2$; $b = -1$; $\alpha = \pm 3$, and compare the results for estimating the accuracy of the coefficients.

(f) Compute the Maclaurin expansions for $(1-x^2)^{-1/2}$ and $\arcsin(x)$, and for $y = 2a rcsinh\,(x/2)$. Expand also dy/dx and y^2. Convert the coefficients to rational numbers, as long as they seem to be reliable. Save the results, or make it easy to reproduce them, for comparisons with the results of Problem 12(b).

Comment: The last three series are fundamental for the expansions of differential operators in powers of central difference operators, which lead to highly accurate formulas for numerical differentiation.

(g) Two power series that generate the Bernoulli numbers are given in Example 3.1.5, namely

$$x \equiv \Bigl(\sum_{i=1}^{\infty}\frac{x^i}{i!}\Bigr)\Bigl(\sum_{j=0}^{\infty}\frac{B_j x^j}{j!}\Bigr); \qquad \frac{x}{2}\,\frac{e^{x/2}+e^{-x/2}}{e^{x/2}-e^{-x/2}} = \sum_{j=0}^{\infty}\frac{B_{2j}x^{2j}}{(2j)!}.$$

Compute B_{2j} for (say) $j \le 30$ in floating point, using each of these formulas, and compute the difference of the results, which are influenced by rounding errors. Try to find, whether one of the sequences is more accurate than the other, by means of the formula in (3.1.18) for (say) $j > 4$. Then convert the results to rational numbers. Use several tolerances in the function rat and compare with [1, Table 23.2]. Some of the results are likely to disagree. Why?

(h) The Kummer confluent hypergeometric function $M(a, b, x)$ is defined by the power series (3.1.14). Kummer's first identity, i.e.

$$M(a, b, -x) = e^{-x}M(b-a, b, x),$$

is important, e.g., because the series on the left hand side is ill-conditioned if $x \gg 1$, $a > 0$, $b > 0$, while the expression on the right hand side is well-conditioned. Check the identity experimentally by computing the difference between the series on the left hand side and on the right for a few values of a, b, The computed coefficients are afflicted by rounding errors. Are the differences small enough to convince you of the validity of the formula?

11. Read about expansions of composite functions in Sec. 3.1.4
(a) Write the Horner recurrence for the evaluation of the matrix polynomial $f(g(S_N))$ according to (3.1.25). Then show that the following MATLAB function evaluates the first row of $h(S_N) = f(g(S_N))$, if $g(0) = 0$.

```
function h1 = comp(f1,g1,N);
%
% INPUT: the integer N and the rows f1, g1, with the
% first N Maclaurin coefficients for the analytic functions
% f(z), g(z).
% OUTPUT: The row h1 with the first N Maclaurin coefficients
% for the composite function h(z)=f(g(z)), where g1(1)=g(0)=0.
% computed according to the algorithm for a composite function.
% Error message if g(0)\ne 0.
```

```
if g1(1) ~= 0,
   error('g(0) ~= 0 in a composite function f(g(z))')
end
e1t = zeros(1,N);  e1t(1)=1;
r = f1(N)*e1t;
gs = toep(g1,N);
for j = N-1:-1:1,
     r = r*gs + f1(j)*e1t;
end
h1 = r;
```

(b) Matrix functions in Matlab : For $h(z) = e^{g(z)}$ it is convenient to use the matrix function `expm(g(SN))` or, on the vector level, `h1 = e1t*expm(g(SN))`, rather than to use `h1 = comp(f1,g1)`. If $f(0) \neq 0$, you can analogously use the functions `logm` and `sqrtm`. They may be slower and less accurate than `h1 = comp(f1,g1)`, but they are typically fast and accurate enough.
Compare computing time and accuracy in the use of `expm(k * logm(eye(N) + SN)` and `toeplib('bin',N,k)` for a few values of N and k.

Comment: *The* Matlab *function* `funm` *should, however, not be used*, because it uses an eigenvalue method that does not work well for matrices that have multiple eigenvalues. For triangular Toeplitz matrices the diagonal elements are multiple eigenvalues. The functions `expm, logm` and `sqrtm` should not be confused with the functions `exp, log` and `sqrt`, which operate on the matrix elements.

(c) Expand $e^{\sin(z)}$ in powers of z in two ways: *first* using the function in Problem 11(a); *second* using the matrix functions of Matlab. Show that the latter can be written

`HN = expm(imag(expm(i*SN))).`

Do not be surprised if you find a dirty imaginary part of H_N. Kill it!
Compare the results of the two procedures. If you have done the runs appropriately, the results should agree excellently.

(d) Treat the series $h(z) = \sqrt{(1+e^z)}$ in three different ways, and compare the results, with respect to validity, accuracy and computing time.

(i) Set $ha(z) = h(z)$, and determine $f(z)$, $g(z)$, analytic at $z = 0$, so that $g(0) = 0$. Compute `ha1 = comp(f1,g1,N)`. Do you trust the result?

(ii) Set $h(z) = H(z)$. Compute `HN = sqrtm(eye(N) + expm(SN))`.
In the first test, i.e. for $N = 6$, display the matrix H_N, and check that H_N is an upper triangular Toeplitz matrix. For larger values of N, display the first row only, and compare it to $ha1$. If you have done all this correctly, the agreement should be extremely good, and we can practically conclude that both are very accurate.

(iii) Try the "natural", although "illegal', decomposition $hb(z) = f(g(z))$, with $f(x) = (1+x)^{0.5}$, $g(z) = e^z$. Remove temporarily the error stop. Demonstrate by numerical experiment that $hb1$ is very wrong. If this is a surprise, read Sec. 3.1.4 once more.

(e) Suppose that you perform matrix level operations on $f(S_M)$, $g(S_M), \ldots,$

where $M < N$. Show that the results are exactly equal to the principal $M \times M$ submatrices of the results obtained with the matrices $f(S_N)$, $g(S_N), \ldots$

12. *Reversion of series.* Let

$$y = f(x) = \sum_{j=1}^{\infty} f1(j)x^{j-1},$$

where $f1(1) = f(0) = 0$, $f1(2) = f'(0) = 1$ (with the notation used in Sec. 3.1.5 and in the previous problems). Power series reversion is to find the power series for the inverse function

$$x = g(y) = \sum_{j=1}^{\infty} g1(j)y^{j-1},$$

where $g1(1) = g(0) = 0$. Read also the last paragraphs of Sec. 3.1.5.

We work with truncated series with N terms in the Toeplitz matrix representation. The inverse function relationship gives the matrix equation $f(g(S_N)) = S_N$. Because $g(0) = 0$, we have, by (3.1.25),

$$f(g(S_N)) = \sum_{j=1}^{N} f1(j)(g(S_N)^{j-1}.$$

Now Horner's scheme can be used for computing the polynomial *and its derivative*, the latter is obtained by algorithmic differentiation; see Sec. 1.3.1.

The first row of this matrix equation is treated by Newton's method in the code breku listed below. The Horner algorithms are adapted to the first row.[16] The notations in the code is almost the same as in the theoretical description, although lower case letters are used, e.g., the matrix $g(S_N)$ is denoted gs, and $fgs1$ is the first row of the matrix $f(g(S_N))$. The equation reads $fgs1 - s1 = 0$.

(a) Convince yourself that the following MATLAB function implements power series reversion under a certain condition. Describe in a few words the main features of the method.

```
function g1 = breku(f1,N);
%
% INPUT: The row vector f1 that represents a (truncated)
% Maclaurin series
% OUTPUT: The row g1, i.e. the first N terms of the series
% x = g(y) where y = f(x).
% Note that f1(1) = 0, f1(2) = 1; if not, there will
% be an error message. The integer N, i.e. the length
% of the truncated series wanted in the output, is optional
% input; by default N = length(f1).
```

[16] The name"breku" comes from Brent and Kung, who were probably the first mathematicians to apply Newton's method to series reversion, although with a different formulation of the equation than ours (no Toeplitz matrices).

```
% If length(f1) < N, f1 is extended to length N by zeros.
% Uses the function toep(r,N) (see Problem~9) for expanding
% a row to an upper triangular Toeplitz matrix.
%
if ~ (f1(1) ~= 0|f1(2) ~= 1),
    error('wrong f1(1) or f1(2)');
end
lf1 =  length(f1);
if (nargin == 1|lf1 > N), N = lf1; end
if lf1 < N, f1 = [f1 zeros(1, N-lf1)]  end
maxiter = floor(log(N)/log(2));
e1t = [1, zeros(1,N-1)];
s1 = [0 1 zeros(1,N-2)];  gs1 = s1;
for iter = 0:maxiter
    gs  = toep(gs1,N);
% Horner's scheme for computing the first rows
% of f(gs) and f'(g(s)):
    fgs1 = f1(N)*e1t;  der1 = zeros(1,N);
    for j  =  N-1:-1:1
        ofgs1 = fgs1;   %ofgs1 means "old" fgs1
        fgs1 = ofgs1*gs + f1(j)*e1t ;
        der1 = ofgs1 + der1*gs ;
    end
    % A Newton iteration for the equation fgs1 - s1 = 0:
    gs1 = gs1 - (fgs1 - s1)/toep(der1,N);
end
g1 = gs1;
```

Comment: The radius of convergence depends on the singularities of $g(y)$, which are typically related to the singularities of $f(x)$ and to the zeros of $f'(x)$, (why?). There are other cases, e.g., if $f'(x) \to 0$ as $x \to \infty$ then $\lim f(x)$ may be a singularity of $g(y)$.

(b) Apply the code breku to the computation of $g(y)$ for $f(x) = \sin x$ and for $f(x) = 2\sinh(x/2)$. Compare with the results of Problem 10 (d). Then reverse the two computed series $g(y)$, and study how you return to the original expansion of $f(x)$, more or less accurately. Use tic toc to take the time, for a few values of N.

(c) Apply the code breku to compute $g(y)$ for $f(x) = \ln(1+x)$, $f(x) = e^x - 1$, $f(x) = x + x^2$, $f(x) = x + x^2 + x^3$.
If you know an analytic expression for $g(y)$, find the Maclaurin expansion for this, and compare with the expansions obtained from breku .
Apply breku to the computed expansions of $g(y)$, and study how accurately you return to the expansion of $f(x)$.

(d)$f(x) = xe^x$; the inverse function $g(y)$ is known as the Lambert W function.[17] Determine $g(y)$. Then reverse the power series for $g(y)$, and compare with the expansion of $f(x)$.

[17] Johann Heinrich Lambert (1728–1777), German mathematician, physicist and astronomer, and colleague of Euler and Lagrange at the Berlin Academy of Sciences. He is best known for

(e) Estimate for $f(x) = xe^x$ the radius of convergence approximately, by means of the ratios of the coefficients computed in (d), and exactly; see the comment after the code above.

(f)* Set $y = f(x)$. Suppose that $y(0) \neq 0$, $y'(0) \neq 0$. Show that the code breku can be used for expanding the inverse function in powers of $(y - y(0))/y'(0)$. Construct some good test example.

(g)* For the equation $\sin x - (1 - y)x = 0$, express $x^2 = g(y)$ (why x^2?), with $N = 12$. Then express x in the form $x \approx \pm y^{1/2}P(y)$, where $P(y)$ is a truncated power series with (say) 11 terms.

(h)* Make simple temporary changes in the code breku, so that you can follow the iterations on the screen.

3.2 More About Series

3.2.1 Laurent and Fourier Series

A **Laurent series** is a series of the form

$$\sum_{n=-\infty}^{\infty} c_n z^n. \tag{3.2.1}$$

Its convergence region is the intersection of the convergence regions of the expansions

$$\sum_{n=0}^{\infty} c_n z^n \quad \text{and} \quad \sum_{m=1}^{\infty} c_{-m} z^{-m},$$

the interior of which are determined by conditions of the form $|z| < r_2$ and $|z| > r_1$. The convergence region can be void, e.g., if $r_2 < r_1$.

If $0 < r_1 < r_2 < \infty$ the convergence region is an *annulus*, $r_1 < |z| < r_2$. The series defines an analytic function in the annulus. Conversely, if $f(z)$ is a **single-valued analytic function** in this annulus, it is there represented by a Laurent series, that converges uniformly in every closed subdomain of the annulus.

The coefficients are determined by the following formula, due to Cauchy[18]

$$c_n = \frac{1}{2\pi i} \int_{|z|=r} z^{-n-1} f(z)dz, \quad r_1 < r < r_2, -\infty < n < \infty, \tag{3.2.2}$$

and

$$|c_n| \leq r^{-n} \max_{|z|=r} |f(z)|. \tag{3.2.3}$$

The extension to the case when $r_2 = \infty$ is obvious; the extension to $r_1 = 0$ depends on whether there are any terms with negative exponents or not. In the extension of *formal* power series to *formal Laurent series*, however, only a finite number of

his illumination laws and for the continued fraction expansions of elementary functions, Sec. 3.5.3. His W function was "rediscovered" a few years ago, [14].

[18] Augustin Cauchy (1789–1857) is the father of modern analysis. He is the creator of complex analysis, a centerpiece of which this formula plays a fundamental role.

terms with negative indices are allowed to be different from zero; see Henrici loc. cit. Sec. 1.8. If you substitute z for z^{-1} an infinite number of negative indices is allowed, if the number of positive indices is finite.

Example 3.2.1.

A function may have several Laurent expansions (with different regions of convergence), e.g.,

$$(z-a)^{-1} = \begin{cases} -\sum_{n=0}^{\infty} a^{-n-1} z^n & \text{if } |z| < |a| \\ \sum_{m=1}^{\infty} a^{m-1} z^{-m} & \text{if } |z| > |a|. \end{cases}$$

The function $1/(z-1) + 1/(z-2)$ has three Laurent expansions, with validity conditions $|z| < 1$, $1 < |z| < 2$, $2 < |z|$, respectively. The series contains both positive and negative powers of z in the middle case only. The details are left for Problem 4 (a).

Lemma 3.2.2 can, with some modifications, be generalized to Laurent series (and to complex Fourier series), e.g.,(3.2.17) becomes

$$\tilde{c}_n - c_n = \dots c_{n-2N} r^{-2N} + c_{n-N} r^{-N} + c_{n+N} r^N + c_{n+2N} r^{2N} \dots \tag{3.2.4}$$

Remark 3.2.1. The restriction to *single-valued* analytic functions is important in this subsection. In this book we cannot entirely avoid to work with **multi-valued** functions such as $\sqrt{z}$, $\ln z, z^\alpha$, (α non-integer). We always work with such a function, however, in some region where one branch of it, determined by some convention, is single-valued. In the examples mentioned, the natural conventions are to require the function to be positive when $z > 1$, and to forbid z to cross the negative real axis. In other words, the complex plane has a **cut** along the negative real axis. The annulus mentioned above is in these cases incomplete; its intersection with the negative real axis is missing, and we cannot use a Laurent expansion.

For a function like $\ln\left(\frac{z+1}{z-1}\right)$, we can, depending on the context, cut out either the interval $[-1, 1]$ or the complement of this interval with respect to the real axis. We then use an expansion into negative or into positive powers of z, respectively.

If $r_1 < 1 < r_2$, we set $F(t) = f(e^{it})$. Note that $F(t)$ is a periodic function; $F(t+2\pi) = F(t)$. By (3.2.1) and (3.2.2), the Laurent series then becomes for $z = e^{it}$ a **Fourier series**:

$$F(t) = \sum_{n=-\infty}^{\infty} c_n e^{int}, \quad c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-int} F(t)\, dt. \tag{3.2.5}$$

Note that $c_{-m} = O(r_1^m)$ for $m \to +\infty$, and $c_n = O(r_2^{-n})$ for $n \to +\infty$. *The formulas in* (3.2.5), *however, are valid in much more general situations*, where $c_n \to 0$ much more slowly, and where $F(t)$ cannot be continued to an analytic function $f(z)$, $z = re^{it}$, in an annulus. (In such a case $r_1 = 1 = r_2$, typically.)

A Fourier series is often written in the following form,

$$F(t) = \tfrac{1}{2}a_0 + \sum_{k=1}^{\infty}(a_k \cos kt + b_k \sin kt). \tag{3.2.6}$$

Consider $c_k e^{ikt} + c_{-k} e^{-ikt} \equiv a_k \cos kt + b_k \sin kt$. Since $e^{\pm ikt} = \cos kt \pm i \sin kt$, we obtain for $k \geq 0$:

$$a_k = c_k + c_{-k} = \frac{1}{\pi}\int_{-\pi}^{\pi} F(t)\cos kt\, dt; \quad b_k = i(c_k - c_{-k}) = \frac{1}{\pi}\int_{-\pi}^{\pi} F(t)\sin kt\, dt. \tag{3.2.7}$$

Also note that $a_k - ib_k = 2c_k$. If $F(t)$ is real for $t \in \mathbf{R}$ then $c_{-k} = \bar{c}_k$.

We mention without proof the important **Riemann–Lebesgue theorem**,[19] by which the Fourier coefficients c_n *tend to zero as* $n \to \infty$ *for any function that is integrable* (in the sense of Lebesgue), a fortiori for any periodic function that is continuous everywhere. A finite number of finite jumps in each period are also allowed.

A function $F(t)$ is said to be of bounded variation in an interval if, in this interval, it can expressed in the form $F(t) = F_1(t) - F_2(t)$, where F_1 and F_2 are non-decreasing bounded functions. A finite number of jump discontinuities are allowed. The variation of F over the interval $[a, b]$ is denoted $\int_a^b |dF(t)|$. If F is differentiable the variation of F equals $\int_a^b |F'(t)|\, dt$.

Another classical result in the theory of Fourier series reads: *If $F(t)$ is of* **bounded variation**. *in the closed interval $[-\pi, \pi]$ then $c_n = O(n^{-1})$*; see Titchmarsh [45, § 13.21, §13.73]. This can be generalized. *Suppose that $F^{(p)}$ is of bounded variation on $[-\pi, \pi]$, and that $F^{(j)}$ is continuous everywhere for $j < p$. Denote the Fourier coefficients of $F^{(p)}(t)$ by $c_n^{(p)}$. Then*

$$c_n = (in)^{-p} c_n^{(p)} = O(n^{-p-1}). \tag{3.2.8}$$

This follows from the above classical result, after the integration of the formula for c_n in (3.2.2) by parts p times. Bounds for the truncation error of a Fourier series can also be obtained from this. The details are left for Problem 4 (d), together with a further generalization. A similar result is that $c_n = o(n^{-p})$ if $F^{(p)}$ is integrable, hence a fortiori if $F \in C^p$.

In particular, we find for $p = 1$ (since $\sum n^{-2}$ is convergent) that the Fourier series (3.2.2) *converges absolutely and uniformly* in $\mathbf{R}$. It can also be shown that *the Fourier series is valid*, i.e. the sum is equal to $F(t)$.

3.2.2 The Cauchy–FFT Method

An alternative method for deriving coefficients of power series, when many terms are needed is based on the following classic result. Suppose that the value $f(z)$

[19] Jean Baptist Joseph Fourier (1768–1830), French mathematician and physicist. In a pioneering publication about the flow of heat (1822), he utilized series of the type of equation (3.2.6). B. Riemann (1826-1866), German mathematician, made fundamental contributions to Analysis and Geometry. H. Lebesgue (1875-1941), French mathematician, created path-breaking general concepts of measure and integral.

of an analytic function can be computed at any point inside and on the circle $C_r = \{z : |z-a| = r\}$, and set

$$M(r) = \max|f(z)|, \quad z \in C_r, \quad z = a + re^{i\theta}, \quad z' = a + r'e^{i\theta},\ (r' < r).$$

Then the coefficients of the Taylor expansion around a are determined by Cauchy's formula,

$$a_n = \frac{1}{2\pi i}\int_{C_r} \frac{f(z)}{(z-a)^{(n+1)}}\,dz = \frac{r^{-n}}{2\pi}\int_0^{2\pi} f(a+re^{i\theta})e^{-ni\theta}\,d\theta. \tag{3.2.9}$$

For a derivation, multiply the Taylor expansion (3.1.3) by $(z-a)^{-n-1}$, integrate term by term over C_r, and note that

$$\frac{1}{2\pi i}\int_{C_r}(z-a)^{j-n-1}\,dz = \frac{1}{2\pi}\int_0^{2\pi} r^{j-n}e^{(j-n)i\theta}\,d\theta = \begin{cases} 1, & \text{if } j = n; \\ 0, & \text{if } j \neq n. \end{cases} \tag{3.2.10}$$

The following inequalities are useful consequences of the definitions and of (3.2.9).

$$|a_n| \le r^{-n}M(r), \tag{3.2.11}$$

$$|R_n(z')| \le \sum_{j=n}^{\infty}|a_j(z'-a)^j| \le \sum_{j=n}^{\infty} r^{-j}M(r)(r')^j = \frac{M(r)(r'/r)^n}{1-r'/r}, \quad 0 \le r' < r.$$

This form of the remainder term of a Taylor series is useful in theoretical studies, and also for practical purpose, if the maximum modulus $M(r)$ is easier to estimate than the nth derivative.

Set $z = a + re^{i\theta}$, $\Delta\theta = 2\pi/N$, and apply the trapezoidal rule to the second integral in (3.2.9). Then[20]

$$a_n \approx \tilde{a}_n \equiv \frac{1}{Nr^n}\sum_{k=0}^{N-1} f(a+re^{ik\Delta\theta})e^{-ink\Delta\theta}, \quad n = 0 : N-1. \tag{3.2.12}$$

The approximate Taylor coefficients $\tilde{a}_n$, or rather the numbers $a_n^\star = \tilde{a}_n Nr^n$, are here expressed as a case of the (direct) **Discrete Fourier Transform**. More generally, this transform maps an *arbitrary* sequence $\{\alpha_k\}_0^{N-1}$ to a sequence $\{a_n^\star\}_0^{N-1}$, by the following equations:

$$a_n^\star = \sum_{k=0}^{N-1}\alpha_k e^{-ink\Delta\theta}, \quad n = 0 : N-1. \tag{3.2.13}$$

The transform will be studied more systematically in Sec. 4.6.

If N is a power of 2, it is shown in Sec. 4.6 that, given the N values α_k, $k = 0 : N-1$, and $e^{-i\Delta\theta}$, *no more than* $N\log_2 N$ *complex multiplications and additions are*

[20] See (1.2.6). Note that the integrand has the same value for $\theta = 2\pi$ as for $\theta = 0$. The terms $\frac{1}{2}f_0$ and $\frac{1}{2}f_N$ that appear in the general trapezoidal rule can therefore in this case be replaced by f_0.

needed for the computation of all the N coefficients $a_n^\star$, if an implementation of the discrete Fourier transform known as the **Fast Fourier Transform (FFT)** is used. This makes our theoretical considerations very practical. (Packages for interactive mathematical computation usually contain commands related to FFT.)

It is also shown in Sec. 4.6 that the **inverse** of the discrete Fourier transform (3.2.13) is given by the formulas,

$$\alpha_k = (1/N)\sum_{n=0}^{N-1} a_n^\star e^{ink\Delta\theta}, \quad k = 0 : N-1. \tag{3.2.14}$$

It looks almost like the direct Discrete Fourier Transform (3.2.13), except for the sign of i and the factor $1/N$. It can therefore also be performed by means of $O(N\log N)$ elementary operations, instead of the $O(N^3)$ operations that the most obvious approach to this task would require, (i.e. by solving the linear system (3.2.13)).

In our context, i.e. the computation of Taylor coefficients, we have, by (3.2.12) and the line after that equation,

$$\alpha_k = f(a + re^{ik\Delta\theta}), \qquad a_n^\star = \tilde{a}_n N r^n. \tag{3.2.15}$$

Set $z_k = a + re^{ik\Delta\theta}$. Using (3.2.15), the inverse transformation then becomes,[21]

$$f(z_k) = \sum_{n=0}^{N-1} \tilde{a}_n (z_k - a)^n, \quad k = 0 : N-1. \tag{3.2.16}$$

Since the Taylor coefficients are equal to $f^{(n)}(a)/n!$, this is de facto a method for the accurate *numerical differentiation of an analytic function.* If r and N are chosen appropriately, it is more well-conditioned than most methods for *numerical differentiation*, such as the difference approximations mentioned in Chapter 1; see also Sec. 3.3 and Chapter 4. It requires, however, complex arithmetic for a convenient implementation. We call this the **Cauchy–FFT method** for Taylor coefficients and differentiation.

The question arises, how to choose N and r. Theoretically, any r less than the radius of convergence ρ would do, but there may be trouble with cancellation if r is small. On the other hand, the truncation error of the numerical integration usually increases with r. "*Scylla and Charybdis situations*" [22] like this are very common with numerical methods.

It is typically the rounding error that sets the limit for the accuracy; it is usually not expensive to choose r and N, so that the truncation error becomes much smaller. A rule of thumb for this situation is to guess a value of $\hat{n}$, i.e. how

[21] One interpretation of these equations is that the polynomial $\sum_{n=0}^{N-1} \tilde{a}_n (z-a)^n$ is the solution of a special, although important, interpolation problem for the function f, analytic inside a circle in **C**.

[22] According to American Heritage Dictionary Scylla is a rock on the Italian side of the Strait of Messina, opposite to the whirlpool Charybdis, personified by Homer (Ulysses) as a female sea monster who devoured sailors. The problem is to navigate safely between them.

many terms will be needed in the expansion, and then to try two values for N (powers of 2) larger than $\hat{n}$. *If ρ is finite* try $r = 0.9\rho$ and $r = 0.8\rho$, and compare the results. They may or may not indicate that some other values of N and r (and perhaps also $\hat{n}$) should also be tried. On the other hand, *if $\rho = \infty$*, try, e.g., $r = 1$ and $r = 3$, and compare the results. Again the results indicate whether or not more experiments should be made.

One can also combine numerical experimentation with a theoretical analysis of a more or less simplified model, including a few elementary optimization calculations. The authors take the opportunity to exemplify below this type of "hard analysis" on this question.

We first derive two lemmas, which are important also in many other contexts. First we have a discrete analogue of equation (3.2.10).

Lemma 3.2.1. *Let p, N be integers. Then $\sum_{k=0}^{N-1} e^{2\pi ipk/N} = 0$, unless $p = 0$ or p is a multiple of N. In these exceptional case every term equals 1, and the sum equals N.*

Proof. If p is neither 0 nor a multiple of N, the sum is a geometric series, the sum of which is equal to $(e^{2\pi ip} - 1)/(e^{2\pi ip/N} - 1) = 0$. The rest of the statement is obvious. □

Lemma 3.2.2. *Suppose that $f(z) = \sum_0^\infty a_n(z-a)^n$ is analytic in the disc $|z-a| < \rho$. Let $\tilde{a}_n$ be defined by (3.2.12), where $r < \rho$. Then*

$$\tilde{a}_n - a_n = a_{n+N}\, r^N + a_{n+2N}\, r^{2N} + a_{n+3N}\, r^{3N} + \dots, \qquad 0 \le n < N. \tag{3.2.17}$$

Proof. Since $\Delta\theta = 2\pi/N$,

$$\tilde{a}_n = \frac{1}{Nr^n}\sum_{k=0}^{N-1} e^{-2\pi ink/N} \sum_{m=0}^{\infty} a_m \left(re^{2\pi ik/N}\right)^m = \frac{1}{Nr^n}\sum_{m=0}^{\infty} a_m r^m \sum_{k=0}^{N-1} e^{2\pi i(-n+m)k/N}.$$

By the previous lemma, the inner sum of the last expression is zero, unless $m - n$ is a multiple of N. Hence (recall that $0 \le n < N$),

$$\tilde{a}_n = \frac{1}{Nr^n}\left(a_n r^n N + a_{n+N}\, r^{n+N} N + a_{n+2N}\, r^{n+2N} N + \dots\right),$$

from which equation (3.2.17) follows. □

Let $M(r)$ be the maximum modulus for the function $f(z)$ on the circle C_r, and denote by $M(r)U$ an upper bound for the error of a computed function value $f(z)$, $|z| = r$, where $U \ll 1$. Assume that rounding errors during the computation of $\tilde{a}_n$ are of minor importance.

Then, by (3.2.12), $M(r)U/r^n$ is a bound for the *rounding error* of $\tilde{a}_n$. (The rounding errors during the computation can be included by a redefinition of U.)

Next we shall consider the *truncation error* of (3.2.12). First we *estimate* the coefficients that occur in (3.2.17) by means of $\max|f(z)|$ on a circle with radius r'; $r' > r$, where r is the radius of the circle used in the *computation* of the first N coefficients. So, in (3.2.9) we substitute r', j for r, n, respectively, and obtain the inequality

$$|a_j| \le M(r')(r')^{-j}, \qquad 0 < r < r' < \rho.$$

The actual choice of r', strongly depends on the function f.[23] Put this inequality into (3.2.17), where we shall choose $r < r' < \rho$. Then

$$\begin{aligned}|\tilde{a}_n - a_n| &\le M(r')\big((r')^{-n-N}r^N + (r')^{-n-2N}r^{2N} + (r')^{-n-3N}r^{3N} + \ldots\big)\\ &= M(r')(r')^{-n}\big((r/r')^N + (r/r')^{2N} + (r/r')^{3N} + \ldots\big) = \frac{M(r')(r')^{-n}}{(r'/r)^N - 1}.\end{aligned}$$

We make a digression here, because *this is an amazingly good result.* The trapezoidal rule that was used in the calculation of the Taylor coefficients is typically expected to have an error that is $O\big((\Delta\theta)^2\big) = O\big(N^{-2}\big)$. (As before, $\Delta\theta = 2\pi/N$.) This application is, however, *a very special situation: a periodic analytic function is integrated over a full period.* We shall return to results like this several times. In this case, for fixed values of r, r', the truncation error is $O\big((r/r')^N\big) = O\big(e^{-\eta/\Delta\theta}\big)$, where $\eta > 0$, $\Delta\theta \to 0+$. This tends to zero faster than any power of $\Delta\theta$.

It follows that a bound for the total error of $\tilde{a}_n$, i.e. the sum of the bounds for the rounding and the truncation errors, is given by

$$UM(r)r^{-n} + \frac{M(r')(r')^{-n}}{(r'/r)^N - 1}, \qquad r < r' < \rho. \tag{3.2.18}$$

Example 3.2.2. "Scylla and Charybdis" in the Cauchy–FFT.

We shall discuss how to choose the parameters r and N, so that the *absolute error bound* of a_n, given in (3.2.18) becomes uniformly small for (say) $n = 0 : \hat{n}$. $1 + \hat{n} \gg 1$ is thus the number of Taylor coefficients requested. The parameter r' does not belong to the Cauchy–FFT method, but it has to be chosen well in order to make the *bound* for the truncation error realistic.

The discussion is rather technical, and you may omit it at a first reading. It may, however, be useful to study this example later, because similar technical subproblems occur in many serious discussions of numerical methods that contain parameters that should be appropriately chosen.

First consider the *rounding error.* By the maximum modulus theorem, $M(r)$ is an increasing function, hence, for $r > 1$, $\max_n M(r)r^{-n} = M(r) > M(1)$. On the other hand, for $r \le 1$, $\max_n M(r)r^{-n} = M(r)r^{-\hat{n}}$.[24] Let r^* be the value of r, for which this maximum is minimal. Note that $r^* = 1$ unless $M'(r)/M(r) = \hat{n}/r$ for some $r \le 1$.

Then try to determine N and $r' \in [r^*, \rho)$ so that, for $r = r^*$, the bound for the the second term of (3.2.18) becomes much smaller than the first term, i.e. *the*

[23] In rare cases we may choose $r' = \rho$.

[24] $\hat{n}$ was introduced in the beginning of this example.

truncation error is made negligible compared to the rounding error. This works well if $\rho \gg r^*$. In such cases, we may therefore choose $r = r^*$, and the total error is then just a little larger than $UM(r^*)(r^*)^{-\hat{n}}$.

For example, if $f(z) = e^z$ then $M(r) = e^r$, $\rho = \infty$. In this case $r^* = 1$ (since $\hat{n} \gg 1$). Then we shall choose N and $r' = N$, so that $e^{r'}/((r')^N - 1) \ll eU$. One can show that it is sufficient to choose $N \gg |\ln U/\ln|\ln U||$. For instance, if $U = 10^{-15}$, this is satisfied with a wide margin by $N = 32$. On a computer, the choice $r = 1$, $N = 32$, gave (with 53 bits floating point arithmetic) an error less than $2\cdot 10^{-16}$. The results were much worse for $r = 10$, and for $r = 0.1$; the maximum error of the first 32 coefficients became $4\cdot 10^{-4}$ and $9\cdot 10^{13}$(!), respectively. In the latter case the errors of the first 8 coefficients did not exceed 10^{-10}, but the rounding error of a_n, due to cancellations, increase rapidly with n.

If ρ is not much larger than r^*, the procedure described above may lead to a value of N that is much larger than $\hat{n}$. In order to avoid this, we now set $\hat{n} = \alpha N$. We now confine the discussion to the case that $r < r' < \rho \le 1$, $n = 0 : \hat{n}$. Then, with all other parameters fixed, the bound in (3.2.18) is maximal for $n = \hat{n}$. We simplify this bound; $M(r)$ is replaced by the larger quantity $M(r')$, and the denominator is replaced by $(r'/r)^N$.

Then, for given r', α, N, we set $x = (r/r')^N$ and determine x so that

$$M(r')(r')^{-\alpha N}(Ux^{-\alpha} + x)$$

is minimized. The minimum is obtained for $x = (\alpha U)^{1/(1+\alpha)}$, i.e. for $r = r'x^{1/N}$, and the minimum is equal to[25]

$$M(r')(r')^{-n}U^{1/(1+\alpha)}c(\alpha), \quad \text{where} \quad c(\alpha) = (1+\alpha)\alpha^{-\alpha/(1+\alpha)}.$$

We see that the error bound contains the factor $U^{1/(1+\alpha)}$. This is, e.g., proportional to $2U^{1/2}$ for $\alpha = 1$, and to $1.65U^{4/5}$ for $\alpha = \frac{1}{4}$. The latter case is thus much more accurate, but, for the same $\hat{n}$, one has to choose N four times as large, which leads to more than four times as many arithmetic operations. In practice, $\hat{n}$ is usually given, and the order of magnitude of U can be estimated. Then α is to be chosen to make a compromise between the requirements for a good accuracy and for a small volume of computation. *If ρ is not much larger than r^**, we may choose

$$N = \hat{n}/\alpha, \quad x = (\alpha U)^{1/(1+\alpha)}, \quad r = r'x^{1/N}.$$

Experiments were made with $f(z) = \ln(1-z)$. Then $\rho = 1$, $M(1) = \infty$. Take $\hat{n} = 64$, $U = 10^{-15}$, $r' = 0.999$. Then $M(r') = 6.9$. For $\alpha = 1, 1/2, 1/4$, we have $N = 64, 128, 256$, respectively. The above theory suggests $r = 0.764, 0.832, 0.894$, respectively. The theoretical estimates of the absolute errors become, $10^{-9}, 2.4\,10^{-12}, 2.7\,10^{-14}$, respectively. The smallest errors obtained in experiments with these three values of α are, $6\,10^{-10}, 1.8\,10^{-12}, 1.8\,10^{-14}$, which were obtained for $r = 0.766, 0.838, 0.898$, respectively. So, the theoretical predictions of these experimental results are very satisfactory.

[25] This is a rigorous upper bound of the error for this value of r, in spite of the simplifications in the formulation of the minimization.

3.2.3 Chebyshev Polynomials

The **Chebyshev**[26] **polynomials** of the first kind are

$$T_n(z) = \cos n\phi, \quad z = \cos\phi, \tag{3.2.19}$$

Note that $T_0(z) = 1$, $T_1(z) = z$. That $T_n(z)$ is an nth degree polynomial follows, by induction, from the important **recurrence relation**,

$$T_{n+1}(z) = 2zT_n(z) - T_{n-1}(z), \quad (n \geq 1), \tag{3.2.20}$$

which follows from the well known trigonometric formula

$$\cos(n+1)\phi + \cos(n-1)\phi = 2\cos\phi\cos n\phi.$$

We obtain,

$$\begin{aligned} &T_2(z) = 2z^2 - 1; \quad T_3(z) = 4z^3 - 3z; \quad T_4(z) = 8z^4 - 8z^2 + 1, \\ &T_5(z) = 16z^5 - 20z^3 + 5z; \quad T_7(z) = 64z^7. - 112z^5 + 56z^3 - 7z \end{aligned}$$

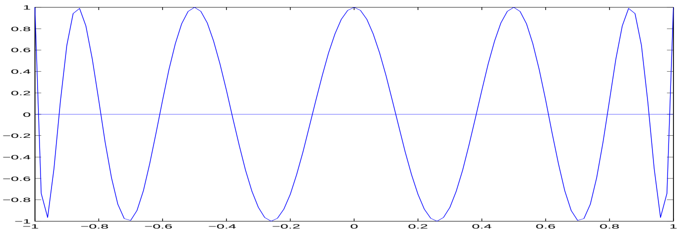

Figure 3.2.1. *The Chebyshev polynomial* $T_{12}(x), x \in [-1, 1]$.

The **Chebyshev polynomials** of the second kind are

$$U_{n-1}(z) = \frac{\sin n\phi}{\sin\phi}, \quad \text{where} \quad z = \cos\phi, \tag{3.2.21}$$

satisfies the same recurrence relation, with the initial conditions $U_{-1}(z) = 0$, $U_0(z) = 1$; its degree is $n-1$. (When we write just Chebyshev polynomial we refer to the first kind.)

The Chebyshev polynomial $T_n(x)$ has n zeros in $[-1, 1]$ given by

$$x_k = \cos\left(\frac{2k-1}{n}\frac{\pi}{2}\right), \quad k = 1 : n, \tag{3.2.22}$$

[26] Pafnuti Lvovich Chebyshev (1821–1894), Russian mathematician, pioneer in approximation theory and the constructive theory of functions. His name has many different transcriptions, e.g., Tschebyscheff. This may explain why the polynomials that bear his name are denoted $T_n(x)$. He also gave important contributions to probability theory and number theory.

the **Chebyshev points**, and $n+1$ *extrema*

$$x'_k = \cos\left(\frac{k\pi}{n}\right), \quad k = 0 : n. \tag{3.2.23}$$

These results follow directly from the fact that $\cos(n\phi) = 0$ for $\phi = (2k+1)\pi/(2n)$, and that $\cos(n\phi)$ has maxima for $\phi = k\pi/n$.

Note that from (3.2.19) it follows that $|T_n(x)| \leq 1$ for $x \in [-1, 1]$, in spite that its leading coefficient is as large as 2^{n-1}. The Chebyshev polynomials have a unique **minimax property**: (For a use of this property; see, Example 3.2.4.)

Example 3.2.3.

Figure 3.2.1 shows a plot of the Chebyshev polynomial $T_{12}(x)$ for $x \in [-1, 1]$. Setting $z = 1$ in the recurrence relation (3.2.20) and using $T_0(1) = T_1(1) = 1$, it follows that $T_n(1) = 1$, $n \geq 0$. From $T'_0(1) = 0$ an $T'_1(1) = 1$ and differentiating the recurrence relation we get

$$T'_{n+1}(z) = 2(zT'_n(z) + T_n(z)) - T'_{n-1}(z), (n \geq 1).$$

It follows easily by induction that $T'_n(1) = n^2$, that is *outside the interval* $[-1, 1]$ *the Chebyshev polynomials grow rapidly.*

Lemma 3.2.3.

The Chebyshev polynomials have the following **minimax property***: Of all* n*th degree polynomials with leading coefficient* 1*, the polynomial* $2^{1-n}T_n(x)$ *has the smallest magnitude* 2^{1-n} *in* $[-1, 1]$.

Proof. Suppose there were a polynomial $p_n(x)$, with leading coefficient 1 such that $|| < 2^{1-n}$ for all $x \in [-1, 1]$. Let x'_k, $k = 0 : n$, be the abscissae of the extrema of $T_n(x)$. Then we would have

$$p_n(x_0) < 2^{1-n}T_n(x'_0), \quad p_n(x_1) > 2^{1-n}T_n(x'_1), \quad p_n(x_2) < 2^{1-n}T_n(x'_2), \ldots,$$

etc., up to x'_n. From this it follows that the polynomial

$$p_n(x) - 2^{1-n}T_n(x)$$

changes sign in each of the n intervals (x'_k, x'_{k+1}), $k = 0 : n-1$. This is impossible, since the polynomial is of degree $n-1$. This proves the minimax property. □

Expansions in terms of Chebyshev polynomials are an important aid in studying functions on the interval $[-1, 1]$. If one is working with a function $f(t)$, $t \in [a, b]$, then one should make the substitution

$$t = \tfrac{1}{2}(a+b) + \tfrac{1}{2}(b-a)x, \tag{3.2.24}$$

which maps the interval $[-1, 1]$ onto $[a, b]$.

Consider the approximation to the function $f(x) = x^n$ on $[-1, 1]$ by a polynomial of lower degree. From the minimax property of Chebyshev polynomials it follows that the maximum magnitude of the error is minimized by the polynomial

$$p(x) = x^n - 2^{1-n}T_n(x). \tag{3.2.25}$$

From the symmetry property $T_n(-x) = (-1)^n T_n(x)$, it follows that this polynomial has in fact degree $n-2$. The error $2^{1-n}T_n(x)$ assumes its extrema 2^{1-n} in a sequence of $n + 1$ points, $x_i = \cos(i\pi/n)$. The sign of the error alternates at these points.

Suppose that one has obtained, e.g., by Taylor series, a truncated power series approximation to a function $f(x)$. By repeated use of (3.2.25), the series can be replaced by a polynomial of lower degree with a moderately increased bound for the truncation error. This process, called **economization of power series** often yields a useful polynomial approximation to $f(x)$ with a considerably smaller number of terms than the original power series.

Example 3.2.4.

If the series expansion $\cos x = 1 - x^2/2 + x^4/24 - \cdots$ is truncated after the x^4-term, the maximum error is 0.0014 in $[-1, 1]$. Since $T_4(x) = 8x^4 - 8x^2 + 1$, it holds that

$$x^4/24 \approx x^2/24 - 1/192$$

with an error which does not exceed $1/192 = 0.0052$. Thus the approximation

$$\cos x = (1 - 1/192) - x^2(1/2 - 1/24) = 0.99479 + 0.45833x^2$$

has an error whose magnitude does not exceed $0.0052 + 0.0014 < 0.007$. This is less than one-sixth of the error 0.042, which is obtained if the power series is truncated after the x^2-term.

Note that for the economized approximation cos(0) is not approximated by 1. It may not be acceptable that such an exact relation is lost. In this example one could have asked for a polynomial approximation to $(1 - \cos x)/x^2$ instead.

The Chebyshev polynomials are perhaps the most important example of a family of **orthogonal polynomials**; see Sec. 4.5.5. The **Chebyshev expansion** of a function $f(z)$,

$$f(z) = \sum_{j=0}^{\infty} c_j T_j(z), \tag{3.2.26}$$

have many useful properties. Set $e^{i\phi} = w$; ϕ and z may be complex. Then

$$z = \tfrac{1}{2}(w + w^{-1}), \quad T_n(z) = \tfrac{1}{2}(w^n + w^{-n}), \tag{3.2.27}$$
$$w = z \pm \sqrt{z^2 - 1}, \quad (z + \sqrt{z^2 - 1})^n = T_n(z) + U_{n-1}(z)\sqrt{z^2 - 1}.$$

It follows that the Chebyshev expansion (3.2.26) formally corresponds to a symmetric Laurent expansion,

$$g(w) = f\left(\tfrac{1}{2}(w + w^{-1})\right) = \sum_{-\infty}^{\infty} a_j w^j; \quad a_{-j} = a_j = \begin{cases} \tfrac{1}{2}c_j, & \text{if } j > 0; \\ c_0, & \text{if } j = 0. \end{cases}$$

It can be shown, e.g., by the parallelogram law, that $|z+1|+|z-1| = |w|+|w|^{-1}$, Hence, if $R > 1$, $z = \frac{1}{2}(w + w^{-1})$ maps the annulus $\{w : R^{-1} < |w| < R\}$, twice onto an ellipse $\mathcal{E}_R$, determined by the relation,

$$\mathcal{E}_R = \{z : |z-1| + |z+1| \le R + R^{-1}\}, \tag{3.2.28}$$

with foci at 1 and -1. The axes are, respectively, $R+R^{-1}$ and $R-R^{-1}$, and hence *R is the sum of the semi-axes.*

Note that, as $R \to 1$, the ellipse degenerates into the interval $[-1, 1]$. As $R \to \infty$, it becomes close to the circle $|z| < \frac{1}{2}R$. It follows from (3.2.27) etc. that this family of confocal ellipses are level curves of $|w| = |z \pm \sqrt{z^2-1}|$. In fact, we can also write,

$$\mathcal{E}_R = \left\{ z : 1 \le |z + \sqrt{z^2-1}| \le R \right\}. \tag{3.2.29}$$

Theorem 3.2.4.

Let $f(z)$ be real-valued for $z \in [-1, 1]$, analytic and single-valued for $z \in \mathcal{E}_R$, $R > 1$. Assume that $|f(z)| \le M$ for $z \in \mathcal{E}_R$. Then[27]

$$\Big| f(x) - \sum_{j=0}^{n-1} c_j T_j(x) \Big| \le \frac{2MR^{-n}}{1 - 1/R} \quad \textit{for } x \in [-1, 1].$$

Proof. Set as before, $z = \frac{1}{2}(w + w^{-1})$, $g(w) = f\big(\frac{1}{2}(w + w^{-1})\big)$. Then $g(w)$ is analytic in the annulus $R^{-1} + \epsilon \le |w| \le R - \epsilon$, and hence the Laurent expansion (1.2) converges there. In particular it converges for $|w| = 1$, hence the Chebyshev expansion for $f(x)$ converges when $x \in [-1, 1]$.

Set $r = R - \epsilon$. By Cauchy's formula, we obtain, for $j > 0$,

$$|c_j| = 2|a_j| = \Big| \frac{2}{2\pi i} \int_{|w|=r} g(w) w^{-(j+1)} dw \Big| \le \frac{2}{2\pi} \int_0^{2\pi} M r^{-j-1} r d\phi = 2Mr^{-j}.$$

We then obtain, for $x \in [-1, 1]$,

$$\Big| f(x) - \sum_{j=0}^{n-1} c_j T_j(x) \Big| = \Big| \sum_{n}^{\infty} c_j T_j(x) \Big| \le \sum_{n}^{\infty} |c_j| \le 2M \sum_{n}^{\infty} r^{-j} \le 2M \frac{r^{-n}}{1 - 1/r}.$$

This holds for any $\epsilon > 0$. We can here let $\epsilon \to 0$ and thus replace r by R. □

If a Chebyshev expansion converges rapidly, the truncation error is, by and large, determined by the first few neglected terms. As indicated by Figures 3.2.1 and 3.2.5 the error curve is oscillating with slowly varying amplitude in $[-1, 1]$. In contrast, the truncation error of a power series is proportional to a power of x.

Note that $f(z)$ is allowed to have a singularity arbitrarily close to the interval $[-1, 1]$, and the convergence of the Chebyshev expansion will still be exponential, although the exponential rate deteriorates, as $R \downarrow 1$.

[27] A generalization to complex values of x is formulated in Problem 6.

The numerical value of a truncated Chebyshev expansion can be computed by means of **Clenshaw's algorithm** which holds for any sum of the form $S = \sum_{k=1}^{n} c_k \phi_k$, where $\{\phi_k\}$ satisfies a **three term recurrence relation**

Theorem 3.2.5. Clenshaw's algorithm [13]

Suppose that a sequence $\{p_k\}$ satisfies the three term recurrence relation

$$p_{k+1} = \gamma_k p_k - \beta_k p_{k-1}, \quad k = 0 : n-1, \tag{3.2.30}$$

where $p_{-1} = 0$. Then

$$S = \sum_{k=0}^{n} c_k p_k = y_0 p_0$$

where y_0 is obtained by the recursion

$$\begin{aligned} &y_{n+1} = 0, \qquad y_n = c_n, \\ &y_k = c_k + \gamma_{k-1} y_{k+1} - \beta_k y_{k+2}, \quad k = n-1 : -1 : 0. \end{aligned} \tag{3.2.31}$$

Proof. Write the recursion (3.2.30) in matrix form as

$$\begin{pmatrix} 1 & & & & \\ -\gamma_0 & 1 & & & \\ \beta_1 & -\gamma_1 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{n-1} & -\gamma_{n-1} & 1 \end{pmatrix} \begin{pmatrix} p_0 \\ p_1 \\ \vdots \\ p_{n-1} \\ p_n \end{pmatrix} = \begin{pmatrix} p_0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix},$$

or $Lp = g$, $g = p_0 e_1$, where L is unit lower triangular and e_1 is the first column of the unit matrix. Then

$$S = c^T p = c^T L^{-1} g = g^T (L^T)^{-1} c = g^T y,$$

where y is the solution to the upper triangular system $L^T y = c$. Solving this by backsubstitution we get the recursion (3.2.31). □

It can proved that Clenshaw's algorithm is componentwise backward stable with respect to the data γ_k and β_k; see Smoktunowicz [41].

Clenshaw's algorithm can also be applied to series of Legendre functions, Bessel functions, Coulomb wave functions etc., because they satisfy recurrence relations of this type, where the α_k, γ_k depend on x; see the Handbook [1] or any text on special functions. Other applications are the case when the ϕ_k are the denominators or numerators of the approximants of a *continued fraction*; see Sec. 3.5.1

Important properties of trigonometric functions and Fourier series can be reformulated in the terminology of Chebyshev polynomials. For example, they satisfy certain orthogonality relations; see Sec. 4.5.5. Also results like (3.2.8) concerning how the rate of decrease of the coefficients or the truncation error of a Fourier series, is related to the smoothness properties of its sum, can be translated to Chebyshev expansions. So, even *if F is not analytic, a Chebyshev expansion converges under amazingly general conditions* (unlike a power series), but the convergence is much slower than exponential. A typical result reads: *if $f \in C^k[-1,1]$, $k > 0$, there exists a bound for the truncation error that decreases uniformly like $O(n^{-k}\log n)$.* Sometimes convergence acceleration can be successfully applied to such series.

3.2.4 Perturbation Expansions

In the equations of applied mathematics it is often possible to identify a small dimensionless parameter (say) ϵ, $\epsilon \ll 1$. The case when $\epsilon = 0$ is called the *reduced problem* or the unperturbed case, and one asks for a **perturbation expansion**, i.e. an expansion of the solution of the perturbed problem into powers of the perturbation parameter ϵ. In many cases it can be proved that the expansion has the form $c_0 + c_1\epsilon + c_2\epsilon^2 + \ldots$, but there are also important cases, where the expansion contains fractional or a few negative powers.

In this subsection, we consider an analytic equation $\phi(z,\epsilon) = 0$ and seek expansions for the roots $z_i(\epsilon)$ in powers of ϵ. This has some practical interest in its own right, but it is mainly to be considered as a preparation for more interesting applications of perturbation methods to more complicated problems. A simple perturbation example for a *differential equation* is given in Problem 10.

If $z_i(0)$ is a simple root, i.e. if $\partial\phi/\partial z \neq 0$, for $(z,\epsilon) = (z_i(0), 0)$, then a theorem of complex analysis tells us that $z_i(\epsilon)$ is an analytic function in a neighborhood of the origin, hence the expansion

$$z_i(\epsilon) - z_i(0) = c_1\epsilon + c_2\epsilon^2 + \ldots$$

has a positive (or infinite) radius of convergence. We call this a **regular perturbation problem**. The techniques of power series reversion, presented in Sec. 3.1.4, can often be applied after some preparation of the equation. Computer algebra systems are also used in perturbation problems, if expansions with many terms are needed.

Example 3.2.5.

We shall expand the roots of $\phi(z,\epsilon) \equiv \epsilon z^2 - z + 1 = 0$ into powers of ϵ. The reduced problem $-z + 1 = 0$ has only one finite root $z_1(0) = 1$. Set $z = 1 + x\epsilon$, $x = c_1 + c_2\epsilon + c_3\epsilon^2 + \ldots$. Then $\phi(1 + x\epsilon,\ \epsilon)/\epsilon = (1 + x\epsilon)^2 - x = 0$, i.e.

$$(1 + c_1\epsilon + c_2\epsilon^2 + \ldots)^2 - (c_1 + c_2\epsilon + c_3\epsilon^2 + \ldots) = 0.$$

Matching the coefficients of ϵ^0, ϵ^1, ϵ^2, we obtain the system

$$\begin{aligned} 1 - c_1 = 0 \ &\Rightarrow\ c_1 = 1; \\ 2c_1 - c_2 = 0 \ &\Rightarrow\ c_2 = 2; \\ 2c_2 + c_1^2 - c_3 = 0 \ &\Rightarrow\ c_3 = 5; \end{aligned}$$

hence $z_1(\epsilon) = 1 + \epsilon + 2\epsilon^2 + 5\epsilon^3 + \ldots$.

Now, the easiest way to obtain the expansion for the second root $z_2(\epsilon)$, is to use the fact that the sum of the roots of the quadratic equation equals ϵ^{-1}, hence $z_2(\epsilon) = \epsilon^{-1} - 1 - \epsilon - 2\epsilon^2 + \dots$.

Note the appearance of the term ϵ^{-1}. This is due to a characteristic feature of this example. The degree of the polynomial is lower for the reduced problem than it is for $\epsilon \neq 0$; one of the roots escapes to ∞ as $\epsilon \to 0$. This is an example of a **singular perturbation** problem, an important type of problem for differential equations; see Problem 10.

If $\partial\phi/\partial z = 0$, for some z_i, the situation is more complicated; z_i is a multiple root, and the expansions look differently. If $z_i(0)$ is a k-fold root then there may exist an expansion of the form

$$z_i(\epsilon) = c_0 + c_1\epsilon^{1/k} + c_2(\epsilon^{1/k})^2 + \dots$$

for each of the k roots of ϵ, but this is not always the case. See (3.2.32) below, where the expansions are of a different type. *If one tries to determine the coefficients in an expansion of the wrong form, one usually runs into contradictions,* but the question about the right form of the expansions still remains.

The answers are given by the classical theory of *algebraic functions*, where Riemann surfaces and Newton polygons are two of the key concepts, see, e.g., Bliss [5]. We shall, for several reasons, not use this theory here. One reason is that it seems hard to generalize some of the methods of algebraic function theory to more complicated equations, such as differential equations. We shall instead use a general **balancing procedure**, recommended in Lin and Segel [31, Sec. 9.1], where it is applied to singular perturbation problems for differential equations too.

The basic idea is very simple: each term in an equation behaves like some power of ϵ. *The equation cannot hold, unless there is a β, such that a pair of terms of the equation behave like $A\epsilon^\beta$, (with different values of A), and the ϵ-exponents of the other terms are larger than or equal to β.* (Recall that larger exponents make smaller terms.)

Let us return to the previous example. Although we have already determined the expansion for $z_2(\epsilon)$ (by a trick that may not be useful for other problems than single analytic equations), we shall use this task to illustrate the balancing procedure. Suppose that

$$z_2(\epsilon) \sim A\epsilon^\alpha,\ (\alpha < 0).$$

The three terms of the equation $\epsilon z^2 - z + 1 = 0$ then get the exponents

$$1 + 2\alpha,\ \alpha,\ 0.$$

Try the first two terms as the candidates for being the dominant pair. Then $1+2\alpha = \alpha$, hence $\alpha = -1$. The three exponents become $-1,\ -1,\ 0$. Since the third exponent is larger than the exponent of the candidates, this choice of pair seems possible, but we have not shown that it is the only possible choice.

Now try the first and the third terms as candidates. Then $1 + 2\alpha = 0$, hence $\alpha = -\frac{1}{2}$. The exponent of the non-candidate is $-\frac{1}{2} \leq 0$; this candidate pair is thus impossible. Finally, try the second and the third terms. Then $\alpha = 0$, but we are only interested in negative values of α.

The conclusion is that we can try coefficient matching in the expansion $z_2(\epsilon) = c_{-1}\epsilon^{-1} + c_0 + c_1\epsilon + \ldots$. We don't need to do it, since we know the answer already, but it indicates how to proceed in more complicated cases.

Example 3.2.6.

First consider the equation $z^3 - z^2 + \epsilon = 0$. The reduced problem $z^3 - z^2 = 0$ has a single root, $z_1 = 1$, and a double root, $z_{2,3} = 0$. No root has escaped to ∞. By a similar coefficient matching as in the previous example we find that $z_1(\epsilon) = 1 - \epsilon - 2\epsilon^2 + \ldots$. For the double root, set $z = A\epsilon^\beta$, $\beta > 0$. The three terms of the equation obtain the exponents $3\beta,\ 2\beta, 1$. Since 3β is dominated by 2β we conclude that $2\beta = 1$, i.e.$\beta = 1/2$,

$$z_{2,3}(\epsilon) = c_0\epsilon^{1/2} + c_1\epsilon + c_2\epsilon^{3/2} + \ldots.$$

By matching the coefficients of $\epsilon,\ \epsilon^{3/2},\ \epsilon^2$, we obtain the system

$$\begin{aligned} -c_0^2 + 1 = 0 &\ \Rightarrow c_0 = \pm 1, \\ -2c_0c_1 + c_0^3 = 0 &\ \Rightarrow c_1 = \tfrac{1}{2}, \\ -2c_0c_2 - c_1^2 + 2c_0^2c_1 + c_1c_0^2 = 0 &\ \Rightarrow c_2 = \pm\tfrac{5}{8}, \end{aligned}$$

hence $z_{2,3}(\epsilon) = \pm\epsilon^{1/2} + \frac{1}{2}\epsilon \pm \frac{5}{8}\epsilon^{3/2} + \ldots$.

There are, however, equations with a double root, where the perturbed pair of roots do not behave like $\pm c_0\epsilon^{1/2}$ as $\epsilon \to 0$. In such cases the balancing procedure may help. Consider the equation

$$(1 + \epsilon)z^2 + 4\epsilon z + \epsilon^2 = 0. \tag{3.2.32}$$

The reduced problem reads $z^2 = 0$, with a double root. Try $z \sim A\epsilon^\alpha$, $\alpha > 0$. The exponents of the three terms become $2\alpha,\ \alpha + 1,\ 2$. We see that $\alpha = 1$ makes the three exponents all equal to 2; this is fine. So, set $z = \epsilon y$. The equation reads, after division by ϵ^2, $(1 + \epsilon)y^2 + 4y + 1 = 0$, hence $y(0) = a \equiv -2 \pm \sqrt{3}$. Coefficient matching yields the result

$$z = \epsilon y = a\epsilon + (-a + a^2/2)\epsilon^2 + \ldots, \quad a = -2 \pm \sqrt{3},$$

where all exponents are natural numbers.

If ϵ is small enough, the last term included can serve as an error estimate. A more reliable error estimate (or even an error bound) can be obtained by inserting the truncated expansion into the equation. It shows that *the truncated expansion satisfies a modified equation exactly.* The same idea was indicated for a differential equation in Example 3.1.2; see also Problem 10, and it can be applied to equations of many other types.

3.2.5 Ill-Conditioned Series

Slow convergence is not the only numerical difficulty that occurs in connection with infinite series. There are also series with oscillating terms and a complicated type of catastrophic cancellation. The size of some terms are many orders of magnitude larger than the sum of the series. Small relative errors in the computation of the large terms lead to a large relative error in the result. We call such a series **ill-conditioned**.

An important class of sequences $\{c_n\}$, are known as **completely monotonic**.

Definition 3.2.6.

A sequence $\{u_n\}$ *is completely monotonic for* $n \geq a$ *iff*

$$u_n \geq 0, \quad (-\Delta)^j u_n \geq 0, \quad \forall j \geq 0, \ n \geq a, \ \text{(integers)}.$$

Such series have not been subject to many systematic investigations. One simply tries to avoid them. For the important "special functions" of Applied Mathematics, such as Bessel Functions, confluent hypergeometric functions etc., there usually exists *expansions into descending powers of* z that can be useful, when $|z| \gg 1$ and the usual series, in *ascending* powers, are divergent or ill-conditioned. Another possibility is to use *multiple precision* in computations with ill-conditioned power series; this is relatively expensive and laborious (but the difficulties should not be exaggerated). There are, however, also other, less known, possibilities that will now be exemplified. The subject is still open for new fresh ideas, and we hope that the following pages and the related problems at the end of the section will stimulate some readers to thinking about it.

First, we shall consider power series of the form

$$\sum_{n=0}^{\infty} \frac{(-x)^n c_n}{n!}, \tag{3.2.33}$$

where $x \gg 1$, although not so large that there is risk for overflow. We assume that the coefficients c_n are positive and slowly varying (relatively to $(-x)^n/n!$). The ratio of two consecutive terms is

$$\frac{c_{n+1}}{c_n} \frac{-x}{n+1} \approx \frac{-x}{n+1}.$$

We see that the series converges for all x, and that the magnitude increases iff $n + 1 < |x|$. *The term of largest magnitude is thus obtained for* $n \approx |x|$. Denote its magnitude by $M(x)$. Then, for $x \gg 1$, the following type of approximations can be used, e.g., for crude estimates of the number of terms needed, the arithmetic precision that is to be used etc. in computations related to ill-conditioned power series:

$$M(x) \approx c_x e^x (2\pi x)^{-1/2}, \quad \text{i.e.,} \ \log_{10} M(x)/c_0 \approx 0.43x - \tfrac{1}{2}\log_{10}(2\pi x). \tag{3.2.34}$$

This follows from the classical **Stirling's formula**,

$$x! \sim (x/e)^x \sqrt{2\pi x}, \quad x \gg 1, \tag{3.2.35}$$

that gives $x!$ with a relative error that is about $\frac{1}{12x}$. You find a proof of this in most textbooks on calculus. It will often be used in the rest of this book. A more accurate and general version is given in Example 3.3.12 together with a few more facts about the gamma function, $\Gamma(z)$, an analytic function that interpolates the factorial, $\Gamma(n+1) = n!$ if n is a natural number. Sometimes the notation $z!$ is used instead of $\Gamma(z+1)$ also if z is not an integer.

There exist **preconditioners**, i.e. transformations that can convert classes of ill-conditioned power series (with accurately computable coefficients) to more well-conditioned problems. One of the most successful preconditioners known to the authors is the following:[28]

$$\sum_{n=0}^{\infty} \frac{(-x)^n c_n}{n!} = e^{-x} \sum_{m=0}^{\infty} \frac{x^m (-\Delta)^m c_0}{m!}. \tag{3.2.36}$$

This identity is proved in Example . A hint to a shorter proof is given in Problem 3.21.

Example 3.2.7.

Consider the function

$$F(x) = \frac{1}{x} \int_0^x \frac{1-e^{-t}}{t}\, dt = 1 - \frac{x}{2^2 \cdot 1!} + \frac{x^2}{3^2 \cdot 2!} - \dots,$$

i.e. $F(x)$ is a particular case of (3.2.33) with $c_n = (n+1)^{-2}$. We shall look at three methods of computing $F(x)$ for $x = 10 : 10 : 50$, named A, B, C. $F(x)$ decreases smoothly from 0.2880 to 0.0898. The computed values of $F(x)$ are denoted $FA(x), FB(x), FC(x)$.

The coefficients c_n, $n = 0 : 119$, are given in IEEE floating point, double precision. The table of results show that, except for $x = 50$, 120 terms is much more than necessary for the rounding of the coefficients to become the dominant error source.

x	10	20	30	40	50
$F(x) \approx$	0.2880	0.1786	0.1326	0.1066	0.0898
lasttermA	$1 \cdot 10^{-82}$	$8 \cdot 10^{-47}$	$7 \cdot 10^{-26}$	$6 \cdot 10^{-11}$	$2 \cdot 10^{1}$
$M(x; A)$	$3 \cdot 10^{1}$	$1 \cdot 10^{5}$	$9 \cdot 10^{8}$	$1 \cdot 10^{13}$	$1 \cdot 10^{17}$
$\lvert FA(x) - F(x)\rvert$	$2 \cdot 10^{-15}$	$5 \cdot 10^{-11}$	$2 \cdot 10^{-7}$	$3 \cdot 10^{-3}$	$2 \cdot 10^{1}$
lasttermB	$4 \cdot 10^{-84}$	$1 \cdot 10^{-52}$	$4 \cdot 10^{-36}$	$2 \cdot 10^{-25}$	$2 \cdot 10^{-18}$
$M(x; B)$	$4 \cdot 10^{-2}$	$2 \cdot 10^{-2}$	$1 \cdot 10^{-2}$	$7 \cdot 10^{-3}$	$5 \cdot 10^{-3}$
$\lvert FC(x) - FB(x)\rvert$	$7 \cdot 10^{-9}$	$2 \cdot 10^{-14}$	$6 \cdot 10^{-17}$	0	$1 \cdot 10^{-16}$

[28] The notation $\Delta^m c_n$ for high order differences was introduced in Sec. 1.1.3.

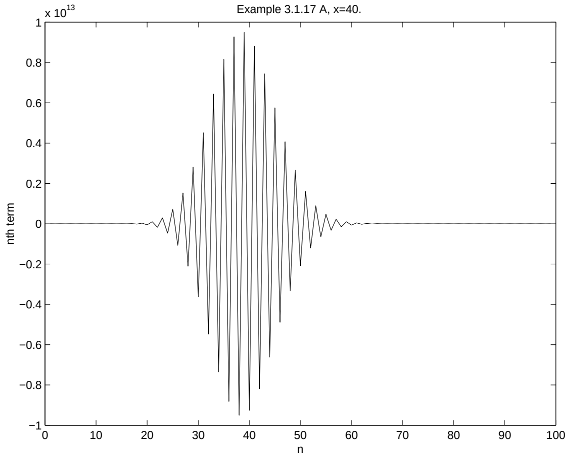

Figure 3.2.2. *Example* 3.2.5 *A: Terms of* (3.2.33), $c_n = (n+1)^{-2}$, $x = 40$, *no preconditioner. Note the scale, and look also in the table. Since the largest term is* 10^{13}, *it is no surprise that the relative error of the sum is not better than* 0.03, *in spite that double precision floating point has been used.*

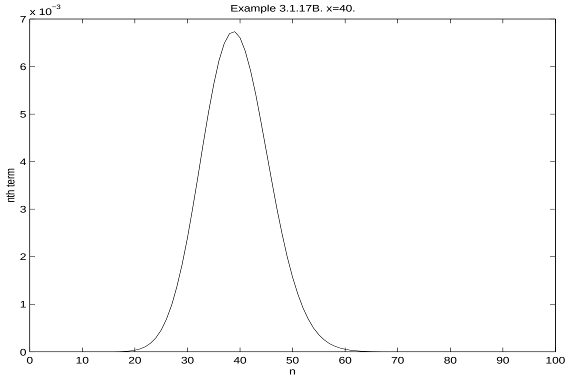

Figure 3.2.3. *Example* 3.2.5 *B:* $c_n = (n+1)^{-2}$, $x = 40$, *with the preconditioner in* (3.2.36). *The terms of the right hand side, including the factor* e^{-x}, *becomes a so-called* **bell sum***; the largest term is about* $7\,10^{-3}$. *The computed sum is correct to 16 decimal places.*

A We use (3.2.33) without preconditioner. $M(x; A)$ is the largest magnitude of the terms of the expansion. $M(x; A) \cdot 10^{-16}$ tells the order of magnitude of the effect of the rounding errors on the computed value $FA(x)$. Similarly, the truncation error is crudely estimated by lasttermA. See also Figure 3.1.6.

B. We use the preconditioner (3.2.36). In this example $c_n = (n+1)^{-2}$. In Problem 3.2.2(c) we find the following explicit expressions, related to the series on the right hand side of the preconditioner for this example.

$$(-\Delta)^m c_0 = (-\Delta)^m c_n|_{n=0} = c_0(-\Delta)^m x^{-2}|_{x=1} = \frac{c_0}{m+1}\sum_{k=0}^{m}\frac{1}{k+1},$$

$$F(x) = c_0 e^{-x}\sum_{m=0}^{\infty}\frac{x^m}{m!}\frac{1}{(m+1)}\sum_{k=0}^{m}\frac{1}{k+1}. \tag{3.2.37}$$

Note that $(-\Delta)^m c_0$ is positive and smoothly decreasing; (This is not a special feature only for this example, but it holds for sequences $\{c_n\}$, which are completely monotonic.)

The largest term is thus smaller than the sum, and the series (3.2.37) is **well-conditioned**. It can be shown that, if $x \gg 1$, the mth term is approximately proportional to the value at m of the normal probability density with mean x and standard deviation equal to $\sqrt{x}$; note the resemblance to a Poisson distribution. Multiple precision is not needed here.

$M(x; B)$ and lasttermB are defined analogously to $M(x; A)$ and lasttermA, The B-values are very different from the A-values. In fact they indicate that *all values of $FB(x)$, referred to in the table, give $F(x)$ to full accuracy.*

C. The following expression for $F(x)$,

$$xF(x) \equiv \sum_{n=1}^{\infty}\frac{(-x)^n}{nn!} = -\gamma - \ln x - E_1(x); \quad E_1(x) = \int_x^{\infty}\frac{e^{-t}}{t}\,dt, \tag{3.2.38}$$

is valid for all $x > 0$; see [1, **5.1.11**]. $E_1(x)$ is known as the **exponential integral**, and

$$\gamma = 0.57721\,56649\,01532\,86061\ldots$$

is the well known **Euler's constant**. In the next section, an asymptotic expansion for $E_1(x)$ for $x \gg 1$ is derived, the first two terms of which are used here in the computation of $F(x; C)$ for the table above.

$$E_1(x) \approx e^{-x}(x^{-1} - x^{-2}), \quad x \gg 1.$$

This approximation is the dominant part of the error of $F(x; C)$; it is less than $e^{-x}2x^{-4}$. $F(x; C)$ gives full accuracy for (say) $x > 25$.

More examples of sequences, for which rather simple explicit expressions for the high order differences are known, are given in Problem 3.21. The **Kummer confluent hypergeometric function** $M(a, b, x)$ was defined in (3.1.14). We have

$$M(a,b,x) = 1 + \sum_{n=1}^{\infty}\frac{(-x)^n c_n}{n!}, \quad c_n = c_n(a,b) = \frac{a(a+1)\ldots(a+n-1)}{b(b+1)\ldots(b+n-1)}.$$

In our context $b > a > 0$, $n > 0$. The oscillatory series for $M(a, b, -x)$, $x > 0$, is ill-conditioned if $x \gg 1$.

By Problem 3.21, $(-\Delta)^n c_0(a,b) = c_n(b-a,b) > 0$, $n > 0$, hence the preconditioner (3.2.36) yields the equation

$$M(a,b,-x) = e^{-x} M(b-a,b,x), \tag{3.2.39}$$

where the series on the right hand side has positive terms, because $b-a > 0$, $x > 0$, and is a well-conditioned *bell sum.* The mth term has typically a sharp maximum for $m \approx x$; compare Figure 3.2.7. Equation (3.2.39) is in the theory of the confluent hypergeometric functions known as **Kummer's first identity**. It is emphasized here, because several functions with famous names of their own are particular cases of the Kummer function. These share the numerous useful properties of Kummer's function, e.g., the above identity; see the theory in Lebedev [30, Secs. 9.9–9.14][29] and the formulas in [1, Ch. 13] in particular Table 13.6 of special cases. An important example is the error function (see Example 1.2.3) that can be expressed in terms of Kummer's confluent hypergeometric as .

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt = \frac{2x}{\sqrt{\pi}} M\left(\frac{1}{2}, \frac{3}{2}, -x^2\right). \tag{3.2.40}$$

If we cannot find explicit expressions for high order differences, we can make a *difference scheme* by the recurrence $\Delta^{m+1} c_n = \Delta^m c_{n+1} - \Delta^m c_n$. Unfortunately the computation of a difference scheme suffers from numerical instability. Suppose that the absolute errors of the c_n are bounded by ϵ. Then the absolute errors can become as large as 2ϵ in the first differences, 4ϵ in the second differences etc. More generally, the absolute errors of $(-\Delta)^m c_n$ can become as large as $2^m \epsilon$. (You find more about this in Examples 3.2.2 and 3.2.3.) In connection with ill-conditioned series, this instability is much more disturbing than in the traditional applications of difference schemes to interpolation etc., where m is seldom much larger than 10. Recall that $m \approx x$ for the largest term of the preconditioned series. So, if $x > 53$ even this term may not have any correct bit if IEEE double precision is used, and many terms are needed after this.

So, during the computation of the new coefficients, $(-\Delta)^m c_n$, (only once for the function F, and with double accuracy in the results), the old coefficients c_n must be available with multiple accuracy, and multiple precision must be used in the computation of their difference scheme. Otherwise, we cannot evaluate the series with a decent accuracy for much larger values of x than we could have done without preconditioning. Note, however, that if satisfactory coefficients have been obtained for the preconditioned series, double precision is sufficient when the series is evaluated for large values of x. (It is different for method A above.)

Let $F(x)$ be the function that we want to compute for $x \gg 1$, where it is defined by an ill-conditioned power series $F_1(x)$. A more general preconditioner can be described as follows. Try to find a power series $P(x)$ with positive coefficients such that the power series $P(x)F_1(x)$ has less severe cancellations than than $F_1(x)$.

In order to distinguish between the algebraic manipulation and the numerical evaluation of the functions defined by these series, we introduce the indeterminate

[29] Unfortunately, the formulation of Kummer's first identity in [30, Eqn. (9.11.2)] contains a serious sign error.

$\mathbf{x}$ and describe a **more general preconditioner** as follows:

$$\mathbf{F_2^*(x)} = \mathbf{P(x)} \cdot \mathbf{F_1(x)}; \qquad F_2(x) = F_2^*(x)/P(x). \tag{3.2.41}$$

The second statement is a usual scalar evaluation (no bold-face). Here $P(x)$ may be evaluated by some other method than the power series, if it is more practical. If $P(x) = e^x$, and $F_1(x)$ is the series defined by (3.2.33), then it can be shown that $F_2(x)$ is mathematically equivalent to the right hand side of (3.2.36); see Example 3.2.1. In these cases $F_2(x)$ has positive coefficients.

If, however, $F_1(x)$ has a positive zero, this is also a zero of $F_2^*(x)$, and hence it is impossible that all coefficients of the series $\mathbf{F_2^*(x)}$ have the same sign. Nevertheless, the following example shows that the preconditioner (3.2.41) can sometimes be successfully used in such a case too.

Table 3.2.1.

1	x	10	20	30	40	50
2	$J_0(x) \approx$	$-2 \cdot 10^{-1}$	$2{\cdot}10^{-1}$	$-9 \cdot 10^{-2}$	$7 \cdot 10^{-3}$	$6 \cdot 10^{-2}$
3	$N1(x)$	26	41	55	69	82
4	$J(x; N1) - J_0(x)$	$9 \cdot 10^{-14}$	$3 \cdot 10^{-10}$	$-2 \cdot 10^{-6}$	$-1 \cdot 10^{-1}$	$-2 \cdot 10^{2}$
5	$N2(x)$	16	26	36	46	55
6	$IJ(x; N2) \approx$	$-7 \cdot 10^{2}$	$7 \cdot 10^{6}$	$-7 \cdot 10^{10}$	$1 \cdot 10^{14}$	$2 \cdot 10^{19}$
7	$I_0(x) \approx$	$3 \cdot 10^{3}$	$4 \cdot 10^{7}$	$8 \cdot 10^{11}$	$1 \cdot 10^{16}$	$3 \cdot 10^{20}$
8	$IJ(x)/I_0(x) - J_0(x)$	$3 \cdot 10^{-17}$	$2 \cdot 10^{-14}$	$3 \cdot 10^{-13}$	$-5 \cdot 10^{-12}$	$2 \cdot 10^{-10}$

Example 3.2.8.

The two functions

$$J_0(x) = \sum_{n=0}^{\infty} (-1)^n \frac{(x^2/4)^n}{(n!)^2}, \qquad I_0(x) = \sum_{n=0}^{\infty} \frac{(x^2/4)^n}{(n!)^2},$$

are examples of Bessel functions of the first kind; I_0 is nowadays called a modified Bessel function. $J_0(x)$ is oscillatory and bounded, while $I_0(x) \sim e^x/\sqrt{2\pi x}$ for $x \gg 1$. Since all coefficients of I_0 are positive, we shall set $P = I_0$, $F_1 = J_0$, and try

$$\mathbf{F_2^*(x)} = \mathbf{IJ(x)} \equiv \mathbf{I_0(x)} \cdot \mathbf{J_0(x)}, \quad F_2(x) = F_2^*(x)/I_0(x),$$

as a preconditioner for the power series for $J_0(x)$, which is ill-conditioned if $x \gg 1$. In Table 3.2.1 line 2 and line 7 are obtained from the fully accurate built-in functions for $J_0(x)$ and $I_0(x)$. $J(x; N1)$ is computed in IEEE double precision from $N1$ terms of the above power series for $J_0(x)$. $N1 = N1(x)$ is obtained by a termination criterion that should give full accuracy or, if the estimate of the effect of the rounding error is bigger than 10^{-16}, the truncation error should be smaller than this estimate. We omit the details; see also Problem 12 (d).

The coefficients of $\mathbf{IJ(x)}$ are obtained from the second expression for γ_m given in Problem 12 (c). $N2 = N2(x)$ is the number of terms used in the expansion of $\mathbf{IJ(x)}$, by a termination criterion, similar to the one described for $J(x;N1)$. Compared to line 4, line 8 is a remarkable improvement, obtained without the use of multiple precision.

For series of the form

$$\sum_{n=0}^{\infty} a_n \frac{(-x^2)^n}{(2n)!}$$

one can generate a preconditioner from $P(x) = \cosh x$. This can also be applied to $J_0(x)$ and other Bessel functions; see Problem 12 (e).

3.2.6 Divergent or Semiconvergent Series

That a series is convergent is no guarantee that it is numerically useful. In this section, we shall see examples of the reverse situation: a divergent series can be of use in numerical computations. This sounds strange, but it refers to series where the size of the terms decreases rapidly at first and increases later, and where an error bound (see Figure 3.2.4), can be obtained in terms of the first neglected term. Such series are sometimes called **semiconvergent**. An important subclass are the *asymptotic* series; see below.

Example 3.2.9.

We shall derive a semiconvergent series for the computation of Euler's function

$$f(x) = e^x E_1(x) = e^x \int_x^{\infty} e^{-t} t^{-1}\, dt = \int_0^{\infty} e^{-u}(u+x)^{-1}\, du$$

for large values of x. (The second integral was obtained from the first by the substitution $t = u + x$.) The expression $(u+x)^{-1}$ should first be expanded in a geometric series with remainder term, valid even for $u > x$,

$$(u+x)^{-1} = x^{-1}(1+x^{-1}u)^{-1} = x^{-1}\sum_{j=0}^{n-1}(-1)^j x^{-j} u^j + (-1)^n (u+x)^{-1}(x^{-1}u)^n$$

We shall frequently use the well known formula

$$\int_0^{\infty} u^j e^{-u}\, du = j! = \Gamma(j+1).$$

We write $f(x) = S_n(x) + R_n(x)$, where

$$S_n(x) = x^{-1}\sum_{j=0}^{n-1}(-1)^j x^{-j}\int_0^{\infty} u^j e^{-u} du = \frac{1}{x} - \frac{1!}{x^2} + \frac{2!}{x^3} - \ldots + (-1)^{n-1}\frac{(n-1)!}{x^n},$$

$$R_n(x) = (-1)^n \int_0^{\infty} (u+x)^{-1}\Big(\frac{u}{x}\Big)^n e^{-u} du.$$

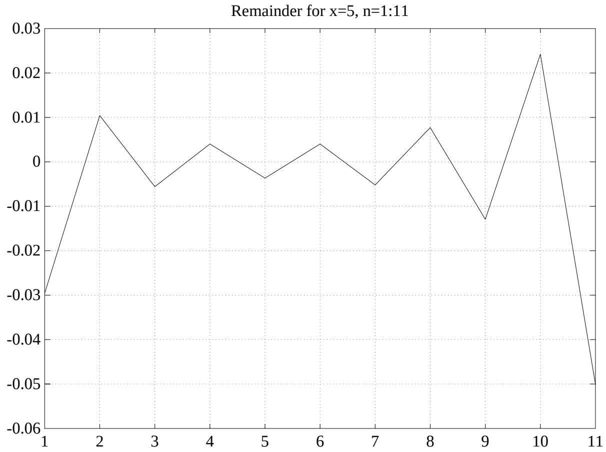

Figure 3.2.4. *The first 11 error estimates of the semiconvergent series of Example* 3.2.7*; see* (3.2.43)*. The smallest actual error is only 5% of the smallest error estimate.*

The terms in $S_n(x)$ qualitatively behave as in Figure 3.2.4. The ratio between the last term in S_{n+1} and the last term in S_n is

$$-\frac{n!}{x^{n+1}}\frac{x^n}{(n-1)!} = -\frac{n}{x}, \tag{3.2.42}$$

and since the absolute value of that ratio for fixed x is unbounded as $n \to \infty$, the sequence $\{S_n(x)\}_{n=1}^{\infty}$ *diverges for every positive* x. But since $sign\, R_n(x) = (-1)^n$ for $x > 0$, it follows from Theorem 3.1.4 that

$$f(x) = \frac{1}{2}\Big(S_n(x) + S_{n+1}(x)\Big) \pm \frac{1}{2}\frac{n!}{x^{n+1}}. \tag{3.2.43}$$

The idea is now to choose n so that the estimate of the remainder is as small as possible. According to (3.2.42), this happens when n is equal to the integer part of x. For $x = 5$ we choose $n = 5$,

$$S_5(5) = 0.2 - 0.04 + 0.016 - 0.0096 + 0.00768 = 0.17408,$$
$$S_6(5) = S_5(5) - 0.00768 = 0.16640,$$

which gives $f(5) = 0.17024 \pm 0.00384$. The correct value is 0.17042, so the actual error is only 5% of the error bound.

For larger values of x the accuracy attainable increases. One can show that the bound for the *relative* error using the above computational scheme decreases approximately as $(\pi\cdot x/2)^{1/2}e^{-x}$; an extremely good accuracy for large values of x, if one stops at the smallest term. It can even be improved further, by the use of the convergence acceleration techniques presented in Sec. 3.4, notably the *repeated averages* algorithm, also known as the **Euler transformation**; see Sec. 3.4.3. The algorithms for the transformation of a power series into a rapidly convergent continued fraction, mentioned in Sec. 3.5.1, can also be successfully applied to this example and to many other divergent expansions.

One can derive the same series expansion as above by repeated integration by parts. This is often a good way to derive numerically useful expansions, convergent or semi-convergent, with a remainder in the form of an integral. For convenient reference, we formulate this as a lemma that is easily proved by induction and the mean value theorem of integral calculus. See Problem 13 for applications.

Lemma 3.2.7. *Repeated Integration by Parts.*

Let $F \in C^p(a,b)$, let G_0 be a piecewise continuous function, and let $G_0, G_1, \dots$ be a sequence of functions such that $G'_{j+1}(x) = G_j(x)$ with suitably chosen constants of integration. Then

$$\int_a^b F(t)G_0(t)\,dt = \sum_{j=0}^{p-1}(-1)^j F^{(j)}(t)G_{j+1}(t)\Big|_{t=a}^{b} + (-1)^p \int_a^b F^{(p)}(t)G_p(t)\,dt.$$

The sum is the "expansion", and the last integral is the "remainder". If $G_p(t)$ has a constant sign in (a,b), the remainder term can also be written in the form

$$(-1)^p F^{(p)}(\xi)(G_{p+1}(b) - G_{p+1}(a)), \quad \xi \in (a,b).$$

The expansion in Lemma 3.2.7 is valid as an *infinite* series, if and only if the remainder tends to 0 as $p \to \infty$. Even if the sum converges as $p \to \infty$, it may converge to the wrong result.

The series in Example 3.2.9 is an expansion in *negative* powers of x, with the property that for all n, the remainder, when $x \to \infty$, approaches zero faster than the last included term. Such an expansion is said to **represent** $f(x)$ **asymptotically** as $x \to \infty$. Such an **asymptotic series** can be either convergent or divergent (semiconvergent). In many branches of applied mathematics, divergent asymptotic series are an important aid, though they are often needlessly surrounded by an air of mysticism.

It is important to appreciate that *an asymptotic series does not define a sum uniquely.* For example $f(x) = e^{-x}$ is asymptotically represented by the series $\sum 0x^{-j}$, as $x \to \infty$. So e^{-x}, (and many other functions), can therefore be added to the function, for which the expansion was originally obtained.

Asymptotic expansions are not necessarily expansions into negative powers of x. An expansion into *positive* powers of $x - a$,

$$f(x) \sim \sum_{\nu=0}^{n-1} c_\nu (x-a)^\nu + R_n(x),$$

represents $f(x)$ asymptotically when $x \to a$ if

$$\lim_{x\to a}(x-a)^{-(n-1)}R_n(x)=0.$$

Asymptotic expansions of the error of a numerical method into positive powers of a step length h are of great importance in the more advanced study of numerical methods. Such expansions form the basis of simple and effective acceleration methods for improving numerical results; see Sec. 3.4.

Review Questions

1. Give the Cauchy formula for the coefficients of Taylor and Laurent series, and describe the Cauchy–FFT method. Give the formula for the coefficients of a Fourier series. For which of the functions in Table 3.1.1 does also another Laurent expansion exist?

2. Describe by an example the balancing procedure that was mentioned in the subsection about perturbation expansions.

3. Define the Chebyshev polynomials, and tell some interesting properties of these and of Chebyshev expansions. For example, what do you know about the speed of convergence of a Chebyshev expansion for various classes of functions? (The detailed expressions are not needed.)

4. Describe and exemplify, what is meant by an ill-conditioned power series and a preconditioner for such a series.

5. Define what is meant, when one says that the series $\sum_0^\infty a_n x^{-n}$

(a) converges to a function $f(x)$ for $x \ge R$;

(b) represents a function $f(x)$ asymptotically as $x \to \infty$.

(c) Give an example of a series that represents a function asymptotically as $x \to \infty$, although it diverges for every finite positive x.

(d) What is meant by semi-convergence? Say a few words about termination criteria and error estimation.

Problems and Computer Exercises

1. Some of the functions appearing in Table 3.1.1, in Problem 3.1.6, and in other examples and problems are *not single-valued* in the complex plane. Brush up your Complex Analysis, and find out how to define the branches, where these expansions are valid, and (if necessary) define cuts in the complex plane that must not be crossed. It turns out not to be necessary for these expansions. Why?

(a) If you have access to programs for functions of complex variables (or to commands in some package for interactive computation), find out the conventions used for functions like square root, logarithm, powers, arctan etc.

If the manual does not give enough detail, invent numerical tests, both with strategically chosen values of z and with random complex numbers in some appropriate domain around the origin. For example, do you obtain

$$\ln\left(\frac{z+1}{z-1}\right) - \ln(z+1) + \ln(z-1) = 0, \quad \forall z?$$

Or, what values of $\sqrt{z^2-1}$ do you obtain for $z = \pm i$? What values should you obtain, if you want the branch which is positive for $z > 1$?

(b) What do you obtain, if you apply Cauchy's coefficient formula or the Cauchy–FFT method to find a Laurent expansion for $\sqrt{z}$? Note that $\sqrt{z}$ is analytic everywhere in an annulus, but that does not help. The expansion is likely to become weird. Why?

2. (a) Apply (on a computer) the Cauchy–FFT method to find the Maclaurin coefficients a_n of (say) e^z, $\ln(1-z)$ and $(1+z)^{1/2}$. Make experiments with different values of r and N, and compare with the exact coefficients. This presupposes that you have access to good programs for complex arithmetic and FFT.

Try to summarize your experiences of how the error of a_n depends on r, N. You may find some guidance in Example 3.2.2.

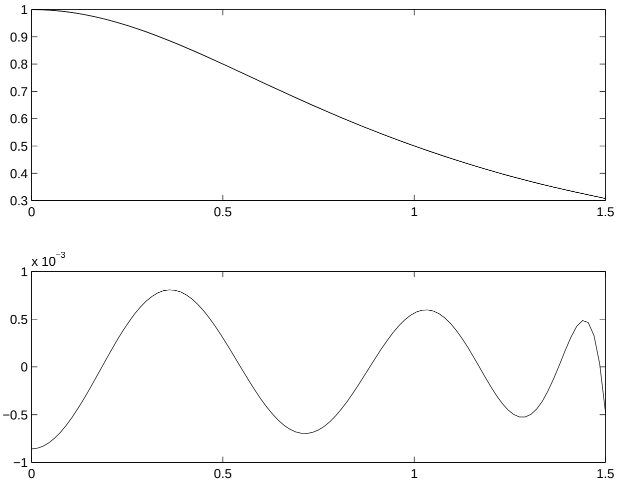

Figure 3.2.5. *Illustrations to Problem* 3c. Upper part: *The function* $f(x) = \frac{1}{1+x^2}$, $x \in [0, 1.5]$. Lower part: *The error of the expansion of* $f(x)$ *in a sum of Chebyshev polynomials* $\{T_n(x/1.5)\}$, $n \le 10$. *The scale is* 10^{-3} *in the lower curve.*

3. (a) Suppose that r is located inside the unit circle; t is real. Show that

$$\frac{1-r^2}{1-2r\cos t + r^2} = 1 + 2\sum_{n=1}^{\infty} r^n \cos nt,$$

$$\frac{2r\sin t}{1-2r\cos t+r^2} = 2\sum_{n=1}^{\infty} r^n \sin nt.$$

Hint: First suppose that r is real. Set $z = re^{it}$. Show that the two series are the real and imaginary parts of $(1+z)/(1-z)$. Finally make analytic continuation of the results.

(b) Let a be positive, $x \in [-a,a]$, while w is complex, $w \notin [-a,a]$. Let $r = r(w)$, $|r| < 1$ be a root of the quadratic $r^2 - (2w/a)r + 1 = 0$. Show that (with an appropriate definition of the square root)

$$\frac{1}{w-x} = \frac{1}{\sqrt{w^2-a^2}} \cdot \left(1 + 2\sum_{n=1}^{\infty} r^n T_n\left(\frac{x}{a}\right)\right), \quad (w \notin [-a,\ a],\ x \in [-a,a]).$$

(c) Find the expansion of $1/(1+x^2)$ for $x \in [-1.5, 1.5]$ into the polynomials $T_n(x/1.5)$. Explain the order of magnitude of the error and the main features of the error curve in Figure 3.2.5.

Hint: Set $w = i$, and take the imaginary part. Note that r becomes imaginary.

4. (a) Find the Laurent expansions for

$$f(z) = 1/(z-1) + 1/(z-2).$$

(b) How do you use the Cauchy–FFT method for finding Laurent expansions? Test your ideas on the function in the previous subproblem (and on a few other functions). There may be some pitfalls with the interpretation of the output from the FFT program, related to so-called **aliasing**; see Sec. 4.6.4 and Strang [44].

(c) As in Sec. 3.2.1, suppose that $F^{(p)}$ is of bounded variation in $[-\pi,\pi]$ and denote the Fourier coefficients of $F^{(p)}$ by $c_n^{(p)}$. Derive the following generalization of (3.2.8):

$$c_n = \frac{(-1)^{n-1}}{2\pi} \sum_{j=0}^{p-1} \frac{F^{(j)}(\pi) - F^{(j)}(-\pi)}{(in)^{j+1}} + \frac{c_n^{(p)}}{(in)^p},$$

and show that if we add the condition that $F \in C^j[-\infty,\infty]$, $j < p$, then the asymptotic results given in (and after) (3.2.8) hold.

(d) Let $z = \frac{1}{2}(w + w^{-1})$. Show that $|z-1| + |z+1| = |w| + |w|^{-1}$.

Hint: Use the parallelogram law, $|p-q|^2 + |p+q|^2 = 2(|p|^2 + |q|^2)$.

5. (a) The expansion of arcsinh t into powers of t, truncated after t^7, is obtained from Problem 1.6 (b). Using economization of a power series construct from this a polynomial approximation of the form $c_1t + c_3t^3$ in the interval $-\frac{1}{2} \le t \le \frac{1}{2}$. Give bounds for the truncation error for the original truncated expansion and for the economized expansion.

(b) The graph of $T_{12}(x)$ for $x \in [-1,1]$ is shown in Figure 3.2.1. Draw the graph of $T_{12}(x)$ for (say) $x \in [-1.1, 1.1]$.

6. Show the following generalization of Theorem 3.2.4. Assume that $|f(z)| \le M$ for $z \in \mathcal{E}_R$. Let $|\zeta| \in \mathcal{E}_\rho$, $1 < \rho < r \le R - \epsilon$. Then the Chebyshev expansion of $f(\zeta)$ satisfies the inequality

$$\left| f(\zeta) - \sum_{j=0}^{n-1} c_j T_j(\zeta) \right| \le \frac{2M(\rho/R)^n}{1 - \rho/R}.$$

Hint: Set $\omega = \zeta + \sqrt{\zeta^2 - 1}$, and show that $|T_j(\zeta)| = |\frac{1}{2}(\omega^j + \omega^{-j})| \le \rho^j$.

7. Compute a few terms of the expansions into powers of ϵ or k of each of the roots of the following equations, so that the error is $O(\epsilon^2)$ or $O(k^{-2})$ (ϵ is small and positive; k is large and positive). Note that some terms may have fractional or negative exponents. Also try to fit an expansion of the wrong form in some of these examples, and see what happens.

(a) $(1+\epsilon)z^2 - \epsilon = 0$; (b) $\epsilon z^3 - z^2 + 1 = 0$; (c) $\epsilon z^3 - z + 1 = 0$;
(d) $z^4 - (k^2+1)z^2 - k^2 = 0$, $(k^2 \gg 1)$.

8. Modify Clenshaw's algorithm to a formula for the derivative of an orthogonal expansion.

9. (a) Let α_j, $j = 1 : n$ be the zeros of the Chebyshev polynomial $T_n(x)$, $n \ge 1$. (There are, of course, simple trigonometric expressions for them.) Apply Clenshaw's algorithm to compute $\sum_{m=0}^{n-1} T_m(\alpha_1)T_m(x)$, for $x = \alpha_j$, $j = 1 : n$. It turns out that the results are remarkably simple. (An explanation to this will be found in Sec. 4.5.

(b) Show that $S = \sum_{k=0}^{n-1} c_k\phi_k$ can be computed by a forward version of Clenshaw's algorithm that reads

```
$y_{-2} = 0;$   $y_{-1} = 0;$
for $k = 0 : n-1,$
    $y_k = (-y_{k-2} + \alpha_k y_{k-1} + c_k)/\gamma_{k+1};$
end
$S = c_n\phi_n + \gamma_n y_{n-1}\phi_{n-1} - y_{n-2}\phi_n.$
```

Add this version as an option to your program, and study Numerical Recipes [36, Sec. 5.4], from which this formula is quoted (with adaptation to our notation etc.). Make some test example of your own choice.

10. The solution of the boundary value problem

$$(1+\epsilon)y'' - \epsilon y = 0, \quad y(0) = 0, \quad y(1) = 1,$$

has an expansion of the form $y(t;\epsilon) = y_0(t) + y_1(t)\epsilon + y_2(t)\epsilon^2 + \dots$.

(a) By coefficient matching, set up differential equations and boundary conditions for y_0, y_1, y_2, and solve them. You naturally use the boundary conditions of the original problem for y_0. Make sure you use the right boundary conditions for y_1, y_2.

(b) Set $R(t) = y_0(t) + \epsilon y_1(t) - y(t;\epsilon)$. Show that $R(t)$ satisfies the (modified) differential equation

$$(1+\epsilon)R'' - \epsilon R = \epsilon^2(7t - t^3)/6, \quad R(0) = 0, \quad R(1) = 0.$$

11. (a) Apply Kummer's first identity (3.2.39) to the error function $\operatorname{erf}(x)$, to show that

$$\operatorname{erf}(x) = \frac{2x}{\sqrt{\pi}} e^{-x^2} M\Big(1, \frac{3}{2}, x^2\Big) = \frac{2x}{\sqrt{\pi}} e^{-x^2}\Big(1 + \frac{2x^2}{3} + \frac{(2x^2)^2}{3\cdot 5} + \frac{(2x^2)^3}{3\cdot 5\cdot 7} + \dots\Big).$$

Why is this series well conditioned? (Note that it is a bell sum; compare Figure 3.2.7.) Investigate the largest term, rounding errors, truncation errors and termination criterion etc. in the same way as in (a).

(b) $\operatorname{erfc}(x)$ has a semi-convergent expansion for $x \gg 1$ that begins

$$\operatorname{erfc}(x) = 1 - \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2}\,dt = \frac{e^{-x^2}}{x\sqrt{\pi}}\Big(1 - \frac{1}{2x^2} + \frac{3}{4x^4} - \frac{15}{8x^6} + \dots\Big).$$

Give an explicit expression for the coefficients, and show that the series diverges for every x. Where is the smallest term? Estimate its size.

Hint: Set $t^2 = x^2 + u$, and proceed analogously to Example 3.2.8. See Problem 3.1.7 (c), $\alpha = \frac{1}{2}$, about the remainder term. Alternatively, apply repeated integration by parts. It may be easier to find the remainder in this way.

12. Other notations for series, with application to Bessel functions.

(a) Set

$$f(x) = \sum_{n=0}^\infty \frac{a_n x^n}{n!}; \qquad g(x) = \sum_{n=0}^\infty \frac{b_n x^n}{n!}; \qquad h(x) = \sum_{n=0}^\infty \frac{c_n x^n}{n!};$$

$$\phi(w) = \sum_{n=0}^\infty \frac{\alpha_n w^n}{n!n!}; \qquad \psi(w) = \sum_{n=0}^\infty \frac{\beta_n w^n}{n!n!}; \qquad \chi(w) = \sum_{n=0}^\infty \frac{\gamma_n w^n}{n!n!}.$$

Let $h(x) = f(x)\cdot g(x)$; $\chi(w) = \phi(w)\cdot\psi(w)$. Show that

$$c_n = \sum_{j=0}^n \binom{n}{j} a_j b_{n-j}; \qquad \gamma_n = \sum_{j=0}^n \binom{n}{j}^2 \alpha_j \beta_{n-j}.$$

Derive analogous formulas for series of the form $\sum_{n=0}^\infty a_n w^n/(2n)!$ etc.. Suggest how to divide two power series in these notations.

(b) Let $a_j = (-1)^j a'_j$; $g(x) = e^x$. Show that

$$c_n = \sum_{j=0}^n \binom{n}{j} (-1)^j a'_j.$$

Comment: By (3.2.1), this can can also be written $c_n = (-1)^n \Delta^n a_0$. This proves the mathematical equivalence of the preconditioners (3.1.55) and (3.1.59) if $P(x) = e^x$.

(c) Set, according to Example 3.2.8 and (a) (of this problem), $w = -x^2/4$,

$$J_0(x) = \sum_{n=0}^{\infty} \frac{(-1)^n w^n}{n!n!}; \quad I_0(x) = \sum_{n=0}^{\infty} \frac{w^n}{n!n!}; \quad IJ(x) \equiv I_0(x)J_0(x) = \sum_{n=0}^{\infty} \frac{\gamma_n w^n}{n!n!}.$$

Show that

$$\gamma_n = \sum_{j=0}^{n} (-1)^j \binom{n}{j} \binom{n}{n-j} = \begin{cases} (-1)^m \binom{2m}{m}, & \text{if } n = 2m; \\ 0, & \text{if } n = 2m+1. \end{cases}$$

Hint: The first expression for γ_n follows from (a). It can be interpreted as the coefficient of t^n in the product $(1-t)^n(1+t)^n$. The second expression for γ_n is the same coefficient in $(1-t^2)^n$.

(d) The second expression for γ_n in (c) is used in Example 3.2.8.[30] Reconstruct and extend the results of that example. Design a termination criterion. Where is the largest modulus of a term of the preconditioned series, and how large is it approximately? Make a crude guess in advance of the rounding error in the preconditioned series.

*(e) Show that the power series of $J_0(x)$ can be written in the form

$$\sum_{n=0}^{\infty} a_n \frac{(-x^2)^n}{(2n)!},$$

where a_n is positive and decreases slowly and smoothly.

Hint: Compute a_{n+1}/a_n.

Try preconditioning with $P(x) = \cosh x$. At the time of writing the authors do not know whether this is useful without multiple precision or not.

*(f) It is known; see Lebedev [30, (9.13.11)], that

$$J_0(x) = e^{-ix} M\left(\tfrac{1}{2}, 1; 2ix\right),$$

where $M(a, b, c)$ is Kummer's confluent hypergeometric function, this time with an imaginary argument. Show that Kummer's first identity is unfortunately of no use here for preconditioning the power series.

Comment: Most of the formulas and procedures in this problem can be generalized to the series for the Bessel functions of the first kind of general integer order, $(z/2)^{-n}J_n(x)$. These belong to the most studied functions of Applied Mathematics, and there exist more efficient methods for computing them; see, e.g., Numerical Recipes [36, Chapter 6]. This problem shows, however, that *preconditioning can work well* for a non-trivial power series, and it is worth to be tried, e.g., for other power series that may occur in connection with new applications.

[30] It is much better conditioned than the first expression. This may be one reason why multiple precision is not needed here.

13. (a) Derive the expansion of Example 3.2.5 by repeated integration by parts.

(b) Derive the Maclaurin expansion with the remainder according to (3.1.5) by the application of repeated integration by parts to the equation

$$f(z) - f(0) = z \int_0^1 f'(zt)\, d(t-1).$$

3.3 Difference Operators and Operator Expansions

3.3.1 Properties of Difference Operators

Difference operators are handy tools for the derivation, analysis, and practical application of numerical methods for many problems for interpolation, differentiation, and quadrature of a function in terms of its values at equidistant arguments. The simplest notations for difference operators and applications to derivatives, were mentioned in Sec. 1.2.3.

Let y denote a sequence $\{y_n\}$. Then we define the **shift operator** E (or translation operator) and the **forward difference operator** Δ by the relations

$$Ey = \{y_{n+1}\}, \qquad \Delta y = \{y_{n+1} - y_n\},$$

(see Sec. 1.2). E and Δ are thus operators which map one sequence to another sequence. Note, however, that if y_n is defined for $a \le n \le b$ only, then Ey_b is not defined, and the sequence Ey has fewer elements than the sequence y. (It is therefore sometimes easier to extend the sequences to infinite sequences, e.g., by adding zeros in both directions outside the original range of definition.)

These operators are **linear**, i.e. if α, β are real or complex constants and if y, z are two sequences, then $E(\alpha y + \beta z) = \alpha Ey + \beta Ez$, and similarly for Δ.

Powers of E and Δ are defined recursively, i.e.

$$E^k y = E(E^{k-1}y), \qquad \Delta^k y = \Delta(\Delta^{k-1}y).$$

By induction, the first relation yields $E^k y = \{y_{n+k}\}$. We extend the validity of this relation to $k = 0$ by setting $E^0 y = y$ and to negative values of k. $\Delta^k y$ is called the kth difference of the sequence y. We make the convention that $\Delta^0 = 1$. There will be little use of Δ^k for negative values of k in this book, although Δ^{-1} can be interpreted as a summation operator.

Note that $\Delta y = Ey - y$, and $Ey = y + \Delta y$ for any sequence y. It is therefore convenient to express these as equations between operators:

$$\Delta = E - 1, \qquad E = 1 + \Delta.$$

The identity operator is in this context traditionally denoted by 1. It can be shown that all formulas derived from the axioms of commutative algebra can be used for these operators, for example, the binomial theorem for positive integral k.

$$\Delta^k = (E-1)^k = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} E^j, \qquad E^k = (1+\Delta)^k = \sum_{j=0}^{k} \binom{k}{j} \Delta^j, \quad (3.3.1)$$

giving

$$(\Delta^k y)_n = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} y_{n+j}, \qquad y_{n+k} = (E^k y)_n = \sum_{j=0}^{k} \binom{k}{j} (\Delta^j y)_n. \tag{3.3.2}$$

We abbreviate the notation further and write, for example, $Ey_n = y_{n+1}$ instead of $(Ey)_n = y_{n+1}$, and $\Delta^k y_n$ instead of $(\Delta^k y)_n$. However, it is important to remember that Δ *operates on sequences* and not on elements of sequences. Thus, strictly speaking, this abbreviation is incorrect, though convenient. The formula for E^k will, in next subsection, be extended to an infinite series for non-integral values of k, but that is beyond the scope of algebra.

A **difference scheme** consists of a sequence and its difference sequences, arranged in the following way:

$$\begin{array}{ccccc}
y_0 & & & & \\
 & \Delta y_0 & & & \\
y_1 & & \Delta^2 y_0 & & \\
 & \Delta y_1 & & \Delta^3 y_0 & \\
y_2 & & \Delta^2 y_1 & & \Delta^4 y_0 \\
 & \Delta y_2 & & \Delta^3 y_1 & \\
y_3 & & \Delta^2 y_2 & & \\
 & \Delta y_3 & & & \\
y_4 & & & &
\end{array}$$

A difference scheme is best computed by successive subtractions; the formulas in (3.3.1) are used mostly in theoretical contexts.

In many applications the quantities y_n are computed in increasing order $n = 0, 1, 2, \ldots$, and it is natural that a difference scheme is constructed by means of the quantities previously computed. One therefore introduces the **backward difference operator** $\nabla y_n = y_n - y_{n-1} = (1 - E^{-1}) y_n$. For this operator we have

$$\nabla^k = (1 - E^{-1})^k, \qquad E^{-k} = (1 - \nabla)^k. \tag{3.3.3}$$

Note the **reciprocity** in the relations between ∇ and E^{-1}.

Any linear combination of the elements y_n, $y_{n-1}, \ldots$ y_{n-k} can also be expressed as a linear combination of y_n, $\nabla y_n, \ldots, \nabla^k y_n$, and vice versa[31] . For example,

$$y_n + y_{n-1} + y_{n-2} = 3y_n - 3\nabla y_n + \nabla^2 y_n,$$

because $1 + E^{-1} + E^{-2} = 1 + (1 - \nabla) + (1 - \nabla)^2 = 3 - 3\nabla + \nabla^2$. By the reciprocity, we also obtain $y_n + \nabla y_n + \nabla^2 y_n = 3y_n - 3y_{n-1} + y_{n-2}$.

[31] An analogous statement holds for the elements y_n, $y_{n+1}, \ldots, y_{n+k}$ and forward differences.

In this notation the difference scheme reads

$$\begin{array}{ccccc}
y_0 & & & & \\
 & \nabla y_1 & & & \\
y_1 & & \nabla^2 y_2 & & \\
 & \nabla y_2 & & \nabla^3 y_3 & \\
y_2 & & \nabla^2 y_3 & & \nabla^4 y_4 \\
 & \nabla y_3 & & \nabla^3 y_4 & \\
y_3 & & \nabla^2 y_4 & & \\
 & \nabla y_4 & & & \\
y_4 & & & &
\end{array}$$

In the backward difference scheme the subscripts are constant along diagonals directed upwards (backwards) to the right, while, in the forward difference scheme, subscripts are constant along diagonals directed downwards (forwards). Note, e.g., that $\nabla^k y_n = \Delta^k y_{n-k}$. In a computer, a backward difference scheme is preferably stored as a lower triangular matrix.

Example 3.3.1.
Part of the difference scheme for the sequence $y = \{\dots, 0, 0, 0, 1, 0, 0, 0, \dots\}$ is given below.

$$\begin{array}{ccccccccc}
 & & & 0 & & 1 & & -7 & \\
 & & 0 & & 1 & & -6 & & 28 \\
 & 0 & & 1 & & -5 & & 21 & \\
0 & & 1 & & -4 & & 15 & & -56 \\
 & 1 & & -3 & & 10 & & -35 & \\
1 & & -2 & & 6 & & -20 & & 70 \\
 & -1 & & 3 & & -10 & & 35 & \\
0 & & 1 & & -4 & & 15 & & -56 \\
 & 0 & & -1 & & 5 & & -21 & \\
 & & 0 & & 1 & & -6 & & 28 \\
 & & & 0 & & -1 & & 7 &
\end{array}$$

This example shows the *effect of a disturbance in one element* on the sequence of the higher differences. Because the effect broadens out and grows quickly, difference schemes are useful in the investigation and correction of computational and other errors, so-called **difference checks**. Notice that, since the differences are *linear* functions of the sequence, a **superposition principle** holds. The effect of errors can thus be estimated by studying simple sequences such as the one above.

Example 3.3.2.
The following is a difference scheme for a 5 decimal table of the function $f(x) = \tan x$, $x \in [1.30, 1.36]$, with step $h = 0.01$. The differences are given with

10^{-5} as unit.

x	y	∇y	$\nabla^2 y$	$\nabla^3 y$	$\nabla^4 y$	$\nabla^5 y$	$\nabla^6 y$
1.30	3.60210						
		14498					
1.31	3.74708		1129				
		15627		140			
1.32	3.90335		1269		26		
		16896		166		2	
1.33	4.07231		1435		28		9
		18331		194		11	
1.34	4.25562		1629		39		
		19960		233			
1.35	4.45522		1862				
		21822					
1.36	4.67344						

We see that the differences decrease roughly by a factor of 0.1—that indicates that the step size has been chosen suitably for the purpose of interpolation, numerical quadrature etc.—until the last two columns, where the rounding errors of the function values have a visible effect.

Example 3.3.3.

For the sequence $y_n = (-1)^n$ one finds easily that

$$\nabla y_n = 2y_n, \quad \nabla^2 y_n = 4y_n, \ldots, \quad \nabla^k y_n = 2^k y_n.$$

If the error in the elements of the sequence are bounded by ϵ, it follows that the errors of the kth differences are bounded by $2^k\epsilon$. A rather small reduction of this bound is obtained if the errors are assumed to be independent random variables (Problem 3.4.25).

It is natural also to consider *difference operations on functions* not just on sequences. E and Δ map the function f onto functions whose values at the point x are

$$E\,f(x) = f(x+h), \qquad \Delta f(x) = f(x+h) - f(x), \tag{3.3.4}$$

where h is the *step size*. Of course, Δf depends on h; in some cases this should be indicated in the notation. One can, for example, write $\Delta_h f(x)$, or $\Delta f(x;h)$. If we set $y_n = f(x_0 + nh)$, the difference scheme of the function with step size h is the same as for the sequence $\{y_n\}$. Again it is important to realize that, in this case, the operators act on *functions*, not on the values of functions. It would be more correct to write $f(x_0+h) = (Ef)(x_0)$. Actually, the notation $(x_0)Ef$ would be even more logical, since the insertion of the value of the argument x_0 is the last operation to be done, and the convention for the order of execution of operators proceeds from right to left, but this notation would be too revolutionary.[32]

[32] The notation $[x_0]f$ occurs, however, naturally in connection with divided differences, Sec. 4.2.

Note that *no new errors are introduced during the computation of the differences, but the effects of the original irregular errors of y grow exponentially.* We emphasize the word **irregular errors**, e.g., rounding errors in y, since systematic errors, e.g., the truncation errors in the numerical solution of a differential equation, often have a smooth difference scheme. For example, if the values of y have been produced by the iterative solution of an equation, where x is a parameter, with the same number of iterations for every x and y and the same algorithm for the first approximation, then the truncation error of y is likely to be a smooth function of x.

Difference operators are in many respects similar to differentiation operators. Let f be a polynomial. By Taylor's formula,

$$\Delta f(x) = f(x+h) - f(x) = hf'(x) + \frac{1}{2}h^2 f''(x) + \dots .$$

We see from this that $deg\,\Delta f = deg\, f - 1$. Similarly for differences of higher order; *if f is a polynomial of degree less than k, then*

$$\Delta^{k-1} f(x) = constant, \quad \Delta^p f(x) = 0, \ \forall p \geq k.$$

The same holds for backward differences.

The following important result can be derived directly from Taylor's theorem with the integral form of the remainder. Assume that all derivatives of f up to kth order are continuous. If $f \in C^k$,

$$\Delta^k f(x) = h^k f^{(k)}(\zeta), \quad \zeta \in [x, x+kh]. \tag{3.3.5}$$

Hence $h^{-k}\Delta^k f(x)$ is an approximation to $f^{(k)}(x)$; the error of this approximation approaches zero as $h \to 0$ (i.e. as $\zeta \to x$). As a rule, the error is approximately proportional to h. We postpone the proof to Sec. 4.2.1, where it appears as a particular case of a theorem concerning divided differences.

Even though difference schemes do not have the same importance today that they had in the days of hand calculations or calculation with desk calculators, they are still important conceptually, and we shall also see how they are still useful also in practical computing. In a computer it is more natural to store a difference scheme as an array, e.g. with y_n, ∇y_n, $\nabla^2 y_n, \dots$, $\nabla^k y_n$ in a row (instead of along a diagonal).

Many formulas for differences are analogous to formulas for derivatives, though usually more complicated. The following results are among the most important.

Lemma 3.3.1.

It holds that

$$\Delta^k(a^x) = (a^h - 1)^k a^x, \qquad \nabla^k(a^x) = (1 - a^{-h})^k a^x. \tag{3.3.6}$$

For sequences, i.e. if h=1,

$$\Delta^k\{a^n\} = (a-1)^k\{a^n\}, \qquad \Delta^k\{2^n\} = \{2^n\}. \tag{3.3.7}$$

Proof. Let c be a given constant. For $k = 1$ we have

$$\Delta(ca^x) = ca^{x+h} - ca^x = ca^x a^h - ca^x = c(a^h - 1)a^x$$

The general result follows easily by induction. The backward difference formula is derived in the same way. □

Lemma 3.3.2. *Summation by Parts*

$$\sum_{n=0}^{N-1} u_n \Delta v_n = u_N v_N - u_0 v_0 - \sum_{n=0}^{N-1} \Delta u_n \, v_{n+1}. \tag{3.3.8}$$

Proof. (Compare the rule for integration by parts and its proof!) Notice that

$$\sum_{n=0}^{N-1} \Delta w_n = (w_1 - w_0) + (w_2 - w_1) + \ldots + (w_N - w_{N-1}) = w_N - w_0.$$

Use this on $w_n = u_n v_n$. From the result in Lemma 3.3.1 one gets after summation,

$$u_N v_N - u_0 v_0 = \sum_{n=0}^{N-1} u_n \Delta v_n + \sum_{n=0}^{N-1} \Delta u_n v_{n+1},$$

and the result follows. (For an extension; see Problem 1d.) □

Lemma 3.3.3. *The Difference of a Product*

$$\Delta(u_n v_n) = u_n \Delta v_n + \Delta u_n \, v_{n+1}. \tag{3.3.9}$$

Proof. We have

$$\Delta(u_n v_n) = u_{n+1} v_{n+1} - u_n v_n = u_n(v_{n+1} - v_n) + (u_{n+1} - u_n)v_{n+1}.$$

Compare the above result with the formula for differentials, $d(uv) = udv + vdu$. Note that we have v_{n+1} (not v_n) on the right-hand side. □

3.3.2 The Calculus of Operators

Formal calculations with operators, using the rules of algebra and analysis, are often an elegant means of assistance in *finding approximation formulas that are exact for all polynomials of degree less than (say) k*, and they should therefore be useful for functions that can be accurately approximated by such a polynomial.

Our calculations often lead to divergent (or semi-convergent) series, but the way we handle them can usually be justified by means of the theory of formal power series, of which a brief introduction was given at the end of Sec. 3.1.5. The operator calculations also provide error estimates, asymptotically valid as the step size $h \to 0$. Strict error bounds can be derived by means of Peano's remainder theorem,Sec. 3.3.3.

Operator techniques are sometimes successfully used (see, e.g., Sec. 3.3.4) in a way that it is hard, or even impossible, to justify by means of formal power series. It is then not trivial to formulate appropriate conditions for the success and to derive satisfactory error bounds and error estimates, but it can sometimes be done.

We make a digression about terminology. More generally, *the word* **operator** *is in this book used for a function that maps a linear space $\mathcal{S}$ into another linear space $\mathcal{S}'$.* $\mathcal{S}$ can, for example, be a space of functions, a coordinate space, or a space of sequences. The dimension of these spaces can be finite or infinite. For example, the differential operator D maps the infinite-dimensional space $C^1[a,b]$ of functions with a continuous derivative, defined on the interval $[a,b]$, into the space $C[a,b]$ of continuous functions on the same interval.

In the following we denote by $\mathcal{P}_k$ the set of polynomials of degree *less than* k.[33] Note that $\mathcal{P}_k$ is a k-dimensional linear space, for which $\{1, x, x^2, \ldots, x^{k-1}\}$ is a basis called the *power basis*; the coefficients $(c_1, c_2, \ldots, c_k)$ are then the *coordinates* of the polynomial p defined by $p(x) = \sum_{i=1}^{k} c_i x^{i-1}$.

For simplicity, we shall assume that the space of functions on which the operators are defined is $C^\infty(-\infty,\infty)$, i.e. the functions are infinitely differentiable on $(-\infty,\infty)$. This sometimes requires (theoretically) a modification of a function outside the bounded interval where it is interesting. There are techniques for achieving this, but they are beyond the scope of this book. Just imagine that they have been applied.

We define the following operators:

$Ef(x) = f(x+h)$	Shift (or translation) operator
$\Delta f(x) = f(x+h) - f(x)$	Forward difference operator
$\nabla f(x) = f(x) - f(x-h)$	Backward difference operator
$Df(x) = f'(x)$	Differentiation operator
$\delta f(x) = f(x+\frac{1}{2}h) - f(x-\frac{1}{2}h)$	Central difference operator
$\mu f(x) = \frac{1}{2}\big(f(x+\frac{1}{2}h) + f(x-\frac{1}{2}h)\big)$	Averaging operator

Suppose that the values of f are given on an equidistant grid only, e.g., $x_j = x_0 + jh$, $j = -M : N$, (j is integer). Set $f_j = f(x_j)$. Note that δf_j, $\delta^3 f_j \ldots$, (odd powers) and μf_j *cannot* be exactly computed; they are available halfway between the grid points. (A way to get around this is given later; see (3.3.47)) The even powers $\delta^2 f_j$, $\delta^4 f_j \ldots$, and $\mu\delta f_j$, $\mu\delta^3 f_j \ldots$, *can* be exactly computed. This follows from the formulas

$$\mu\delta f(x) = \frac{1}{2}\big(f(x+h) - f(x-h)\big), \quad \mu\delta = \tfrac{1}{2}(\Delta + \nabla), \quad \delta^2 = \Delta - \nabla. \qquad (3.3.10)$$

[33] Some authors use similar notations to denote the set of polynomials of degree less than or equal to k.

Several other notations are in use, e.g., at the study of difference methods for partial differential equations D_{+h}, D_{0h}, D_{-h} are used instead of $\Delta, \mu\delta, \nabla$, respectively.

An operator P is said to be a **linear operator** if

$$P(\alpha f + \beta g) = \alpha P f + \beta P g$$

holds for arbitrary complex constants α, β and arbitrary functions f, g. The above six operators are all linear. The operation of multiplying by a constant α, is also a linear operator.

If P and Q are two operators, then their sum, product, etc., can be defined in the following way:

$$\begin{aligned} (P+Q)f &= Pf + Qf, \\ (P-Q)f &= Pf - Qf, \\ (PQ)f &= P(Qf), \\ (\alpha P)f &= \alpha(Pf), \\ P^n f &= P \cdot P \cdots Pf, \quad n \text{ factors.} \end{aligned}$$

Two operators are equal, $P = Q$ if $Pf = Qf$, for all f in the space of functions considered. Notice that $\Delta = E - 1$. One can show that the following rules hold for all linear operators:

$$\begin{aligned} P + Q &= Q + P, \qquad & P + (Q + R) &= (P + Q) + R, \\ P(Q + R) &= PQ + PR, \qquad & P(QR) &= (PQ)R. \end{aligned}$$

The above six operators, E, Δ, ∇, hD, δ, and μ, and the combinations of them by these algebraic operations make a *commutative ring*, so $PQ = QP$ holds for these operators, and any algebraic identity that is generally valid in such rings can be used.

If $\mathcal{S} = \mathbf{R^n}$, $\mathcal{S}' = \mathbf{R^m}$, and the elements are *column* vectors, then the linear operators are matrices of size $[m, n]$. They do generally not commute.

If $\mathcal{S}' = \mathbf{R}$ or $\mathbf{C}$, the operator is called a **functional**. Examples of functionals are, if x_0 denotes a fixed (though arbitrary) point,

$$Lf = f(x_0), \quad Lf = f'(x_0), \quad Lf = \int_0^1 e^{-x} f(x)dx, \quad \int_0^1 |f(x)|^2 dx;$$

all except the last one are **linear functionals**.

There is a subtle distinction here. For example, E is a linear operator that maps a function to a function. Ef is the function whose value at the point x is $f(x+h)$. If we consider a fixed point, e.g. x_0, then $(Ef)(x_0)$ is a scalar. This is therefore a *linear functional.* We shall allow ourselves to simplify the notation and to write $Ef(x_0)$, but it must be understood that E operates on the function f, not on the function value $f(x_0)$. This was just one example; simplifications like this will be made with other operators than E, and similar simplifications in notation were suggested earlier in this chapter. There are, however, situations, where it is, for the sake of clarity, advisable to return to the more specific notation with a larger number of parentheses.

If we represent the vectors in $\mathbf{R}^n$ by *columns* y, the linear functionals in $\mathbf{R}^n$ are the scalar products $a^T x = \sum_{i=1}^n a_i y_i$; every *row* a^T thus defines a linear functional.

Examples of linear functionals in $\mathcal{P}_k$ are linear combinations of a finite number of function values, $Lf = \sum a_j f(x_j)$. If $x_j = x_0 + jh$ the same functional can be expressed in terms of differences, e.g., $\sum a'_j \Delta^j f(x_0)$; see Problem 3. The main topic of this section is to show how operator methods can be used for finding approximations of this form to linear functionals in more general function spaces. First, we need a general theorem.

Theorem 3.3.4.

Let $x_1, x_2, \ldots, x_k$ be k distinct real (or complex) numbers. Then no non-trivial relation of the form

$$\sum_{j=1}^k a_j f(x_j) = 0 \tag{3.3.11}$$

can hold for all $f \in \mathcal{P}_k$. If we add one more point (x_0), there exists only one non-trivial relation of the form $\sum_{j=0}^k a'_j f(x_j) = 0$, (except that it can be multiplied by an arbitrary constant). In the equidistant case, i.e. if $x_j = x_0 + jh$, then

$$\sum_{j=0}^k a'_j f(x_j) \equiv c\Delta^k f(x_0), \quad c \neq 0.$$

Proof. If (3.3.11) were valid for all $f \in \mathcal{P}_k$, then the linear system $\sum_{j=1}^k x_j^{i-1} a_j = 0$, $i = 1 : k$,would have a non-trivial solution $(a_1, a_2, \ldots, a_k)$. The matrix of the system, however, is a so called **Vandermonde matrix**

$$V = [x_j^{i-1}]_{i,j=1}^k = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_k \\ \vdots & \vdots & \cdots & \vdots \\ x_1^{k-1} & x_2^{k-1} & \cdots & x_k^{k-1} \end{pmatrix}. \tag{3.3.12}$$

Its determinant is known to equal the product of all differences $(x_i - x_j)$, $i > j$, $1 < i \leq k$, which is nonzero.

Now we add the point x_0. Suppose that there exist two relations,

$$\sum_{j=0}^k b_j f(x_j) = 0, \qquad \sum_{j=0}^k c_j f(x_j) = 0.$$

with linearly independent coefficient vectors. Then we can find a (non-trivial) linear combination, where x_0 has been eliminated, but this contradicts the result that we have just proved. Hence the hypothesis is wrong; the two coefficient vectors must be proportional. We have seen above that, in the equidistant case, $\Delta^k f(x_0) = 0$ is such a relation. More generally, we shall see in Chapter 4 that, for $k+1$ arbitrary distinct points, the kth order *divided difference* is zero for all $f \in \mathcal{P}_k$. □

Corollary 3.3.5.

Suppose that a formula for interpolation, numerical differentiation or integration etc. has been derived, for example by an operator technique. If it is a linear combination of the values of $f(x)$ at k given *distinct points x_j, $j = 1 : k$, and is exact for all $f \in \mathcal{P}_k$, this formula is unique. (If it is exact for all $f \in \mathcal{P}_m$, $m < k$, only, it is not unique.)*

In particular, for any $\{c_j\}_{j=1}^{k}$, a unique polynomial $P \in \mathcal{P}_k$ is determined by the interpolation conditions $P(x_j) = c_j$, $j = 1 : k$.

Proof. The difference between two formulas that use the same function values would lead to a relation that is impossible, by the theorem. □

Now we shall go outside of polynomial algebra and consider also *infinite series of operators*. The Taylor series

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + \frac{h^3}{3!}f'''(x) + \dots$$

can be written symbolically as

$$Ef = \Big(1 + hD + \frac{(hD)^2}{2!} + \frac{(hD)^3}{3!} + \dots\Big)f.$$

We can here treat hD like an algebraic indeterminate, and consider the series inside the parenthesis (without the operand) as a *formal power series*[34]

For a formal power series the concepts of convergence and divergence do not exist. When the operator series acts on a function f, and is evaluated at a point c, we obtain an ordinary numerical series, related to the linear functional $Ef(c) = f(c+h)$. We know that this Taylor series may converge or diverge, depending on f, c, and h.

Roughly speaking, the last part of Sec. 3.1.5 tells that, with some care, "analytic functions" of one indeterminate can be handled with the same rules as analytic functions of one complex variable.

Theorem 3.3.6.

$$\begin{aligned} e^{hD} &= E = 1 + \Delta, \qquad e^{-hD} = E^{-1} = 1 - \nabla, \\ 2\sinh \tfrac{1}{2}hD &= e^{hD/2} - e^{-hD/2} = \delta, \\ (1+\Delta)^{\theta} &= (e^{hD})^{\theta} = e^{\theta hD}, \quad (\theta \in \mathbf{R}). \end{aligned}$$

Proof. The first formula follows from the previous discussion. The second and the third formulas are obtained in a similar way. (Recall the definition of δ.) The last

[34] We now abandon the bold-type notation for indeterminates and formal power series used in Sec. 3.1.5 for the function e^{hD}, which is defined by this series. The reader is advised to take a look again at the last part of Sec. 3.1.5.

formula follows from the first formula together with Lemma 3.1.9 (in Sec. 3.1.3).
□ It follows from the power series expansion that

$$(e^{hD})^\theta f(x) = e^{\theta hD} f(x) = f(x + \theta h),$$

when it converges. Since $E = e^{hD}$ it is natural to *define*

$$E^\theta f(x) = f(x + \theta h),$$

and we extend this definition also to such values of θ that the power series for $e^{\theta hD} f(x)$ is divergent. Note that, e.g., the formula $E^{\theta_2} E^{\theta_1} f(x) = E^{\theta_2+\theta_1} f(x)$, follows from this definition.

When one works with operators or functionals it is advisable to avoid notations like Δx^n, $De^{\alpha x}$, where the variables appear in the operands. For two important functions we therefore set

$$F_\alpha : F_\alpha(x) = e^{\alpha x}; \qquad f_n : f_n(x) = x^n. \tag{3.3.13}$$

Let P be any of the operators mentioned above. When applied to F_α it acts like a scalar that we shall call **the scalar of the operator** [35] and denote it by $\mathrm{sc}(P)$,

$$PF_\alpha = \mathrm{sc}(P) F_\alpha.$$

We may also write $\mathrm{sc}(P; h\alpha)$ if it is desirable to emphasize its dependence on $h\alpha$. (We normalize the operators so that this is true, e.g., we work with hD instead of D.) Note that

$$\mathrm{sc}(\beta P + \gamma Q) = \beta \mathrm{sc}(P) + \gamma \mathrm{sc}(Q),\ (\beta, \gamma \in \mathbf{C}), \quad \mathrm{sc}(PQ) = \mathrm{sc}(P)\mathrm{sc}(Q),$$

For our most common operators we obtain

$$(E^\theta) = e^{\theta h\alpha}; \qquad \mathrm{sc}(\nabla) = \mathrm{sc}(1 - E^{-1}) = 1 - e^{-h\alpha}; \tag{3.3.14}$$

$$\mathrm{sc}(\Delta) = \mathrm{sc}(E - 1) = e^{h\alpha} - 1; \tag{3.3.15}$$

$$\mathrm{sc}(\delta) = \mathrm{sc}(E^{1/2} - E^{-1/2}) = e^{h\alpha/2} - e^{-h\alpha/2}.$$

Let Q_h be one of the operators hD, Δ, δ, ∇. It follows from the last formulas that

$$\mathrm{sc}(Q_h) \sim h\alpha,\ (h \to 0); \quad |\mathrm{sc}(Q_h)| \le |h\alpha| e^{|h\alpha|}$$

The main reason for grouping these operators together is that each of them has the important property (3.3.5), i.e. $Q_h^k f(c) = h^k f^{(k)}(\zeta)$, where ζ lies in the smallest interval that contains all the arguments used in the computation of $Q_h^k f(c)$. Hence,

$$f \in \mathcal{P}_k \quad \Rightarrow \quad Q_h^n f = 0, \quad \forall n \ge k. \tag{3.3.16}$$

This property [36] makes each of these four operators well suited to be the indeterminate in a formal power series that, hopefully, will be able to generate a sequence of

[35] In applied Fourier analysis this scalar is, for $\alpha = i\omega$, often called the *symbol of the operator*.

[36] The operators E and μ do *not* possess this property.

approximations, L_1, L_2, $L_3\ldots$, to a given linear operator L. L_n is the nth partial sum of a formal power series for L. Then

$$f \in \mathcal{P}_k \quad \Rightarrow \quad L_n f = L_k f, \quad \forall n \geq k. \tag{3.3.17}$$

We shall see in the next theorem that, for expansion into powers of Q_h,

$$\lim_{n\to\infty} L_n f(x) = Lf(x)$$

if f is a polynomial. This is not quite self-evident, because it is not true for all functions f, and we have seen in Sec. 3.1.5 Sec. 3.1.3 that it can happen that an expansion converges to a "wrong result". We shall see more examples of that later. Convergence does not necessarily imply validity.

Suppose that z is a complex variable, and that $\phi(z)$ is analytic at the origin, i.e. $\phi(z)$ is equal to its Maclaurin series, (say)

$$\phi(z) = a_0 + a_1 z + a_2 z^2 + \ldots,$$

if $|z| < \rho$ for some $\rho > 0$. For multivalued functions we always refer to the principal branch. The operator function $\phi(Q_h)$ is usually defined by the *formal* power series,

$$\phi(Q_h) = a_0 + a_1 Q_h + a_2 Q_h^2 + \ldots,$$

where Q_h is treated like an algebraic indeterminate.

Table 3.3.1. *Bickley's table of relations between difference operators*

	E	Δ	δ	∇	hD
E	E	$1+\Delta$	$1+\frac{1}{2}\delta^2+\delta\sqrt{1+\frac{1}{4}\delta^2}$	$\dfrac{1}{1-\nabla}$	e^{hD}
Δ	$E-1$	Δ	$\delta\sqrt{1+\frac{1}{4}\delta^2}+\frac{1}{2}\delta^2$	$\dfrac{\nabla}{1-\nabla}$	$e^{hD}-1$
δ	$E^{1/2}-E^{-1/2}$	$\Delta(1+\Delta)^{-1/2}$	δ	$\nabla(1-\nabla)^{-1/2}$	$2\sinh\frac{1}{2}hD$
∇	$1-E^{-1}$	$\dfrac{\Delta}{1+\Delta}$	$\delta\sqrt{1+\frac{1}{4}\delta^2}-\frac{1}{2}\delta^2$	∇	$1-e^{-hD}$
hD	$\ln E$	$\ln(1+\Delta)$	$2\sinh^{-1}\frac{1}{2}\delta$	$-\ln(1-\nabla)$	hD
μ	$\frac{1}{2}(E^{1/2}+E^{-1/2})$	$\dfrac{1+\frac{1}{2}\Delta}{(1+\Delta)^{1/2}}$	$\sqrt{1+\frac{1}{4}\delta^2}$	$\dfrac{1-\frac{1}{2}\nabla}{(1-\nabla)^{1/2}}$	$\cosh\frac{1}{2}hD$

The operators E, hD, Δ, δ, ∇ and μ are related to each others. See Table 3.3.1 that is adapted from an article by the eminent blind British mathematician W. G. Bickley (1948). Some of these formulas follow almost directly from the definitions, others are derived in this section, and the rest are left for Problem 5e. We find the value sc$(\cdot)$ for each of these operators by *substituting α for D in the last column of the table.* (Why?)

Example 3.3.4. *Express E in terms of ∇.*

The definition of ∇ reads in operator form $E^{-1} = 1 - \nabla$. This can be looked upon as a formal power series (with only two non-vanishing terms) for the reciprocal of E with ∇ as the indeterminate. By the rules for formal power series mentioned in Sec. 3.1.5, we obtain *uniquely*

$$E = (E^{-1})^{-1} = (1 - \nabla)^{-1} = 1 + \nabla + \nabla^2 + \dots .$$

We find in the table an equivalent expression containing a fraction line. Suppose that we have proved the last column of the table. So, $\mathrm{sc}(\nabla) = 1 - e^{-h\alpha}$, hence

$$\mathrm{sc}((1-\nabla)^{-1}) = (e^{-h\alpha})^{-1} = e^{h\alpha} = \mathrm{sc}(E).$$

Example 3.3.5.

Suppose that we have proved the first and the last columns of Bickley's table (except for the equation $hD = \ln E$). We shall prove one of the formulas in the second column, namely the equation $\delta = \Delta(1+\Delta)^{-1/2}$. By the first column, the right hand side is equal to $(E-1)E^{-1/2} = E^{1/2} - E^{-1/2} = \delta$, Q.E.D.

We shall also compute $\mathrm{sc}(\Delta(1+\Delta)^{-1/2})$. Since $\mathrm{sc}(\Delta) = e^{h\alpha} - 1$ we obtain

$$\begin{aligned}\mathrm{sc}(\Delta(1+\Delta)^{-1/2}) &= (e^{h\alpha} - 1)(e^{h\alpha})^{-1/2} = e^{h\alpha/2} - e^{-h\alpha/2}\\ &= 2\sinh \tfrac{1}{2}h\alpha = \mathrm{sc}(\delta).\end{aligned}$$

By the aid of Bickley's table, we are in a position to transform L into the form $\phi(Q_h)R_h$. (A sum of several such expressions with different indeterminates can also be treated.)

- Q_h is the one of the four operators, hD, Δ, δ, ∇, which we have chosen to be the "indeterminate".

$$Lf \simeq \phi(Q_h)f = (a_0 + a_1 Q_h + a_2 Q_h^2 + \dots)f. \tag{3.3.18}$$

  The coefficients a_j are the same as the Maclaurin coefficients of $\phi(z)$, $z \in \mathbf{C}$ if $\phi(z)$ is analytic at the origin. They can be determined by the techniques described in Sec. 3.1.4 and Sec. 3.1.5. The meaning of the relation $\simeq$ will hopefully be clear from the following theorem.

- R_h is, e.g., $\mu\delta$ or E^k, k integer, or more generally any linear operator with the properties that $R_h F_\alpha = \mathrm{sc}(R_h)F_\alpha$, and that the values of $R_h f(x_n)$ on the grid $x_n = x_0 + nh$, n integer, are determined by the values of f on the same grid.

Theorem 3.3.7. *Recall the notation Q_h for either of the operators Δ, δ, ∇, hD, and the notations $F_\alpha(x) = e^{\alpha x}$, $f_n(x) = x^n$. Note that*

$$F_\alpha(x) = \sum_{n=0}^{\infty} \frac{\alpha^n}{n!} f_n(x). \tag{3.3.19}$$

Also recall the scalar of an operator and its properties, e.g.,

$$LF_\alpha = \mathrm{sc}(L)F_\alpha, \qquad Q_h^j F_\alpha = (\mathrm{sc}(Q_h))^j F_\alpha;$$

for the operators under consideration the scalar depends on $h\alpha$.
Assumptions:

(i) *A formal power series equation* $L = \sum_{j=0}^{\infty} a_j Q_h^j$ *has been derived.*[37] *Furthermore,* $|\mathrm{sc}(Q_h)| < \rho$, *where* ρ *is the convergence radius of the series* $\sum a_j z^j$, $z \in \mathbf{C}$, *and*

$$\mathrm{sc}(L) = \sum_{j=0}^{\infty} a_j (\mathrm{sc}(Q_h))^j. \tag{3.3.20}$$

(ii)

$$L\frac{\partial^n}{\partial\alpha^n}F_\alpha(x) = \frac{\partial^n}{\partial\alpha^n}(LF_\alpha)(x)$$

at $\alpha = 0$, *or equivalently,*

$$L\int_C \frac{F_\alpha(x)\,d\alpha}{\alpha^{n+1}} = \int_C \frac{(LF_\alpha)(x)\,d\alpha}{\alpha^{n+1}}. \tag{3.3.21}$$

where C *is any circle with the origin as center.*

(iii) *The domain of* x *is a bounded interval* I_1 *in* $\mathbf{R}$.

Then

$$LF_\alpha = \Big(\sum_{j=0}^{\infty} a_j Q_h^j\Big)F_\alpha, \quad \text{if } |\mathrm{sc}(Q_h)| < \rho, \tag{3.3.22}$$

$$Lf(x) = \sum_{j=0}^{k-1} a_j Q_h^j f(x), \quad \text{if } f \in \mathcal{P}_k, \tag{3.3.23}$$

for any positive integer k.

A **strict error bound** *for* (3.3.23), *if* $f \notin \mathcal{P}_k$, *is obtained in Peano's Theorem* 3.3.8.

An **asymptotic error estimate** *(as* $h \to 0$ *for fixed* k*) is given by the first neglected non-vanishing term* $a_r Q_h^r f(x) \sim a_r (hD)^r f(x)$, $r \geq k$, *if* $f \in C^r[I]$, *where the interval* I *must contain all the points used in the evaluation of* $Q_h^r f(x)$.

Proof. By Assumption 1,

$$LF_\alpha = \mathrm{sc}(L)F_\alpha = \lim_{J\to\infty}\sum_{j=0}^{J-1} a_j \mathrm{sc}(Q_h^j)F_\alpha = \lim_{J\to\infty}\sum_{j=0}^{J-1} a_j Q_h^j F_\alpha = \lim_{J\to\infty}\Big(\sum_{j=0}^{J-1} a_j Q_h^j\Big)F_\alpha,$$

[37] To simplify the writing, the operator R_h is temporarily neglected. See one of the comments below.

hence $LF_\alpha = (\sum_{j=0}^{\infty} Q_h^j)F_\alpha$. This proves the first part of the theorem.
By (3.3.19), Cauchy's formula (3.2.9) and Assumption 2,

$$\begin{aligned}\frac{2\pi i}{n!} Lf_n(x) &= L\int_C \frac{F_\alpha(x)\,d\alpha}{\alpha^{n+1}} = \int_C \frac{(LF_\alpha)(x)\,d\alpha}{\alpha^{n+1}} \\ &= \int_C \sum_{j=0}^{J-1} \frac{a_j Q_h^j F_\alpha(x)\,d\alpha}{\alpha^{n+1}} + \int_C \sum_{j=J}^{\infty} \frac{a_j \mathrm{sc}(Q_h)^j F_\alpha(x)\,d\alpha}{\alpha^{n+1}}.\end{aligned}$$

Let ϵ be any positive number. Choose J so that the modulus of the last term becomes $\epsilon\theta_n 2\pi/n!$, where $|\theta_n| < 1$. This is possible, since $|\mathrm{sc}(Q_h)| < \rho$; see Assumption (i). Hence, for every $x \in I_1$,

$$Lf_n(x) - \epsilon\theta_n = \frac{n!}{2\pi i}\sum_{j=0}^{J-1} a_j Q_h^j \int_C \frac{F_\alpha(x)\,d\alpha}{\alpha^{n+1}} = \sum_{j=0}^{J-1} a_j Q_h^j f_n(x) = \sum_{j=0}^{k-1} a_j Q_h^j f_n(x).$$

The last step holds if $J \ge k > n$, because, by (3.3.16), $Q_h^j f_n = 0$ for $j > n$. It follows that $\left|Lf_n(x) - \sum_{j=0}^{k-1} a_j Q_h^j f_n(x)\right| < \epsilon$ for every $\epsilon > 0$, hence $Lf_n = \sum_{j=0}^{k-1} a_j Q_h^j f_n$.

If $f \in \mathcal{P}_k$, f is a linear combination of f_n, $n = 0 : k-1$. Hence $Lf = \sum_{j=0}^{k-1} a_j Q_h^j f$ if $f \in \mathcal{P}_k$. This proves the second part of the theorem.

The error bound is derived in Sec. 3.3.1. Recall the important formula (3.3.5) that expresses the kth difference as the value of the k'th derivative in a point located in an interval that contains all the points used in the in the computation of the k'th difference. i.e. the ratio of the error estimate $a_r(hD)^r f(x)$ to the true truncation error tends to 1, as $h \to 0$. □

Remark 3.3.1. This theorem is concerned with series of powers of the four operators collectively denoted Q_h. One may try to use operator techniques also to *find* a formula involving, e.g., an infinite expansion into powers of the operator E. Then one should try afterwards to find sufficient conditions for the validity of the result. This procedure will be illustrated in connection with Euler–Maclaurin's formula in Sec. 3.4.4.

Sometimes, operator techniques which are not covered by this theorem can, after appropriate restrictions, be justified (or even replaced) by *transform methods*, e.g., z-transforms, Laplace or Fourier transforms.

The operator R_h that was introduced just before the theorem, was neglected in the proof, in order to simplify the writing. We now have to multiply the operands by R_h in the proof and in the results. This changes practically nothing for F_α, since $R_hF_\alpha = \mathrm{sc}(R_h)F_\alpha$. In (3.3.23) there is only a trivial change, because the polynomials f and R_hf may not have the same degree. For example, if $R_h = \mu\delta$ and $f \in P_k$ then $R_hf \in P_{k-1}$. The verification of the assumptions typically offers no difficulties.

It follows from the linearity of (3.3.22) that *it is satisfied also if F_α is replaced by a linear combination of exponential functions F_α with different α*, provided that $|\mathrm{sc}(Q_h)| < \rho$ for all the occurring α. With some care, one can let the linear combination be an infinite series or an integral.

There are two things to note in connection with the asymptotic error estimates. First, the step size should be small enough; this means in practice that, in the beginning, the magnitude of the differences should decrease rapidly, as their order increases. When the order of the differences becomes large, it often happens that the moduli of the differences also become increasing. This can be due to two causes: semi-convergence (see the next comment) and/or rounding errors.

The *rounding errors* of the data may have so large effects on the high order differences[38] that the error estimation does not make sense. One should then use a smaller value of the order k, where the rounding errors have a smaller influence. An advantage with the use of a difference scheme is that it is relatively easy to choose the order k adaptively, and sometimes also the step size h.

This comment is of particular importance for numerical differentiation. Numerical illustrations and further comments are given below in Example 3.3.6 and Problem 6b, and in several other places.

The sequence of approximations to Lf may converge or diverge, depending on f and h. It is also often *semiconvergent*, recall Sec. 3.2.6, but in practice the rounding errors mentioned in the previous comment, have often, though not always, taken over already, when the truncation error passes its minimum. See Problem 6b.

Example 3.3.6. *The Backwards Differentiation Formula.*

By Theorem 3.3.6, $e^{-hD} = 1 - \nabla$. We look upon this as a formal power series; the indeterminate is $Q_h = \nabla$. By Example 3.1.11,

$$L = hD = -\ln(1-\nabla) = \nabla + \frac{1}{2}\nabla^2 + \frac{1}{3}\nabla^3 + \dots \tag{3.3.24}$$

Verification of the assumptions of Theorem 3.3.7: [39]

(i) $\mathrm{sc}(\nabla) = 1 - e^{-h\alpha}$; the convergence radius is $\rho = 1$.

$$\mathrm{sc}(L) = \mathrm{sc}(hD) = h\alpha; \quad \sum_{j=1}^{\infty} \mathrm{sc}(\nabla)^j/j = -\ln(1-(1-e^{-h\alpha})) = h\alpha.$$

The convergence condition $|\mathrm{sc}(\nabla)| < 1$ reads $h\alpha > -\ln 2 = -0.69$ if α is real, $|h\omega| < \pi/3$ if $\alpha = i\omega$.

(ii) For $\alpha = 0$, $D\dfrac{\partial^n}{\partial\alpha^n}(e^{\alpha x}) = Dx^n = nx^{n-1}$. By Leibniz' rule:

$$\frac{\partial^n}{\partial\alpha^n}(\alpha e^{\alpha x}) = 0x^n + nx^{n-1}.$$

By the theorem, we now obtain *a formula for numerical differentiation that is exact for all $f \in \mathcal{P}_k$.*

$$hf'(x) = \Big(\nabla + \frac{1}{2}\nabla^2 + \frac{1}{3}\nabla^3 + \dots + \frac{1}{k-1}\nabla^{k-1}\Big)f(x) \tag{3.3.25}$$

[38] Recall Example 3.3.2

[39] Recall the definition of the scalar sc(·), after (3.3.13).

By Theorem 3.3.4, this is the *unique* formula of this type that uses the values of $f(x)$ at the k points $x_n : -h : x_{n-k+1}$. The same approximation can be derived in many other ways, perhaps with a different appearance; see Chapter 4. This derivation has several advantages; the same expansion yields approximation formulas for every k, and *if $f \in C^k$, $f \notin \mathcal{P}_k$, the first neglected term, i.e. $\frac{1}{k}\nabla_h^k f(x_n)$, provides an* **asymptotic error estimate**, if $f^{(k)}(x_n) \neq 0$.

We now apply this formula to the table in Example 3.3.2, where $f(x) = \tan x$, $h = 0.01$, $k = 6$,

$$0.01 f'(1.35) \approx 0.1996 + \frac{0.0163}{2} + \frac{0.0019}{3} + \frac{0.0001}{4} - \frac{0.0004}{5},$$

i.e. we obtain a sequence of approximate results,

$$f'(1.35) \approx 19.96, \quad 20.78, \quad 20.84, \quad 20.84, \quad 20.83.$$

The correct value to 3D is $(\cos 1.35)^{-2} = 20.849$. Note that the last result is worse than the next to last. Recall the last comments to the theorem. In this case this is due to the rounding errors of the data. Upper bounds for their effect of the sequence of approximate values of $f'(1.35)$ is, by Example 3.3.3, shown in the series

$$10^{-2}\Big(1 + \frac{2}{2} + \frac{4}{3} + \frac{8}{4} + \frac{16}{5} + \dots\Big).$$

A larger version of this problem was run on a computer with the machine unit $2^{-53} \approx 10^{-16}$; $f(x) = \tan x$, $x = 1.35 : -0.01 : 1.06$. In the beginning the error decreases rapidly, but after 18 terms the rounding errors take over, and the error then grows almost exponentially (with constant sign). The eighteenth term and its rounding error have almost the same modulus (but opposite sign). The smallest error equals $5\,10^{-10}$, and is obtained after 18 terms; after 29 terms the actual error has grown to $2\,10^{-6}$. Such a large number of terms is seldom used in practice, unless a very high accuracy is demanded. See also Problem 6b, a computer exercise that offers both similar and different experiences.

Equation (3.3.24)—or its variable step size variant in Chapter 4—is called the **Backwards Differentiation Formula**. It is the basis of the important BDF method for the numerical integration of ordinary differential equations.

Coefficients for *Backwards differentiation formulas for higher derivatives*, are obtained from the equations

$$(hD/\nabla)^k = (-\ln(1-\nabla)/\nabla)^k.$$

The following formulas were computed by means of the matrix representation of a truncated power series:

$$\begin{pmatrix} hD/\nabla \\ (hD/\nabla)^2 \\ (hD/\nabla)^3 \\ (hD/\nabla)^4 \\ (hD/\nabla)^5 \end{pmatrix} = \begin{pmatrix} 1 & 1/2 & 1/3 & 1/4 & 1/5 \\ 1 & 1 & 11/12 & 5/6 & 137/180 \\ 1 & 3/2 & 7/4 & 15/8 & 29/15 \\ 1 & 2 & 17/6 & 7/2 & 967/240 \\ 1 & 5/2 & 25/6 & 35/6 & 1069/144 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ \nabla \\ \nabla^2 \\ \nabla^3 \\ \nabla^4 \end{pmatrix}. \qquad (3.3.26)$$

The rows of the matrix are the first rows taken from the matrix representation of each of the expansions $(hD/\nabla)^k$, $k = 1 : 5$.

When the effect of the *irregular errors* of the data on a term becomes larger in magnitude than the term itself, the term should, of course, be neglected; it does more harm than good. This happens relatively early for the derivatives of high order; see Problem 6. When these formulas are to be used inside a program (rather than during an interactive post-processing of results of an automatic computation), some rules for *automatic truncation* have to be designed; an interesting kind of detail in scientific computing.

The *forwards differentiation formula*, which is analogously based on the operator series,

$$hD = \ln(1+\Delta) = \Delta - \frac{1}{2}\Delta^2 + \frac{1}{3}\Delta^3 \pm \dots \tag{3.3.27}$$

is sometimes useful too. We obtain the coefficients for derivatives of higher order by inserting minus signs in the second and fourth columns of the matrix in (3.3.26).

A grid (or a table) may be too sparse to be useful for numerical differentiation and for the computation of other linear functionals. For example, we saw above that the successive backward differences of $e^{i\omega x}$ increase exponentially if $|\omega h| > \pi/3$. In such a case the grid, where the values are given, gives insufficient information about the function. One also says that "the grid does not *resolve* the function". This is often indicated by a strong variation in the higher differences. However, even this indication can sometimes be absent. An extreme example is, $f(x) = \sin(\pi x/h)$, on the grid $x_j = jh$, $j = 0, \pm 1, \pm 2, \dots$. All the values, all the higher differences, and thus the estimates of $f'(x)$ at all grid points are zero, but the correct values of $f'(x_j)$ are certainly not zero. So, this is an example where the expansion (trivially) converges, but it is not valid! (Recall the discussion of a Maclaurin expansion for a non-analytic function at the end of Sec. 3.1.3. Now a similar trouble can occur also for an analytic function.)

A less trivial example is given by the functions

$$f(x) = \sum_{n=1}^{20} a_n \sin(2\pi n x), \quad g(x) = \sum_{n=1}^{10} (a_n + a_{10+n}) \sin(2\pi n x).$$

$f(x) = g(x)$ on the grid, hence they have the same difference scheme, but $f'(x) \neq g'(x)$ on the grid, and typically $f(x) \neq g(x)$ between the grid points.

3.3.3 The Peano Theorem

One can often, by a combination of theoretical and numerical evidence, rely on asymptotic error estimates. Since there are exceptions, it is interesting that there are two general methods for deriving strict error bounds. We call one of them the **norms and distance formula**. It is not restricted to polynomial approximation, and it is typically easy to use, but it requires some advanced concepts and it often overestimates the error. We therefore postpone the presentation of that method to a later chapter.

We shall now present another method, due to Peano[40]. Consider a linear functional $\tilde{L}$, e.g., $\tilde{L}f = \sum_{j=1}^{p} \; b_j f(x_j)$, suggested for the approximate computation of another linear functional L, e.g., $Lf = \int_0^1 \sqrt{x} f(x)dx$. Suppose that it is exact, when it is applied to any polynomial of degree less than k: In other words, $\tilde{L}f = Lf$, for all $f \in \mathcal{P}_k$. The remainder is then itself a linear functional, $R = L - \tilde{L}$, with the special property that

$$Rf = 0 \quad \text{if} \quad f \in \mathcal{P}_k.$$

Next theorem gives a representation for such functionals, which provides a universal device for deriving error bounds for approximations of the type that we are concerned with. Let $f \in C^n[a, b]$. In order to make the discussion less abstract we confine it to functionals of the following form, $0 \le m < n$,

$$Rf = \int_a^b \phi(x)f(x)dx + \sum_{j=1}^{p} \big(b_{j,0}f(x_j) + b_{j,1}f'(x_j) + \ldots + b_{j,m}f^{(m)}(x_j)\big), \quad (3.3.28)$$

where the function ϕ is integrable, and the points x_j lie in the bounded real interval $[a, b]$, and $b_{j,m} \neq 0$ for at least one value of j. Moreover, *we assume that*

$$Rp = 0 \quad \text{for all} \quad p \in \mathcal{P}_k. \quad (3.3.29)$$

We define the function[41]

$$t_+ = \max(t, 0); \quad t_+^j = \big(t_+\big)^j; \quad t_+^0 = \frac{1 + sign\, t}{2}; \quad (3.3.30)$$

The function t_+^0 is often denoted $H(t)$ an is known as the **Heaviside**[42] **unit step function**. The function sign is defined in Definition def3.1.sign. Note that $t_+^j \in C^{j-1}$, $(j \ge 1)$. The **Peano kernel** $K(u)$ of the functional R is defined by the equation,

$$K(u) = \frac{1}{(k-1)!} R_x (x-u)_+^{k-1}, \quad x \in [a, b], \quad u \in (-\infty, \infty). \quad (3.3.31)$$

The subscript in R_x indicates that R acts on the variable x (not u).

The function $K(u)$ *vanishes outside* $[a, b]$, because:

- if $u > b$ then $u > x$, hence $(x-u)_+^{k-1} = 0$ and $K(u) = 0$,
- if $u < a$ then $x > u$. It follows that $(x-u)_+^{k-1} = (x-u)^{k-1} \in \mathcal{P}_k$, hence $K(u) = 0$, by (3.3.31) and (3.3.29).

If $\phi(x)$ is a polynomial then $K(u)$ becomes a piecewise polynomial; the points x_j are the joints of the pieces. In this case $K \in C^{k-m-2}$; the order of differentiability may be lower, if ϕ has singularities.

We are now in a position to prove an important theorem.

[40] Giuseppe Peano (1858-1932), Italian mathematician and logician.

[41] We use the neutral notation t here for the variable, to avoid to tie up the function too closely with the variables x and u which play a special role in the following.

[42] Oliver Heaviside (1850–1925) English physicist.

Theorem 3.3.8. *Peano's Remainder Theorem.*

Suppose that $Rp = 0$ for all $p \in \mathcal{P}_k$. Then [43] *, for all $f \in C^k[a,b]$,*

$$Rf = \int_{-\infty}^{\infty} f^{(k)}(u)K(u)du. \tag{3.3.32}$$

The definition and some basic properties of the Peano kernel $K(u)$ were given above.

Proof. By Taylor's formula,

$$f(x) = \sum_{j=1}^{k-1} \frac{f^{(j)}(a)}{j!}(x-a)^j + \int_a^x \frac{f^{(k)}(u)}{(k-1)!}(x-u)^{k-1}du.$$

This follows from putting $n = k$, $z = x - a$, $t = (u-a)/(x-u)$ into (3.1.5). We rewrite the last term as $\int_a^\infty f^{(k)}(u)(x-u)_+^{k-1}du$. Then apply the functional $R = R_x$ to both sides. Since we can allow the interchange of the functional R with the integral, for the class of functionals that we are working with, this yields

$$Rf = 0 + R\int_a^\infty \frac{f^{(k)}(u)(x-u)_+^{k-1}}{(k-1)!}du = \int_a^\infty \frac{f^{(k)}(u)R_x(x-u)_+^{k-1}}{(k-1)!}du,$$

The theorem then follows from (3.3.31). □

Corollary 3.3.9.

Suppose that $Rp = 0$ for all $p \in \mathcal{P}_k$. Then

$$R_x(x-a)^k = k!\int_{-\infty}^{\infty} K(u)du. \tag{3.3.33}$$

For any $f \in C^k[a,b]$, $Rf = \frac{f^{(k)}(\xi)}{k!}R_x((x-a)^k)$, holds for some $\xi \in (a,b)$, if and only if $K(u)$ does not change its sign.

If $K(u)$ changes its sign, the best possible error bound *reads*

$$|Rf| \le \sup_{u\in[a,b]} |f^{(k)}(u)| \int_{-\infty}^{\infty} |K(u)|du;$$

a formula with $f^{(k)}(\xi)$ is not generally true in this case.

Proof. First suppose that $K(u)$ does not change sign. Then, by (3.3.32) and the mean value theorem of Integral Calculus, $Rf = f^{(k)}(\xi)\int_{-\infty}^{\infty} K(u)du$, $\xi \in [a,b]$. For $f(x) = (x-a)^k$ this yields (3.3.33). The "if" part of the corollary follows from the combination of these formulas for Rf and $R(x-a)^k$.

If $K(u)$ changes its sign, the "best possible bound" is approached by a sequence of functions f chosen so that (the continuous functions) $f^{(k)}(u)$ approach (the discontinuous function) sign $K(u)$. The "only if" part follows. □

[43] The definition of $f^{(k)}(u)$ for $u \notin [a,b]$ is arbitrary.

Example 3.3.7.

The remainder of the *trapezoidal rule* (one step of length h) reads

$$Rf = \int_0^h f(x)dx - \frac{h}{2}(f(h) + f(0)).$$

We know that $Rp = 0$ for all $p \in \mathcal{P}_2$. The Peano kernel is zero for $u \notin [0, h]$, while for $u \in [0, h]$,

$$K(u) = \int_0^h (x-u)_+ dx - \frac{h}{2}((h-u)_+ + 0)) = \frac{(h-u)^2}{2} - \frac{h(h-u)}{2} = \frac{-u(h-u)}{2} < 0$$

We also compute

$$\frac{Rx^2}{2!} = \int_0^h \frac{x^2}{2}dx - \frac{h \cdot h^2}{2 \cdot 2} = \frac{h^3}{6} - \frac{h^3}{4} = -\frac{h^3}{12}.$$

Since the Peano kernel does not change sign, we conclude that

$$Rf = -\frac{h^3}{12}f''(\xi), \quad \xi \in (0, h).$$

Example 3.3.8. *Peano kernels for difference operators.*

Let $Rf = \Delta^3 f(a)$, and set $x_i = a + ih$, $i = 0:3$. Note that $Rp = 0$ for $p \in \mathcal{P}_3$. Then

$$\begin{aligned} Rf &= f(x_3) - 3f(x_2) + 3f(x_1) - f(x_0), \\ 2K(u) &= (x_3 - u)_+^2 - 3(x_2 - u)_+^2 + 3(x_1 - u)_+^2 - (x_0 - u)_+^2, \end{aligned}$$

i.e.

$$2K(u) = \begin{cases} 0, & \text{if } u > x_3; \\ (x_3 - u)^2, & \text{if } x_2 \le u \le x_3; \\ (x_3 - u)^2 - 3(x_2 - u)^2, & \text{if } x_1 \le u \le x_2; \\ (x_3 - u)^2 - 3(x_2 - u)^2 + 3(x_1 - u)^2 \equiv (u - x_0)^2, & \text{if } x_0 \le u \le x_1; \\ (x_3 - u)^2 - 3(x_2 - u)^2 + 3(x_1 - u)^2 - (x_0 - u)^2 \equiv 0, & \text{if } u < x_0. \end{cases}$$

For the simplification of the last two lines we used that $\Delta_u^3 (x_0 - u)^2 \equiv 0$. Note that $K(u)$ is a piecewise polynomial in $\mathcal{P}_3$ and that $K''(u)$ is discontinuous at $u = x_i$, $i = 0:3$.

It can be shown (numerically or analytically) that $K(u) > 0$ in the interval (u_0, u_3). This is no surprise, for, by (3.3.5), $\Delta^n f(x) = h^n f^{(n)}(\xi)$ for any integer n, and, by the above corollary, this could not be generally true if $K(u)$ changes its sign. These calculations can be generalized to $\Delta^k f(a)$ for an arbitrary integer k. This example will be generalized in Sec. 4.2.5 to divided differences of non-equidistant data.

In general it is rather laborious to determine a Peano kernel. Sometimes one can show that the kernel is piecewise a polynomial, that it has a symmetry, and that has a simple form in the intervals near the boundaries. All this can simplify the computation, and might have been used in these examples.

It is usually much easier to compute $R((x-a)^k)$, and an *approximate error estimate* is often given by

$$Rf \sim \frac{f^{(k)}(a)}{k!} R\big((x-a)^k\big), \quad f^{(k)}(a) \neq 0. \tag{3.3.34}$$

For example, suppose that $x \in [a,b]$, where $b-a$ is of the order of magnitude of a step size parameter h, and that f is analytic in $[a,b]$. By Taylor's formula,

$$f(x) = p(x) + \frac{f^{(k)}(a)}{k!}(x-a)^k + \frac{f^{(k+1)}(a)}{(k+1)!}(x-a)^{k+1} + \ldots, \quad f^{(k)}(a) \neq 0,$$

where $p \in \mathcal{P}_k$, hence $Rp = 0$. Most of the common functionals can be applied term by term. Then

$$Rf = 0 + \frac{f^{(k)}(a)}{n!} R_x(x-a)^k + \frac{f^{(k+1)}(a)}{(k+1)!} R_x(x-a)^{k+1} + \ldots .$$

Assume that, for some c, $R_x(x-a)^k = O(h^{k+c})$, for $k = 1, 2, 3, \ldots$. (This is often the case.) Then (3.3.34) becomes an **asymptotic error estimate** as $h \to 0$. It was mentioned above that for formulas derived by operator methods, an asymptotic error estimate is directly available anyway, but if a formula is derived by other means (see Chapter 4) this error estimate is important.

Asymptotic error estimates are frequently used in computing, because they are often much easier to to derive and apply than strict error bounds. The question is, however, how to know (or feel), that "the computation is in the asymptotic regime", where an asymptotic estimate is practically reliable. Much can be said about this central question of Applied Mathematics. Let us her just mention that a difference scheme displays well the quantitative properties of a function needed for the judgment.

If $Rp = 0$ for $p \in \mathcal{P}_k$, then a fortiori $Rp = 0$ for $p \in \mathcal{P}_{k-i}$, $i = 0 : k$. We may thus obtain a Peano kernel for each i, which is temporarily denoted by $K_{k-i}(u)$. They are obtained by integration by parts,

$$R_k f = \int_{-\infty}^{\infty} K_k(u) f^{(k)}(u)\,du = \int_{-\infty}^{\infty} K_{k-1}(u) f^{(k-1)}(u)\,du \tag{3.3.35}$$

$$= \int K_{k-2}(u) f^{(k-2)}(u)\,du \ldots, \tag{3.3.36}$$

where $K_{k-i} = (-D)^i K_k$, $i = 1, 2, \ldots$, *as long as* K_{k-i} *is integrable. The lower order kernels are useful, e.g., if the actual function* f *is not as smooth as the usual remainder formula requires.*

For the trapezoidal rule we obtained in Example 3.3.7

$$K_1(u) = \frac{h}{2}u_+^0 + \frac{h}{2} - u + \frac{h}{2}(u-h)_+^0.$$

A second integration by parts can only be performed within the framework of Dirac's delta functions (distributions); K_0 is not integrable. A reader, who is familiar with these generalized functions, may enjoy the following formula:

$$Rf = \int_{-\infty}^{\infty} K_0(u)f(u)du \equiv \int_{-\infty}^{\infty} \Big(-\frac{h}{2}\delta(u) + 1 - \frac{h}{2}\delta(u-h)\Big)f(u)du.$$

This is for one step of the trapezoidal rule, but many functionals can be expressed analogously.

3.3.4 Approximation Formulas by Operator Methods

We shall now demonstrate how operator methods are very useful for deriving approximation formulas. For example, in order to find interpolation formulas we consider the operator expansion

$$f(b-\gamma h) = E^{-\gamma}f(b) = (1-\nabla)^{\gamma}f(b) = \sum_{j=0}^{\infty}\binom{\gamma}{j}(-\nabla)^j f(b).$$

The verification of the assumptions of Theorem 3.3.7 offers no difficulties, and we omit the details. Truncate the expansion before $(-\nabla)^k$. By the theorem we obtain, for every γ, an approximation formula for $f(b-\gamma h)$ that uses the function values $f(b-jh)$ for $j = 0 : k-1$; it is exact if $f \in \mathcal{P}_k$, and is unique in the sense of Theorem 3.3.4; We also obtain an asymptotic error estimate if $f \notin \mathcal{P}_k$, namely the first neglected term of the expansion, i.e.

$$\binom{\gamma}{k}(-\nabla)^k f(b) \sim \binom{\gamma}{k}(-h)^k f^{(k)}(b)$$

Note that the binomial coefficients are polynomials in the variable γ, and hence also in the variable $x = b - \gamma h$.

It follows that the approximation formula yields **a unique polynomial** $P_B \in \mathcal{P}_k$, that solves the **interpolation problem**: $P_B(b-hj) = f(b-hj),\ j = 0 : k-1$; ($B$ stands for Backward). If we set $x = b - \gamma h$, we obtain

$$P_B(x) = E^{-\gamma}f(b) = (1-\nabla)^{\gamma}f(a) = \sum_{j=0}^{k-1}\binom{\gamma}{j}(-\nabla)^j f(b) \qquad (3.3.37)$$
$$= f(b-\gamma h) + O(h^k f^{(k)}).$$

Due to the uniqueness; see the corollary of Theorem 3.3.4, the approximation to $f'(b)$ obtained by the first $k-1$ terms in Example 3.2.4 for $x_n = b$ is exactly the derivative $P_B'(b)$.

Similarly, the interpolation polynomial $P_F \in \mathcal{P}_k$ that uses *forward* differences based on the values of f at $a, a+h, \ldots, a+(k-1)h$, reads, if we set $x = a + \theta h$,

$$P_F(x) = E^{\theta}f(a) = (1+\Delta)^{\theta}f(a)\sum_{j=0}^{k-1}\binom{\theta}{j}\Delta^j f(a) \qquad (3.3.38)$$
$$= f(a+\theta h) + O(h^k f^{(k)}).$$

These formulas are known as **Newton's interpolation formulas for constant step size**, backwards and forwards. The generalization to variable step size will be found in Sec. 4.2.1.

There exists a similar expansion for *central differences.* Set

$$\phi_0(\theta) = 1, \quad \phi_1(\theta) = \theta, \quad \phi_j(\theta) = \frac{\theta}{j}\binom{\theta + \frac{1}{2}j - 1}{j-1}, \quad (j > 1). \tag{3.3.39}$$

ϕ_j is an even function if j is even, and an odd function if j is odd. It can be shown that $\delta^j\phi_k(\theta) = \phi_{k-j}(\theta)$, and $\delta^j\phi_k(0) = \delta_{j,k}$, (Kronecker's delta). The functions ϕ_k have thus an analogous relation to the operator δ as, e.g., the functions $\theta^j/j!$ and $\binom{\theta}{j}$ have to the operators D and Δ, respectively. We obtain the following expansion, analogous to Taylor's formula and Newton's forward interpolation formula. The proof is left for Problem 4(b). Then

$$E^\theta f(a) = \sum_{j=0}^{k-1} \phi_j(\theta)\delta^j f(a) = f(a + \theta h) + O(h^k f^{(k)}). \tag{3.3.40}$$

The direct practical importance of this formula is small, since $\delta^j f(a)$ cannot be expressed as a linear combination of the given data when j is odd. There are several formulas, where this drawback has been eliminated by various transformations. They were much in use before the computer age; each formula had its own group of fans. We shall derive only one of them, by a short break-neck application of the formal power series techniques.[44] Note that

$$\begin{aligned}
E^\theta &= e^{\theta hD} = \cosh\theta hD + \sinh\theta hD,\\
\delta^2 &= e^{hD} - 2 + e^{-hD}, \qquad e^{hD} - e^{-hD} = 2\mu\delta,\\
\cosh\theta hD &= \tfrac{1}{2}(E^\theta + E^{-\theta}) = \sum_{j=0}^{\infty}\phi_{2j}(\theta)\delta^{2j},\\
\sinh\theta hD &= \frac{1}{\theta}\frac{d(\cosh\theta hD)}{d(hD)} = \sum_{j=0}^{\infty}\phi_{2j}(\theta)\frac{1}{\theta}\frac{d\delta^{2j}}{d\delta^2}\frac{d\delta^2}{d(hD)}\\
&= \sum_{j=0}^{\infty}\phi_{2j}(\theta)\frac{j\delta^{2(j-1)}}{\theta}(e^{hD} - e^{-hD}) = \sum_{j=0}^{\infty}\phi_{2j}(\theta)\frac{2j}{\theta}\mu\delta^{2j-1}.
\end{aligned}$$

Hence,

$$f(x_0 + \theta h) = f_0 + \theta\mu\delta f_0 + \frac{\theta^2}{2!}\delta^2 f_0 + \sum_{j=2}^{\infty}\phi_{2j}(\theta)\Big(\frac{2j}{\theta}\mu\delta^{2j-1}f_0 + \delta^{2j}f_0\Big). \tag{3.3.41}$$

This is known as **Stirling's interpolation formula**. [45] The first three terms have been taken out from the sum, in order to show their simplicity and their resemblance

[44] Differentiation of a formal power series with respect to an indeterminate has a purely algebraic definition. See the last part of Sec. 3.1.5.

[45] James Stirling (1692–1770), British mathematician, perhaps most famous for his amazing approximation to $n!$.

to Taylor's formula. They yield the most practical formula for quadratic interpolation; it is easily remembered and worth to be remembered. An approximate error bound for this quadratic interpolation reads $|0.016\delta^3 f|$ if $|\theta| < 1$.

Note that

$$\phi_{2j}(\theta) = \theta^2(\theta^2 - 1)(\theta^2 - 4)\cdots(\theta^2 - (j-1)^2)/(2j)!.$$

The expansion yields a true interpolation formula if it is truncated after an *even* power of δ. For $k = 1$ you see that $f_0 + \theta\mu\delta f_0$ is not a formula for linear interpolation; it uses three data points instead of two. It is similar for all odd values of k.

Strict error bounds can be found by means of Peano's theorem, but the remainder terms given in Sec. 4.2.1 for Newton's general interpolation formula (that does not require equidistant data) typically give the answer easier. Both are typically of the form $c_{k+1}f^{(k+1)}(\xi)$ and require a bound for a derivative of high order. The assessment of such a bound typically costs much more work than performing interpolation in one point.

A more practical approach is to estimate a bound for this derivative by means of a bound for the differences of the same order. (Recall the important formula in (3.3.5).) This is not a rigorous *bound*, but it typically yields a quite reliable error *estimate*, in particular if you put a moderate safety factor on the top of it. There is much more to be said about the choice of step size and order; we shall return to this kind of questions in later chapters.

You can make error estimates during the run; it can happen sooner or later that it does not decrease, when you increase the order. You may just as well stop there, and accept the most recent value as the result. This event is most likely due to the influence of irregular errors, e.g. rounding errors, but it can also indicate that the interpolation process is semi-convergent only.

The attainable accuracy of polynomial interpolation applied to a table with n equidistant values of an analytic function, depends strongly on θ; the results are much poorer near the boundaries of the data set than near the center. This question will be illuminated in Sec. 4.8 by means of complex analysis.

Example 3.3.9.

The continuation of the difference scheme of a polynomial is a classical application of a difference scheme for obtaining a smooth extrapolation of a function outside its original domain. Given the values $y_{n-i} = f(x_n - ih)$ for $i = 1 : k$ and the backward differences, $\nabla^j y_{n-1}$, $j = 1 : k-1$. Recall that $\nabla^{k-1}y$ is a constant for $y \in \mathcal{P}_k$. Consider the algorithm

```
$\nabla^{k-1}y_n = \nabla^{k-1}y_{n-1};$
for $j = k-1 : -1 : 1,$
    $\nabla^{j-1}y_n = \nabla^{j-1}y_{n-1} + \nabla^j y_n;$        (3.3.42)
end
$y_n = \nabla^0 y_n;$
```

It is left for Problem 7a to show that the result y_n is the value at $x = x_n$ of the interpolation polynomial which is determined by y_{n-i}, $i = 1 : k$. This is a kind of inverse use of a difference scheme; there are additions from right to left along a diagonal, instead of subtractions from left to right.

This algorithm, which needs additions only, was used long ago for the production of mathematical tables, e.g., for logarithms. Suppose that one knows, e.g., by means of a series expansion, a relatively complicated polynomial approximation to (say) $f(x) = \ln x$, that is accurate enough in (say) the interval $[a, b]$, and that this has been used for the computation of k very accurate values $y_0 = f(a)$, $y_1 = f(a+h), \dots y_{k-1}$, needed for starting the difference scheme. The algorithm is then used for $n = k,\ k+1,\ k+2, \dots,\ (b-a)/h.$ $k-1$ additions only are needed for each value y_n. Some analysis must have been needed for the choice of the step h to make the tables useful with (say) linear interpolation, and for the choice of k to make the basic polynomial approximation accurate enough over a substantial number of steps. The precision used was higher, when the table was produced than when it was used. When $x = b$ was reached, a new approximating polynomial was needed for continuing the computation over an other interval; at least a new value of $\nabla^{k-1} y_n$.

This procedure was the basis of the unfinished Difference Engine project of the great 19th century British computer pioneer Charles Babbage. He abandoned it after a while in order to spend more time on his huge "Analytic Engine" project, which was also unfinished, but he documented a lot of ideas, where he was (say) 100 years ahead of his time. "Difference engines" based on Babbage's ideas were, however, constructed in Babbage's own time, by the Swedish inventors Scheutz (father and son) 1834 and by Wiberg 1876, and they were applied, among other things, to the automatic calculation and printing of tables of logarithms. See, e.g., Goldstine [21].

The algorithm in (3.3.42) can be generalized to the case of non-equidistant with the use of divided differences; see Sec. 4.2.1.

We now derive some central difference formulas for numerical differentiation. From the definition and from Bickley's table (Table 3.2.1)

$$\delta \equiv E^{1/2} - E^{-1/2} = 2\sinh\Big(\frac{1}{2}hD\Big). \tag{3.3.43}$$

We may therefore put $x = \frac{1}{2}hD$, $\sinh x = \frac{1}{2}\delta$ into the following expansion (see Problem 3.1.7),

$$x = \sinh x - \frac{1}{2}\frac{\sinh^3 x}{3} + \frac{1\cdot 3}{2\cdot 4}\frac{\sinh^5 x}{5} - \frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}\frac{\sinh^7 x}{7} \pm \dots,$$

with the result

$$hD = 2\text{arcsinh}\frac{\delta}{2} = \delta - \frac{\delta^3}{24} + \frac{3\delta^5}{640} - \frac{5\delta^7}{7,168} + \frac{35\delta^9}{294,912} - \frac{63\delta^{11}}{2,883,584} \pm \dots. \tag{3.3.44}$$

The verification of the assumptions of Theorem 3.3.7 follows the pattern of Example 3.3.6, and we omit the details. Since arcsinh$z$, $z \in \mathbf{C}$ has the same singularities as its derivative $(1+z^2)^{-1/2}$, namely $z = \pm i$, it follows that the expansion in (3.3.44), if $\text{sc}(\delta/2)$ is substituted for $\delta/2$, converges if $\text{sc}(\delta/2) < 1$, hence $\rho = 2$.

By squaring the above relation, we obtain

$$(hD)^2 = \delta^2 - \frac{\delta^4}{12} + \frac{\delta^6}{90} - \frac{\delta^8}{560} + \frac{\delta^{10}}{3,150} - \frac{\delta^{12}}{16,632} \pm \ldots,$$

$$f''(x_0) \approx \left(1 - \frac{\delta^2}{12} + \frac{\delta^4}{90} - \frac{\delta^6}{560} + \frac{\delta^8}{3,150} - \frac{\delta^{10}}{16,632} \pm \ldots\right) \frac{\delta^2 f_0}{h^2}. \quad (3.3.45)$$

By Theorem 3.3.7 (3.3.45) holds for all polynomials. Since the first neglected non-vanishing term of (3.3.45) when applied to f, is (asymptotically) $c\delta^{12} f''(x_0)$, the formula for $f''(x)$ is exact if $f'' \in \mathcal{P}_{12}$, i.e. if $f \in \mathcal{P}_{14}$, although only 13 values of $f(x)$ are used. We thus gain one degree and, in the application to other functions than polynomials, one order of accuracy, compared to what we may have expected by counting unknowns and equations only; see Theorem 3.3.4. *This is typical for a problem that has a symmetry with respect to the hull of the data points.*

Suppose that the values $f(x)$ are given on the grid $x = x_0 + nh$, n integer. Since (3.3.44) contains odd powers of δ, it cannot be used to compute f'_n on the same grid. as pointed out in the beginning of Sec. 3.3.2. This difficulty can be overcome by means of another formula given in Bickley's table, namely

$$\mu = \sqrt{1 + \delta^2/4}. \quad (3.3.46)$$

This is derived as follows. The formulas

$$\mu = \cosh \frac{hD}{2}, \qquad \frac{\delta}{2} = \sinh \frac{hD}{2}$$

follow rather directly from the definitions; the details are left for Problem 5a. The formula $(\cosh hD)^2 - (\sinh hD)^2 = 1$ holds also for formal power series. Hence

$$\mu^2 - \frac{1}{4}\delta^2 = 1, \quad \text{or} \quad \mu^2 = 1 + \frac{1}{4}\delta^2,$$

from which the relation (3.3.46) follows.

If we now multiply the right hand side of equation (3.3.44) by the expansion

$$1 = \mu\Big(1 + \frac{1}{4}\delta^2\Big)^{-1/2} = \mu\Big(1 - \frac{\delta^2}{8} + \frac{3\delta^4}{128} - \frac{5\delta^6}{1,024} + \frac{35\delta^8}{32,768} + \ldots\Big). \quad (3.3.47)$$

we obtain

$$hD = \Big(1 - \frac{\delta^2}{6} + \frac{\delta^4}{30} - \frac{\delta^6}{140} \pm \ldots\Big)\mu\delta. \quad (3.3.48)$$

This leads to a useful central difference formula for the first derivative (where we have used more terms than we displayed in the above derivation).

$$f'(x_0) = \Big(1 - \frac{\delta^2}{6} + \frac{\delta^4}{30} - \frac{\delta^6}{140} + \frac{\delta^8}{630} - \frac{\delta^{10}}{2772} \pm \ldots\Big)\frac{f_1 - f_{-1}}{2h}. \quad (3.3.49)$$

If you truncate the operator expansion in (3.3.49) after the δ^{2k} term, you obtain exactly the derivative of the interpolation polynomial of degree $2k+1$ for $f(x)$ that is determined by the $2k+2$ values f_i, $i = \pm 1, \pm 2, \ldots, \pm(k+1)$. Note that all the neglected terms in the expansion vanish when $f(x)$ is any polynomial of degree $2k+2$, independent of the value of f_0. (Check the statements first for $k = 0$; you will recognize a familiar property of the parabola.) So, although we search for a formula that is exact in $\mathcal{P}_{2k+2}$, we actually find a formula that is exact in $\mathcal{P}_{2k+3}$.

By the multiplication of the expansions in (3.3.45) and (3.3.48), we obtain the following formulas, which have applications in other sections

$$\begin{aligned} (hD)^3 &= \Big(1 - \frac{1}{4}\delta^2 + \frac{7}{120}\delta^4 + \ldots\Big)\mu\delta^3 \\ (hD)^5 &= \Big(1 - \frac{1}{3}\delta^2 + \ldots\Big)\mu\delta^5 \\ (hD)^7 &= \mu\delta^7 + \ldots \end{aligned} \tag{3.3.50}$$

Another valuable feature typical for δ^2*-expansions*, i.e. for expansions in powers of δ^2, is the rapid convergence. It was mentioned earlier that $\rho = 2$, hence $\rho^2 = 4$, (while $\rho = 1$ for the backwards differentiation formula). The error constants of the differentiation formulas obtained by (3.3.45) and (3.3.49) are thus relatively small.

All this is typical for the symmetric approximation formulas which are based on central differences; see, e.g., the above formula for $f''(x_0)$, or the next example. In view of this, can we forget the forward and backward difference formulas altogether? Well, this is not quite the case, since one must often deal with data that are unsymmetric with respect to the point where the result is needed. For example, given f_{-1}, f_0, f_1, how would you compute $f'(x_1)$? Asymmetry is also typical for the application to *initial value problems* for differential equations; see Volume III. In such applications methods based on symmetric rules for differentiation or integration have sometimes inferior properties of numerical stability.

When a problem has a symmetry around some point x_0, you are advised to try to derive a δ^2-expansion. The first step is to express the relevant operator in the form $\Phi(\delta^2)$, where the function Φ is analytic at the origin.

To find a δ^2-expansion for $\Phi(\delta^2)$ is algebraically the same thing as expanding $\Phi(z)$ into powers of a complex variable z. So, the methods for the manipulation of power series mentioned in Sec. 3.2.2 and Problem 3.1.8 are available, and so is the Cauchy–FFT method (Sec. 3.1.4). *For suitably chosen r, N you evaluate*

$$\Phi(re^{2\pi ik/N}), \quad k = 0 : N-1,$$

and obtain the coefficients of the δ^2-expansion by the FFT! You can therefore derive a long expansion, and later truncate it as needed. You also obtain error estimates for all these truncated expansions for free.

Suppose that you have found a truncated δ^2-expansion, (say)

$$A(\delta^2) \equiv a_1 + a_2\delta^2 + a_3\delta^4 + \ldots + a_{k+1}\delta^{2k},$$

but you want instead an equivalent symmetric expression of the form

$$B(E) \equiv b_1 + b_2(E + E^{-1}) + b_3(E^2 + E^{-2}) + \ldots + b_{k+1}(E^k + E^{-k}).$$

Note that $\delta^2 = E - 2 + E^{-1}$. The transformation $A(\delta^2) \mapsto B(E)$ can be performed in several ways. Since it is linear it can be expressed by a matrix multiplication of the form $b = M_{k+1}a$, where a, b are column vectors for the coefficients, and M_{k+1} is the $k+1 \times k+1$ upper triangular submatrix in the northwest corner of a matrix M that turns out to be

$$M = \begin{pmatrix} 1 & -2 & 6 & -20 & 70 & -252 & 924 & -3432 \\ & 1 & -4 & 15 & -56 & 210 & -792 & 3003 \\ & & 1 & -6 & 28 & -120 & 495 & -2002 \\ & & & 1 & -8 & 45 & -220 & 1001 \\ & & & & 1 & -10 & 66 & -364 \\ & & & & & 1 & -12 & 91 \\ & & & & & & 1 & -14 \\ & & & & & & & 1 \end{pmatrix}. \qquad (3.3.51)$$

Note that the matrix elements are binomial coefficients that can be generated recursively (Sec. 3.1.2). It is therefore easy to extend the matrix; this 8×8 matrix is sufficient for a δ^2-expansion up to the term $a_8\delta^{14}$.

The operator D^{-1} is defined by the relation $(D^{-1}f)(x) = \int^x f(t)\,dt$. The lower limit is not fixed, so $D^{-1}f$ contains an arbitrary integration constant. Note that $DD^{-1}f = f$, while $D^{-1}Df = f + C$, where C is the integration constant. A difference expression like $D^{-1}f(b) - D^{-1}f(a) = \int_a^b f(t)\,dt$ is uniquely defined. So is also $\delta D^{-1}f$, but $D^{-1}\delta f$ has an integration constant.

A right-hand inverse can be defined also for the operators Δ, ∇, δ. For example, $(\nabla^{-1}u)_n = \sum^{j=n} u_j$ has an arbitrary summation constant but, e.g., $\nabla\nabla^{-1} = 1$, and $\Delta\nabla^{-1} = E\nabla\nabla^{-1} = E$ are uniquely defined.

One can make the inverses unique by restricting the class of sequences (or functions). For example, if we require that $\sum_{j=0}^{\infty} u_j$ is convergent, and make the convention that $(\Delta^{-1}u)_n \to 0$ as $n \to \infty$, then $\Delta^{-1}u_n = -\sum_{j=n}^{\infty} u_j$; notice the minus sign. Also notice that this is consistent with the following formal computation:

$$(1 + E + E^2 + \ldots)u_n = (1 - E)^{-1}u_n = -\Delta^{-1}u_n.$$

We recommend, however, some extra care with infinite expansions into powers of operators like E that is not covered by Theorem 3.3.7, but the finite expansion

$$1 + E + E^2 + \ldots + E^{n-1} = (E^n - 1)(E - 1)^{-1} \qquad (3.3.52)$$

is valid.

In Chapter 5 we will use operator methods together with the Cauchy–FFT method for finding the **Newton–Cotes'** formulas for symmetric numerical integration.

Operator techniques can also be extended to *functions of several variables.* The basic relation is again the operator form of Taylor's formula, which in the case of two variables reads,

$$\begin{aligned} u(x_0 + h, y_0 + k) &= \exp\left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right) u(x_0, y_0) \\ &= \exp\left(h\frac{\partial}{\partial x}\right) \exp\left(k\frac{\partial}{\partial y}\right) u(x_0, y_0). \end{aligned} \qquad (3.3.53)$$

3.3.5 Single Linear Difference Equations

Historically, the term **difference equation** was probably first used in connection with an equation of the form

$$b_0\Delta^k y_n + b_1\Delta^{k-1}y_n + \dots b_{k-1}\Delta y_n + b_k y_n = 0, \quad n = 0, 1, 2, \dots$$

which reminds of a linear homogeneous differential equation. It follows, however, from the discussion after (3.3.1) and (3.3.3) that this equation can also be written in the form

$$y_{n+k} + a_1 y_{n+k-1} + \dots + a_k y_n = 0, \tag{3.3.54}$$

and nowadays this is what one usually means by a single homogeneous linear difference equation of kth order with *constant coefficients*; a difference equation without differences. More generally, if we let the coefficients a_i depend on n; we have a linear difference equation with *variable coefficients.* If we replace the zero on the right hand side with some known quantity r_n, we have a *nonhomogeneous* linear difference equation.

These types of equations are the main topic of this section. The coefficients and the unknown are real or complex numbers. We shall occasionally see examples of more general types of difference equations, e.g., a nonlinear difference equation $F(y_{n+k}, y_{n+k-1}, \dots, y_n) = 0)$, and we shall, in Volume III, deal with *first order systems* of difference equations, i.e. $y_{n+1} = A_n y_n + r_n$, where r_n, y_n, etc. are vectors while A_n is a square matrix. Finally, *partial difference equations* where you have two (or more) subscripts in the unknown, occur often as numerical methods for partial differential equations, but they have many other important applications too.

A difference equation can be viewed as a *recurrence relation.* With given values of $y_0,\ y_1, \dots,\ y_{k-1}$, called the **initial values** or the **seed** of the recurrence, we can successively compute $y_k,\ y_{k+1},\ y_{k+2}, \dots$; we see that *the general solution of a k'th order difference equation contains k arbitrary constants*, just like the general solution of the k'th order differential equation. There are other important similarities between difference and differential equations, for example the following superposition result.

Lemma 3.3.10. *The general solution of a nonhomogeneous linear difference equation (also with variable coefficients) is the sum of one particular solution of it, and the general solution of the corresponding homogeneous difference equation.*

In practical computing, the recursive computation of the solution of a difference equations is most common. It was mentioned at the end of Sec. 3.2.3 that many important functions, e.g., Bessel functions and orthogonal polynomials, satisfy second order linear difference equations with variable coefficients, (although this terminology was not used there). Other important applications are the multistep methods for ordinary differential equations.

In such an application you are usually interested in the solution for one particular initial condition, but due to rounding errors in the initial values you obtain another solution. It is therefore of interest to know the behaviour of the solutions of the corresponding homogeneous difference equation. The questions are:

- *Can we use a recurrence to find the wanted solution accurately?*
- *How shall we use a recurrence, forward or backward?*

Forward recurrence is the type we described above. In backward recurrence we choose some large integer N, and give (almost) arbitrary values of y_{N+i}, $i = 0 : k-1$ as seed, and compute y_n for $n = N-1 : -1 : 0$.

We have seen this already in Example 1.3.3 (and in Problem 10a of Sec. 1.3) for an inhomogeneous first order recurrence relation. It was there found that the forward recurrence was useless, while backward recurrence, with a rather naturally chosen seed, gave satisfactory results; (see Example 1.3.4 and Problem 10b).

It is often like this, though not always. In Problem 9 of Sec. 1.3 it is the other way around: the forward recurrence is useful, and the backward recurrence is useless.

Sometimes **boundary values** are prescribed for a difference equation instead of initial values, (say) p values at the beginning and $q = k-p$ values at the end, e.g., the values of y_0, y_1,..., y_{p-1}, and $y_{N-q}, \ldots,_{N-1}$, y_N are given. Then the difference equation can be treated as a *linear system* with $N-k$ unknown. This also holds for a difference equation with variable coefficients and for an inhomogeneous difference equation. *From the point of view of numerical stability, such a treatment can be better than either recurrence.* The amount of work is somewhat larger, not very much though, for the matrix is a band matrix. We have sees in Example 1.3.2 that for a fixed number of bands *the amount of work to solve such a linear system is proportional to the number of unknown.* An important particular case is when $k = 2$, $p = q = 1$; the linear system is then tridiagonal. An algorithm for such linear systems is described in Example 1.3.2.

Another similarity for differential and difference equations, is that the general solution of a linear equation with constant coefficients has a simple closed form. Although, in most cases, the real world problems have variable coefficients (or are nonlinear), one can often formulate a class of model problems with constant coefficients, with similar features. The analysis of such model problems can give hints, e.g., whether forward or backward recurrence should be used, or other questions related to the design and the analysis of the numerical stability of a numerical method for a more complicated problem.

We shall therefore now study how to solve a *single homogeneous linear difference equation with constant coefficients* (3.3.54), i.e.

$$y_{n+k} + a_1 y_{n+k-1} + \ldots + a_k y_n = 0.$$

It is satisfied by the sequence $\{y_j\}$, where $y_j = cu^j$, $(u \neq 0,\ c \neq 0)$, if and only if $u^{n+k} + a_1 u^{n+k-1} + \ldots + a_k u^n = 0$, that is when

$$\phi(u) \equiv u^k + a_1 u^{k-1} + \ldots + a_k = 0. \tag{3.3.55}$$

Equation (3.3.55) is called the **characteristic equation** of (3.3.54); $\phi(u)$ is called the **characteristic polynomial**.

Theorem 3.3.11.

If the characteristic equation has k different *roots, $u_1, \dots, u_k$, then the general solution of equation* (3.3.54) *is given by the sequences $\{y_n\}$, where*

$$y_n = c_1 u_1^n + c_2 u_2^n + \cdots + c_k u_k^n, \tag{3.3.56}$$

where $c_1, c_2, \dots, c_k$ are arbitrary constants.

Proof. That $\{y_n\}$ satisfies equation (3.3.54) follows from the previous comments and from the fact that the equation is linear. The parameters $c_1, c_2, \dots, c_k$ can be adjusted to arbitrary initial conditions $y_0, y_1, \dots, y_{k-1}$ by solving the system of equations

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ u_1 & u_2 & \cdots & u_k \\ \vdots & \vdots & & \vdots \\ u_1^{k-1} & u_2^{k-1} & \cdots & u_k^{k-1} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{k-1} \end{pmatrix}.$$

The matrix is a Vandermonde matrix and its determinant is thus equal to the product of all differences $(u_i - u_j)$, $i \geq j$, $1 < i \leq k$, which is nonzero; see the proof of Theorem 3.3.4. □

Example 3.3.10.

Consider the difference equation $y_{n+2} - 5y_{n+1} + 6y_n = 0$ with initial conditions $y_0 = 0$, $y_1 = 1$. Forward recurrence yields $y_2 = 5$, $y_3 = 19$, $y_4 = 65, \dots$.

The characteristic equation $u^2 - 5u + 6 = 0$ has roots $u_1 = 3$, $u_2 = 2$. Hence, the general solution is $y_n = c_1 3^n + c_2 2^n$. The initial conditions give the system of equations

$$c_1 + c_2 = 0, \qquad 3c_1 + 2c_2 = 1,$$

with solution $c_1 = 1$, $c_2 = -1$, hence $y_n = 3^n - 2^n$.

As a check we find $y_2 = 5$, $y_3 = 19$ in agreement with the results found by using forward recurrence.

Example 3.3.11.

Consider the difference equation

$$T_{n+1}(x) - 2xT_n(x) + T_{n-1}(x) = 0, \quad n \geq 1, \quad -1 < x < 1,$$

with initial conditions $T_0(x) = 1$, $T_1(x) = x$. We obtain $T_2(x) = 2x^2 - 1$, $T_3(x) = 4x^3 - 3x$, $T_4(x) = 8x^4 - 8x^2 + 1, \dots$. By induction, $T_n(x)$ is an nth degree polynomial in x.

We can obtain a simple formula for $T_n(x)$ by solving the difference equation. The characteristic equation is $u^2 - 2xu + 1 = 0$, with roots $u = x \pm i\sqrt{1 - x^2}$. Set

$x = \cos\phi$, $0 < x < \pi$. Then $u = \cos\phi \pm i \sin\phi$, and thus

$$u_1 = e^{i\phi}, \qquad u_2 = e^{-i\phi}, \quad u_1 \neq u_2.$$

The general solution is $T_n(x) = c_1 e^{in\phi} + c_2 e^{-in\phi}$, and the initial conditions give

$$c_1 + c_2 = 1, \qquad c_1 e^{i\phi} + c_2 e^{-i\phi} = \cos\phi,$$

with solution $c_1 = c_2 = 1/2$. Hence, $T_n(x) = \cos(n\phi)$, $x = \cos\phi$.

These polynomials are thus identical to the important Chebyshev polynomials that were introduced in (3.2.19), and were there in fact denoted by $T_n(x)$.

We excluded the cases $x = 1$ and $x = -1$, i.e. $\phi = 0$ and $\phi = \pi$, respectively. For the particular initial values of this example, there are no difficulties; the solution $T_n(x) = \cos n\phi$ depends continuously on ϕ, and as $\phi \to 0$ or $phi \to \pi$, $T_n(x) = \cos n\phi$ converges to 1 $\forall n$ or $(-1)^n$ $\forall n$, respectively.

When we ask for the general solution of the difference equation, the matters are a little more complicated, because the characteristic equation has in these cases a double root; $u = 1$ for $x = 1$, $u = -1$ for $x = -1$. Although they are thus covered by the next theorem, we shall look at them directly, because they are easy to solve, and they give a good preparation for the general case.

If $x = 1$, the difference equation reads $T_{n+1} - 2T_n + T_{n-1} = 0$, i.e. $\Delta^2 T_n = 0$. We know from before (see, e.g., Theorem 3.3.4) that this is satisfied iff $T_n = an + b$. The solution is no longer built up by exponentials; a linear term is there too.

If $x = -1$, the difference equation reads $T_{n+1} + 2T_n + T_{n-1} = 0$. Set $T_n = (-1)^n V_n$. The difference equation becomes, after division by $(-1)^{n+1}$, $V_{n+1} - 2V_n + V_{n-1} = 0$, with the general solution, $V_n = an + b$, hence $T_n = (-1)^n(an + b)$.

Theorem 3.3.12.

When u_i is an m_i-fold root of the characteristic equation, then the difference equation (12.3.3) *is satisfied by the sequence $\{y_n\}$, where*

$$y_n = P_i(n) u_i^n,$$

and P_i is an arbitrary polynomial in $\mathcal{P}_{m_i}$. The general solution of the difference equation is a linear combination of solutions of this form using all the distinct roots of the characteristic equation.

Proof. We can write the polynomial $P \in \mathcal{P}_{m_i}$ in the form

$$P_i(n) = b_1 + b_2 n + b_3 n(n-1) + \cdots + b_{m_i} n(n-1)\cdots(n - m_i + 2).$$

Thus it is sufficient to show that equation (3.3.54) is satisfied when

$$y_n = n(n-1)\cdots(n-p+1)u_i^n = (u^p \partial^p(u^n)/\partial u^p)_{u=u_i}, \quad p = 1 : m_i - 1. \tag{3.3.57}$$

Substitute this in the left-hand side of equation (3.3.54):

$$u^p \frac{\partial^p}{\partial u^p}\Big(u^{n+k} + a_1 u^{n+k-1} + \cdots + a_k u^n\Big) = u^p \frac{\partial^p}{\partial u^p}\big(\phi(u) u^n\big)$$
$$= u^p \Big(\phi^{(p)}(u) u^n + \binom{p}{1}\phi^{(p-1)}(u) n u^{n-1} + \cdots + \binom{p}{p}\phi(u)\frac{\partial^p}{\partial u^p}(u^n)\Big).$$

The last manipulation was made using Leibniz's rule.

Now ϕ and all the derivatives of ϕ which occur in the above expression are 0 for $u = u_i$, since u_i is an m_i-fold root. Thus the sequences $\{y_n\}$ in equation (3.3.57) satisfy the difference equation. We obtain a solution with $\sum m_i = k$ parameters by the linear combination of such solutions derived from the different roots of the characteristic equation.

It can be shown (see, e.g., Henrici [23, p. 214]) that these solutions are linearly independent. (This also follows from a different proof given in Chapter. 13, where a difference equation of higher order is transformed to a system of first order difference equations. This transformation also leads to other ways of handling inhomogeneous difference equations than those which are presented in this section.) □

Note that the double root cases discussed in the previous example are completely in accordance with this theorem. We take one more example.

Example 3.3.12.

Consider the difference equation $y_{n+3} - 3y_{n+2} + 4y_n = 0$. The characteristic equation is $u^3 - 3u^2 + 4 = 0$ with roots $u_1 = -1$, $u_2 = u_3 = 2$. Hence, the general solution reads

$$y_n = c_1(-1)^n + (c_2 + c_3 n)2^n.$$

For a **nonhomogeneous** linear difference equation of order k, one can often find a *particular solution* by the use of an *"Ansatz"* with undetermined coefficients; thereafter, by Lemma 3.3.10 one can get the general solution by adding the general solution of the homogeneous difference equation.

Example 3.3.13.

Consider the difference equation $y_{n+1} - 2y_n = a^n$, with initial condition $y_0 = 1$. Try the "Ansatz" $y_n = ca^n$. One gets

$$ca^{n+1} - 2ca^n = a^n, \quad c = 1/(a-2), \quad a \neq 2.$$

Thus the general solution is $y_n = a^n/(a-2) + c_1 2^n$. By the initial condition, $c_1 = 1 - 1/(a-2)$, hence

$$y_n = \frac{a^n - 2^n}{a-2} + 2^n. \tag{3.3.58}$$

When $a \to 2$, l'Hospital's rule gives $y_n = 2^n + n2^{n-1}$. Notice how the "Ansatz" must be modified when a is a root of the characteristic equation.

The general rule when the right hand side is of the form $P(n)a^n$ (or a sum of such terms), where P is a polynomial, is that the contribution of this term to y_n is $Q(n)a^n$, where Q is a polynomial. If a does not satisfy the characteristic equation then $\mathrm{d}eg\, Q = \mathrm{d}eg\, P$; if a is a single or a double root of the characteristic equation, then $\mathrm{d}eg\, Q = \mathrm{d}eg\, P + 1$ or $\mathrm{d}eg\, Q = \mathrm{d}eg\, P + 2$, respectively, etc. The coefficients of Q are determined by the insertion of $y_n = Q(n)a^n$ on the left hand side of the equation and matching the coefficients with the right hand side.

Another way to find a particular solution is based on the calculus of operators. Suppose that an inhomogeneous difference equation is given in the form $\psi(Q)y_n = b_n$, where Q is one of the operators Δ, δ and ∇, or an operator easily derived from these, e.g., $\frac{1}{6}\delta^2$, see Problem 24(d).

In Sec. 3.1.5 $\psi(Q)^{-1}$ was defined by the formal power series with the same coefficients as the Maclaurin series for the function $1/\psi(z)$, $z \in \mathbf{C}$, $\psi(0) \neq 0$. In simple cases, e.g., if $\psi(Q) = a_0 + a_1Q$, these coefficients are usually easily found. Then $\psi(Q)^{-1}b_n$ *is a particular solution* of the difference equation $\psi(Q)y_n = b_n$; the truncated expansions approximate this. Note that if $Q = \delta$ or ∇, the infinite expansion demands that b_n is defined also if $n < 0$.

Note that a similar technique, with the operator D, can also be applied to linear differential equations. Today this technique has to a large extent been replaced by the Laplace transform,[46] that yields essentially the same algebraic calculations as operator calculus.

In some branches of applied mathematics it is popular to treat nonhomogeneous difference equations by means of a **generating function**, also called the **z-transform**, since both the definition and the practical computations are analogous to the Laplace transform. The z-transform of the sequence $y = \{y_n\}_0^\infty$ is

$$Y(z) = \sum_{n=0}^{\infty} y_n z^{-n}. \tag{3.3.59}$$

Note that the sequence $\{Ey\} = \{y_{n+1}\}$ has the z-transform $zY(z) - y_0$, $\{E^2y\} = \{y_{n+2}\}$ has the z-transform $z^2Y(z) - y_0z - y_1$, etc.

If $Y(z)$ is available in *analytic* form, it can often be brought to a sum of functions, whose inverse z-transforms are known, by means of various analytic techniques, notably expansion into partial fractions, e.g., if $Y(z)$ is a rational function. On the other hand, if *numerical values* of $Y(z)$ have been computed for complex values of z on some circle in $\mathbf{C}$ by means of an algorithm, then y_n can be determined by an obvious modification of the Cauchy–FFT method described in Sec. 3.1.3 (for expansions into negative powers of z). More information about the z-transform can be found in Strang [44, Sec. 6.3].

We are now in a position to exemplify in more detail the use of linear difference equations to studies of numerical stability, of the type mentioned above.

Theorem 3.3.13.

Necessary and sufficient for boundedness (stability) of all solutions of the difference equation (3.3.54) *for all positive n is the following* **root condition***: (We shall say either that a difference equation or that a characteristic polynomial satisfies the root condition; the meaning is the same.)*

i. *All roots of characteristic equation* (3.3.55) *should be located inside or on the unit circle $|z| \leq 1$;*

ii. *The roots on the unit circle should be simple.*

[46] The Laplace transform is traditionally used for similar problems for linear differential equations, e.g., in electrical engineering.

Proof. Follows directly from Theorem 3.3.12. □

This root condition corresponds to cases, where it is the absolute error that matters. It is basic in the theory of linear multistep methods for ordinary differential equations. Computer Graphics and an algebraic criterion due to Schur are useful for investigations of the root condition in particular if the recurrence relation under investigation contains parameters.

There are important applications of single linear difference equations to the study of the stability of numerical methods. When a recurrence is used one is usually interested in the solution for one particular initial condition, but a rounding error in an initial value produces a different solution, and it is therefore of interest to know the behaviour of other solutions of the corresponding homogeneous difference equation. We have seen this already in Sec. 1.3.3 for an inhomogeneous first order recurrence relation, but it is even more important for recurrence relations of higher order.

The following example is based on a study by J. Todd[47] in 1950 (see [46]).

Example 3.3.14.

Consider the initial-value problem

$$y''(x) = -y, \quad y(0) = 0, \quad y'(0) = 1, \tag{3.3.60}$$

with the exact solution $y(x) = \sin x$. To compute an approximate solution $y_k = y(x_k)$ at equidistant points $x_k = kh$, where h is a step length, we approximate the second derivative according to (3.3.45),

$$h^2 y_k'' = \delta^2 y_k + \frac{\delta^4 y_k}{12} + \frac{\delta^6 y_k}{90} + \dots. \tag{3.3.61}$$

We first use the first term only; the second term shows that the truncation error of this approximation of y_k'' is asymptotically $h^2 y^{(4)}/12$. We then obtain the difference equation $h^{-2}\delta^2 y_k = -y_k$ or, in other words,

$$y_{k+2} = (2 - h^2)y_{k+1} - y_k, \quad y_0 = 0, \tag{3.3.62}$$

where a suitable value of y_1 is to be assigned. In the third column of Table 3.3.2 we show the results obtained using this recursion formula with $h = 0.1$ and $y_1 = \sin 0.1$. All computations in this example were carried out using IEEE double precision.We obtain about 3 digits accuracy at the end of the interval $x = 1.5$.

Since the algorithm was based on a second order accurate approximation of y'' one may expect that the solution of the differential equation is also second order accurate. This turns out to be correct in this case, e.g., if we divide the step size by 2, the errors will be divided by 4, approximately. We shall, however; see that we cannot always draw conclusions of this kind; we also have to take the numerical stability into account.

[47] John Todd, Irish-American numerical analyst that was one of the first studies of the numerical stability of an algorithm for the approximate solution of ordinary differential equations.

Table 3.3.2. *Integrating $y'' = -y$, $y(0) = 0$, $y'(0) = 1$; the letters U and S in the headlines of the last two columns refer to “Unstable” and “Stable’.*

x_k	$\sin x_k$	2nd order	4th order U	4th order S
0.1	0.0998334166	0.0998334	0.0998334166	0.0998334166
0.2	0.1986693308	0.1986685	0.1986693307	0.1986693303
0.3	0.2955202067	0.2955169	0.2955202067	0.2955202050
0.4	0.3894183423	0.3894101	0.3894183688	0.3894183382
0.5	0.4794255386	0.4794093	0.4794126947	0.4794255305
0.6	0.5646424734	0.5646143	0.5643841035	0.5646424593
0.7	0.6442176872	0.6441732	0.6403394433	0.6442176650
0.8	0.7173560909	0.7172903	0.6627719932	0.7173560580
0.9	0.7833269096	0.7832346	0.0254286676	0.7833268635
1.0	0.8414709848	0.8413465	−9.654611899	0.8414709226
1.1	0.8912073601	0.8910450	−144.4011267	0.8912072789
1.2	0.9320390860	0.9318329	−2010.123761	0.9320389830
1.3	0.9635581854	0.9633026	−27834.59620	0.9635580577
1.4	0.9854497300	0.9851393	−385277.6258	0.9854495749
1.5	0.9974949866	0.9971245	−5332730.260	0.9974948015

In the hope to obtain a more accurate solution, we shall now use one more term in the expansion (3.3.61); the third term then shows that the truncation error of this approximation is asymptotically $h^4y^{(6)}/90$. The difference equation now reads

$$\delta^2 y_k - \frac{1}{12}\delta^4 y_k = -h^2 y_k \tag{3.3.63}$$

or,

$$y_{k+2} = 16y_{k+1} - (30 - 12h^2)y_k + 16y_{k-1} - y_{k-2}, \quad k \geq 2, \quad y_0 = 0, \tag{3.3.64}$$

where starting values for y_1, y_2, and y_3 need to be assigned. We choose the correct values of the solution rounded to double precision. The results from this recursion are shown in the fourth column of Table 3.3.2. We see that disaster has struck—the recursion is severely unstable! Already for $x = 0.6$ the results are less accurate than the second order scheme. For $x \geq 0.9$ the errors dominate completely.

We shall now look at these difference equations from the point of view of the root condition. The characteristic equation for (3.3.62) reads $u^2 - (2-h^2)u + 1 = 0$, and since $|2 - h^2| < 2$, direct computation shows that it has simple roots of unit modulus. The root condition is satisfied. By Example 3.3.11, the solution of (3.3.62) is $y_n = T_n(1 - h^2/2)$.

For (3.3.64) the characteristic equation reads $u^4 - 16u^3 + (30 - 12h^2)u^2 - 16u + 1 = 0$. We see immediately that *the root condition cannot be satisfied.* Since the sum of the roots equals 16, it is impossible that all roots are inside or on the unit circle. In fact, the largest root equals 13.94. So, a tiny error at $x = 0.1$ has been multiplied by $13.94^{14} \approx 10^{16}$ at the end.

A stable fourth order accurate method can easily be constructed. Using the differential equation we replace the term $\delta^4 y_k$ in (3.3.63) by $h^2\delta^2 y_k'' = -h^2\delta^2 y_k$. This leads to the recursion formula[48]

$$y_{k+1} = \left(2 - \frac{h^2}{1+h^2/12}\right) y_k - y_{k-1}, \quad y_0 = 0. \tag{3.3.65}$$

This difference equation satisfies the root condition if $h^2 < 6$ (see Problem 23(a)). It requires y_0, $y_1 \approx y(h)$ as seed. The results using this recursion formula with $h = 0.1$ and $y_1 = \sin 0.1$, are shown in the fifth column of Table 3.3.2. The error at the end is about $2\cdot 10^{-7}$, which is much better than $3.7\cdot 10^{-4}$, obtained with the 2nd-order method.

Remark 3.3.2. If the solution of the original problem is itself strongly decreasing or strongly increasing, one should consider the location of the characteristic roots with respect to a circle in the complex plane that corresponds to the interesting solution. For example, if the interesting root is 0.8 then a root equal to -0.9 causes oscillations that may eventually become disturbing, if one is interested in *relative* accuracy also in a long run, even if the oscillating solution is small in the beginning.

Many problems contain homogeneous or nonhomogeneous linear difference equations with variable coefficients, for which the solutions are not known in a simple closed form.

We now confine the discussion to the cases where the original problems are to compute a particular solution of a *second order difference equation with variable coefficients*; several interesting problems of this type were mentioned above, and we formulated the questions: *can* we use a recurrence to find the wanted solution accurately, and *how* shall we use a recurrence, forwards or backwards. Typically the original problem contains some parameter, and one usually wants to make a study for an interval of parameter values.

Such questions are sometimes studied with *frozen coefficients*, i.e. the model problems are in the class of difference equations with constant coefficients in the range of the actual coefficients of the original problem; if one of the types of recurrence is satisfactory (i.e. numerically stable in some sense) for all model problems, one would like to conclude that they are satisfactory also for the original problem, but *the conclusion is not always valid* without further restrictions on the coefficients—see a counterexample in Problem 23c.

The technique with *frozen coefficients provides just a hint* that should always be *checked by numerical experiments* on the original problem. It is beyond the scope of this text to discuss what restrictions are needed. *If the coefficients of the original problem are slowly varying, however, there is a good chance that the numerical tests will confirm the hint*—but again: how slowly is "slowly"? A warning against the use of one of the types of recurrence may also be a valuable result of a study, although it is negative.

[48] This is a special case of Numerov's method (cf. Problem 3.4.28). It can be traced back at least to B. Numerov 1924.

The following lemma exemplifies a type of tool that may be useful in such cases. The proof is left for Problem 22a. Another useful tool is presented in Problem 24a and applied in Problem 24b.

Lemma 3.3.14. *Suppose that the wanted sequence y_n^* satisfies a difference equation (with constant coefficients),*

$$\alpha y_{n+1} + \beta y_n - \gamma y_{n-1} = 0, \quad (\alpha > \gamma > 0,\ \beta > 0),$$

and that y_n^* is known to be positive for all sufficiently large n. Then the characteristic roots can be written $0 < u_1 < 1$, $u_2 < 0$ and $|u_2| > u_1$. Then y_n^* is unique apart from a positive factor c; $y_n^* = cu_1^n$, $c > 0$.

A solution $\bar{y}_n$, called the trial solution *that is approximately of this form can be computed for $n = N : -1 : 0$ by* backward *recurrence starting with the "seed"$y_{N+1} = 0$ $y_N = 1$. If an accurate value of y_0^* is given, the wanted solution is*

$$y_n^* = \bar{y}_n y_0^* / \bar{y}_0,$$

with a relative error approximately proportional to $(u_2/u_1)^{n-N}$. (neglecting a possible error in y_0^*).[49]

The *forward* recurrence is not recommended for finding y_n^* in this case, since the positive term $c_1 u_1^n$ will eventually be drowned by the oscillating term $c_2 u_2^n$ that will be introduced by the rounding errors. The proof is left for Problem 24c. Even if y_0 (in the use of the forward recurrence) has no rounding errors, such errors committed at later stages will yield similar contributions to the numerical results.

Example 3.3.15.

The "original problem" is to compute the parabolic cylinder function $U(a, x)$ which satisfies the difference equation

$$(a + \tfrac{1}{2})U(a+1, x) + xU(a, x) - U(a-1, x) = 0,$$

see Handbook [1, Ch. 19]; in particular Example 19.28.1.

To be more precise, we consider the case $x = 5$. Given $U(3, 5) = 5.2847\,10^{-6}$ (obtained from a table in [1, p. 710], we want to determine $U(a, 5)$ for integer values of a, $a > 3$, as long as $|U(a, 5)| > 10^{-15}$. We guess (a priori) that the discussion can be restricted to the interval (say) $a = [3, 15]$. The above lemma then gives the hint of a backward recurrence, for $a = a' - 1 : -1 : 3$ for some appropriate a' (see below), in order to obtain a trial solution $\bar{U}_a$ with the seed $\bar{U}_{a'} = 1$, $\bar{U}_{a'+1} = 0$. Then the wanted solution becomes, by the Lemma, (with changed notation),

$$U(a, 5) = \bar{U}_a U(3, 5)/\bar{U}_3.$$

The positive characteristic root of the frozen difference equation varies from 0.174 to 0.14 for $a = 5 : 15$; while the modulus of the negative root is between 6.4 and

[49] If y_n^* is defined by some other condition, one can proceed analogously.

3.3 times as large. This motivates a choice of $a' \approx 4 + (-9 - \log 5.3)/\ln 0.174 \approx 17$ for the backward recursion; it seems advisable to choose a' (say) 4 units larger than the value where U becomes negligible.
Forward recurrence with correctly rounded starting values $U(3,5) = 5.2847\,10^{-6}$, $U(4,5) = 9.172\,10^{-7}$, gives oscillating (absolute) errors of relatively slowly decreasing amplitude, approximately 10^{-11}, that gradually drowns the exponentially decreasing true solution; the estimate of $U(a,5)$ itself became negative for $a = 10$, and then the results oscillated with approximate amplitude 10^{-11}, while the correct results decrease from the order of 10^{-11} to 10^{-15} as $a = 10 : 15$. The details are left for Problem 23b.

It is conceivable that this procedure can be used for all x in some interval around 5, but we refrain from presenting the properties of the parabolic cylinder function needed for determining the interval.

If the problem is nonlinear, one can instead solve the original problem with two seeds, (say) y'_N, y''_N, and study how the results deviate. The seeds should be so close that a linearization like $f(y'_n) - f(y''_n) \approx r_n(y'_n - y''_n)$ is acceptable, but $y'_n - y''_n$ should be well above the rounding error level. A more recent and general treatment of these matters is found in [17, Chapter 6].

Review Questions

1. Give expressions for the shift operator E^k in terms of Δ, ∇, and hD, and expressions for the central difference operator δ^2 in terms of E and hD.

2. Derive the best upper bound for the error of $\Delta^n y_0$, if we only know that the absolute value of the error of y_i, $i = 0, \dots, n$ does not exceed ϵ.

3. There is a theorem (and a corollary) about existence and uniqueness of approximation formulas of a certain type that are exact for polynomials of certain class. Formulate these results, and sketch the proofs.

4. What bound can be given for the k'th difference of a function in terms of a bound for the k'th derivative of the same function?

5. Formulate the basic theorem concerning the use of operator expansions for deriving approximation formulas for linear operators.

6. Formulate Peano's Remainder Theorem, and compute the Peano kernel for a given symmetric functional (with at most four subintervals).

7. Express polynomial interpolation formulas in terms of forward and backward difference operators.

8. Give Stirling's interpolation formula for quadratic interpolation with approximate bounds for truncation error and irregular error.

9. Derive central difference formulas for $f'(x_0)$ and $f''(x_0)$ that are exact for $f \in \mathcal{P}_4$. They should only use function values at x_j, $j = 0, \pm 1, \pm 2, \dots$, as many as needed. Give asymptotic error estimates.

10. Derive the formula for the general solution of the difference equation $y_{n+k} +$

$a_1 y_{n+k-1} + \ldots + a_k y_n = 0$, when the characteristic equation has simple roots only. What is the general solution, when the characteristic equation has multiple roots?

11. What is the general solution of the difference equation $\Delta^k y_n = an + b$?

Problems and Computer Exercises

1. (a) Show that

$$(1+\Delta)(1-\nabla) = 1, \qquad \Delta - \nabla = \Delta\nabla = \delta^2 = E - 2 + E^{-1},$$
$$\delta^2 y_n = y_{n+1} - 2y_n + y_{n-1}.$$

(b) Let $\Delta^p y_n, \nabla^p y_m, \delta^p y_k$ all denote the same quantity. How are n, m, k connected? Along which lines in the difference scheme are the subscripts constant?

(c) Given the values of $y_n,\ \nabla y_n, \ldots,\ \nabla^k y_n$, for a particular value of n. Find a recurrence relation for computing $y_n,\ y_{n-1}, \ldots,\ y_{n-k}$, by simple additions only. On the way you obtain the full difference scheme of this sequence.

(d) *Repeated summation by parts.* Show that if $u_1 = u_N = v_1 = v_N = 0$, then

$$\sum_{n=1}^{N-1} u_n \Delta^2 v_{n-1} = -\sum_{n=1}^{N-1} \Delta u_n \Delta v_n = \sum_{n=1}^{N-1} v_n \Delta^2 u_{n-1}.$$

(e) Show that if $\Delta^k v_n \to 0$, as $n \to \infty$, then $\sum_{n=m}^{\infty} \Delta^k v_n = -\Delta^{k-1} v_m$.

(f) Show that $(\mu\delta^3 + 2\mu\delta) f_0 = f_2 - f_{-2}$

(g) Prove, e.g., by means of summation by parts, that $\sum_{n=0}^{\infty} u_n z^n,\ |z| = 1,\ z \neq 1$, is convergent if $u_n \to 0$ monotonically. Formulate similar results for real cosine and sine series.

2. (a) Prove, e.g., by induction, the following two formulas:

$$\Delta_x^j \binom{x}{k} = \binom{x}{k-j}, \quad j \le k,$$

where Δ_x means differencing with respect to x, with $h = 1$.

$$\Delta^j x^{-1} = \frac{(-h)^j j!}{x(x+h)\cdots(x+jh)}.$$

Find the analogous expression for $\nabla^j x^{-1}$.

(b) What formulas with derivatives instead of differences are these formulas analogous to?

(c) Show the following formulas, if $x,\ a$ are integers:

$$\sum_{n=a}^{x-1} \binom{n}{k-1} = \binom{x}{k} - \binom{a}{k},$$

$$\sum_{n=x}^{\infty} \frac{1}{n(n+1)\cdots(n+j)} = \frac{1}{j}\cdot\frac{1}{x(x+1)\cdots(x+j-1)}.$$

Modify these results for non-integer x; $x-a$ is still an integer.

(d) Suppose that $b \neq 0, -1, -2, \ldots$, and set

$$c_0(a,b) = 1, \quad c_n(a,b) = \frac{a(a+1)\ldots(a+n-1)}{b(b+1)\ldots(b+n-1)}, \quad n = 1, 2, 3, \ldots$$

Show, e.g., by induction that $(-\Delta)^k c_n(a,b) = c_k(b-a,b)c_n(a,b+k)$, hence $(-\Delta)^n c_0(a,b) = c_n(b-a,b)$.

(e) Compute for $a = e$, $b = \pi$ (say), $c_n(a,b)$, $n = 1:100$. How do you avoid overflow? Compute $\Delta^n c_0(a,b)$, both numerically by the difference scheme, and according to the formula in (d). Compare the results and formulate your experiences. Do the same with $a = e$, $b = \pi^2$.

Do the same with $\Delta^j x^{-1}$ for various values of x, j and h.

3. Set

$$\mathrm{Y}_{ord} = (y_{n-k}, y_{n-k+1}, \ldots, y_{n-1}, y_n),$$
$$\mathrm{Y}_{dif} = (\nabla^k y_n,\ \nabla^{k-1} y_n, \ldots, \nabla y_n,\ y_n).$$

Note that the results of this problem also hold if the y_j are column vectors.

(a) Find a matrix P, such that $\mathrm{Y}_{dif} = \mathrm{Y}_{ord}P$. Show that

$$\mathrm{Y}_{ord} = \mathrm{Y}_{dif}P \quad \text{hence} \quad P^{-1} = P.$$

How do you generate this matrix by means of a simple recurrence relation?

Hint: P is related to the Pascal matrix, but do not forget the minus signs in this triangular matrix. Compare Problem 3 of Sec. 1.2.

(b) Suppose that $\sum_{j=0}^{k} \alpha_j E^{-j}$ and $\sum_{j=0}^{k} a_j \nabla^j$ represent the same operator. Set $\alpha = (\alpha_k, \alpha_{k-1}, \ldots, \alpha_0)^T$, and $a = (a_k, a_{k-1}, \ldots, a_0)^T$, i.e. $\mathrm{Y}_{ord} \cdot \alpha \equiv \mathrm{Y}_{dif} \cdot a$. Show that $Pa = \alpha$, $P\alpha = a$.

(c) The matrix P depends on the integer k. Is it true that the matrix which is obtained for a certain k is a submatrix of the matrix you obtain for a larger value of k?

(d) Compare this method of performing the mapping $\mathrm{Y}_{ord} \mapsto \mathrm{Y}_{dif}$ with the ordinary construction of a difference scheme. Consider the number of arithmetic operations, the kind of arithmetic operations, rounding errors, convenience of programming in a language with matrix operations as primary operations etc. Compare in the same way this method of performing the inverse mapping with the algorithm in Problem 1c.

4. (a) Set $f(x) = \tan x$. Compute by the use of the table of $\tan x$ (in Example 3.3.2), and the interpolation and differentiation formulas given in the above examples (almost) as accurately as possible

$$f'(1.35), \quad f(1.322), \quad f'(1.325), \quad f''(1.32).$$

Estimate the influence of rounding errors of the function values and estimate the truncation errors.

(b) Write a program for computing a difference scheme. Use it for computing the difference scheme for more accurate values of $\tan x$, $x = 1.30 : 0.01 : 1.35$, and calculate improved values of the functionals in (a). Compare the error estimates with the true errors.

(c) Verify the assumptions of Theorem 3.3.7 for one of the three interpolation formulas in Sec. 3.3.4.

(d) It is rather easy to find the values at $\theta = 0$ of the first two derivatives of Stirling's interpolation formula. You find thus explicit expressions for the coefficients in the formulas for $f'(x_0)$ and $f''(x_0)$ in (3.3.49) and (3.3.45), respectively. Check numerically a few coefficients in these equations, and explain why they are reciprocals of integers. Also note that each coefficient in (3.3.49) has a simple relation to the corresponding coefficient in (3.3.45).

5. (a) Study Bickley's table (Table 3.2.1), and derive some of the formulas, in particular the expressions for δ and μ in terms of hD, and vice versa.

(b) Show that $h^{-k}\delta^k - D^k$ has an expansion into *even* powers of h, when k is even. Find an analogous result for $h^{-k}\mu\delta^k - D^k$ when k is odd.

6. (a) Compute

$$f'(10)/12, \quad f^{(3)}(10)/720, \quad f^5(10)/30240,$$

by means of (3.3.26), given values of $f(x)$ for integer values of x. (This is asked for, e.g., in applications of Euler–Maclaurin's formula, Sec. 3.4.4.) Do this for $f(x) = x^{-3/2}$. Compare with the correct derivatives. Then do the same also for $f(x) = (x^3 + 1)^{-1/2}$.

(b) Study the backwards differentiation formula, see Example 3.3.6, on a computer. Compute $f'(1)$ for $f(x) = 1/x$, for $h = 0.02$ and $h = 0.03$, and compare with the exact result. Make a semi-logarithmic plot of the total error after n terms, $n = 1 : 29$. Study also the sign of the error. For each case, try to find out whether the achievable accuracy is set by the rounding errors or by the semi-convergence of the series.

Hint: A formula mentioned in Problem 2(a) can be helpful. Also note that this problem is both similar and very different from the function $\tan(x)$ that was studied in Example 3.3.6.

(c) Set $x_i = x_0 + ih$, $t = (x - x_2)/h$. Show that

$$y(x) = y_2 + t\Delta y_2 + \frac{t(t-1)}{2}\Delta^2 y_2 + \frac{t(t-1)(t-2)}{6}\Delta^3 y_1$$

equals the interpolation polynomial in $\mathcal{P}_4$ determined by the values (x_i, y_i), $i = 1 : 4$. (Note that $\Delta^3 y_1$ is used instead of $\Delta^3 y_2$ which is located outside the scheme. Is this OK?)

7. (a) Show the validity of the algorithm in (3.3.42).

(b) A well known formula reads

$$P(D)(e^{\alpha t}u(t)) = e^{\alpha t}P(D + \alpha)u(t),$$

where P is an arbitrary polynomial. Prove this, as well as the following analogous formulas:

$$P(E)(a^n u_n) = a^n P(aE)u_n,$$
$$P(\Delta/h)\big((1+\alpha h)^n u_n\big) = (1+\alpha h)^n P((1+\alpha h)\Delta/h+\alpha)u_n.$$

Can you find a more beautiful or more practical variant?

8. Find the Peano kernel $K(u)$ for the functional $\Delta^2 f(x_0)$. Compute $\int_{\mathbf{R}} K(u)\,du$ both by direct integration of $K(u)$, and by computing $\Delta^2 f(x_0)$ for a suitably chosen function f.

9. Set $y_j = y(t_j)$, $y'_j = y'(t_j)$. The following relations are of great interest in the numerical integration of the differential equations $y' = f(y)$:

(a) The **implicit Adams formula**:

$$y_{n+1} - y_n = h\left(a_0 y'_{n+1} + a_1\nabla y'_{n+1} + a_2\nabla^2 y'_{n+1} + \cdots\right).$$

Show that $\nabla = -\ln(1-\nabla)\sum a_i\nabla^i$, and find a recurrence relation for the coefficients. The coefficients a_i, $i = 0:6$, read as follows. Check a few of them.

$$a_i = 1,\quad -\frac{1}{2},\quad -\frac{1}{12},\quad -\frac{1}{24},\quad -\frac{19}{720},\quad -\frac{3}{160},\quad -\frac{863}{60480}.$$

Alternatively, derive the coefficients by means of the matrix representation, of a truncated power series.

(b) The **explicit Adams formula**:

$$y_{n+1} - y_n = h\left(b_0 y'_n + b_1\nabla y'_n + b_2\nabla^2 y'_n + \cdots\right).$$

Show that $\sum b_i\nabla^i E^{-1} = \sum a_i\nabla^i$, and show that

$$b_n - b_{n-1} = a_n, \quad (n \ge 1).$$

The coefficients $b_i,\ i = 0:6$, read as follows. Check a few of them.

$$b_i = 1,\quad \frac{1}{2},\quad \frac{5}{12},\quad \frac{3}{8},\quad \frac{251}{720},\quad \frac{95}{288},\quad \frac{19087}{60480}.$$

(c) Apply the the second order explicit Adams formula, i.e.

$$y_{n+1} - y_n = h(y'_n + \tfrac{1}{2}\nabla y'_n),$$

to the differential equation $y' = -y^2$, with initial condition $y(0) = 1$ and step size $h = 0.1$. Two initial values are needed for the recurrence; $y_0 = y(0) = 1$, of course, and we choose[50] $y_1 = 0.9090$. Then compute $y'_0 = -y_0^2$, $y'_1 = -y_1^2$. Then the explicit Adams formula yields y_2, and so on. Compute a few steps, and compare with the exact solution.[51]

[50] There are several ways of obtaining $y_1 \approx y(h)$, e.g., by one step of Runge's 2nd order method, see Sec. 1.4.3, or by a series expansion, like in Example 3.1.1.

[51] For an *implicit* Adams formula it is necessary, in this example, to solve a quadratic equation in each step.

10. Let $y_j = y_0 + jh$. Find the asymptotic behavior as $h \to 0$ of

$$(5(y_1 - y_0) + (y_2 - y_1))/(2h) - y_0' - 2y_1'.$$

Comment: This is of interest in the analysis of cubic spline interpolation in Sec. 4.4.4.

11. It sometimes happens that the values of some function $f(x)$ can be computed by some very time-consuming algorithm only, and that one therefore computes it much sparser than is needed for the application of the results. It was common in the pre-computer age to compute sparse tables that needed interpolation by polynomials of a high degree; then one needed a simple procedure for **subtabulation**, i.e. to obtain a denser table for some section of the table. Today a similar situation may occur in connection with the graphical output of the results of (say) a numerical solution of a differential equation.

Define the operators ∇ and ∇_k by the equations

$$\nabla f(x) = f(x) - f(x-h), \quad \nabla_k f(x) = f(x) - f(x-kh), \; (k < 1),$$

and set

$$\nabla_k^r = \sum_{s=r}^{\infty} c_{rs}(k)\nabla^s.$$

(a) In order to compute the coefficients c_{rs}, $r \le s \le m$, you are advised to use a subroutine for finding the coefficients in the product of two polynomials, truncate the result, and apply the subroutine $m-1$ times.

(b) Given

f_n	∇f_n	$\nabla^2 f_n$	$\nabla^3 f_n$	$\nabla^4 f_n$
1	0.181269	0.032858	0.005956	0.001080

Compute for $k = \frac{1}{2}$, $f_n = f(x_n)$, $\nabla_k^j f_n$ for $j = 1:4$. Compute $f(x_n - h)$ and $f(x_n - 2h)$, by means of both $\{\nabla^j f_n\}$ and $\{\nabla_k^j f_n\}$ and compare the results. How big difference of the results did you expect, and how big difference do you obtain?

12. Solve the following difference equations. A solution in complex form should be transformed to real form. As a check, compute (say) y_2 both by recurrence and by your closed form expression.

(a) $y_{n+2} - 2y_{n+1} - 3y_n = 0$, $y_0 = 0$, $y_1 = 1$;

(b) $y_{n+2} - 4y_{n+1} + 5y_n = 0$, $y_0 = 0$, $y_1 = 2$;

(c) There exist problems with two-point boundary conditions for difference equations, as for differential equations. $y_{n+2} - 2y_{n+1} - 3y_n = 0$, $y_0 = 0$, $y_{10} = 1$;

(d) $y_{n+2} + 2y_{n+1} + y_n = 0$, $y_0 = 1$, $y_1 = 0$;

(e) $y_{n+1} - y_n = 2^n$, $y_0 = 0$;

(f) $y_{n+2} - 2y_{n+1} - 3y_n = 1 + \cos\frac{\pi n}{3}$, $y_0 = y_1 = 0$;

Hint: The right hand side is $\Re(1 + a^n)$, where $a = e^{\pi i/3}$.

(g) $y_{n+1} - y_n = n$, $y_0 = 0$;

(h) $y_{n+1} - 2y_n = n2^n$, $y_0 = 0$;

13. (a) Prove Lemma 3.3.10.

(b) Consider the difference equation $y_{n+2} - 5y_{n+1} + 6y_n = 2n + 3(-1)^n$. Determine a particular solution of the form $y_n = an + b + c(-1)^n$.

(c) Solve also the difference equation $y_{n+2} - 6y_{n+1} + 5y_n = 2n + 3(-1)^n$. Why and how must you change the form of the particular solution?

14. (a) Show that the difference equation $\sum_{i=0}^{k} b_i \Delta^i y_n = 0$ has the characteristic equation: $\sum_{i=0}^{k} b_i (u-1)^i = 0$.

(b) Solve the difference equation $\Delta^2 y_n - 3\Delta y_n + 2y_n = 0$, with initial condition $\Delta y_0 = 1$.

(c) Find the characteristic equation for the equation $\sum_{i=0}^{k} b_i \nabla^i y_n = 0$?

15. The influence of wrong boundary slopes for cubic spline interpolation (with equidistant data)—see Sec. 4.4—is governed by the difference equation

$$e_{n+1} + 4e_n + e_{n-1} = 0, \quad 0 < n < m,$$

e_0, e_m given. Show that $e_n \approx u^n e_0 + u^{m-n} e_m$, $u = \sqrt{3} - 2 \approx -0.27$. More precisely

$$\left| e_n - (u^n e_0 + u^{m-n} e_m) \right| \le \frac{2|u^{3m/2}|}{1 - |u|^m} \max(|e_0|, |e_m|).$$

Generalize the simpler of these results to other difference and differential equations.

16. The Fibonacci sequence is defined by the recurrence relation

$$y_n = y_{n-1} + y_{n-2}, \quad y_0 = 0, \quad y_1 = 1.$$

(a) Calculate $\lim_{n\to\infty} y_{n+1}/y_n$.

(b) The error of the secant method (see Sec. 6.2.2) satisfies approximately the difference equation $\epsilon_n = C\epsilon_{n-1}\epsilon_{n-2}$. Solve this difference equation. Determine p, such that $\epsilon_{n+1}/\epsilon_n^p$ tends to a finite nonzero limit as $n \to \infty$. Calculate this limit.

17. For several algorithms using the "divide and conquer strategy", such as the Fast Fourier Transform and some sorting methods, one can find that the work $W(n)$ for the application of them to data of size n satisfies a recurrence relation of the form:

$$W(n) = 2W(n/2) + kn,$$

where k is a constant. Find $W(n)$.

18. When the recursion

$$x_{n+2} = (32x_{n+1} - 20x_n)/3, \quad x_0 = 3, \ x_1 = 2,$$

was solved numerically in low precision (23 bits mantissa), one obtained for x_i, $i = 2 : 12$ the (rounded) values

$$1.33,\ 0.89,\ 0.59,\ 0.40,\ 0.26,\ 0.18,\ 0.11,\ 0.03,\ -0.46,\ -5.05,\ -50.80.$$

Explain the difference from the exact values $x_n = 3(2/3)^n$.

19. (a) k, N are given integers $0 \le k <\le N$. A "discrete Green's function" $G_{n,k}$, $0 \le n \le N$ for the central difference operator $-\Delta\nabla$ together with the boundary conditions given below, is defined as the solution $u_n = G_{n,k}$ of the difference equation with boundary conditions, a

$$-\Delta\nabla u_n = \delta_{n,k}, \quad u_0 = u_N = 0;$$

($\delta_{n,k}$ is Kronecker's delta). Derive a fairly simple expression for $G_{n,k}$.

(b) Find (by computer) the inverse of the tridiagonal matrix

$$A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}.$$

What is the relation between Problems (a) and (b)? Find a formula for the elements of A^{-1}. Express the solution of the inhomogeneous difference equation $-\Delta\nabla u_n = b_n$, $u_0 = u_N = 0$, both in terms of the Green function $G_{n,k}$ and in terms of A^{-1} (for general N).

(c) Try to find an analogous formula[52] for the solution of an inhomogeneous boundary value problem for the *differential* equation $-u'' = f(x)$, $u(0) = u(1) = 0$.

20. (a) Demonstrate the formula

$$\sum_0^\infty \frac{(-x)^n c_n}{n!} = e^{-x} \sum_0^\infty \frac{x^n(-\Delta)^n c_0}{n!}. \qquad (3.3.66)$$

Hint: Use the relation $e^{-xE} = e^{-x(1+\Delta)} = e^{-x}e^{-x\Delta}$.

(b) For completely monotonic sequences $\{c_n\}$ and $\{(-\Delta)^n c_0\}$ are typically positive and decreasing sequences. For such sequences, the left hand side becomes extremely ill-conditioned for large x, (say) $x = 100$, while the graph of the terms on the right hand side (if exactly computed) are bell-shaped, almost like the normal probability density with mean x and standard deviation $\sqrt{x}$. We have called such a sum a *bell sum.* Such positive sums can be computed with little effort and no trouble with rounding errors, *if their coefficients are accurate.*

[52] In a *differential* equation, analogous to Problem 21(a), the Kronecker delta is to be replaced by the Dirac delta function. Also note that the inverse of the differential operator here can be described as an integral operator with the Green's function as the "kernel".

Compute the left hand side of (3.3.66), for $c_n = 1/(n+1)$, $x = 10 : 10 : 100$, and compute the right hand side, both with numerically computed differences and with exact differences; the latter are found in Problem 2a. (In this particular case you can also find the exact sum.)

Suppose that the higher differences $\{(-\Delta)^n c_0\}$ have been computed recursively from rounded values of c_n. Explain why one may fear that the right hand side of (3.3.66) does not provide much better results than the left hand side.

(c) Use (3.3.66) to derive the second expansion for $\operatorname{erf}(x)$ in Problem 11 of Sec. 3.2 from the first expansion.

Hint: Use one of the results of Problem 2 a.

(d) If $c_n = c_n(a, b)$ is defined as in Problem 2d, then the left hand side becomes the Maclaurin expansion of the Kummer function $M(a, b, -x)$; see the Hanbook [1, Ch. 13]; Show that

$$M(a, b, -x) = e^{-x} M(b - a, b, x)$$

by means of the results of Problems 23a and 2d.

21. (a) The difference equation $y_n + 5y_{n-1} = n^{-1}$ was discussed in Sec. 1.3.3. It can also be written thus: $(6 + \Delta)y_{n-1} = n^{-1}$. The expansion of $(6 + \Delta)^{-1} n^{-1}$ into powers of $\Delta/6$ provides a particular solution of the difference equation. Compute this numerically for a few values of n. Try to prove the convergence, with or without the expression in Problem 2b. Is this the same as the particular solution $I_n = \int_0^1 x^n (x+5)^{-1} dx$ that was studied in Example 1.2.1?
Hint: What happens as $n \to \infty$? Can more than one solution of this difference equation be bounded as $n \to \infty$?

(b) Make a similar study to the difference equation related to the integral in Problem 6 of Sec.1.2. Why does the argument suggested by the hint of (a) not work in this case? Try another proof.

22. (a) Prove Lemma 3.3.14. How is the conclusion to be changed, if we do not suppose that $\gamma < \alpha$, though the coefficients are still positive? Show that a backward recurrence is still to be recommended.

(b) Work out on a computer the numerical details of Example 3.3.15, and compare with the Handbook [1, Example 19.28.1]. (Some deviations are to be expected, since Miller used other rounding rules.) Try to detect the oscillating component by computing the difference scheme of the the computed $U(a, 5)$, and estimate roughly the error of the computed values.

23. (a) For which constant real a does the difference equation

$$y_{n+1} - 2ay_n + y_{n-1} = 0$$

satisfy the root condition?
For which values of the real constant a does there exist a solution, such that $\lim_{n\to\infty} y_n = 0$? For these values of a, how do you construct a solution $y_n = y_n^*$ by a recurrence and normalization, so that this condition as well as

the condition $y_0^* + 2\sum_{m=1}^{\infty} y_{2m}^* = 1$ are satisfied. Is y_n^* unique? Give also an explicit expression for y_n^*.

For the other real values of a, show that y_n^* does not exist, but that for any given y_0, y_1 a solution can be accurately constructed by forward recurrence. Give an explicit expression for this solution in terms of Chebyshev polynomials (of the first and the second kind). Is it true that backward recurrence is also stable, though more complicated than forward recurrence?

(b) The Bessel function $J_k(z)$ satisfies the difference equation,

$$J_{k+1}(z) - (2k/z)J_k(z) + J_{k-1}(z) = 0, \quad k = 1,\ 2,\ 3, \dots,$$

and the identities,

$$J_0(z) + 2J_2(z) + 2J_4(z) + 2J_6(z) + \ldots = 1;$$

$$J_0(z) - 2J_2(z) + 2J_4(z) - 2J_6(z) + \ldots = \cos z;$$

see Abramowitz and Stegun [1], 9.1.27, 9.1.46 and 9.1.47.

Show how one of the identities can be used for normalizing the trial sequence obtained by a backwards recurrence. Under what condition does Problem 26(a) give the hint to use the backwards recurrence for this difference equation?

Study the section on Bessel functions of integer order in Numerical Recipes. Apply this technique for $z = 10,\ 1,\ 0.1$ (say). The asymptotic formula (see [1, **9.3.1**])

$$J_k(z) \sim \frac{1}{\sqrt{2\pi k}}\Big(\frac{ez}{2k}\Big)^k, \quad k \gg 1,\ z \text{ fixed.}$$

may be useful for your decision where to start the backward recurrence. Use at least two starting points, and subtract the results (after normalization).

Comment: The above difference equation for $J_k(z)$ is also satisfied by a function denoted $Y_k(z)$,

$$Y_k(z) \sim \frac{-2}{\sqrt{2\pi k}}\Big(\frac{ez}{2k}\Big)^{-k}, \quad (k \gg 1).$$

How do these two solutions disturb each other, when forward or backward recurrence is used?

(c) *A counterexample* to the technique with frozen coefficients. Consider the difference equation $y_{n+1} - (-1)^n y_n + y_{n-1} = 0$. The technique with frozen coefficients leads to the consideration of the difference equations

$$z_{n+1} - 2az_n + z_{n-1} = 0, \quad a \in [-0.5, 0.5];$$

all of them have only bounded solutions. Find by numerical experiment that, nevertheless, there seems to exist unbounded solutions y_n of the first difference equation.

Comment: A theoretical proof of this is found by noting that the mapping $(y_{2n},\ y_{2n+1}) \mapsto (y_{2n+2},\ y_{2n+3})$ is represented by a matrix that is independent of n and has an eigenvalue that is less than -1.

24. Let $\{b_n\}_{-\infty}^{\infty}$ be a given sequence, and consider the difference equation,

$$y_{n-1} + 4y_n + y_{n+1} = b_n,$$

which can also be written in the form $(6 + \delta^2)y_n = b_n$.

(a) Show that the difference equation has at most one solution that is bounded for $-\infty < n < +\infty$. Find a particular solution in the form of an expansion into powers of the operator $\delta^2/6$. (This is hopefully bounded.)

(b) Apply it numerically to the sequence $b_n = (1+n^2h^2)^{-1}$, for a few values of the step size h, e.g., $h = 0.1, 0.2, 0.5, 1$. Study for $n = 0$ the rate of decrease (?) of the terms in the expansion. Terminate when you estimate that the error is (say) 10^{-6}. Check how well the difference equation is satisfied by the result.

(c) Study theoretically bounds for the terms when $b_n = \exp(i\omega hn)$ $\omega \in \mathbf{R}$. Does the expansion converge? Compare your conclusions with numerical experiments. Extend to the case when $b_n = B(nh)$, where $B(t)$ can be represented by an absolutely convergent Fourier integral, $B(t) = \int_{-\infty}^{\infty} e^{i\omega t}\beta(\omega)d\omega$. Note that $B(t) = (1+t^2)^{-1}$ if $\beta(\omega) = \frac{1}{2}e^{-|\omega|}$. Compare the theoretical results with the experimental results in (b).

(d) Put $Q = \delta^2/6$. Show that $\tilde{y}_n \equiv (1 - Q + Q^2 + \ldots \pm Q^{k-1})b_n/6$ satisfies the difference equation $(1 + Q)(\tilde{y}_n - y_n) = Q^k b_n/6$.

Comment: This procedure is worthwhile if the sequence b_n is so smooth that (say) 2 or 3 terms give satisfactory accuracy.

3.4 Acceleration of Convergence

3.4.1 Introduction

If a sequence $\{s_n\}_0^{\infty}$ converges slowly towards a limit s, but has a sort of regular behavior when n is large, it can under certain conditions be transformed into another infinite sequence $\{s'_n\}$, that converges much faster to the same limit. Here s'_n usually depends on the first n elements of the original sequence only. This is called **convergence acceleration**. Such a *sequence* transformation may be iterated, to yield a sequence of infinite sequences, $\{s''_n\}$, $\{s'''_n\}$ etc., hopefully with improved convergence towards the same limit s. For an *infinite series* convergence acceleration means the convergence acceleration of its sequence of partial sums. Some algorithms are most easily discussed in terms of sequences, others in terms of series.

Several transformations, linear as well as nonlinear, have been suggested and are successful, under various conditions. Some of them, like Aitken, repeated averages, and Euler's transformation, are most successful on *oscillating sequences* (alternating series or series in a complex variable). Others, like variants of Aitken acceleration, Euler–Maclaurin and Richardson, work primarily on *monotonic sequences* (series with positive terms). Some techniques for convergence acceleration transform a power series into a sequence of rational functions, e.g., continued fractions, Padé approximation, and the ϵ-algorithm

Convergence acceleration cannot be applied to "arbitrary sequences"; some sort of conditions are necessary that restrict the variation of the future elements of the sequence, i.e. the elements which are not computed numerically. In this section, these conditions are of a rather general type, in terms of *monotonicity, analyticity or asymptotic behavior* of simple and usual types. For the class of completely monotonic functions and some related classes of analytic functions the techniques of convergence acceleration can be put on a relatively solid theoretical basis.

Definition 3.4.1.

A function $u(s)$ is completely monotonic for $s \geq a$, $s \in \mathbf{R}$, iff

$$u(s) \geq 0, \quad (-1)^{(j)} f^{(j)}(s) \geq 0, \quad s \geq a \;\; \forall \;\; j \geq 0 \text{ (integer)}, \; \forall \; s \geq a, \text{ (real)}.$$

Nevertheless some of these techniques may even sometimes be successfully applied to *semi-convergent sequences*. Several of them can also use a limited number of coefficients of a power series for the computation of values of an *analytic continuation* of a function, outside the circle of convergence of the series that defined it.

In addition to the "general purpose" techniques to be discussed in this chapter, there are other techniques of convergence acceleration based on the use of more specific knowledge about a problem. For example, Poisson summation formula

$$\sum_{n=-\infty}^{\infty} f(n) = \sum_{j=-\infty}^{\infty} \hat{f}(j), \quad \hat{f}(\omega) = \int_{-\infty}^{\infty} f(\omega) e^{-2\pi i \omega x} \, dx; \tag{3.4.1}$$

($\hat{f}$ is the Fourier transform of f). This can be amazingly successful to a certain class of series $\sum a(n)$, namely if $a(x)$ has a rapidly decreasing Fourier Transform. The Poisson formula is also an invaluable tool for the design and analysis of numerical methods for several problems; see Theorem 3.4.4.

Irregular errors are very disturbing when these techniques are used. They sometimes set the limit for the reachable accuracy. For the sake of simplicity we therefore use IEEE double precision, in most examples.

3.4.2 Comparison Series and Aitken Acceleration

Suppose that the terms in the series $\sum_{j=1}^{\infty} a_j$ behave, for large j, like the terms of a series $\sum_{j=1}^{\infty} b_j$, i.e. $\lim_{j\to\infty} a_j/b_j = 1$. Then if the sum $s = \sum_{j=1}^{\infty} b_j$ is known one can write

$$\sum_{j=1}^{\infty} a_j = s + \sum_{j=1}^{\infty} (a_j - b_j),$$

where the series on the right hand side converges more quickly than the given series. We call this making use of a simple **comparison problem**. The same idea is used in many other contexts—for example, in the computation of integrals where the integrand has a singularity. Usual comparison series are

$$\sum_{j=1}^{\infty} n^{-2} = \pi^2/6, \qquad \sum_{j=1}^{\infty} n^{-4} = \pi^4/90, \quad etc.$$

A general expression for $\sum_{j=1}^{\infty} n^{-2r}$, is given by (3.4.27). No simple closed form is known for $\sum_{j=1}^{\infty} n^{-3}$.

Example 3.4.1.

The term $a_j = (j^4+1)^{-1/2}$ behaves, for large j, like $b_j = j^{-2}$, whose sum is $\pi^2/6$. Thus

$$\sum_{j=1}^{\infty} a_j = \pi^2/6 + \sum_{j=1}^{\infty} \big((j^4+1)^{-1/2} - j^{-2})\big) = 1.64493 - 0.30119 = 1.3437.$$

Five terms on the right hand side are sufficient for four-place accuracy in the final result. Using the series on the left hand side, one would not get four-place accuracy until after 20,000 terms.

This technique is unusually successful in this example. The reader is advised to find out that and why it is less successful for $a_j = (j^4 + j^3 + 1)^{-1/2}$.

An important comparison sequence is a geometric sequence

$$y_n = a + bk^n,$$

for which

$$\nabla s_n = y_n - y_{n-1} = bk^{n-1}(k-1).$$

It this is fitted to the three most recently computed terms of a given sequence, $y_n = s_n$ for (say) $n = j, j-1, j-2$, then $\nabla y_j = \nabla s_j$, $\nabla y_{j-1} = \nabla s_{j-1}$, and

$$k = \nabla s_j / \nabla s_{j-1}.$$

Hence

$$bk^j = \frac{\nabla s_j}{1 - 1/k} = \frac{\nabla s_j}{1 - \nabla s_{j-1}/\nabla s_j} = \frac{(\nabla s_j)^2}{\nabla^2 s_j}.$$

This yields a comparison sequence for each j. Suppose that $|k| < 1$. Then the comparison sequence has the $s'_j = \lim_{n\to\infty} y_n = a = y_j - bk^j$, i.e.

$$s \approx s'_j = s_j - \frac{(\nabla s_j)^2}{\nabla^2 s_j}. \tag{3.4.2}$$

This is called **Aitken acceleration**[53] and is the most popular *nonlinear* acceleration methods.

[53] Alexander Craig Aitken (1895–1967), Scotch mathematician born in New Zealand.

If $\{s_n\}$ is exactly a geometric sequence, i.e. if $s_n - a = k(s_{n-1} - a)\ \ \forall\ \ n$, then $s'_j = s\ \forall j$. Otherwise it can be shown (Henrici [24, 1964]) that under the assumptions

$$\lim_{j\to\infty} s_j = s, \quad \text{and} \quad \lim \frac{s_{j+1} - s_j}{s_j - s_{j-1}} = k^*, \quad |k^*| < 1, \tag{3.4.3}$$

the sequence $\{s'_j\}$ converges faster than does the sequence $\{s_j\}$. The above assumptions can often be verified for sequences arising from iterative processes and for many other applications.

If you want the sum of slowly convergent *series*, it may seem strange to compute the sequence of partial sums, and then compute the first and second differences of rounded values of this sequence in order to apply Aitken acceleration. The *a-version* of Aitken acceleration works on the terms a_j of an infinite series instead of on its partial sums s_j.

Clearly we have $a_j = \nabla s_j$, $j = 1 : N$. The a-version of Aitken acceleration thus reads $s'_j = s_j - a_j^2/\nabla a_j$, $j = 1 : N$. We want to determine a'_j so that

$$\sum_{k=1}^{j} a'_k = s'_j, \quad j = 1 : N.$$

Then

$$a'_1 = 0, \quad a'_j = a_j - \nabla(a_j^2/\nabla a_j), \quad j = 2 : N,$$

and $s'_N = s_N - a_N^2/\nabla a_N$ (show this). We may expect that this a-version of Aitken acceleration handles rounding errors better.

The condition $|k^*| < 1$ is a *sufficient* condition only. In practice, Aitken acceleration seems *most efficient if* $k^* = -1$. Indeed, it often converges even if $k^* < -1$; see Problem 7. It is *much less successful if* $k^* \approx 1$, e.g., for slowly convergent series with positive terms.

The Aitken acceleration process can often be *iterated*, to yield sequences, $\{s''_n\}_0^\infty$, $\{s'''_n\}_0^\infty$, etc., defined by the formulas

$$s''_j = s'_j - \frac{(\nabla s'_j)^2}{\nabla^2 s'_j}, \qquad s'''_j = s''_j - \frac{(\nabla s''_j)^2}{\nabla^2 s''_j} \dots \tag{3.4.4}$$

Example 3.4.2.

By (3.1.10), it follows for $x = 1$ that

$$1 - 1/3 + 1/5 - 1/7 + 1/9 - \ldots = \arctan 1 = \pi/4 \approx 0.7853981634.$$

This series converges very slowly. Even after 500 terms there still occur changes in the third decimal. Consider the partial sums $s_j = \sum_{n_0}^{j} (-1)^j (2n + 1)^{-1}$, with $n_0 = 5$, and compute the **iterated Aitken** sequences as indicated above.

The (sufficient) theoretical condition mentioned above is not satisfied, since $\nabla s_n/\nabla s_{n-1} \to -1$ as $n \to \infty$. Nevertheless, we shall see that the Aitken acceleration works well, and that the iterated accelerations converge rapidly. One gains two digits for every pair of terms, in spite of the slow convergence of the original series. The results in the table below were obtained using IEEE double precision. The errors of s'_j, s''_j, ... are denoted e'_j, e''_j, ...

j	s_j	e_j	e'_j	e''_j	e'''_j
5	0.744012	−4.1387e−2			
6	0.820935	3.5536e−2			
7	0.754268	−3.1130e−2	−1.7783e−4		
8	0.813092	2.7693e−2	1.1979e−4		
9	0.760460	−2.4938e−2	−8.4457e−5	−1.3332e−6	
10	0.808079	2.2681e−2	6.1741e−5	7.5041e−7	
11	0.764601	−2.0797e−2	−4.6484e−5	−4.4772e−7	−1.0289e−8

Example 3.4.3.

Set $a_n = e^{-\sqrt{n+1}}$, $n \geq 0$. As before, we denote by s_n the partial sums of $\sum a_n$, $s = \lim s_n = 1.67040681796634$, and use the same notations as above. Note that

$$\nabla s_n / \nabla s_{n-1} = a_n/a_{n-1} \approx 1 - \tfrac{1}{2} n^{-1/2}, \quad (n \gg 1],$$

so this series is slowly convergent. Computations with plain and iterated Aitken in IEEE double precision gave the results below:

j	e_{2j}	$e_{2j}^{(j)}$
0	−1.304	−1.304
1	−0.882	−4.10e−1
2	−0.640	−1.08e−1
3	−0.483	−3.32e−2
2	−0.374	−4.41e−3
5	−0.295	−7.97e−4
6	−0.237	−1.29e−4
7	−0.192	−1.06e−5
8	−0.158	−1.13e−5

The sequence $\{e_{2j}^{(j)}\}$ is monotonic until $j = 8$. After this $|e_{2j}^{(j)}|$ is mildly fluctuating around 10^{-5} (at least until $j = 24$), and the differences $\nabla s_{2j}^{(j)} = \nabla e_{2j}^{(j)}$ are sometimes several powers of 10 smaller than the actual errors and are misleading as error estimates. The rounding errors have taken over, and it is almost no use to compute more terms.

It is possible to use more terms for obtaining higher accuracy by applying iterated Aitken acceleration to a **thinned sequence** e.g., s_4, s_8, $s_{12}, \ldots$, Problem 4. Note the thinning is performed on a *sequence* that converges to the limit to be computed, e.g., the partial sums of a series. Only in so-called *bell sums* (see Problem 30) we shall do *a completely different kind of thinning*, namely a thinning of the *terms* of a series.

The convergence ratios of the thinned sequence are much smaller; for the series of the previous example they become approximately

$$\left(1 - \tfrac{1}{2}n^{-1/2}\right)^4 \approx 1 - 2n^{-1/2}, \quad n \gg 1.$$

The most important point is, though, that the rounding errors become more slowly amplified, so that terms far beyond the eighth number of the un-thinned sequence can be used in the acceleration, resulting in a much improved final accuracy.

How to realize the thinning depends on the sequence; a different thinning will be used in the next example.

Example 3.4.4.

We shall compute,

$$s = \sum_{n=1}^{\infty} n^{-3/2} = 2.612375348685488.$$

If all partial sums are used in Aitken acceleration, it turns out that the error $|e_{2j}^{(j)}|$ is decreasing until $j = 5$, when it is 0.07, and it remains on approximately this level for a long time.

j	0	1	2	3	4	5
E_{2j+1}	−1.61	−0.94	−4.92e−1	−2.49e−1	−1.25e−1	−6.25e−2
$E_{2j+1}^{(j)}$	−1.61	−1.85	−5.06e−2	−2.37e−4	−2.25e−7	2.25e−10

A much better result is obtained by means of thinning, but since the convergence is much slower here than in the previous case, we shall try “geometric” thinning rather than the “arithmetic” thinning used above, i.e. we now set $S_m = s_{2^m}$. Then

$$\nabla S_m = \sum_{1+2^{m-1}}^{2^m} a_n, \quad S_j = S_0 + \sum_{m=1}^{j} \nabla S_m, \quad E_j = S_j - s.$$

(If maximal accuracy is wanted, it may be advisable to use the ”divide and conquer technique” for computing these sums; see Problem 2.3.5, but it has not been used here.) By the approximation of the sums by integrals one can show that $\nabla S_m / \nabla S_{m-1} \approx 2^{-1/2}$, $m \gg 1$. The table above shows the errors of the first thinned sequence and the results after iterated Aitken acceleration. The last result has used 1024 terms of the original series, but since

$$s_n - s = -\sum_{j=n}^{\infty} j^{-3/2} \approx -\int_n^{\infty} t^{-3/2}\,dt = -\frac{2}{3}n^{-1/2}, \tag{3.4.5}$$

10^{20} terms would have been needed for obtaining this accuracy without convergence acceleration.

For sequences such that

$$s_n - s = c_0 n^{-p} + c_1 n^{-p-1} + O(n^{-p-2}), \quad p > 0,$$

where s, c_0, c_1 are unknown, the following variant of Aitken acceleration, (Bjørstad et al. [3]) is more successful:

$$s'_n = s_n - \frac{p+1}{p} \frac{\Delta s_n \nabla s_n}{\Delta s_n - \nabla s_n}. \tag{3.4.6}$$

It turns out that s'_n is two powers of n more accurate than s_n, $s'_n - s = O(n^{-p-2})$; see Problem 12. More generally, suppose that there exists a longer (unknown) asymptotic expansion of the form

$$s_n = s + n^{-p}(c_0 + c_1 n^{-1} + c_2 n^{-2} + \ldots), \quad n \to \infty. \tag{3.4.7}$$

This is a rather common case. Then we can extend this to an to an *iterative variant*, where p is to be increased by 2 in each iteration; $i = 0, 1, 2, \ldots$ is a superscript, i.e.

$$s_n^{i+1} = s_n^i - \frac{p+2i+1}{p+2i} \frac{\Delta s_n^i \nabla s_n^i}{\Delta s_n^i - \nabla s_n^i}. \tag{3.4.8}$$

If p is also unknown, it can be estimated by means of the equation,

$$\frac{1}{p+1} = -\Delta \frac{\Delta s_n}{\Delta s_n - \nabla s_n} + O(n^{-2}). \tag{3.4.9}$$

Example 3.4.5.

We consider the same series as in the previous example, i.e. $s = \sum n^{-3/2}$. We use (3.4.8) without thinning. Here $p = -1/2$, see Problem 13. As usual, the errors are denoted $e_j = s_j - s$, $e_{2j}^i = s_{2j}^i - s$. In the right column of the table below, we show the errors from a computation with 12 terms of the original series,

j	e_{2j}	e_{2j}^j
0	−1.612	−1.612
1	−1.066	−8.217e−3
2	−0.852	−4.617e−5
3	−0.730	+2.528e−7
4	−0.649	−1.122e−9
5	−0.590	−0.634e−11
6	−0.544	−1.322e−9

From this point the errors were around 10^{-10} or a little below. The rounding errors have taken over, and the differences are, as in Example 3.3.4, misleading for error estimation. If needed, higher accuracy can be obtained by "arithmetic thinning" with more terms.

In this computation only 12 terms were used. In the previous example a less accurate result was obtained by means of 1024 terms of the same series, but we must appreciate that the technique of Example 3.3.5 did not require the existence of an asymptotic expansion for s_n and may therefore have a wider range of application.

There are not yet so many theoretical results that give justice to the practically observed efficiency of iterated Aitken accelerations for oscillating sequences. One reason for this can be that the transformation (3.4.2), which the algorithms are based on, is *nonlinear*). For methods of convergence acceleration that are based on *linear* transformations, theoretical estimates of convergence rates and errors are closer to the practical performance of the methods.

In a generalization of Aitken acceleration one considers a transformation that is exact for sequences satisfying

$$a_0(s_n - a) + \cdots + a_k(s_{n-k} - a) = 0, \quad \forall \ \ n. \tag{3.4.10}$$

Shanks considered the sequence transformation

$$e_k(s_n) = \frac{\begin{vmatrix} s_n & s_{n+1} & \cdots & s_{n+k} \\ s_{n+1} & s_{n+2} & \cdots & s_{n+k+1} \\ \vdots & \vdots & \cdots & \vdots \\ s_{n+k} & s_{n+k+1} & \cdots & s_{n+2k} \end{vmatrix}}{\begin{vmatrix} \Delta^2 s_n & \cdots & \Delta^2 s_{n+k-1} \\ \vdots & \cdots & \vdots \\ \Delta^2 s_{n+k-1} & \cdots & \Delta^2 s_{n+2k-2} \end{vmatrix}}, \quad k = 1, 2, 3, \ldots \tag{3.4.11}$$

For $k = 1$ Shanks' transformation reduces to Aitken's Δ^2 process. It can be proved that $e_k(s_n) = a$ if and only if s_n satisfies (3.4.10). The determinants in the definition of $e_k(s_n)$ have a very special structure and are called **Hankel determinants**[54]. Such determinants satisfy a recurrence relationship, which can be used for implementing the transformation. An elegant recursive procedure to compute $e_k(s_n)$ directly, the **epsilon algorithm**, will be discussed further in Sec. sec3.5.3 in connection with continued fraction and Padé approximants.

3.4.3 Euler's Transformation

In 1755 Euler gave the first version of what is now called **Euler's transformation**. Let

$$S = \sum_{j=0}^{\infty} (-1)^j u_j, \tag{3.4.12}$$

[54] Named after the German mathematician Hermann Hankel (1839–1873).

be an alternating series ($u_j \geq 0$). Then Euler showed that

$$S = \sum_{k=0}^{\infty} \frac{1}{2^k} \Delta^k u_k, \tag{3.4.13}$$

Often it is better to apply Euler's transformation to the tail of a series.

We shall now apply another method of acceleration based on **repeated averaging** of the partial sums. Consider again the same series as in Example 3.4.2, i.e..

$$\sum_{j=0}^{\infty} (-1)^j (2j+1)^{-1} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \ldots = \frac{\pi}{4}. \tag{3.4.14}$$

Let S_N be the sum of the first N terms. The columns to the right of the S_N-column in the scheme given in Table 3.4.1 are formed by building averages.

Each number in a column is the mean of the two numbers which stand to the left and upper left of the number itself. In other words, each number is the mean of its "west" and "northwest" neighbor. The row index of M tells how many terms are used from the original series, while the column index -1 equals the number of repeated averagings. Only the digits which are different from those in the previous column are written out.

Table 3.4.1. *Summation by repeated averaging.*

N	S_N	M_2	M_3	M_4	M_5	M_6	M_7
6	0.744 012						
7	0.820 935	782 474					
8	0.754 268	787 602	5038				
9	0.813 092	783 680	5641	340			
10	0.760 460	786 776	5228	434	387		
11	0.808 079	784 270	5523	376	405	396	
12	0.764 601	786 340	5305	414	395	400	398

Notice that the values in each column oscillate. In general, for an alternating series, it follows from the next theorem together with (3.2.4) that *if the absolute value of the* jth *term, considered as a function of* j*, has a* kth *derivative which approaches zero monotonically for* $j > N_0$*, then every other value in column* M_{k+1} *is larger than the sum, and every other is smaller.* The above premise is satisfied here, since if $f(j) = (2j+1)^{-1}$ then $f^{(k)}(j) = c_k(2j+1)^{-1-k}$, which approaches zero monotonically.

If round-off is ignored, it follows from column M_6 that $0.785396 \leq \pi/4 \leq 0.785400$. To take account of round-off error, we set $\pi/4 = 0.785398 \pm 3 \cdot 10^{-6}$. The actual error is only $1.6\,10^{-7}$. In Example3.4.2 iterated Aitken accelerations gave about one decimal digit more with the same data.

It is evident how the above method can be applied to any *alternating series.* The diagonal elements are equivalent to the results from using Euler's transformation.

Euler's transformation and the averaging method, can be generalized for the convergence acceleration of a general complex power series

$$S(z)=\sum_{j=1}^{\infty}u_j z^{j-1}. \tag{3.4.15}$$

The alternating series obtained for $z=-1$. Other applications include *Fourier series.* They can be brought to this form, with $z=e^{i\phi}$, $-\pi\le\phi\le\pi$; see Problem 14 and Example 3.4.7. The irregular errors of the coefficients play a big role if $|\phi|\ll\pi$, and it is important to reduce their effects by means of a variant of the thinning technique, described (for Aitken acceleration) in the previous section. Another interesting application is the *analytic continuation* of the power series outside its circle of convergence; see Example 3.4.8.

Theorem 3.4.2.

The tail of the power series in (3.4.15) *can formally be transformed into the expansion, ($z\ne 1$).*

$$S(z)-\sum_{j=1}^{n}u_j z^{j-1}=\sum_{j=n+1}^{\infty}u_j z^{j-1}=\frac{z^n}{1-z}\sum_{s=0}^{\infty}P^s u_{n+1},\quad P=\frac{z}{1-z}\Delta. \tag{3.4.16}$$

Set $N=n+k-1$, and set

$$M_{n,1}=\sum_{j=1}^{n}u_j z^{j-1};\quad M_{N,k}=M_{n,1}+\frac{z^n}{1-z}\sum_{s=0}^{k-2}P^s u_{n+1};\quad n=N-k+1. \tag{3.4.17}$$

These quantities can be computed by the following recurrence formula that yields several estimates based on N terms from the original series.[55] *This is called the* **generalized Euler transformation**.

$$M_{N,k}=\frac{M_{N,k-1}-zM_{N-1,k-1}}{1-z},\quad k=2:N. \tag{3.4.18}$$

For $z=-1$, this is the repeated average algorithm described above, and $P=-\frac{1}{2}\Delta$.

Assume that $|z|\le 1$, that $\sum u_j z^{j-1}$ converges, and that $\Delta^s u_N\to 0$, $s=0:k$ as $N\to\infty$. Then $M_{N,k}\to S(z)$, as $N\to\infty$. If, moreover, $\Delta^{k-1}u_j$ has a constant sign for $j\ge N-k+2$, then the following strict error bounds are obtained:

$$|M_{N,k}-S(z)|\le|z(M_{N,k}-M_{N-1,k-1})|=|M_{N,k}-M_{N,k-1}|,\quad (k\ge 2). \tag{3.4.19}$$

Proof. We first note that, as $N\to\infty$, $P^s u_N\to 0$, $s=0:k$, and hence, by (3.4.17), $\lim M_{N,k}=\lim M_{N,0}=S(z)$.

[55] See Algorithm 3.3.1 for an adaptive choice of a kind of optimal output.

Euler's transformation can be formally derived by operators as follows:

$$\begin{aligned}S(z) - M_{n,1} &= z^n \sum_{i=0}^{\infty} (zE)^i u_{n+1} = \frac{z^n}{1-zE} u_{n+1} \\ &= \frac{z^n}{1-z-z\Delta} u_{n+1} = \frac{z^n}{1-z} \sum_{s=0}^{\infty} P^s u_{n+1}.\end{aligned}$$

In order to derive (3.4.18), note that this relation can equivalently be written thus,

$$M_{N,k} - M_{N,k-1} = z(M_{N,k} - M_{N-1,k-1}), \tag{3.4.20}$$

$$M_{N,k-1} - M_{N-1,k-1} = (1-z)(M_{N,k} - M_{N-1,k-1}). \tag{3.4.21}$$

Remembering that $n = N - k + 1$, we obtain, by (3.4.17),

$$M_{N,k} - M_{N-1,k-1} = \frac{z^{N-k+1}}{1-z} P^{k-2} u_{N-k+2}, \tag{3.4.22}$$

and it can be shown (Problem 17) that

$$M_{N,k-1} - M_{N-1,k-1} = z^n P^{k-2} u_{n+1} = z^{N-k+1} P^{k-2} u_{N-k+2}. \tag{3.4.23}$$

By (3.4.22) and (3.4.23), we now obtain (3.4.21) and hence also the equivalent equations (3.4.20) and (3.4.18).

Now substitute j for N into (3.4.23), and add the p equations obtained for $j = N+1, \ldots, N+p$. We obtain:

$$M_{N+p,k-1} - M_{N,k-1} = \sum_{j=N+1}^{N+p} z^{j-k+1} P^{k-2} u_{j-k+2}.$$

Then substitute $k+1$ for k, and $N+1+i$ for j. Let $p \to \infty$, while k is fixed. It follows that

$$S(z) - M_{N,k} = \sum_{j=N+1}^{\infty} z^{j-k} P^{k-1} u_{j-k+1} = \frac{z^{N-k+1} \cdot z^{k-1}}{(1-z)^{k-1}} \sum_{i=0}^{\infty} z^i \Delta^{k-1} u_{N-k+2+i}, \tag{3.4.24}$$

hence

$$|S(z) - M_{N,k}| \le \left|(z/(1-z))^{k-1} z^{N-k+1}\right| \sum_{i=0}^{\infty} \left|\Delta^{k-1} u_{N-k+2+i}\right|.$$

We now use the assumption that $\Delta^{k-1} u_j$ has constant sign for $j \ge N-k+2$. Since $\sum_{i=0}^{\infty} \Delta^{k-1} u_{N-k+2+i} = -\Delta^{k-2} u_{N-k+2}$, it follows that

$$|S(z) - M_{N,k}| \le \left| z^{N-k+1} \frac{z^{k-1} \Delta^{k-2} u_{N-k+2}}{(1-z)^{k-1}} \right| = \left| \frac{z \cdot z^{N-k+1}}{1-z} P^{k-2} u_{N-k+2} \right|.$$

Now, by (3.4.22), $|S(z) - M_{N,k}| \leq |z| \cdot |M_{N,k} - M_{N-1,k-1}|$. This is the first part of (3.4.19). The second part then follows from (3.4.20). □

Comments: Note that the elements $M_{N,k}$ become rational functions of z for fixed N, k. If the term u_n, as a function of n, belongs to $\mathcal{P}_k$, then the classical Euler transformation (for $n = 0$) yields the exact value of $S(z)$ after k terms, if $|z| < 1$. This follows from (3.4.16), because $\sum u_j z^j$ is convergent, and $P^s u_{n+1} = 0$ for $s \geq k$. In this particular case, $S(z) = Q(z)(1 - z)^{-k}$, where Q is a polynomial; in fact the Euler transformation gives $S(z)$ correctly for all $z \neq 1$.

The advantage of the recurrence formula (3.4.18), instead of a more direct use of (3.4.16), is that it provides a whole lower triangular matrix of estimates, so that one can, by means of a simple test, decide when to stop. This yields a result with strict error bound, if $\Delta^{k-1} u_j$ has a constant sign (for all j with a given k), and if the effect of rounding errors is evidently smaller than Tol. If these conditions are not satisfied, there is a small risk that the algorithm may terminate if the error estimate is incidentally small, e.g., near a sign change of $\Delta^{k-1} u_j$.

The irregular errors of the initial data are propagated to the results. In the long run, they are multiplied by approximately $|z/(1 - z)|$ from a column to the next—this is less than one if $\Re z < 1/2$—but in the beginning this growth factor can be as large as $(1 + |z|)/|1 - z|$. It plays no role for alternating series; its importance when $|1 - z|$ is smaller will be commented in Example 3.4.7.

The following algorithm is mainly based on the above theorem, but the possibility for the irregular errors to become dominant has been taken into account (somewhat) in the third alternative of the termination criterion.

Algorithm 3.4.1 The Generalized Euler Transformation

This algorithm is based on Theorem 3.4.2, with a tolerance named Tol, and a termination criterion based on (3.4.19), by the computation and inspection of the elements of M in a certain order, until it finds a pair of neighboring elements that satisfies the criterion.

The classical Euler transformation would only consider the diagonal elements M_{NN}, $N = 1, 2, ...$ and the termination would have been based on $|M_{NN} - M_{N-1,N-1}|$. The strategy used in this algorithm is superior for an important class of series.

```
function [sum,errest,N,kk] = euler(z,u,Tol)
%
% EULER applies the generalized Euler transform to a power
% series with terms u(j)z^j. The elements of M are inspected
% in a certain order, until a pair of neighboring elements
% are found that satisfies a termination criterion.
% Input are .....
%
Nmax = length(u);
errest = Inf;  olderrest = errest;
N = 1;  kk = 2; M(1,1) = u(1);
```

```
while (errest > Tol) & (N < Nmax) & (errest <= olderrest)
    N = N+1;  olderrest = errest;
    M(N,1) = M(N-1,1)+ u(N)*z^(N-1); % New partial sum
    for k = 2:N,
       M(N,k) = (M(N,k-1) - z*M(N-1,k-1))/(1-z);
       temp = abs(M(N,k) - M(N,k-1))/2;
       if temp < errest,
          kk = k; errest = temp;
       end
    end
 end
 sum = (M(N,kk) + M(N,kk-1))/2;
```

An oscillatory behavior of the values $|M_{N,k} - M_{N,k-1}$ in the same row, indicates that the irregular errors have become dominant. The smallest error estimates may then become unreliable.

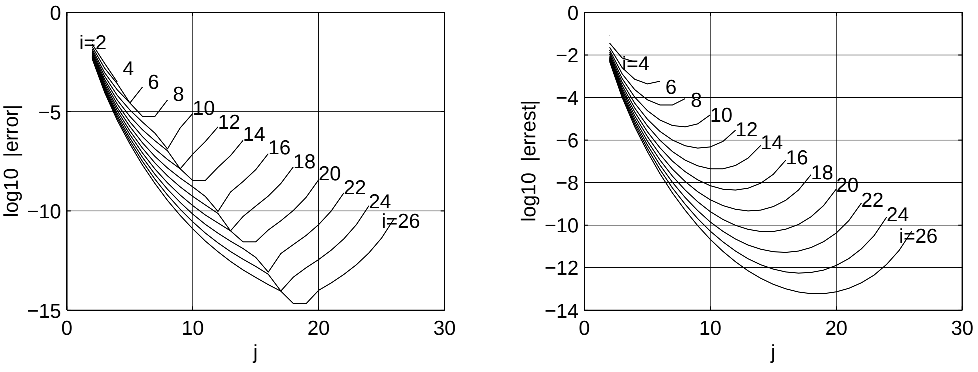

Figure 3.4.1. *Logarithms of the actual errors and the error estimates for $M_{N,k}$ in a more extensive computation for the alternating series in (3.4.14) with completely monotonic terms. The tolerance is here set above the level, where the irregular errors become important; for a smaller tolerance parts of the lowest curves may become less smooth in some parts.*

The above algorithm gives a strict error bound if, in the notation used in the theorem, $\Delta^{k-1}u_i$ has a constant sign for $i \geq N - k + 2$ (in addition to the other conditions of the theorem). We recall that a sequence, for which this condition is satisfied *for every k*, is called completely monotonic; see Definition 3.2.6.

It may seem difficult to check if this condition is satisfied. It turns out that many sequences that can be formed from sequences like $\{n^{-\alpha}\}$, $\{e^{-\alpha n}\}$ by simple operations and combinations, belong to this class. The generalized Euler transformation yields a sequence that converges at least as fast as a geometric series. The convergence ratio depends on z; it is less than one in absolute value for any complex z, except for $z > 1$ on the real axis. *So, the generalized Euler transformation often provides an analytic continuation of a power series outside its circle of convergence.*

For *alternating series*, with completely monotonic terms, i.e. for $z = -1$, the convergence ratio typically becomes $\frac{1}{3}$. This is in good agreement with Figure 3.4.1. Note that the minimum points for the errors lie almost on a straight line in Figure 3.5.1, and that the optimal value of k/N is approximately $\frac{2}{3}$, if $N \gg 1$, and if there are no irregular errors.

Example 3.4.6.
A program, essentially the same as Algorithm 3.4.3, is applied to the series

$$\sum_{j=1}^{\infty}(-1)^j j^{-1} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \ldots = \ln 2 = 0.69314\,71805\,599453.$$

with $\text{Tol}= 10^{-6}$, It stops when $N = 12$, $kk = 9$. The errors $e_k = M_{N,k} - \ln 2$ and the differences $\frac{1}{2}\nabla_k M_{N,k}$ along the last row of M read:

k	1	2	3	...	9	10	11	12
e_k	−3.99e-2	1.73e-3	−1.64e-4	...	−4.51e-7	5.35e-7	−9.44e-7	2.75e-6
$\nabla/2$		2.03e-2	−9.47e-4	...	−4.93e-7	4.93e-7	−7.40e-7	1.85e-6

Note that $|errest| = 4.93\,10^{-7}$ and $sum - \ln 2 = \frac{1}{2}(e_9 + e_8) = 4.2\,10^{-8}$. Almost full accuracy is obtained for $\text{Tol} = 10^{-16}$, $maxN = 40$. The results are $N = 32$, $kk = 22$, $errest = 10^{-16}$, $|error| = 2\,10^{-16}$. Note that $errest < |error|$; this can happen when we ask for such a high accuracy that the rounding errors are not negligible.

Example 3.4.7. *Application to Fourier series.*
Consider a complex power series

$$S(z) = \sum_{n=1}^{\infty} u_n z^{n-1}, \quad z = e^{i\phi}.$$

A Fourier series that is originally of the form $\sum_{-\infty}^{\infty}$ or in trigonometric form, can easily be brought to this form; see Problem 14. As we shall see, the results can often be improved considerably by the application of thinning. Let THIN be a positive integer. The thinned form of $S(z)$ reads

$$S(z) = \sum_{p=1}^{\infty} u_p^* z^{\text{THIN}\cdot(p-1)}, \quad u_p^* = \sum_{j=1}^{thin} u_{j+thin\cdot(p-1)}\, z^{j-1}.$$

For example, if $z = e^{i\pi/3}$ and THIN =3, the series becomes an alternating series, perhaps with complex coefficients. It does not matter in the numerical work that u_p^* depends on z.

We consider the case $S(z) = -\ln(1-z)/z = \sum z^{n-1}/n$, which is typical for a power series with completely monotonic terms. (The rates of convergence are the same for almost all series of this class.) Numerical computation, essentially by the above algorithm, gave the following results. The coefficients u_j are computed in IEEE double precision. We make the rounding errors during the computations less important by subtracting the first row of partial sums by its last element; it is, of course, added again to the final result.[56] The first table shows, for various ϕ, the most accurate result that can be obtained without thinning. These limits are due to the rounding errors; we can make the pure truncation error arbitrarily small by choosing N large enough.

ϕ	π	$2\pi/3$	$\pi/2$	$\pi/3$	$\pi/4$	$\pi/6$	$\pi/8$	$\pi/12$	$\pi/180$
|error|	2e-16	8e-16	1e-14	6e-12	1e-9	7e-8	5e-7	3e-5	2e-1
N	30	33	36	36	36	36	40	40	100
kk	21	22	20	21	20	14	13	10	(3)

Note that a rather good accuracy is obtained also for $\phi = \pi/8$ and $\phi = \pi/12$, where the algorithm is "unstable", since $|\frac{z}{1-z}| > 1$. In this kind of computations "instability" does not mean that the algorithm is hopeless, but it shows the importance of a good termination criterion. The question is to navigate safely between Scylla and Charybdis. For a small value like $\phi = \pi/180$, the sum is approximately $4.1 + 1.5i$. The smallest error with 100 terms (or less) is 0.02; it is obtained for $k = 3$. Also note that kk/N increases with ϕ.

By *thinning*, much better results are obtained for $\phi \ll \pi$, in particular for $\phi = \pi/180$. This series that has "essentially positive" terms originally can become "essentially alternating" by thinning. We present the errors obtained for four values of the parameter THIN, with different amount of work. Compare $|error|$, kk, etc. with appropriate values in the table above. We see that, by thinning, it is possible to calculate the Fourier series very accurately also for small values of ϕ.

THIN	80	120	90	15
$thin \cdot \phi$	π	$2\pi/3$	$\pi/2$	$\pi/12$
|error|	2e-14	1e-14	3e-13	3e-5
N	28	31	33	41
kk	20	22	18	10
total no. terms	5040	3720	2970	615

Roughly speaking, the optimal convergence rate of the Euler Transformation depends on z in the same way for all power series with completely monotonic coefficients; independently of the rate of convergence of the original series. The above tables from a particular example can therefore—with some safety margin—be used as a guide for the application of the Euler transformation with thinning to any series of this class.

Say that you want the sum of a series $\sum u_n z^n$ for $z = e^{i\phi}$, $\phi = \pi/12$, with relative $|error| < 10^{-10}$. You see in the first table that $|error| = 6\,10^{-12}$ for $\phi = \pi/3 = 4\pi/12$ without thinning. The safety margin is hopefully large enough. Therefore, try $Thin = 4$. We make two tests with completely monotonic terms:

[56] Tricks like this can often be applied in linear computations with a slowly varying sequence of numbers. See, e.g., the discussion of rounding errors in Richardson extrapolation in Sec. 3.3.5.

$u_n = n^{-1}$ and $u_n = \exp(-\sqrt{n})$. $Tol = 10^{-10}$ is hopefully large enough to make the irregular errors relatively negligible. In both tests the actual $|error|$ turns out to be $4\,10^{-11}$, and the total number of terms is $4 \cdot 32 = 128$. The values of *errest* are $6\,10^{-11}$ and $7\,10^{-11}$; both slightly overestimate the actual errors and are still smaller than TOL.

Example 3.4.8. *Application to a divergent power series, (analytic continuation).*
Consider a complex power series

$$S(z) = \sum_{n=1}^{\infty} u_n z^{n-1}, \quad |z| > 1.$$

As in the previous example we study in detail the case of $u_n = 1/n$. It was mentioned above that the generalized Euler transformation theoretically converges in the z-plane, cut along the interval $[1, \infty]$. The limit is $-z^{-1}\ln(1-z)$, a single-valued function in this region. For various z outside the unit circle, we shall see that rounding causes bigger problems here than for Fourier series. The error estimate of Algorithm 3.3.1, usually underestimated the error, sometimes by a factor of ten. The table reports some results from experiments without thinning.

z	-2	-4	-10	-100	-1000	$2i$	$8i$	$1+i$	$2+i$
|error|	2e-12	2e-8	4e-5	3e-3	5e-2	8e-11	1e-3	1e-7	2e-2
N	38	41	43	50	51	40	39	38	39
kk	32	34	39	50	51	28	34	22	24

Thinning can be applied also in this application, but here not only the argument ϕ is increased (this is good), but also $|z|$ (this is bad). Nevertheless, for $z = 1 + i$, the error becomes 10^{-7}, $3\,10^{-9}$, 10^{-9}, $4\,10^{-8}$, for $thin = 1,\ 2,\ 3,\ 4$, respectively. For $z = 2 + i$, however, thinning improved the error only from 0.02 to 0.01. All this is for IEEE double precision.

We shall encounter other methods for alternating series and complex power series, which are even more efficient than the generalized Euler transformation; see the epsilon algorithm in Sec. 3.5.3.

3.4.4 Euler–Maclaurin's Formula

In the summation of series with essentially positive terms the tail of the sum can be approximated by an integral by means of the trapezoidal rule.

As an example, consider the sum $S = \sum_{j=1}^{\infty} j^{-2}$. The sum of the first nine terms is, to four decimal places, 1.5398. It immediately occurs to one to compare the tail of the series with the integral of x^{-2} from 10 to ∞. We approximate the integral according to the trapezoidal rule; see Sec. 1.2

$$\int_{10}^{\infty} x^{-2}\,dx \approx T_1 + T_2 + T_3 + \ldots = \frac{1}{2}(10^{-2} + 11^{-2}) + \frac{1}{2}(11^{-2} + 12^{-2}) + \ldots$$

$$= \sum_{j=10}^{\infty} j^{-2} - \frac{1}{2}10^{-2}.$$

Hence it follows that

$$\sum_{j=1}^{\infty} j^{-2} \approx 1.5398 + [-x^{-1}]_{10}^{\infty} + 0.0050 = 1.5398 + 0.1050 = 1.6448.$$

The correct answer is $\pi^2/6 = 1.64493406684823$. We would have needed about 10,000 terms to get the same accuracy by direct addition of the terms!

The above procedure is not a coincidental trick, but a very useful method. A further systematic development of the idea leads to the important Euler–Maclaurin summation formula. We first derive this heuristically by operator techniques and exemplify its use, including a somewhat paradoxical example that shows that a strict treatment with the consideration of the remainder term is necessary for very practical reasons. Since this formula has several other applications, e.g., in numerical integration, we formulate it more generally than needed for the summation of infinite series.

Consider to begin with a rectangle sum on the finite interval $[a,b]$, with n steps of equal length h, $a+nh=b$; with the operator notation introduced in Sec. 3.2.2.

$$h\sum_{i=0}^{n-1} f(a+ih) = h\sum_{i=0}^{n-1} E^i f(a) = h\frac{E^n-1}{E-1}f(a) = \frac{(E^n-1)}{D}\frac{hD}{e^{hD}-1}f(a).$$

We apply, to the second factor, the expansion derived in Example 3.1.5, with the Bernoulli numbers B_ν. (Recall that $a+nh=b$, $E^n f(a) = f(b)$, etc.)

$$h\sum_{i=0}^{n-1} f(a+ih) = \frac{(E^n-1)}{D}\left(1+\sum_{\nu=1}^{\infty}\frac{B_\nu (hD)^\nu}{\nu!}\right)f(a) \tag{3.4.25}$$

$$= \int_a^b f(x)\,dx + \sum_{\nu=1}^{k}\frac{h^\nu B_\nu}{\nu!}\big(f^{(\nu-1)}(b) - f^{(\nu-1)}(a)\big) + R_{k+1}.$$

Here R_{k+1} is a remainder term that will be discussed thoroughly in Theorem 3.4.4. Set $h=1$, and assume that $f(b)$, $f'(b)$, $\ldots$ tend to zero as $b\to\infty$. Recall that $B_1 = -\frac{1}{2}$, $B_{2j+1}=0$ for $j>0$, and set $k=2r+1$. This yields **Euler–Maclaurin's summation formula**[57]

$$\sum_{i=0}^{\infty} f(a+i) = \int_a^{\infty} f(x)\,dx + \frac{f(a)}{2} - \sum_{j=1}^{r}\frac{B_{2j}f^{(2j-1)}(a)}{(2j)!} + R_{2r+2} \tag{3.4.26}$$

$$= \int_a^{\infty} f(x)\,dx + \frac{f(a)}{2} - \frac{f'(a)}{12} + \frac{f^{(3)}(a)}{720} - \ldots$$

[57] Leonhard Euler (1707–1783), incredibly prolific Swiss mathematician. He gave fundamental contributions to many branches of mathematics and to the mechanics of rigid and deformable bodies as well as to fluid mechanics. Colin Maclaurin (1698–1764), British mathematician. They apparently discovered the summation formula independently; see Goldstine [21, p. 84]. Euler's publication came 1738.

in a form suitable for the convergence acceleration of series of essentially positive terms. We tabulate a few coefficients related to the Bernoulli and the Euler numbers.

Table 3.4.2. *Bernoulli and Euler numbers;* $B_1 = -1/2, E_1 = 1$.

$2j$	0	2	4	6	8	10	12
B_{2j}	1	$\frac{1}{6}$	$-\frac{1}{30}$	$\frac{1}{42}$	$-\frac{1}{30}$	$\frac{5}{66}$	$-\frac{691}{2730}$
$\frac{B_{2j}}{(2j)!}$	1	$\frac{1}{12}$	$-\frac{1}{720}$	$\frac{1}{30240}$	$-\frac{1}{1209600}$	$\frac{1}{47900160}$	
$\frac{B_{2j}}{2j(2j-1)}$	1	$\frac{1}{12}$	$-\frac{1}{360}$	$\frac{1}{1260}$	$-\frac{1}{1680}$	$\frac{1}{1188}$	$-\frac{691}{360360}$
E_{2j}	1	-1	5	-61	1385	-50521	2702765

There are some obscure points in this operator derivation, but we shall consider it as a heuristic calculation only and shall not try to legitimate the various steps of it. With an appropriate interpretation, a more general version of this formula will be proved by other means in Theorem 3.4.4. A general remainder term is obtained there, if you let $b \to \infty$ in (3.4.32). You do not need it often, because the following much simpler error bound is usually applicable—but there are exceptions.

The Euler–Maclaurin expansion (on the right hand side) is typically semi-convergent only. Nevertheless a few terms of the expansion often gives startlingly high accuracy with simple calculations. For example, if $f(x)$ is completely monotonic, i.e. if

$$(-1)^j f^{(j)}(x) \geq 0, \quad x \geq a, \quad j \geq 0,$$

then the partial sums oscillate strictly around the true result; the first neglected term is then a strict error bound. (This statement also follows from the theorem below.)

Before we prove the theorem we shall exemplify how the summation formula is used in practice.

Example 3.4.9.

We return to the case of computing $S = \sum_{j=1}^{\infty} j^{-2}$. and treat it with more precision and accuracy. With $f(x) = x^{-2}$, $a = 10$, we find $\int_a^{\infty} f(x)dx = a^{-1}$, $f'(a) = -2a^{-3}$, $f'''(a) = -24a^{-5}, \ldots$. By (3.4.26), $(r = 2)$,

$$\begin{aligned}\sum_{x=1}^{\infty} x^{-2} &= \sum_{x=1}^{9} x^{-2} + \sum_{i=0}^{\infty} (10+i)^{-2} \\ &= 1.53976\,7731 + 0.1 + 0.005 + 0.00016\,6667 - 0.00000\,0333 + R_6 \\ &= 1.64493\,4065 + R_6.\end{aligned}$$

Since $f(x) = x^{-2}$ is completely monotonic (see Definition 3.2.6), the first neglected term is a strict error bound; it is less than $720\,10^{-7}/30240 < 3 \cdot 10^{-9}$. (The actual error is approximately $2 \cdot 10^{-9}$.)

Although the Euler–Maclaurin expansion, in this example; seems to converge rapidly, it is in fact, only semi-convergent for any $a > 0$, and this is rather typical. We have namely $f^{(2r-1)}(a) = -(2r)!a^{-2r-1}$, and, by Example 3.1.5, $B_{2r}/(2r)! \approx (-1)^{r+1}2(2\pi)^{-2r}$. The ratio of two successive terms is thus $-(2r+2)(2r+1)/(2\pi a)^2$, hence the modulus of terms increase when $2r + 1 > 2\pi a$.

The "rule" that one should terminate a semi-convergent expansion at the term of smallest magnitude, is in general no good for Euler–Maclaurin applications, since the high order derivatives (on the right hand side) are typically much more difficult to obtain than a few more terms in the expansion on the left hand side. Typically, you first choose r, $r \leq 3$, depending on how tedious the differentiations are, and then you choose a in order to meet the accuracy requirements.

In this example we were lucky to have access to simple closed expressions for the derivatives and the integral of f. In other cases, one may use the possibilities for the numerical integration on an infinite interval mentioned in Chapter 5. In Problem 20 (a) you find two formulas that result from the substitution of the formulas (3.3.50) that express higher derivatives in terms of central differences into the Euler–Maclaurin expansion.

An expansion of $f(x)$ into negative powers of x is often useful both for the integral and for the derivatives.

Example 3.4.10.

We consider $f(x) = (x^3 + 1)^{-1/2}$, for which the expansion

$$f(x) = x^{-3/2}(1 + x^{-3})^{-1/2} = x^{-1.5} - \frac{1}{2}x^{-4.5} + \frac{3}{8}x^{-7.5} - \dots$$

was derived and applied in Example 3.1.6. It was found that

$$\textstyle\int_{10}^{\infty} f(x)dx = 0.632410375,$$

correctly rounded, and that $f'''(10) = -4.13 \cdot 10^{-4}$ with less than 1% error. The $f'''(10)$ term in the Euler–Maclaurin expansion is thus $-5.73\,10^{-7}$, with absolute error less than $6 \cdot 10^{-9}$. Inserting this into Euler–Maclaurin's summation formula, together with the numerical values of $\sum_{n=0}^{9} f(n)$ and $\frac{1}{2}f(10) - \frac{1}{12}f'(10)$, we obtain $\sum_{n=0}^{\infty} f(n) = 3.7941\,1570 \pm 10^{-8}$. The reader is advised to work out the details as an exercise.

Example 3.4.11.

Let $f(x) = e^{-x^2}$, $a = 0$. Since all derivatives of odd order vanish at $a = 0$, then the expansion (3.4.26) may give the impression that $\sum_{j=0}^{\infty} e^{-j^2} = \int_0^{\infty} e^{-x^2}\,dx + 0.5 = 1.386\,2269$, but the sum (that is easily computed without any convergence acceleration) is actually $1.386\,3186$, hence the remainder R_{2r+2} cannot tend to zero as $r \to \infty$. The infinite Euler–Maclaurin expansion, where all terms but two are zero, *is convergent but is not valid.* Recall the distinction between the convergence and the validity of an infinite expansion, made in Sec. 3.1.2.

In this case $f(x)$ is not completely monotonic; for example, $f''(x)$ changes sign at $x = 1$. With appropriate choice of r, the general error bound (3.4.32) will tell that the error is very small, but it cannot be used for proving that it is zero—because this is not true.

The mysteries of these examples have hopefully raised the appetite for a more substantial theory, including an error bound for the Euler–Maclaurin formula. We first need some tools that are interesting in their own right.

The **Bernoulli polynomial** $B_n(t)$ is an nth degree polynomial defined by the **symbolic** relation $B_n(t) = (B + t)^n$, where the exponents of B become subscripts after the expansion according to the binomial theorem. The Bernoulli numbers B_j were defined in Example 3.1.5. Their recurrence relation (3.1.16) can be written in the form

$$\sum_{j=0}^{n-1} \binom{n}{j} B_j = 0, \quad n \geq 2,$$

or "symbolically" $(B + 1)^n = B^n = B_n$, (for the computation of B_{n-1}), $n \neq 1$, hence $B_0(t) = 1$, $B_1(t) = t + B_1 = t - 1/2$ and

$$B_n(1) = B_n(0) = B_n, \quad n \geq 2,$$

The **Bernoulli function** $\hat{B}_n(t)$ is a *piecewise polynomial* defined for $t \in \mathbf{R}$ by the equation $\hat{B}_n(t) = B_n(t - \lfloor t \rfloor)$.[58] (Note that $\hat{B}_n(t) = B_n(t)$ if $0 \leq t < 1$.)

Lemma 3.4.3.

(a) $\hat{B}'_{n+1}(t)/(n + 1)! = \hat{B}_n(t)/n!$, $(n > 0)$,
$\hat{B}_n(0) = B_n$. *(For $n = 1$ this is the limit from the right.)*

$$\int_0^1 \frac{B_n(t)}{n!}\, dt = \begin{cases} 1, & \textit{if } n = 0; \\ 0, & \textit{otherwise.} \end{cases}$$

(b) *The piecewise polynomials $\hat{B}_p(t)$ are periodic; $\hat{B}_p(t + 1) = \hat{B}_p(t)$. $\hat{B}_1(t)$ is continuous, except when t is an integer. For $n \geq 2$, $\hat{B}_n \in C^{n-2}(-\infty, \infty)$.*

(c) *The Bernoulli functions have the following (modified) Fourier expansions, $(r \geq 1)$,*

$$\frac{\hat{B}_{2r-1}(t)}{(2r-1)!} = (-1)^r 2 \sum_{n=1}^{\infty} \frac{\sin 2n\pi t}{(2n\pi)^{2r-1}}, \qquad \frac{\hat{B}_{2r}(t)}{(2r)!} = (-1)^{r-1} 2 \sum_{n=1}^{\infty} \frac{\cos 2n\pi t}{(2n\pi)^{2r}}.$$

Note that $\hat{B}_n(t)$ is an even (odd) function, when n is (even odd).

(d) $|\hat{B}_{2r}(t)| \leq |B_{2r}|$.

[58] The function $\lfloor t \rfloor$ is the floor function defined as the largest integer $\leq t$, i.e., the interger part of t. In many older and current works the symbol $[t]$ is used instead, but this should be avoided.

Proof. Statement (a) follows directly from the symbolic binomial expansion of the Bernoulli polynomials.

The demonstration of statement (b) is left for a problem. The reader is advised to draw the graphs of a few low order Bernoulli functions.

The Fourier expansion for $\hat{B}_1(t)$ follows from the Fourier coefficient formulas (3.2.7), (modified for the period 1 instead of 2π). The expansions for $\hat{B}_p(t)$, are then obtained by repeated integrations, term by term, with the use of (a). Statement (d) then follows from the Fourier expansion, because $\hat{B}_{2r}(0) = B_{2r}$. □

Remark 3.4.1. For $t = 0$ we obtain an interesting classical formula, together with a useful asymptotic approximation that was obtained in a different way in Sec. 3.1.2.

$$\sum_{n=1}^{\infty} \frac{1}{n^{2r}} = \frac{|B_{2r}|(2\pi)^{2r}}{2(2r)!}; \qquad \frac{|B_{2r}|}{(2r)!} \sim \frac{2}{(2\pi)^{2r}}. \tag{3.4.27}$$

Also note, how the rate of decrease of the Fourier coefficients is related to the type of singularity of the Bernoulli function at the integer points. (It does not help that the functions are smooth in the interval $[0, 1]$.)

The Bernoulli polynomials have a generating function that is elegantly obtained by means of the following "symbolic" calculation.

$$\sum_{0}^{\infty} \frac{B_n(y)x^n}{n!} = \sum_{0}^{\infty} \frac{(B+y)^n x^n}{n!} = e^{(B+y)x} = e^{Bx}e^{yx} = \frac{xe^{yx}}{e^x - 1}. \tag{3.4.28}$$

If the series is interpreted as a power series in the complex variable x, the convergence radius is 2π.

Theorem 3.4.4. The Euler–Maclaurin Formula.

Set $x_i = a + ih$, $x_n = b$, suppose that $f \in C^{2r+2}(a, b)$, and let $\hat{T}(a : h : b)f$ be the trapezoidal sum

$$\hat{T}(a : h : b)f = \sum_{i=1}^{n} \frac{h}{2}\big(f(x_{i-1}) + f(x_i)\big) = h\Bigg(\sum_{i=0}^{n-1} f(x_i) + \tfrac{1}{2}(f(b) - f(a))\Bigg). \tag{3.4.29}$$

Then

$$\begin{aligned}\hat{T}(a : h : b)f - \int_a^b f(x)\,dx = \frac{h^2}{12}\big(f'(b) - f'(a)\big) - \frac{h^4}{720}\big(f'''(b) - f'''(a)\big) \\ + \ldots + \frac{B_{2r}h^{2r}}{(2r)!}\big(f^{(2r-1)}(b) - f^{(2r-1)}(a)\big) + R_{2r+2}(a, h, b)f.\end{aligned} \tag{3.4.30}$$

The remainder $R_{2r+2}(a, h, b)f$ is $O(h^{2r+2})$. It is represented by an integral with a kernel of constant sign in (3.4.31). *An upper bound for the remainder is given in* (3.4.32). *The estimation of the remainder is very simple in certain important particular cases:*

- *If $f^{(2r+2)}(x)$ does not change sign in the interval $[a,b]$ then $R_{2r+2}(a,h,b)f$ has the same sign as the first neglected*[59] *term.*

- *If $f^{(2r+2)}(x)$ and $f^{(2r)}(x)$ have the same constant sign in $[a,b]$, then the value of the left hand side of* (3.4.30) *lies between the values of the partial sum of the expansion displayed in* (3.4.30) *and the partial sum with one term less.*[60]*.*

In the limit, as $b \to \infty$, these statements still hold—also for the summation formula (3.4.26)*—provided that the left hand side of* (3.4.30) *and the derivatives $f^{(\nu)}(b)$ ($\nu = 1 : 2r+1$) tend to zero, if it is also assumed that*

$$\int_a^\infty |f^{(2r+2)}(x)|\,dx < \infty.$$

Proof. To begin with we consider a single term of the trapezoidal sum, and set $x = x_{i-1} + ht$, $t \in [0,1]$, $f(x) = F(t)$. Suppose that $F \in C^p[0,1]$, where p is an *even* number.

We shall apply *repeated integration by parts*, Lemma 3.2.7, to the integral $\int_0^1 F(t)\,dt = \int_0^1 F(t)B_0(t)\,dt$. Use statement (a) of Lemma 3.4.3 in the equivalent form, $\int B_j(t)/j!\,dt = (B_{j+1}(t)/(j+1)!$

Consider the first line of the expansion in the next equation. Recall that $B_\nu = 0$ if ν is odd and $\nu > 1$. Since $B_{j+1}(1) = B_{j+1}(0) = B_{j+1}$, j will thus be odd in all non-zero terms, except for $j = 0$. Then, with no loss of generality, we assume that p is even.

$$\int_0^1 F(t)\,dt = \sum_{j=0}^{p-1} (-1)^j F^{(j)}(t) \frac{B_{j+1}(t)}{(j+1)!}\Big|_{t=0}^{1} + (-1)^p \int_0^1 F^{(p)}(t)\frac{B_p(t)}{p!}\,dt$$

$$= \frac{F(1)+F(0)}{2} + \sum_{j=1}^{p-1} \frac{-B_{j+1}}{(j+1)!}\big(F^{(j)}(1) - F^{(j)}(0)\big) + \int_0^1 F^{(p)}(t)\frac{B_p(t)}{p!}\,dt$$

$$= \frac{F(1)+F(0)}{2} - \sum_{j=1}^{p-3} \frac{B_{j+1}}{(j+1)!}\big(F^{(j)}(1) - F^{(j)}(0)\big) - \int_0^1 F^{(p)}(t)\frac{B_p - B_p(t)}{p!}\,dt.$$

The upper limit of the sum is reduced to $p-3$, since the last term (with $j = p-1$) has been moved under the integral sign, and all values of j are odd. Set $j+1 = 2k$ and $p = 2r+2$. Then k is an integer that runs from 1 to r. Hence

$$\sum_{j=1}^{p-3} \frac{B_{j+1}}{(j+1)!}(F^{(j)}(1) - F^{(j)}(0)) = \sum_{k=1}^{r} \frac{B_{2k}}{(2k)!}(F^{(2k-1)}(1) - F^{(2k-1)}(0)).$$

[59] If $r = 0$ all terms of the expansion are "neglected".

[60] Formally this makes sense for $r \geq 2$ only, but if we interpret $f^{(-1)}$ as "the empty symbol", it makes sense also for $r = 1$. If f is completely monotonic *the statement holds for every* $r \geq 1$. This is easy to apply, because simple criteria for complete monotonicity etc. are given in Sec. 3.3.6

Now set $F(t) = f(x_{i-1} + ht)$, $t \in [0,1]$. Then $F^{(2k-1)}(t) = h^{2k-1} f^{(2k-1)}(x_{i-1} + ht)$, and make abbreviations like $f_i = f(x_i)$, $f_i^{(j)} = f^{(j)}(x_i)$ etc..

$$\int_{x_{i-1}}^{x_i} f(x)\,dx = h\int_0^1 F(t)\,dt = \frac{h(f_{i-1}+f_i)}{2} - \sum_{k=1}^{r} \frac{B_{2k}h^{2k}}{(2k)!}(f_i^{(2k-1)} - f_{i-1}^{(2k-1)}) - R,$$

where R is the local remainder that is now an integral over $[x_{i-1}, x_i]$. Adding these equations, for $i = 1 : n$, yields a result equivalent to (3.4.30), namely

$$\int_a^b f(x)\,dx = \hat{T}(a:h:b)f - \sum_{k=1}^{r} \frac{B_{2k}h^{2k}}{(2k)!} f^{(2k-1)}(x)\Big|_{x=a}^{b} - R_{2r+2}(a,h,b)f,$$

$$R_{2r+2}(a,h,b)f = h^{2r+2}\int_a^b \Big(B_{2r+2} - \hat{B}_{2r+2}((x-a)/h)\Big)\frac{f^{(2r+2)}(x)}{(2r+2)!}\,dx. \quad (3.4.31)$$

By Lemma 3.4.3, $|\hat{B}_{2r+2}(t)| \le |B_{2r+2}|$, hence the kernel $B_{2r+2} - \hat{B}_{2r+2}((x-a)/h)$ has the same sign as B_{2r+2}. Suppose that $f^{(2r+2)}(x)$ does not change sign on (a,b). Then

$$sign\, f^{(2r+2)}(x) = sign\,\big(f^{(2r+1)}(b) - f^{(2r+1)}(a)\big),$$

hence $R_{2r+2}(a,h,b)f$ has the same sign as the first neglected term.
The second statement about "simple estimation of the remainder" then follows from Theorem 3.1.3, since the Bernoulli numbers (with even subscripts) have alternating signs.

If $sign\, f^{(2r+2)}(x)$ is not constant, then we note instead that

$$|B_{2r+2} - \hat{B}_{2r+2}((x-a)/h)| \le |2B_{2r+2}|,$$

and hence

$$\begin{aligned} |R_{2r+2}(a,h,b)f| &\le h^{2r+2}\frac{|2B_{2r+2}|}{(2r+2)!}\int_a^b |f^{(2r+2)}(x)|dx \\ &\approx 2\Big(\frac{h}{2\pi}\Big)^{2r+2}\int_a^b |f^{(2r+2)}(x)|dx. \end{aligned} \quad (3.4.32)$$

If $\int_a^\infty |f^{(2r+2)}(x)|dx < \infty$ this holds also in the limit as $b \to \infty$. □

Note that there are (at least) three parameters here that can be involved in *different* natural limit processes: For example, one of the parameters can tend to its limit, while the two others are kept fixed. The remainder formula (3.4.32) contains all you need for settling various questions about convergence.

- $b \to \infty$; natural when Euler–Maclaurin's formula is used as a summation formula, or for deriving an approximation formula valid when b is large.
- $h \to 0$; natural when Euler–Maclaurin's formula is used in connection with numerical integration. You see how the values of derivatives of f at the

endpoints a, b can highly improve the estimate of the integral of f, obtained by the trapezoidal rule with constant step size. Euler–Maclaurin's formula is also useful for the design and analysis of other methods for numerical integration; see Romberg's method Sec. 5.3.2.

- $r \to \infty$; $\lim_{r\to\infty} R_{2r+2}(a,h,b)f = 0$ can be satisfied only if $f(z)$ is an entire function, such that $|f^{n)}(a)| = o((2\pi/h)^n)$ as $n \to \infty$. Fortunately, this type of convergence is rarely needed in practice. With appropriate choice of b and h, the expansion is typically rapidly semi-convergent. Since the derivatives of are typically more expensive to compute than the values of f, one frequently reduces h (in integration) or increases b (in summation or integration over an infinite interval), and truncates the expansion several terms before one has reached the smallest term that is otherwise the standard procedure with alternating semi-convergent expansion.

Variations of the Euler–Maclaurin summation formula, with *finite differences instead of derivatives* in the expansion, are given in Problem 20, where you also find *a more general form of the formula*, and two more variations of it.

Euler–Maclaurin's formula can also be used for finding an algebraic expression for a finite sum; see Problem 32 or, as in the following example, for finding an expansion that determines the asymptotic behavior of a sequence or a function.

Example 3.4.12. *An expansion that generalizes Stirling's formula.*

We shall use Euler–Maclaurin formula for $f(x) = \ln x$, $a = m > 0$, $h = 1$, $b = n \ge m$. We obtain

$$\hat{T}(m:1:n)f = \sum_{i=m+1}^{n} \ln i - \tfrac{1}{2}\ln n + \tfrac{1}{2}\ln m = \ln(n!) - \tfrac{1}{2}\ln n - \ln(m!) + \tfrac{1}{2}\ln m,$$

$$f^{(2k-1)}(x) = (2k-2)!x^{1-2k}, \qquad \int_m^n f(x)\,dx = n\ln n - n - m\ln m + m.$$

Note that $\hat{T}(m:1:n)f$ and $\int_m^n f(x)\,dx$ are unbounded as $n \to \infty$, but their difference is bounded. Putting these expressions into (3.4.30), and separating the terms containing n from the terms containing m gives

$$\ln(n!) - (n+\tfrac{1}{2})\ln n + n - \sum_{k=1}^{r} \frac{B_{2k}}{2k(2k-1)n^{2k-1}} \tag{3.4.33}$$

$$= \ln(m!) - (m+\tfrac{1}{2})\ln m + m - \sum_{k=1}^{r} \frac{B_{2k}}{2k(2k-1)m^{2k-1}} - R_{2r+2}(m:1:n).$$

By (3.4.32), after a translation of the variable of integration,

$$\begin{aligned} |R_{2r+2}(m:1:n)| &\le \int_m^n \frac{|2B_{2r+2}|}{(2r+2)x^{2r+2}}\,dx \\ &\le \frac{|2B_{2r+2}|}{(2r+2)(2r+1)|m^{2r+1}|} \approx \frac{(2r)!}{\pi|2\pi m|^{2r+1}}. \end{aligned} \tag{3.4.34}$$

Now let $n \to \infty$ with fixed r, m. First, note that the integral in the error bound converges. Next, in most texts of calculus Stirling's formula is derived in the following form:

$$n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n} \quad (n \to \infty). \tag{3.4.35}$$

If you take the natural logarithm of this, it follows that the left hand side of (3.4.33) tends to $\frac{1}{2}\ln(2\pi)$[61] , and hence

$$\ln(m!) = (m + \tfrac{1}{2}) \ln m - m + \tfrac{1}{2}\ln(2\pi) + \sum_{k=1}^{r} \frac{B_{2k}}{2k(2k-1)m^{2k-1}} + R, \tag{3.4.36}$$

where a bound for R is given by (3.4.34). The numerical values of the coefficients are found in Table 3.4.4.

Almost the same derivation works also for $f(x) = \ln(x + z)$, $m = 0$, where z is a complex number, not on the negative real axis. A few basic facts about the Gamma function are needed; see details in Henrici [26, Sec. 11.11, Example 3].

The result is that *you just replace the integer m by the complex number z in the expansion* (3.4.36). According to the Handbook [1, **6.1.42**] R is to be multiplied by $K(z) = \text{upper bound}_{u \geq 0} |z^2/(u^2 + z^2)|$. For z real and positive, $K(z) = 1$, and since $f'(x) = (z + x)^{-1}$ is completely monotonic, it follows from Theorem 3.4.4 that, *in this case, R is less in absolute value than the first term neglected and has the same sign.*

It is customary to *write* $\ln \Gamma(z + 1)$ *instead of* $\ln(z!)$. The gamma function is one of the most important transcendental functions; see, e.g., the Handbook [1, **6.5**] and Lebedev[30].

This formula (with $m = z$) is useful for the practical computation of $\ln \Gamma(z+1)$. Its semi-convergence is best if $\Re z$ is large and positive. If this condition is not satisfied, the situation can easily be improved by means of logarithmic forms of the

- *reflection formula*: $\Gamma(z)\Gamma(1 - z) = \pi/\sin \pi z$,
- *recurrence formula*: $\Gamma(z + 1) = z\Gamma(z)$.

By simple applications of these formulas the computation of $\ln \Gamma(z + 1)$ for an arbitrary $z \in \mathbf{C}$ is reduced to the computation of the function for a number z', such that $|z'| \geq 17$, $\Re z' > \frac{1}{2}$, for which the total error, if $r = 5$, becomes typically less than 10^{-14}. See Problem 24.

Remark 3.4.2. As you may have noted, we write "the Euler–Maclaurin formula" mainly for (3.4.30) that is used in general theoretical discussions, or if other applications than the summation of an infinite series are the primary issue. The term "the

[61] You may ask why we refer to (3.4.35). Why not? Well, it is not necessary, because it is easy to prove that the left hand side of (3.4.33) increases with n and is bounded; it thus tends to some limit C (say). The proof that $C = \ln \sqrt{2\pi}$ *exactly* is harder, without the Wallis product idea (from 1655) that is probably used in your calculus text, or something equally ingenious or exotic. However, if you compute the right hand side of (3.4.33) for $m = 17$, $r = 5$ (say), and estimate the remainder, you will obtain C to a fabulous guaranteed accuracy, in negligible computer time after a rather short programming time. And you may then replace $\frac{1}{2}\ln 2\pi$ by your own C in (3.4.36), if you like.

Euler–Maclaurin summation formula" is mainly used in connection with (3.4.26), i.e. when the summation of an infinite series is the issue. "The Euler–Maclaurin expansion" denotes both the right hand side of (3.4.30), except for the remainder, and for the corresponding terms of (3.4.26). These distinctions are convenient for us, but they are neither important nor in general use.

Although, in this section, the main emphasis is on the application of the Euler–Maclaurin formula to the computation of sums and limits, we shall comment a little on its possibilities for other applications.

- It shows that the *global truncation error of the trapezoidal rule* for $\int_a^b f(x)\,dx$ with step size h, *has an expansion into powers of* h^2. Note that although the expansion contains derivatives at the boundary points only, the remainder requires that $|f^{(2r+2)}|$ is integrable in the interval $[a, b]$. The Euler–Maclaurin formula is thus the theoretical basis for the application of *repeated Richardson extrapolation* to the results of the trapezoidal rule, known as *Romberg's method*; see Sec5.3.2. Note that *the validity depends on the differentiability properties of* f.

- The Euler–Maclaurin formula can be used for highly accurate numerical integration when the values of some derivatives of f are known at $x = a$ and $x = b$. More about this in Chapter 5.

- Theorem 3.3.3 shows that the trapezoidal rule is second order accurate, unless $f'(a) = f'(b)$, but there exist *interesting exceptions*. Suppose that the function f is infinitely differentiable for $x \in \mathbf{R}$, and that f *has* $[a, b]$ *as an interval of periodicity*, i.e. $f(x + b - a) = f(x), \forall x \in \mathbf{R}$. Then $f^{(k)}(b) = f^{(k)}(a)$, for $k = 0, 1, 2, \ldots$, hence *every term in the Euler–Maclaurin expansion is zero* for the integral over *the whole period* $[a, b]$. One could be led to believe that the trapezoidal rule gives the exact value of the integral, but this is usually not the case; for most periodic functions f, $\lim_{r\to\infty} R_{2r+2}f \neq 0$; *the expansion converges, of course, though not necessarily to the correct result.*

We shall illuminate these amazing properties of the trapezoidal rule from different points of view in several places in this book, e.g. in Sec. 5.3. See also applications to the so-called bell sums in Problem 30.

3.4.5 Repeated Richardson Extrapolation

Let $F(h)$ denote the value of a certain quantity obtained with step length h. In many calculations one wants to know the limiting value of $F(h)$ as the step length approaches zero. However, the work to compute $F(h)$ often increases sharply as $h \to 0$. In addition, the effects of round-off errors often set a practical bound for how small h can be chosen.

Often, one has some knowledge of how the truncation error $F(h) - F(0)$ behaves when $h \to 0$. If

$$F(h) = a_0 + a_1 h^p + O(h^r), \quad h \to 0, \quad r > p,$$

where $a_0 = F(0)$ is the quantity we are trying to compute and a_1 is unknown, then a_0 and a_1 can be estimated if we compute F for two step lengths, h and qh, $q > 1$:

$$F(h) = a_0 + a_1 h^p + O(h^r),$$
$$F(qh) = a_0 + a_1 (qh)^p + O(h^r),$$

from which eliminating a_1 we get

$$F(0) = a_0 = F(h) + \frac{F(h) - F(qh)}{q^p - 1} + O(h^r). \tag{3.4.37}$$

This formula is called **Richardson extrapolation**, or the *deferred approach to the limit.*[62] Examples of this were mentioned in Chapter 1—the application of the above process to the trapezoidal rule for numerical integration (where $p = 2$, $q = 2$), and for differential equations—$p = 1$, $q = 2$ for Euler's method, $p = 2$, $q = 2$ for Runge's 2nd order method.

The term $(F(h) - F(qh))/(q^p - 1)$ is called the *Richardson correction.* It is used in in (3.4.37) for improving the result. Sometimes, however, it is used only for estimating the error. This can make sense, e.g., if the values of F are afflicted by other errors, usually irregular, suspected to be comparable in size to the correction. If the irregular errors are negligible, this error estimate is asymptotically correct. More often, the Richardson correction is used as error estimate for the improved (or extrapolated) value $F(h) + (F(h) - F(qh))/(q^p - 1)$, but this is typically a strong overestimate; the error estimate is $O(h^p)$, while the error is $O(h^r)$, $(r > p)$.

Suppose that a more complete expansion of $F(h)$ in powers of h, is known to exist,

$$F(h) = a_0 + a_1 h^{p_1} + a_2 h^{p_2} + a_3 h^{p_3} + \ldots, \quad 0 < p_1 < p_2 < p_3 < \ldots, \tag{3.4.38}$$

where the exponents are typically known, while the coefficients are unknown. Then one can *repeat the use of Richardson extrapolation* in a way described below. This process is, in many numerical problems—especially in the numerical treatment of integral and differential equations—one of the simplest ways to get results which have tolerable truncation errors. The application of this process becomes especially simple when the step lengths form a geometric series $H, H/q, H/q^2, \ldots$, where $q > 1$ and H is the **basic step length**.

Theorem 3.4.5. *Suppose that an expansion of the form of* (3.4.38)*, where* $0 < p_1 < p_2 < p_3 < \ldots$*, holds for* $F(h)$*, and set* $F_1(h) = F(h)$*,*

$$F_{k+1}(h) = \frac{q^{p_k} F_k(h) - F_k(qh)}{q^{p_k} - 1} = F_k(h) + \frac{F_k(h) - F_k(qh)}{q^{p_k} - 1}, \tag{3.4.39}$$

[62] The idea of a deferred approach to the limit is sometimes used also in the experimental sciences—for example, when some quantity is to be measured in complete vacuum (difficult or expensive to produce). It can then be more practical to measure the quantity for several different values of the pressure. Expansions analogous to equation (3.4.38) can sometimes be motivated by the kinetic theory of gases.

for $k = 1 : (n-1)$, where $q > 1$. Then $F_n(h)$ has an expansion of the form

$$F_n(h) = a_0 + a_n^{(n)} h^{p_n} + a_{n+1}^{(n)} h^{p_{n+1}} + \ldots; \quad a_\nu^{(n)} = \prod_{k=1}^{n-1} \frac{q^{p_k} - q^{p_\nu}}{q^{p_k} - 1} a_\nu. \tag{3.4.40}$$

Note that $a_\nu^{(n)} = 0$ for $\nu < n$.

Proof. Set temporarily $F_k(h) = a_0 + a_1^{(k)} h^{p_1} + a_2^{(k)} h^{p_2} + \ldots + a_\nu^{(k)} h^{p_\nu} + \ldots$. Put this expansion into the first expression on the right hand side of (3.4.39), and, substituting $k+1$ for k, put it into the left hand side. By matching the coefficients for h^{p_ν} we obtain

$$a_\nu^{(k+1)} = a_\nu^{(k)} (q^{p_k} - q^{p_\nu})/(q^{(p_k)} - 1).$$

By (3.4.38), the expansion holds for $k = 1$, with $a_\nu^{(1)} = a_\nu$. The recursion formula then yields the product formula for $a_\nu^{(n)}$. Note that $a_\nu^{(\nu+1)} = 0$, hence $a_\nu^{(n)} = 0, \forall \nu < n$. □

The product formula is for theoretical purpose. The recurrence formula is for practical use. If an expansion of the form of (3.4.38) is known to exist, the above theorem gives a way to compute increasingly better estimates of a_0. The leading term of $F_n(h) - a_0$ is $a_n^{(n)} h^{p_n}$, the exponent of h increases with n. A moment's reflection on equation (3.4.39) will convince the reader that (using the notation of the theorem) $F_{k+1}(h)$ is determined by the $k+1$ values

$$F_1(H), F_1(H/q), \ldots, F_1(H/q^k).$$

With some changes in notation we obtain the following algorithm.

Algorithm 3.4.2 Repeated Richardson extrapolation

For $m = 1 : N$, set $T_{m,1} = F(H/q^{m-1})$, and compute, for $m = 2 : N,\ k = 1 : m-1$,

$$T_{m,k+1} = \frac{q^{p_k} T_{m,k} - T_{m-1,k}}{q^{p_k} - 1} = T_{m,k} + \frac{T_{m,k} - T_{m-1,k}}{q^{p_k} - 1}, \tag{3.4.41}$$

where the second expression usually is preferred,

The computations can be set up in a scheme, where an extrapolated value in the scheme is obtained by using the quantity to its left and the correction diagonally above. (In a computer the results are simply stored in a lower triangular matrix.) According to the argument above, one continues the process, until two values *in the same row* agree to the desired accuracy, i.e.

$$|T_{m,k} - T_{m,k-1}| < Tol - CU,$$

where Tol is the permissible error, and CU is an upper bound of the irregular error, (see below). (Tol should, of course, be chosen larger than CU.) If no other error

Table 3.4.3. *Scheme for repeated Richardson extrapolation*

	$\frac{\Delta}{q^{p_1}-1}$		$\frac{\Delta}{q^{p_2}-1}$		$\frac{\Delta}{q^{p_3}-1}$	
T_{11}						
T_{21}		T_{22}				
T_{31}		T_{32}		T_{33}		
T_{41}		T_{42}		T_{43}		T_{44}

estimate is available, $\min_k |T_{m,k} - T_{m,k-1}| + CU$ is usually chosen as error estimate, even though it is typically a strong overestimate.

Typically $k = m$, and T_{mm} is accepted as the numerical result, but this is not always the case. For instance, if H has been chosen so large that the use of the basic asymptotic expansion is doubtful, then the uppermost diagonal of the extrapolation scheme contains nonsense and should be ignored, except for its element in the first column. Such a case is detected by inspection of the difference quotients in a column. If for some k, where $T_{k+2,k}$ has been computed and the modulus of the relative irregular error of $T_{k+2,k} - T_{k+1,k}$ is less than (say) 20%, *and*, most important, the difference quotient $(T_{k+1,k} - T_{k,k})/(T_{k+2,k} - T_{k+1,k})$ is is very different from its theoretical value q^{p_k}, then the uppermost diagonal is to be ignored (except for its first element). In such a case, one says that H *is outside the asymptotic regime.*

In this discussion a bound for the inherited irregular error is needed. We shall now derive such a bound. Fortunately, it turns out that the numerical stability of the Richardson scheme is typically very satisfactory, (although the total error bound for T_{mk} will never be smaller than the largest irregular error in the first column).

Denote by ϵ_1 the the column vector with the irregular errors of the initial data. We neglect the rounding errors committed during the computations.[63] Then the inherited errors satisfy the same linear recursion formula as the $T_{m,k}$, i.e.

$$\epsilon_{m,k+1} = \frac{q^{p_k}\epsilon_{m,k} - \epsilon_{m-1,k}}{q^{p_k} - 1}.$$

Denote the k'th column of errors by ϵ_k, and set $\|\epsilon_k\|_\infty = \max_m |\epsilon_{m,k}|$. Then

$$\|\epsilon_{k+1}\|_\infty \le \frac{q^{p_k} + 1}{q^{p_k} - 1}\|\epsilon_k\|_\infty.$$

Hence, for every k, $\|\epsilon_{k+1}\|_\infty \le= CU$, where $\|\epsilon_1\|_\infty = U$ and C is the infinite product

$$C = \prod_{k=1}^{\infty} \frac{q^{p_k} + 1}{q^{p_k} - 1} = \prod_{k=1}^{\infty} \frac{1 + q^{-p_k}}{1 - q^{-p_k}}$$

that converges as fast as $\sum q^{-p_k}$; the multiplication of ten factors are thus more than enough for obtaining a sufficiently accurate value of C.

[63] They are usually for various reasons of less importance. One can also *make* them smaller by subtracting a suitable constant from all initial data. This is applicable to all linear methods of convergence acceleration.

Example 3.4.13.

The most common special case is an expansion where $p_k = 2k$,

$$F(h) = a_0 + a_1 h^2 + a_2 h^4 + a_3 h^6 + \dots \tag{3.4.42}$$

this type. The headings of the columns of Table 3.4.3 then become $\Delta/3, \Delta/15, \Delta/63, \dots$. In this case *we find that* $C = \frac{5}{3} \cdot \frac{7}{15} \cdots < 2$ (after less than 10 factors).

For (systems of) ordinary differential equations there exist some general theorems, according to which the form of the asymptotic expansion (3.4.38) of the global error can be found.

- For Numerov's method for ordinary differential equations, discussed in Example 3.3.14 and Problem 28, one can show that we have the same exponents in the expansion for the global error, but $a_1 = 0$. (and the first heading disappears). We thus have the same product as above, except that the first factor disappears, *hence* $C < 2 \cdot \frac{3}{5} = 1.2$.

- For *Euler's method* for ordinary differential equations, presented in Sec. 1.4.2, $p_k = k$; the headings are $\Delta/1, \Delta/3, \Delta/7, \Delta/15, \dots$. *Hence* $C = 3 \cdot \frac{5}{3} \cdot \frac{9}{7} \cdots = 8.25$.

- For *Runge's 2nd order method*, presented in Sec. 1.4.3, the exponents are the same, but $a_1 = 0$ (and the first heading disappears). We thus have the same product as for Euler's method, except that the first factor disappears, *hence* $C = 8.25/3 = 2.75$.

In the special case that $p_j = j \cdot p$, $j = 1, 2, 3, \dots$ in (3.4.38), i.e. for expansions of the form

$$F(h) = a_0 + a_1 h^p + a_2 h^{2p} + a_3 h^{3p} + \dots, \tag{3.4.43}$$

it is not necessary that the step sizes form a geometric progression. We can choose any increasing sequence of integers $q_1 = 1, q_2, \dots, q_k$, set $h_i = H/q_i$, and use an algorithm that looks very similar to repeated Richardson extrapolation. *In cases where both are applicable, i.e. if* $p_k = p \cdot k$, $q_i = q^i$, *they are identical, otherwise they have different areas of application.*

Note that the expansion (3.4.43) is a usual power series in the variable $x = h^p$, which can be approximated by a polynomial in x. Suppose that $k+1$ values $F(H), F(H/q_2), \dots, F(H/q_k)$ are known. Then by the corollary to Theorem 3.2.1, a polynomial $Q \in \mathcal{P}_k$ is uniquely determined by the interpolation conditions

$$Q(x_i) = F(H/q_i), \quad x_i = (H/q_i)^p, \quad i = 1 : k.$$

Our problem is to find $Q(0)$. Many interpolation formulas can be used for this extrapolation. *Neville's algorithm*, which is derived in Sec. 4.2.3, is particularly convenient in this situation. After a change of notation, (4.2.27) yields the following recursion.

Algorithm 3.4.3 Neville's algorithm

For $m = 1 : N$, set $T_{m,1} = F(H/q_m)$, where $1 = q_1 < q_2 < q_3 \ldots$, is any increasing sequence of integers, and compute, for $m = 2 : N$, $k = 1 : m-1$,

$$T_{m,k+1} = T_{m,k} + \frac{T_{m,k} - T_{m-1,k}}{(q_m/q_{m-k})^p - 1} = \frac{(q_m/q_{m-k})^p T_{m,k} - T_{m-1,k}}{(q_m/q_{m-k})^p - 1}. \qquad (3.4.44)$$

The computations can be set up in a triangle matrix as for repeated Richardson extrapolations.

Example 3.4.14. *Computation of π by means of regular polygons.*

The ancient Greeks computed approximate values of the circumference of the unit circle, 2π, by inscribing a regular polygon and computing its perimeter. Archimedes considered the inscribed 96-sided regular polygon, whose perimeter is $6.28206 = 2 \cdot 3.14103$. In general, a regular n-sided polygon inscribed (circumscribed) in a circle with radius 1 has circumference $2c_n$, where $c_n = n\sin(\pi/n)$. If we put $h = 1/n$, then

$$c(h) = c_{1/h} = \frac{1}{h}\sin \pi h = \pi - \frac{\pi^3}{3!}h^2 + \frac{\pi^5}{5!}h^4 - \frac{\pi^7}{7!}h^6 + \ldots,$$

so $c(h)$ satisfy the assumptions for repeated Richardson extrapolation with $p_k = 2k$. In order to use this with $q = 2$, we first derive a recursion formula that leads from c_n to c_{2n}. Using the trigonometric formula $\cos 2x = 1 - 2\sin^2 x$, we have

$$\begin{aligned} c_{2n} = 2n \sin\frac{\pi}{2n} = n\sqrt{2\Big(1 - \cos\frac{\pi}{n}\Big)} &= n\sqrt{2 - 2\sqrt{1 - (c_n/n)^2}} \\ &= 2c_n \Big/ \sqrt{2 + 2\sqrt{1 - (c_n/n)^2}} \end{aligned}$$

(Derive this! The last transformation is made to avoid cancellation and consequential round-off errors.)

Taking $n_1 = 6$, gives $a_6 = 6/2 = 3$, and $b_6 = 6/\sqrt{3} = 3.4641\ldots$. The following table gives c_{n_m} for $n_1 = 6$, $m = 1 : 5$, computed using IEEE double precision using this recursion

m	n_m	c_{n_m}
1	6	**3**.00000000000000
2	12	**3.1**0582854123025
3	24	**3.1**3262861328124
4	48	**3.13**935020304687
5	96	**3.141**03195089051

From this we can deduce that $3.1410 < \pi < 3.1427$, or the famous, slightly weaker, rational lower and upper bounds of Archimedes $3\frac{10}{71} < \pi < 3\frac{1}{7}$. A correctly rounded value of π to twenty digits reads

$$\pi = 3.14159\,26535\,89793\,23846$$

and correct digits in the table are shown in boldface. The next table gives the Richardson scheme using the above values of c_n.

3.14110472164033			
3.14156197063157	**3.141592**45389765		
3.14159073296874	**3.14159265**045789	**3.1415926535**7789	
3.14159253350506	**3.1415926535**4081	**3.141592653589**75	**3.14159265358979**

The errors in successive columns decay as $4^{-2k}, 4^{-3k}, 4^{-4k}$, and the final number is correct to all 14 decimals shown. Hence the accuracy used in computing values in the previous table, which could be thought excessive, has been put to good use! Note that no trigonometric functions were used, only the square root.[64]

Example 3.4.15. *Application to numerical differentiation.*

Bickley's table for difference operators, i.e. Table 3.3.1 in Sec. 3.3.2, we know that

$$\begin{aligned} \frac{\delta}{h} &= \frac{2\sinh(hD/2)}{h} = D + a_2h^2D^3 + a_4h^4D^5 + \dots, \\ \mu &= \cosh(hD/2) = 1 + b_2h^2D^2 + b_4h^4D^4 + \dots, \end{aligned}$$

where the values of the coefficients are now unimportant to us. Hence

$$f'(x) - \frac{f(x+h) - f(x-h)}{2h} = Df(x) - \frac{\mu\delta f(x)}{h} \quad \text{and} \quad f''(x) - \frac{\delta^2 f(x)}{h^2}$$

have expansions into *even* powers of h. Repeated Richardson extrapolation can thus be used with step sizes H, $H/2$, $H/4, \dots$ and headings $\Delta/3$, $\Delta/15$, $\Delta/63, \dots$. For numerical examples, see problems of this section.

Richardson extrapolation can be applied in the same way to the computation of higher derivatives. Because of the division by h^k in the difference approximation of $f^{(k)}$, *irregular errors in the values of $f(x)$ are of much greater importance in numerical differentiation than in interpolation and integration.* It is therefore important to use high order approximations in numerical differentiation, so that larger values of h can be used.

[64] An extension of this example was used as a test problem for Mulprec, a package for (in principle) arbitrarily high precision floating point arithmetic in Matlab. For instance, π was obtained to 203 decimal places with 22 polygons and 21 Richardson extrapolations in less than half a minute. The extrapolations took a small fraction of this time. Nevertheless they increased the number of correct decimals from approximately 15 to 203.

Suppose that the irregular errors of the values of f are bounded in magnitude by ERB, these errors are propagated to $\mu\delta f(x)$, $\delta^2 f(x)$,... with bounds equal to ERB$/h$, 4ERB$/h^2, \ldots$. As mentioned earlier, the Richardson scheme (in the version used here) is no rascal; it multiplies the latter bounds by a factor less than 2.

Review Questions

1. (a) Aitken acceleration is based on fitting three successive terms of a given sequence $\{s_n\}$ to a certain comparison series. Which?

(b) Give sufficent conditions for the accelerated sequence $\{s'_j\}$ to converge faster than $\{s_n\}$.

(c) Aitken acceleration is sometimes applied to a thinned sequence. Why can this give a higher accuracy in the computed limit?

2. (a) State the original version of Euler's transformation for summation of an alternating series $S = \sum_{j=0}^{\infty}(-1)^j uj$, $u_j \geq 0$.

(b) State the modified Euler's transformation for this case and discuss suitable termination criteria. What is the main advantage of the modified algorithm over the classical version?

3. (a) What pieces of information appear in the Euler–Maclaurin formula? Give the generating function for the coefficients. What do you know about the remainder term?

(b) Give at least three important uses of the Euler–Maclaurin formula.

4. The Bernoulli polynomial $B_n(t)$ have a key role in the proof of the Euler–Maclaurin formula. They are defined by the symbolic relation $B_n(t) = (B + t)^n$. How is this relation to interpreted?

5. (a) Suppose that an expansion of $F(h)$

$$F(h) = a_0 + a_1 h^{p_1} + a_2 h^{p_2} + a_3 h^{p_3} + \ldots \quad 0 < p_1 < p_2 < p_3 < \ldots,$$

is known to exist. Describe how $F(0) = a_0$ can be computed by repeated Richardson extrapolation from known values of $F(h)$, $h = H, H/q, H/q^2, \ldots$, for some $q > 1$.

(b) Discuss the choice of q in the procedure in (a). What is the most common case? Give some applications of repeated Richardsonm extrapolation.

Problems and Computer Exercises

1. (a) Compute $\sum_{n=1}^{\infty} \frac{1}{(n+1)^3}$ to eight decimal places by using $\sum_{n=N}^{\infty} \frac{1}{n(n+1)(n+2)}$, for a suitable N, as a comparison series. Estimate roughly how many terms you would have to add without and with the comparison series.

Hint: You find the exact sum of this comparison series in Problem 3.3.2.

(b) Compute the sum also by Euler–Maclaurin's formula or one of its variants in Problem 20(a).

2. Study, or write yourself, programs for some of the following methods: [65]

- iterated Aitken acceleration
- modified iterated Aitken, according to (3.4.8) or an a-version.
- generalized Euler transformation
- one of the central difference variants of Euler–Maclaurin's formula, given in Problem 20(a)

The programs are needed in two slightly different versions.

Version i: For studies of the convergence rate, for a series (sequence) where one knows a sufficiently accurate value *exa* of the sum (the limit). The risk of drowning in figures becomes smaller, if you make graphical output, e.g. like Figure 3.4.1.

Version ii: For a run controlled by a tolerance, like in Algorithm 3.3.1, appropriately modified for the various algorithms. Print also i and, if appropriate, jj. If *exa* is known, it should be subtracted from the result, because it is of interest to compare *errest* with the actual error.

Comment: If you do not know *exa*, find a sufficiently good *exa* by a couple of runs with very small tolerances, before you study the convergence rates (for larger tolerances).

3. The formula for Aitken acceleration is sometimes given in the forms

$$s_n - \frac{(\Delta s_n)^2}{\Delta^2 s_n} \quad \text{or} \quad s_n - \frac{\Delta s_n \nabla s_n}{\Delta s_n - \nabla s_n}.$$

Show that these are equivalent to s'_{n+2} or s'_{n+1}, respectively, in the notations of (3.4.2). Also note that the second formula is $\lim_{p\to\infty} s'_n$ (not s'_{n+1}) in the notation of (3.4.6).

4. (a) Try iterated Aitken with thinning for $\sum_1^\infty e^{-\sqrt{n}}$, according to the suggestions after Example 3.4.3.

(b) Study the effect of small random perturbations to the terms.

5. *Oscillatory series* of the form $\sum_{n=1}^\infty c_n z^n$. Suggested examples:

$$\begin{aligned} c_n = e^{-\sqrt{n}}, \quad 1/(1+n^2), \quad 1/n,\ 1/(2n-1),\\ n/(n^2+n+1) \quad ,1/\sqrt{n}, \quad 1/\ln(n+1), \end{aligned}$$

where $z = -1,\ -0.9,\ e^{i3\pi/4},\ i,\ e^{i\pi/4},\ e^{i\pi/16}$, for the appropriate algorithms mentioned in Problem 2 above. Apply thinning. Try also classical Euler transformation on some of the cases.

Study how the convergence ratio depends on z, and compare with theoretical results. Compare the various methods with each others.

[65] We have Matlab in mind, or some other language with complex arithmetic and graphical output.

6. *Essentially positive series.* of the form $\sum_{n=1}^{\infty} c_n z^n$, where

$$c_n = e^{-\sqrt{n}}, \quad 1/(1+n^2), \quad 1/(5+2n+n^2)), \quad (n \cdot \ln(n+1))^{-2},$$
$$1/\sqrt{n^3+n}, n^{-4/3}, \quad 1/((n+1)(\ln(n+1))^2);$$

$z = 1$, 0.99, 0.9, 0.7, $e^{i\pi/16}$, $e^{i\pi/4}$, i. Use appropriate algorithms from Problem 2.

Try also Euler–Maclaurin's summation formula, or one of its variants, if you can handle the integral with good accuracy. Also try to find a good comparison series; it is not always possible.

Study the convergence rate. Try also *thinning* to the first two methods.

7. *Divergent series.* Apply, if possible, Aitken acceleration and the generalized Euler transformation to the following divergent series $\sum_1^{\infty} c_n z^n$. Compare the numerical results with the results obtained by analytic continuation, using the analytic expression for the sum as a function of z.

(a) $c_n = 1$, $z = -1$; (b) $c_n = n$, $z = -1$;

(c) c_n is an arbitrary polynomial in n; (d) $c_n = 1$, $z = i$;

(e) $c_n = 1$, $z = 2$; (f) $c_n = 1$, $z = -2$.

8. Let y_n be the Fibonacci sequence defined, in Problem 3.3.16 by the recurrence relation,

$$y_n = y_{n-1} + y_{n-2}, \quad y_0 = 0, \quad y_1 = 1.$$

Show that the sequence $\{y_{n+1}/y_n\}_0^{\infty}$ satisfies the sufficient condition for Aitken acceleration, given in the text. Compute a few terms, compute the limit by Aitken acceleration(s), and compare with the exact result.

9. When the current through a galvanometer changes suddenly, its indicator begins to oscillate toward a new stationary value s. The relation between the successive turning points v_0, v_1, v_2, ... is $v_n - s \approx A \cdot (-k)^n$, $0 < k < 1$. Determine from the following series of measurements, Aitken extrapolated values v_2', v_3', v_4' which are all approximations to s:

$$v_0 = 659, \quad v_1 = 236, \quad v_2 = 463, \quad v_3 = 340, \quad v_4 = 406.$$

10. (a) Show that the a-version of Aitken acceleration can be *iterated*, for $i = 0 : N-2$,

$$a_{i+1}^{(i+1)} = 0, \quad a_j^{(i+1)} = a_j^{(i)} - \nabla\left((a_j^{(i)})^2/\nabla a_j^{(i)}\right), \quad j = i+2 : N,$$
$$s_N^{(i+1)} = s_N^{(i)} - (a_N^{(i)})^2/\nabla a_N^{(i)}.$$

(Note that $a_j^{(0)} = a_j$, $s_j^{(0)} = s_j$.) We thus obtain N estimates of the sum s. We cannot be sure that the last estimate $s_N^{(N-1)}$ is the best, due to irregular errors in the terms and during the computations. Accept instead, e.g., the average of a few estimates that are close to each other, or do you have a better suggestion? This also gives you a (not quite reliable) error estimate.

(b) Although we may expect that the a-version of Aitken acceleration handles rounding errors better than the s-version, the rounding errors may set a limit for the accuracy of the result. It is easy to combine *thinning* with this version. How?

(c) Study or write yourself a program for the a-version, and apply it on one or two problems, where you have used the s-version earlier. Also use thinning on a problem, where it is needed. We have here considered N as given. Can you suggest a better termination criterion, or a process for continuing the computation, if the accuracy obtained is disappointing?

11. A function $g(t)$ has the form

$$g(t) = c - kt + \sum_{n=1}^{\infty} a_n e^{-\lambda_n t},$$

where c, k, a_n and $0 < \lambda_1 < \lambda_2 < \ldots < \lambda_n$ are unknown constants and $g(t)$ is known numerically for $t_\nu = \nu h$, $\nu = 0, 1, 2, 3, 4$.

Find out how to eliminate c, in such a way that a sufficient condition for estimating kh by Aitken acceleration is satisfied. Apply this to the following data, where $h = 0.1$, $g_\nu = g(t_\nu)$.

$$g_0 = 2.14789, \quad g_1 = 1.82207, \quad g_2 = 1.59763, \quad g_3 = 1.40680, \quad g_4 = 1.22784.$$

Then, estimate also c.

12. Suppose that the sequence $\{s_n\}$ satisfies the condition $s_n - s = c_0 n^{-p} + c_1 n^{-p-1} + O(n^{-p-2})$, $p > 0$, $n \to \infty$, and set

$$s'_n = s_n - \frac{p+1}{p} \frac{\Delta s_n \nabla s_n}{\Delta s_n - \nabla s_n},$$

It was stated without proof in Sec. 3.3.2 that $s'_n - s = O(n^{-p-2})$.

(a) Design an a-version of this modified Aitken acceleration, or look up in [3].

(b) Since the difference expressions are symmetrical about n one can conjecture that this result would follow from a continuous analogue with derivatives instead of differences. It has been shown [3] that this conjecture is true, but we shall not prove that. Our (easier) problem is just the continuous analogue: suppose that a function $s(t)$ satisfies the condition $s(t) - s = c_0 t^{-p} + c_1 t^{-p-1} + O(t^{-p-2})$, $p > 0$, $t \to \infty$, and set

$$y(t) = s(t) - \frac{p+1}{p} \frac{s'(t)^2}{s''(t)}.$$

Show that $y(t) - s = O(t^{-p-2})$. Formulate and prove the continuous analogue to (3.4.9).

13. (a) Consider as in Example 3.4.5, the sum $\sum n^{-3/2}$. Show that the partial sum s_n has an asymptotic expansion of the form needed in that example, with $p = -1/2$.

Hint: Apply Euler–Maclaurin's formula (theoretically).

(b) Suppose that $\sum a_n$ is convergent, and that $a_n = a(n)$. $a(z)$ is analytic function at $z = \infty$ (for example a rational function), multiplied by some power of $z - c$. Show that such a function has an expansion like (3.4.7), and that the same holds for a product of such functions.

14. *Rewriting a Fourier series for convergence acceleration.*

Consider a *real* function with the Fourier expansion $F(\phi) = \sum_{n=-\infty}^{\infty} c_n e^{in\phi}$.
(a) Show that

$$F(\phi) = c_0 + 2\Re \sum_{n=1}^{\infty} c_n z^n, \quad z = e^{i\phi}.$$

Hint: Show that $c_{-n} = \bar{c}_n$.

(b) Set $c_n = a_n - ib_n$, where a_n, b_n are real. Show that

$$\sum_{n=0}^{\infty} (a_n \cos n\phi + b_n \sin n\phi) = \Re \sum_{n=0}^{\infty} c_n z^n.$$

(c) How would you rewrite the Chebyshev series $\sum_{n=0}^{\infty} T_n(x)/(1+n^2)$?

(d) Consider also how to handle a complex function $F(\phi)$.

15. Compute and plot

$$F(x) = \sum_{n=0}^{\infty} T_n(x)/(1+n^2), \quad x \in [-1, 1].$$

Find out experimentally or theoretically how $F'(x)$ behaves near $x = 1$ and $x = -1$.

16. Compute to (say) 6 decimal places the double sum

$$S = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \frac{(-1)^{m+n}}{(m^2+n^2)} = \sum_{n=1}^{\infty} (-1)^m f(m),$$

where

$$f(m) = \sum_{n=1}^{\infty} (-1)^n (m^2+n^2)^{-1}.$$

Compute, to begin with, $f(m)$ for $m = 1 : 10$, by the generalized Euler transformation. Do you need more values of $f(m)$?

Comment: There exists an explicit formula for $f(m)$ in this case, but you can solve this problem easily without using that.

17. We use the notation of Sec. 3.4.3 (the generalized Euler transformation). Assume that $N \geq k \geq 1$, and set $n = N - k + 1$. A sum is equal to zero, if the upper index is smaller than the lower index.

(a) Prove (3.4.23) that was given without proof in the text, i.e.

$$M_{N,k-1} - M_{N-1,k-1} = z^n P^{k-2} u_{n+1}, \quad (k \geq 2).$$

Hint: By subscript transformations in the definition of $M_{N,k}$, prove that

$$M_{N,k-1} - M_{N-1,k-1} = u_{n+1}z^n + \frac{z^n}{1-z}\sum_{s=0}^{k-3}(zE-1)P^s u_{n+1}.$$

Next, show that $zE-1 = (1-z)(P-1)$, and use this to simplify the expression.

(b) Derive the formulas

$$M_{k-1,k} = \frac{1}{1-z}\sum_{s=0}^{k-2} P^s u_1; \qquad M_{N,k} = M_{k-1,k} + \sum_{j=0}^{n-1} z^j P^{k-1} u_{j+1}.$$

Comment: The first formula gives the partial sums of the classical Euler transformation. The second formula relates the k'th column to the partial sums of the power series with the coefficients $P^{k-1}u_{j+1}$.

18. (a) If $u_j = a^j$, $z = e^{i\phi}$, $\phi \in [0, \pi]$, for which real values of $a \in [0,1]$ does the series on the right of (3.4.16) converge faster than the series on the left?

(b) Find how the classical Euler transformation works if applied to the series

$$\sum z^n, \quad |z| = 1, \quad z \neq 1.$$

Compare how it works on $\sum u_n z^n$, for $u_n = a^n$, $z = z_1$, and for $u_n = 1$, $z = az_1$.

Consider similar questions for other convergence acceleration methods, that are primarily invented for oscillating sequences.

19. Compute $\sum_{k=1}^{\infty} k^{1/2}/(k^2+1)$ with an error of less than 10^{-6}. Sum the first ten terms directly. Then expand the summand in negative powers of k and use Euler–Maclaurin's summation formula. Or try a central difference variant of Euler–Maclaurin's summation formula given in the next problem; then you do not have to compute derivatives.

20. *Variations on the Euler–Maclaurin Theme*

Set $x_i = a + ih$, also for non-integer subscripts, and $x_n = b$.

Two variants with central differences instead of derivatives are interesting alternatives, if the derivatives needed in the Euler–Maclaurin Formula are hard to compute. Check a few of the coefficients on the right hand side of the formula

$$\sum_{j=1}^{\infty} \frac{B_{2j}(hD)^{2j-1}}{(2j)!} \approx \frac{\mu\delta}{12} - \frac{11\mu\delta^3}{720} + \frac{191\mu\delta^5}{60480} - \frac{2497\mu\delta^7}{3628800} + \dots. \qquad (3.4.45)$$

Use the expansion for computing the sum given in the previous problem. This formula is given by Fröberg [19, p. 220], who attributes it to Gauss.

Compare the size of its coefficients with the corresponding coefficients of the Euler–Maclaurin Formula.

Suppose that $h = 1$, and that the terms of the given series can be evaluated also for non-integer arguments. Then another variant is to compute the central differences for (say) $h = 1/2$ in order to approximate *each* derivative needed more accurately by means of (3.3.50). This leads to the formula[66]

$$\sum_{j=1}^{\infty} \frac{B_{2j} D^{2j-1}}{(2j)!} \sim \frac{\mu\delta}{6} - \frac{7\mu\delta^3}{180} + \frac{71\mu\delta^5}{7560} - \frac{521\mu\delta^7}{226800} + \cdots. \tag{3.4.46}$$

($h = 1/2$ for the central differences; $h = 1$ in the series.) Convince yourself of the reliability of the formula, either by deriving it or by testing it for (say) $f(x) = e^{0.1\,x}$.

Show that the rounding errors of the function values cause almost no trouble in the numerical evaluation of these difference corrections.

21. (a) Derive formally in a similar way the following formula for an *alternating series*. Set x_i, $h = 1$, $b = \infty$, and assume that $\lim_{x\to\infty} f(x) = 0$.

$$\sum_{i=0}^{\infty}(-1)^i f(a+i) = \tfrac{1}{2} f(a) - \frac{1}{4} f'(a) + \frac{1}{48} f'''(a) - \cdots - \frac{(2^{2r}-1)B_{2r}}{(2j)!} f^{(2r-1)}(a) - \ldots. \tag{3.4.47}$$

Of course, the integral of f is not needed in this case[67] Compare it with some of the other methods for alternating series on an example of your own choice.

(b) Derive, e.g. by operators (without the remainder R), the following more general form of the Euler–Maclaurin Formula ([1, **23.1.32**]).

$$\sum_{k=0}^{m-1} hf(a + kh + \omega h) = \int_a^b f(t)dt + \sum_{j=1}^{p} \frac{h^j}{j!} B_j(\omega)(f^{(j-1)}(b) - f^{(j-1)}(a)) + R,$$

$$R = -\frac{h^p}{p!} \int_0^1 \hat{B}_p(\omega - t) \sum_{k=0}^{m-1} f^{(p)}(a + kh + th)\, dt.$$

If you use this formula for deriving the midpoint variant in (c), you will find a quite different expression for the coefficients; nevertheless it is the same formula. Tell how this is explained in the Handbook by [1, **23.1.10**], i.e. by the "Multiplication Theorem"[68]

$$B_n(mx) = m^{n-1} \sum_{k=0}^{m-1} B_n(x + k/m), \quad n = 0, 1, 2, \ldots, \quad m = 1, 2, 3, \ldots$$

22. Prove statement (b) of the Lemma 3.4.3. (concerning the periodicity and the regularity of the Bernoulli functions).

[66] The formula is probably very old, but we have not found it in the literature.

[67] Note that the right hand side yields a finite value if f is a constant or, more generally, if f is a polynomial, although the series on the left hand side diverges. The same happens to other summation methods; see comments in the last example of Sec. 3.3.2.

[68] That formula and the remainder R are derived in Nörlund [33], p. 21 and p. 30, respectively.

23. Euler's constant is defined by $\gamma = \lim_{N\to\infty} F(N)$, where

$$F(N) = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{N-1} + \frac{1}{2N} - \ln N.$$

(a) Use the Euler–Maclaurin formula with $f(x) = x^{-1}$, $h = 1$, to show that, for any integer N

$$\gamma = F(N) + \frac{1}{12}N^{-2} - \frac{6}{720}N^{-4} + \frac{120}{30240}N^{-6} - \cdots,$$

where every other partial sum is larger than γ, and every other is smaller.

(b) Compute γ to seven decimal places, using $N = 10$, $\sum_{n=1}^{10} n^{-1} = 2.92896825$, $\ln 10 = 2.30258509$.

(c) Show how repeated Richardson extrapolation can be used to compute γ from the following values:

N	1	2	4	8
$F(N)$	0.5	0.55685	0.57204	0.57592

(d) Extend (c) to a computation, where a larger number of values of $F(N)$ have been computed as accurately as possible, and so that the final accuracy of γ is limited by the effects of rounding errors. Check the result by looking up in an accurate table of mathematical constants, e.g., in [1].

(e) Set

$$S(r) = \sum_{m=1}^{r} \sum_{n=1}^{r} (m^2 + n^2)^{-1}.$$

By a continuous analog, with a double integral instead of a double sum, you may conjecture that $S(R) \sim a \ln R + b$ as $R \to \infty$. You may even suggest a value of the parameter a. Investigate the conjecture, by computing $S(R)$ for a suitable sequence of values of R. If you find support for it, try to estimate a and b.

24. *A digression about the Gamma function.*

(a) The Handbook [1, **6.1.40**] gives an expansion for $\ln \Gamma(z)$ that agrees with formula (3.4.36) for $\ln z!$ (if we substitute z for m), except that the handbook writes $(z - \frac{1}{2}) \ln z$, where we have $(m + \frac{1}{2}) \ln m$. Explain concisely and completely that there is no contradiction here.

(b) An asymptotic expansion for computing $\ln \Gamma(z+1)$, $z \in \mathbf{C}$ is derived in Example 3.4.12. If r terms are used in the asymptotic expansion, the remainder reads:

$$K(z)\frac{(2r)!}{\pi|2\pi z|^{2r+1}} \quad K(z) = \sup_{u\geq 0} \frac{|z^2|}{|u^2 + z^2|}.$$

Set $z = x + iy$. Show the following more useful bound for $K(z)$, valid for $x > 0$,

$$K(z) \leq \begin{cases} 1, & \text{if } x \geq |y|; \\ \frac{1}{2}(x/|y| + |y|/x), & \text{otherwise.} \end{cases}$$

Find a uniform upper bound for the remainder if $r = 5$, $x \geq \frac{1}{2}$, $|z| \geq 17$.

(c) Write a program, e.g., in Matlab, for the computation of $\ln \Gamma(z+1)$. Use the reflection and recurrence formulas to transform the input value z, to another $z = x + iy$ that satisfies $x \geq \frac{1}{2}$, $|z| \geq 17$, for which this asymptotic expansion is to be used with $r = 5$.

Test the program, e.g., by computing the following quantities, and compare with their exact values, e.g.,

$$n!, \quad \Gamma(n+1/2)/\sqrt{\pi}, \quad n = 0, 1, 2, 3, 10, 20.$$

$$\left|\Gamma(\tfrac{1}{2} + iy)\right|^2 = \frac{\pi}{\cosh(\pi y)}, \quad y = \pm 10, \pm 20.$$

If the original input value has a small modulus, there is some cancellation, when when the output from the asymptotic expansion is transformed to $\ln(1 + z_{input})$, resulting in a loss of (say) 1 or 2 decimal digits.

(d) It is often much better to work with $\ln \Gamma(z)$ than with $\Gamma(z)$. For example, one can avoid exponent overflow in the calculation of a binomial coefficient or a value of the beta function, $B(z, w) = \Gamma(z)\Gamma(w)/\Gamma(z + w)$, where (say) the denominator can become too big, even if the final result is of a normal order of magnitude.

Another context where the logarithms are much preferable is in connection with interpolation, numerical differentiation etc.; for $|z| \gg 1$ $\ln \Gamma(z)$ is locally approximated by a polynomial much better than $\Gamma(z)$. The following is an example (for a hand held calculator).

Given $10! = 3628800$; compute $\Gamma(x)$ for $x = 11 : 15$. Compute $\Gamma'(13)$ by using either repeated Richardson extrapolation or the central difference expansion, in two ways:

- Use the values of $\ln \Gamma(x)$, (and multiply the logarithmic derivative by $\Gamma(13)$).
- Use directly the values of $\Gamma(x)$.

The first alternative requires a few more operations. Were they worthwhile?

25. (a) Show that

$$\binom{2n}{n} \sim \frac{2^{2n}}{\sqrt{\pi n}}, \quad n \to \infty,$$

and give an asymptotic estimate of the relative error of this approximation. Check the approximation as well as the error estimate for $n = 5$ and $n = 10$.

(b) *Random errors in a difference scheme.* We know from Example 3.3.3 that if the items y_j of a difference scheme are afflicted with errors less than ϵ in absolute value, then the inherited error of $\Delta^n y_j$ is at most $2^n \epsilon$ in absolute value. If we consider the errors as independent random variables, uniformly distributed in the interval $[-\epsilon, \epsilon]$, show that the error of $\Delta^n y_j$ has the variance $\binom{2n}{n}\frac{1}{3}\epsilon^2$, hence the standard deviation is approximately $2^n \epsilon (9\pi n)^{-1/4}$, if $n \gg 1$. Check the result on a particular case by a Monte Carlo study.

Hint: It is known from Probability theory that the variance of $\sum_{j=0}^{n} a_j \epsilon_j$ is equal to $\sigma^2 \sum_{j=0}^{n} a_j^2$, and that a random variable, uniformly distributed in the interval $[-\epsilon, \epsilon]$, has the variance $\sigma^2 = \epsilon^2/3$. Finally use (3.1.20) with $p = q = n$.

26. (a) The following table of values of a function $f(x)$ is given:

x	0.6	0.8	0.9	1.0	1.1	1.2	1.4
$f(x)$	1.820365	1.501258	1.327313	1.143957	0.951849	0.752084	0.335920

Compute using repeated Richardson extrapolation $f'(1.0)$ and $f''(1.0)$.

27. Compute an approximation to π using Richardson extrapolation with *Neville's algorithm*, based on three simple polygons, with $n = 2, 3$ and 6 sides, not in geometric progression. A 2-sided polygon can be interpreted as a diameter described up and down. Its "circumference" is thus equal to 4. Show that this gives even a little better than the result (3.14103) obtained for the 96-sided polygon without extrapolations.

28. *Numerov's method with Richardson extrapolations*[69]

(a) Show that the formula

$$h^{-2}(y_{n+1} - 2y_n + y_{n-1}) = y_n'' + a(y_{n+1}'' - 2y_n'' + y_{n-1}'')$$

is exact for polynomials of as high degree as possible, if $a = 1/12$. Show that the error has an expansion into *even* powers of h, and determine the first (typically non-vanishing) term of this expansion.

(b) This formula can be applied to the differential equation, $y'' = p(x)y$, with given initial values $y(0)$, $y'(0)$. Show that this yields the recurrence relation

$$y_{n+1} = \frac{(2 + \frac{10}{12}p_n h^2)y_n - (1 - \frac{1}{12}p_{n-1}h^2)y_{n-1}}{1 - \frac{1}{12}p_{n+1}h^2}.$$

Comment: If h is small, information about $p(t)$ is lost by outshifting in the factors $1 - \frac{1}{12}p_{n-1}h^2$ etc. (It is possible to rewrite the formulas in order to reduce the loss of information.) In the application below this causes no trouble with the step sizes suggested, in IEEE double precision. If you must use single precision, however, the outshifting may set a limit to the accuracy in the repeated Richardson extrapolation.

(c) Apply this method, together with two Richardson extrapolations in (d), to the problem of Example 3.1.1, i.e. $y'' = -xy$ with initial values $y(0) = 1$, $y'(0) = 0$, this time over the interval $0 \leq x \leq 4.8$. Denote the numerical solution by $y(x; h)$, i.e. $y_n = y(x_n; h)$.

Compute the seeds $y_1 = y(h, h)$ by the Taylor expansion in Example 3.1.1. The error of $y(0.2, 0, 2$ should be less than 10^{-10}, since we expect that the (global) errors after two Richardson extrapolations can be of that order of magnitude.

[69] See also Example 3.3.14.

Compute $y(x;h)$, $x = 0 : h : 4.8$, for $h = 0.05$, $h = 0.1$, $h = 0.2$. Store these data in a 100×3 matrix (where you must put zeros into some places). Plot $y(x;0.05)$ versus x for $x = 0 : 0.05 : 4.8$.

(d) You proved in (a) that the *local* error has an expansion containing *even powers of h* only. It can be shown that *the same is true for the global error* too. Assume (without proof) that

$$y(x,h) = y(x) + c_1(x)h^4 + c_2(x)h^6 + c_3(x)h^8 + O(h^{10}).$$

Perform the adequate repeated Richardson extrapolations to your stored results. Make semi-logarithmic plots of (the modulus of) the 4th order Richardson corrections for $x = 0 : 0.1 : 4.8$, obtained by means of $y(x;0.05)$ and $y(x;0.1)$. Plot in the same fashion the 6th order corrections for $x = 0 : 0.2 : 4.8$, obtained in the second Richardson extrapolation. The 6th order corrections are used as error estimates for the results from both these Richardson extrapolations.[70]

(e) Express, e.g., by the aid of the Handbook [1, **10.4**], the solution of this initial value problem in terms of Airy functions[71]

$$y(x) = \frac{\mathrm{Ai}(-x) + \mathrm{Bi}(-x)/\sqrt{3}}{2 \cdot 0.3550280539}.$$

Check a few of your results of the repeated Richardson extrapolation by means of [1, Table 10.11] that, unfortunately, gives only 8 decimal places.

Comment: Your results should be more accurate than that. If they are not, the reason can be that the rounding errors have a large influence, but that is not the most probable reason in this case, if IEEE double precision is used. Experience shows that it is hard to avoid programming blunders in this problem. So do not consider the theory or the rounding errors as the primary suspects. Programming errors do not always yield results that are obviously crazy; sometimes the results look reasonable, although the accuracy is much lower than it should be.

29. (a) Determine the Bernoulli polynomials $B_2(x)$ and $B_3(x)$, and find the values and the derivatives at 0 and 1. Factorize the polynomial $B_3(x)$. Draw the graphs of a few periods of $\hat{B}_i(x)$, $i = 1, 2, 3$..

(b) In an old "Cours d'Analyse", we found a "symbolic" formula, essentially

$$h\sum_{j=0}^{n-1} g'(a + jh) = g(b + hB) - g(a + hB). \tag{3.4.48}$$

The expansion of the right hand side into powers of hB, has been followed by the replacement of the powers of B by Bernoulli numbers, the resulting

[70] Although the 6th order correction yields an 8th order accurate result, it is hard to obtain an error estimate of that order without extra assumptions or extra computation.

[71] Airy functions are special functions (related to Bessel functions) with many applications to Mathematical Physics, e.g., the theory of diffraction of radio waves around the earth's surface.

expansion is not necessarily convergent, even if the first power series converges for any complex value of hB.

Show that the second expansion is equivalent to the Euler–Maclaurin formula, and that it is to be interpreted according to Theorem 3.4.4.

(c) If g is a polynomial, the expansion is finite. Show the following important formulas, and check them with known results for $k = 1 : 3$.

$$\sum_{j=0}^{n-1} j^{k-1} = \frac{(B+n)^k - B^k}{k} = \frac{B_k(n) - B_k}{k}. \tag{3.4.49}$$

Also find that (3.4.48) makes sense for $g(x) = e^{\alpha x}$, with the "symbolic" interpretation of the power series for e^{Bx}, if you accept the formula $e^{(B+\alpha)x} = e^{Bx}e^{\alpha x}$.

30. We have called $\sum a_n$ a *bell sum* if a_n as a function of n has a bell-shaped graph, and you must add many terms to get the desired accuracy. Under certain conditions you can get an accurate result by adding (say) every tenth term, and multiply this sum by 10, because both sums can be interpreted as trapezoidal approximations to the same integral, with different step size. Inspired by Euler–Maclaurin's formula, we may hope to be able to obtain high accuracy using an integer stepsize h that is (say) one quarter of the half-width of "the bell". In other words, we do not have to compute and add more than every hth term.

We shall study a class of series

$$S(t) = \sum_{n=0}^{\infty} c_n t^n/n!, \quad t \gg 1, \tag{3.4.50}$$

where $c_n > 0$, $\log c_n$ is rather slowly varying for n large; (say that) $\Delta^p \log c_n = O(n^{-p})$. Let $c(\cdot)$ be a smooth function such that $c(n) = c_n$. We consider $S(t)$ as an approximation to the integral

$$\int_0^{\infty} c(n)t^n/\Gamma(n+1)dn,$$

with a smooth and bell shaped integrand, almost like the normal frequency function, with standard deviation $\sigma \approx k\sqrt{t}$. .

(a) For $p = 1 : 5$, $t = 4^p$, plot $y = \sqrt{2\pi t}e^{-t}t^n/n!$ versus $x = n/t$, $0 \le x \le 3$; all 5 curves on the same picture.

(b) For $p = 1 : 5$, $t = 4^p$, plot $y = \ln(e^{-t}t^n/n!)$ versus $x = (n-t)/\sqrt{t}$, $\max(0, t - 8\sqrt{t}) \le n \le t + 8\sqrt{t}$; all 5 curves on the same picture. Give bounds for the error committed if you neglect the terms of the series $e^{-t}\sum_0^{\infty} t^n/n!$, which are cut out in your picture.

(c) With the same notation as in (b), use Stirling's asymptotic expansion to show theoretically that

$$\frac{e^{-t}t^n}{n!} = \frac{e^{-x^2/2}\big(1 + O(1/\sqrt{t})\big)}{\sqrt{2\pi t}}, \tag{3.4.51}$$

for $t \to \infty$, where the $O(1/\sqrt{t})$-term depends on x. Compare this with the plots.

Comment: If you are familiar with Probability, you recognize that this is related to the normal approximation to the Poisson distribution. It is well known that the mean is t, and the standard deviation is $\sqrt{t}$.

If you are familiar with Mathematical Physics, you see the resemblance to the *saddle point method*, if you interpret the sum of terms like the left hand side (from $n = 0$ to ∞) as an approximation to an integral with stepsize $\Delta n = 1$, i.e.

$$e^{-t}\int_0^\infty t^n/\Gamma(n+1)dn \sim \int_{-\infty}^{\infty} \exp(-x^2/2)/\sqrt{2\pi}dx = 1, \quad (t \to \infty).$$

(Note that $dx = dn/\sqrt{t}$.) A crude approximation for (3.4.50) is $S(t) \approx c(t)e^t$. We aim, however, at higher accuracy than is common when these approximations are used in Probability and Mathematical Physics, We think, for example, of situations where the result is to be used in a calculation where cancellation causes many digits to be lost and a decent relative accuracy is needed in what will be left.

(d) Test these ideas by making numerical experiments with the series

$$e^{-t}\sum_{n\in\mathcal{N}} t^n/n!,$$

where $\mathcal{N} = \{\text{round}(t - 8\sqrt{t}) : h : \text{round}(t + 8\sqrt{t})\}$, for some integers h in the neighborhood of suitable fractions of $\sqrt{t}$, inspired by the outcome of the experiments. Do this for t =1000, 500, 200, 100, 50, 30. Compare with the exact result, and see how the trapezoidal error depends on h, and try to formulate an error estimate that can be reasonably reliable, in cases where the answer is not known. How large must t be, in order that it should be permissible to choose $h > 1$ if you want (say) 6 correct decimals?

(e) Compute, with an error estimate, $e^{-t}\sum_{n=1}^{\infty} t^n/(n\cdot n!)$, with 6 correct decimals for the values of t mentioned in (d). You can also check your result with tables and formulas in the Handbook [1, Ch. 5].

31. If you have a good program for generating primes, denote the nth prime by p_n, and try convergence acceleration to series like

$$\sum \frac{(-1)^n}{p_n}, \quad \sum \frac{1}{p_n^2},$$

or what have you? Due to the irregularity of the sequence of primes, you cannot expect the spectacular accuracy of the previous examples, but it can be fun to see how these methods work, e.g., in combination with some comparison series derived from asymptotic results about primes. The simplest one reads $p_n \sim n \ln n$, $(n \to \infty)$, which is equivalent to the classical prime number theorem.

32. *A summation formula based on the Euler numbers*

(a) The Euler numbers E_n were introduced by (3.1.19). The first values read $E_0 = 1$, $E_2 = -1$, $E_4 = 5$, $E_6 = -61$. They are all integers (Problem 3.1.7c). $E_n = 0$ for odd n, and the sign is alternating for even n. Their generating function reads

$$\frac{1}{\cosh z} = \sum_{j=0}^{\infty} \frac{E_j z^j}{j!}.$$

(a) Show, e.g., by means of operators the following expansion

$$\sum_{k=m}^{\infty} (-1)^{k-m} f(k) \approx \sum_{p=0}^{q} \frac{E_{2p} f^{(2p)}(m - \frac{1}{2})}{2^{2p+1}(2p)!} \tag{3.4.52}$$

Comment: No discussion of convergence etc. is needed; the expansion behaves much like the Euler–Maclaurin expansion, and so does the error estimation; see, e.g., [16].

The coefficient of $f^{(2p)}(m - \frac{1}{2})$ is approximately $2(-1)^p/\pi^{2p+1}$ when $p \gg 1$, e.g., for $p = 3$ the approximation yields $-6.622 \cdot 10^{-4}$, while the exact coefficient is $61/92160 \approx 6.619 \cdot 10^{-4}$.

(b) Apply (3.4.52) for explaining the following curious observation, reported by Borwein et al. [6].

$$\sum_{k=1}^{50} \frac{4(-1)^k}{2k-1} = 3.12159465259\ldots$$
$$(\pi = 3.14159265359\ldots).$$

Note that only three digits disagree. There are several variations on this theme. Borwein et al. actually displayed the case with 40 decimal places based on 50,000 terms. Make “an educated guess” concerning how few digits disagreed.

3.5 Continued Fractions and Padé Approximants

3.5.1 Continued Fractions

Some functions cannot be well approximated by a power series, but can well be approximated by a quotient of power series. In order to study such approximations we first introduce algebraic **continued fractions**. Let r be a number and set

$$r = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 +}}} \ldots = b_0 + \frac{a_1}{b_1 +} \frac{a_2}{b_2 +} \frac{a_3}{b_3 +} \ldots, \tag{3.5.1}$$

where the second expression is a convenient compact notation. If the number of terms is infinite, r is called an *infinite continued fraction.* The terminating fraction

$$r_n = \frac{p_n}{q_n} = b_0 + \frac{a_1}{b_1+}\,\frac{a_2}{b_2+}\cdots\frac{a_n}{b_n} \tag{3.5.2}$$

is called *the nth approximant* of the continued fraction. This can be evaluated *backwards* in n divisions using the recurrence: Set $r = y_0$, where

$$y_n = b_n, \qquad y_{i-1} = b_{i-1} + a_i/y_i, \quad i = n : -1 : 1, \tag{3.5.3}$$

It can happen that in an intermediate step the denominator y_i becomes zero and $y_{i-1} = \infty$. This does no harm if you proceed in the next step when you divide by y_{i-1} the result is set equal to 0. If it happens in the last step, the result is ∞.[72]

A drawback of evaluating an infinite continued fraction expansion by the backwards recursion (3.5.3) is that you have decide where to stop in advance. The following theorem shows how *forwards* (or top down) evaluation can be achieved.

Theorem 3.5.1.

Consider the continued fraction (3.5.1)*. For $n \geq 1$, $r_n = p_n/q_n$, where p_n, q_n satisfies the* **recursion formula**

$$p_n = b_n p_{n-1} + a_n p_{n-2}, \quad p_{-1} = 1, \quad p_0 = b_0, \tag{3.5.4}$$

$$q_n = b_n q_{n-1} + a_n q_{n-2}, \quad q_{-1} = 0, \quad q_0 = 1. \tag{3.5.5}$$

Another useful formula reads

$$p_n q_{n-1} - p_{n-1} q_n = (-1)^{n-1} a_1 a_2 \cdots a_n. \tag{3.5.6}$$

Proof. We prove the recursion formulas by induction. First, for $n = 1$, we obtain

$$\frac{p_1}{q_1} = \frac{b_1 p_0 + a_1 p_{-1}}{b_1 q_0 + a_1 q_{-1}} = \frac{b_1 b_0 + a_1}{b_1 + 0} = b_0 + \frac{a_1}{b_1} = r_1.$$

Next, assume that the formulas are valid up to p_{n-1}, q_{n-1}, for *every* continued fraction. Note that p_n/q_n can be obtained from p_{n-1}/q_{n-1}, by the substitution of $b_{n-1} + a_n/b_n$ for b_{n-1}. Hence

$$\begin{aligned}\frac{p_n}{q_n} &= \frac{(b_{n-1} + a_n/b_n)p_{n-2} + a_{n-1}p_{n-3}}{(b_{n-1} + a_n/b_n)q_{n-2} + a_{n-1}q_{n-3}} = \frac{b_n(b_{n-1}p_{n-2} + a_{n-1}p_{n-3}) + a_n p_{n-2}}{b_n(b_{n-1}q_{n-2} + a_{n-1}q_{n-3}) + a_n q_{n-2}}\\ &= \frac{b_n p_{n-1} + a_n p_{n-2}}{b_n q_{n-1} + a_n q_{n-2}}.\end{aligned}$$

This shows that the formulas are valid also for p_n, q_n. The proof of equation (3.5.6) is left for Problem 2. □

[72] Note that this works automatically in IEEE arithmetic, because of the rules of infinite arithmetic; see Sec. 2.2.3!

Note that since the denominators and numerators of the approximants satisfy a three term recurrence relation they can be evaluated by Clenshaw's algorithm. It is sometimes convenient to write the recursion formulas in matrix form; see Problem 2.

If we substitute $a_n x$ for a_n in (3.5.4)–(3.5.5) then $p_n(x)$ and $q_n(x)$ become polynomials in x of degree n and $n-1$, respectively.

Example 3.5.1.

Consider the following finite continued fraction

$$r(x) = 7 - \frac{3}{x-2-}\,\frac{1}{x-7+}\,\frac{10}{x-2-}\,\frac{2}{x-3}.$$

The algorithm in Theorem 3.5.1 can be used to convert this to rational function form

$$r(x) = \frac{(((7x-101)x+540)x-1204)x+958}{(((x-14)x+72)x-151)x+112}.$$

As indicated, the numerator and denominator can then be evaluated by Horner's rule. The backwards evaluation of the continued fraction form requires fewer operations than of the rational form. However, there at the four points $x = 1, 2, 3, 4$ a division by zero occurs even though $r(x)$ is well defined at these points. However, in IEEE arithmetic the continued fraction evaluates correctly at these points because of the rules of infinite arithmetic! Indeed the continued fraction form can be shown to have smaller errors for $x \in [0, 4]$ and to be immune to overflow; see Higham [27, § 27.1].

In practice the forward recursion for evaluating a continued fraction often generates very large or very small values for the numerators and denominators. There is a risk of *overflow or underflow* with these formulas. We are usually not interested in the p_n, q_n themselves, but in the ratios only. Then we can normalize p_n and q_n by multiplying them by the same factor after they have been computed. If we shall go on and compute p_{n+1}, q_{n+1}, however, *we have to multiply* p_{n-1}, q_{n-1} *by the same factor also*! One must also be careful about the numerical stability of these recurrence relations.

The formula

$$\frac{a_1}{b_1+}\,\frac{a_2}{b_2+}\,\frac{a_3}{b_3+}\cdots = \frac{k_1a_1}{k_1b_1+}\,\frac{k_1k_2a_2}{k_2b_2+}\,\frac{k_2k_3a_3}{k_3b_3+}\cdots, \tag{3.5.7}$$

where the k_i are any non-zero numbers, is known as an **equivalence transformation**. The proof of (3.5.7) is left for Problem 5. .

By the following division algorithm, a rational function can be expressed as a continued fraction that can be evaluated by relatively few arithmetic operations; see Cheney [12, *p.* 151]. Let R_0, R_1 be polynomials, and set $R = R_0/R_1$. The degree of a polynomial R_j is denoted by d_j. By successive divisions (of R_{j-1} by R_j) we obtain quotients Q_j and remainders R_{j+1} as follows. For $j = 1, 2, \ldots,$ until $d_{j+1} = 0$,

$$R_{j-1} = R_jQ_j + R_{j+1}, \quad d_{j+1} < d_j, \tag{3.5.8}$$

hence

$$R = \frac{R_0}{R_1} = Q_1 + \frac{1}{R_1/R_2} = \ldots = Q_1 + \frac{1}{Q_2+}\,\frac{1}{Q_3+}\cdots\frac{1}{Q_k}. \qquad (3.5.9)$$

By means of an equivalence transformation; see (3.5.7), this fraction can be transformed into a slightly more economic form, where the polynomials in the denominators have leading coefficient unity, while the numerators are in general different from 1.

Example 3.5.2. *Best Rational Approximations to a Real Number.*

Every positive number x can be expanded into a continued fraction with integer coefficients of the form,

$$x = b_0 + \frac{1}{b_1+}\,\frac{1}{b_2+}\,\frac{1}{b_3+}\cdots. \qquad (3.5.10)$$

Set $x_0 = x$, $p_{-1} = 1$, $q_{-1} = 0$. For $n = 0, 1, 2, \ldots$ we construct a sequence of numbers,

$$x_n = b_n + \frac{1}{b_{n+1}+}\,\frac{1}{b_{n+2}+}\,\frac{1}{b_{n+3}+}\cdots.$$

Evidently $b_n = \lfloor x_n \rfloor$, the integer part of x_n, and $x_{n+1} = 1/(x_n - b_n)$. Compute p_n, q_n, according to the recursion formulas of Theorem 3.5.1, which can be written in vector form,

$$(p_n, q_n) = (p_{n-2}, q_{n-2}) + b_n(p_{n-1}, q_{n-1}),$$

(since $a_n = 1$). See Figure 4.3.1. Stop when $|x - p_n/q_n| < Tol$ or $n > nmax$. The details are left for Problem 1.

The above algorithm has been used several times in the previous sections, where some coefficients, known to be rational, has been computed in floating point. It is also useful for finding near commensurabilities between events with different periods;[73] see Problem 1c

The German mathematician Felix Klein [28][74] gave the following illuminating description of the sequence $\{(p_n, q_n)\}$ obtained by this algorithm (adapted to our notation):

> "Imagine pegs or needles affixed at all the integral points (p_n, q_n), and wrap a tightly drawn string about the sets of pegs to the right and to the left of the ray, $p = xq$. Then the vertices of the two convex string-polygons which bound our two point sets will be precisely the points $(p_n, q_n)\ldots$, the left polygon having the even convergents, the right one the odd."

Klein also points out that "such a ray makes a cut in the set of integral points" and thus makes Dedekind's definition of irrational numbers very concrete. This

[73] One of the convergents for $\log 2/\log 3$ reads 12/19. This is in a way basic for Western Music, where 13 quints make 7 octaves, i.e. $(3/2)^{12} \approx 2^7$.

[74] Felix Christian Klein (1849–1925). He was born 25/4 1849 and delighted in pointing out that each of the day 5^2, month 2^2, and year 43^2 was the square of a prime.

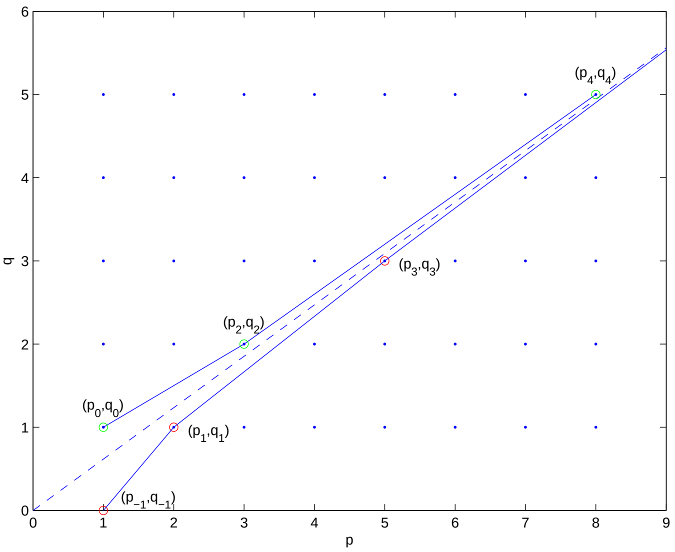

Figure 3.5.1. *Illustration to Example* 3.3.2. *The dashed line is* $\{(p,q) : xq = p\}$ *for* $x = \frac{1}{2}(\sqrt{5}+1)$.

construction; see Figure 3.5.1, illustrates in a concrete way that the successive convergents are closer to x than any numbers with smaller denominators, and that the errors alternate in sign. We omit the details of the proof that this description is correct.

Note that, since $a_j = 1, \forall j$, equation (3.5.6) reads $p_n q_{n-1} - p_{n-1} q_n = (-1)^{n-1}$. This implies that the triangle with vertices at the points $(0,0)$, (q_n, p_n), (q_{n-1}, p_{n-1}) has the smallest possible area, among triangles with integer coordinates, and hence there can be no integer points inside or on the sides of this triangle.

Theorem 3.5.2. (Seidel)[75]

Let all b_n *be positive in the continued fraction*

$$b_0 + \frac{1}{b_1+}\,\frac{1}{b_2+}\,\frac{1}{b_3+}\cdots.$$

Then this converges if and only if the series $\sum b_n$ *diverges.*

Proof. See Cheney [12, p. 184]. □

Figure 3.5.1 corresponds to the example, (see also Problem 3),

$$x = 1 + \frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\cdots \tag{3.5.11}$$

[75] Philipp Ludwig von Seidel (1821–1896) German mathematician and astronomer. In 1846 he submitted his habilitation dissertation entitled "Untersuchungen über die Konvergenz and Divergenz der Kettenbrüche

From Theorem 3.5.2 it follows that this continued fraction is convergent. Then, note that $x = 1 + 1/x$, $x > 0$, hence $x = (\sqrt{5} + 1)/2$. Note also that, by (3.5.6) with $a_j = 1$,

$$\left|x - \frac{p_n}{q_n}\right| \le \left|\frac{p_{n+1}}{q_{n+1}} - \frac{p_n}{q_n}\right| = \frac{|p_{n+1}q_n - p_n q_{n+1}|}{q_{n+1}q_n} = \frac{1}{q_{n+1}q_n} < \frac{1}{q_n^2}. \tag{3.5.12}$$

Comment: If we know or guess that a result x of a computation is a rational number with a reasonably sized denominator, although it was practical to compute it in floating point arithmetic (afflicted by errors of various types), we have a good chance to reconstruct the exact result by applying the above algorithm as a post-processing.

If we just know that the exact x is rational, without any bounds for the number of digits in the denominator and numerator, we must be conservative in claiming that the last fraction that came out of the above algorithm is the exact value of x, even if $|x - p_n/q_n|$ is very small. In fact, the fraction may depend on TOL that is to be chosen with respect to the expected order of magnitude of the error of x. If TOL has been chosen smaller than the error of x, it may, e.g., happen that the last fraction obtained at the termination is wrong, while the correct fraction (with smaller numerator and denominator) may have appeared earlier in the sequence (or it may not be there at all).

So a certain judgment is needed at the application of this algorithm. The smaller the denominator and numerator are, the more likely it is that the fraction is correct. In a serious context, it is advisable to check the result(s) by using exact arithmetic. If x is the root of an equation (or a component of the solution of a system of equations), it is typically much easier to check afterwards that a suggested result is correct than to perform the whole solution process in exact arithmetic.

Continued fractions have also important applications in Analysis; some of the best algorithms for the numerical computation of important analytic functions are based on continued fractions. We shall not give complete proofs but refer to classical books of Perron [35], Wall [47] and Henrici [25, 26].

A continued fraction is said to be equivalent to a given series, iff the sequence of convergents is equal to the sequence of partial sums. There is typically an infinite number of such equivalent fractions. The construction of the continued fraction is particularly simple if we require that the denominators $q_n = 1$, $\forall n \ge 1$. For a power series we shall thus have

$$p_n = c_0 + c_1 x + c_2 x^2 + \ldots c_n x^n, \quad n \ge 1.$$

We must *assume that* $c_j \ne 0\ \forall j \ge 1$.

We shall determine the the elements a_n, b_n by means of the recursion formulas of Theorem 3.5.1 (for $n \ge 2$) with initial conditions. We thus obtain the following equations,

$$\begin{aligned} p_n &= b_n p_{n-1} + a_n p_{n-2}; \quad p_0 = b_0, \quad p_1 = b_0 b_1 + a_1, \\ 1 &= b_n + a_n; \qquad b_1 = 1. \end{aligned}$$

The solution reads $b_0 = p_0 = c_0$, $b_1 = 1$, $a_1 = p_1 - p_0 = c_1 x$, and for $n \geq 2$,

$$\begin{aligned} a_n &= (p_n - p_{n-1})/(p_{n-2} - p_{n-1}) = -xc_n/c_{n-1}; \\ b_n &= 1 - a_n = 1 + xc_n/c_{n-1}; \end{aligned}$$

$$c_0 + c_1 x + \ldots + c_n x^n \ldots = c_0 + \frac{xc_1}{1-} \, \frac{xc_2/c_1}{1 + xc_2/c_1 -} \cdots \frac{xc_n/c_{n-1}}{1 + xc_n/c_{n-1} -} \cdots$$

Of course, an equivalent continued fraction gives by itself *no convergence acceleration, just because it is equivalent.* We shall therefore leave the subject of continued fractions equivalent to a series, after showing two instances of the numerous pretty formulas that can be obtained by this construction.

For

$$f(x) = e^x = 1 + x + x^2/2! + x^3/3! + \ldots$$

and

$$f(x) = \frac{\arctan \sqrt{x}}{\sqrt{x}} = 1 - x/3 + x^2/5 - x^3/7 + \ldots,$$

we obtain for $x = -1$ and $x = 1$, respectively, after simple equivalence transformations,

$$e^{-1} = 1 - \frac{1}{1+} \, \frac{1}{1+y} = \frac{1}{2+y} \;\Rightarrow\; e = 2 + y, \quad \text{where} \quad y = \frac{2}{2+} \, \frac{3}{3+} \, \frac{4}{4+} \, \frac{5}{5+} \ldots;$$

$$\frac{\pi}{4} = \frac{1}{1+} \, \frac{1}{2+} \, \frac{9}{2+} \, \frac{25}{2+} \, \frac{49}{2+} \ldots.$$

There exist, however, other methods to make a correspondence between a power series and a continued fraction. Some of them lead to a considerable convergence acceleration that often makes continued fractions very efficient for the *numerical computation of functions.* We shall return to such methods in Sec. 3.5.2.

Gauss developed a continued fraction for the ratio of two hypergeometric functions (see (3.1.13))

$$\frac{F(a, b+1, c+1; z)}{F(a, b, c; z)} = \frac{1}{1+} \, \frac{a_1 z}{1+} \, \frac{a_2 z}{1+} \, \frac{a_3 z}{1+} \ldots, \tag{3.5.13}$$

$$a_{2n+1} = \frac{(a+n)(c-b+n)}{(c+2n)(c+2n+1)}, \qquad a_{2n} = \frac{(b+n)(c-a+n)}{(c+2n-1)(c+2n)}. \tag{3.5.14}$$

If in (3.5.13) we set $b = 0$, then $F(a, b, c; z) = 1$, and we obtain a continued fraction for $F(a, b+1, c+1; z)$. From this many continued fractions for elementary functions can be derived, such as

$$\ln(1+z) = \frac{z}{1+} \, \frac{z}{2+} \, \frac{z}{3+} \, \frac{2^2 z}{4+} \, \frac{2^2 z}{5+} \, \frac{3^2 z}{6+} \ldots. \tag{3.5.15}$$

$$\frac{1}{2} \ln\left(\frac{1+z}{1-z}\right) = \frac{z}{1-} \, \frac{z^2}{3-} \, \frac{2^2 z^2}{5-} \, \frac{3^2 z^2}{7-} \, \frac{4^2 z^2}{9-} \cdots \tag{3.5.16}$$

$$(3.5.17)$$

$$\arctan z = \frac{z}{1+}\,\frac{z^2}{3+}\,\frac{2^2z^2}{5+}\,\frac{3^2z^2}{7+}\,\frac{4^2z^2}{9+}\cdots \tag{3.5.18}$$

$$\tan z = \frac{z}{1-}\,\frac{z^2}{3-}\,\frac{z^2}{5-}\,\frac{z^2}{7-}\cdots \tag{3.5.19}$$

$$\tanh z = \frac{e^{2z}-1}{e^{2z}+1} = \frac{z}{1+}\,\frac{z^2}{3+}\,\frac{z^2}{5+}\,\frac{z^2}{7+}\cdots. \tag{3.5.20}$$

These expansions can be used also for complex values of z. In fact the fraction for the logarithm can be used in the whole complex plane except in the intervals $(-\infty,-1]$ and $[1,\infty)$. For $\arctan z$, there are similar branch cuts on the imaginary axis. The convergence is slow, when z is near a cut. For an elementary function like these, a program can use some properties of the functions for moving z to a domain, where the continued fraction converges rapidly.

The expansion for $\tan z$ is valid everywhere, except in the poles. In all these cases the region of convergence as well as the speed of convergence is considerably larger than for the power series expansions. For example, the 6'th convergent for $\tan\pi/4$ is almost correct to 11 decimal places.

Example 3.5.3.

Consider the continued fraction for $\ln(1+z)$ and set $z = 1$. The successive approximations to $\ln 2 = 0.69314\,71806$ are:

1/1	2/3	7/10	36/52	208/300	1572/2268	12876/18576
1.000000	0.66667	0.700000	0.692308	0.69333	0.693122	0.693152

Note that the fraction give alternatively upper and lower bounds for $\ln 2$. It can be shown that this is the case when the elements of the continued fraction are positive. To get the accuracy of the last approximation above would require as many as 50,000 terms of the series $\ln 2 = \ln(1+1) = 1 - 1/2 + 1/3 - 1/4 + \cdots$.

Example 3.5.4.

A collection of formulas concerning the important incomplete Gamma function is found in the Handbook [1, **6.5**]. For the sake of simplicity we assume that $x > 0$, although the formulas can be used also in an appropriately cut complex plane. The parameter a may be complex in $\Gamma(a,x)$.[76]

$$\Gamma(a,x) = \int_x^\infty t^{a-1}e^{-t}\,dt, \quad \Gamma(a,0) = \Gamma(a),$$

$$\gamma(a,x) = \Gamma(a) - \Gamma(a,x) = \int_0^x t^{a-1}e^{-t}\,dt, \quad \Re a > 0,$$

$$\Gamma(a,x) = e^{-x}x^a\Big(\frac{1}{x+}\,\frac{1-a}{1+}\,\frac{1}{x+}\,\frac{2-a}{1+}\,\frac{2}{x+}\cdots\Big), \tag{3.5.21}$$

[76] There are plenty of other notations for this function.

$$\gamma(a,x) = e^{-x}x^a\Gamma(a)\sum_{n=0}^{\infty}\frac{x^n}{\Gamma(a+1+n)}.$$

We mention these functions, because they have many applications. Several other important functions can, by simple transformations, be brought to particular cases of this function, e.g., the normal probability function, the chi-square probability function, the exponential integral, and the Poisson distribution.

Continued fractions like these can often be derived by a theorem of Stieltjes[77], which relates continued fractions to orthogonal polynomials that satisfy a recurrence relation of the same type as the one given above. Another method of derivation is the Padé approximation, studied in the next section, that yields a rational function. Both techniques can be looked upon as a *convergence acceleration of an expansion into powers of z or z^{-1}.*

3.5.2 Padé Approximants.

The Padé[78] approximants are a particular type of rational approximations to a function $f(z)$ defined by a power series, The idea is to match the coefficients in the given series as far as possible with a rational approximation $P(x)/Q(x)$. Consider the example (Baker [2])

$$f(x) = \left(\frac{1+2x}{1+x}\right)^{1/2} = 1 + \frac{1}{2}x - \frac{5}{8}x^2 + \frac{13}{16}x^3 - \dots .$$

The first three coefficients are matched by the rational approximation

$$f_{11}(x) = \frac{1+7x/4}{1+5x/4} = 1 + \frac{1}{2}x - \frac{5}{8}x^2 + \frac{25}{32}x^3 - \dots .$$

Note that $f_{11}(x)$ has the value 1.4 for $x=\infty$, which agrees well with the limit $\sqrt{2}$ for $f(x)$. This is in contrast to the behavior of the Taylor series for $f(x)$, which does not converge for $x \geq 1/2$.

We now give a general definition of Padé approximants.

Definition 3.5.3.

The (m,n) Padé approximant associated with

$$f(z) = \sum_{i=0}^{\infty} c_i z^i. \tag{3.5.22}$$

[77] Thomas Jan Stieltjes (1856–1894), mathematician born in the Netherlands. On recommendations from his friend and colleague Charles Hermite in Paris he became docent 1886 and professor 1889 at the university in Toulouse, France.

[78] Henri Eugène Padé (1863–1953) a French mathematician, wrote his thesis under Charles Hermite's supervision.

is, if it exists*, defined to be a rational function*

$$f_{m,n}(z) = \frac{P_{m,n}(z)}{Q_{m,n}(z)} \equiv \frac{\sum_{j=0}^{m} p_j z^j}{\sum_{j=0}^{n} q_j z^j}, \qquad q_0 = 1, \quad (3.5.23)$$

that satisfies

$$r_{m,n}(z) = f(z) - f_{m,n}(z) = Rz^{m+n+1} + O(z^{m+n+2}), \quad z \to 0. \qquad (3.5.24)$$

The Padé approximants to e^z are important because of their relation to methods for solving differential equations. Padé arranged the approximants $f_{m,n}(z)$, $m, n = 0, 1, 2, \ldots.$ in a semi-infinite table. The following is part of the **Padé table** for the exponential function $f(z) = e^z$.

$$\begin{array}{ccc}
\dfrac{1}{1} & \dfrac{1+z}{1} & \dfrac{1+z+\frac{1}{2}z^2}{1} \\
\dfrac{1}{1-z} & \dfrac{1+\frac{1}{2}z}{1-\frac{1}{2}z} & \dfrac{1+\frac{2}{3}z+\frac{1}{6}z^2}{1-\frac{1}{3}z} \\
\dfrac{1}{1-z+\frac{1}{2}z^2} & \dfrac{1+\frac{1}{3}z}{1-\frac{2}{3}z+\frac{1}{6}z^2} & \dfrac{1+\frac{1}{2}z+\frac{1}{12}z^2}{1-\frac{1}{2}z+\frac{1}{12}z^2}
\end{array}$$

The Padé approximants for e^z were given explicitly by Padé (1892) in his thesis. They are

$$P_{m,n}(z) = \sum_{j=0}^{m} \frac{(m+n-j)!\, m!}{(m+n)!\, (m-j)!} \frac{z^j}{j!}, \qquad (3.5.25)$$

$$Q_{m,n}(z) = \sum_{j=0}^{n} \frac{(m+n-j)!\, n!}{(m+n)!\, (n-j)!} \frac{(-z)^j}{j!}, \qquad (3.5.26)$$

with the error

$$r_{m,n}(z) = e^z - \frac{P_{m,n}(z)}{Q_{m,n}(z)} = (-1)^n \frac{m!n!}{(m+n)!(m+n+1)!} z^{m+n+1} + O(z^{m+n+2}). \qquad (3.5.27)$$

Note that $P_{m,n}(z) = Q_{m,n}(-z)$, which reflects the property that $e^{-z} = 1/e^z$. Indeed, the nominator and denominator polynomials can be shown to approximate $e^{z/2}$ and $e^{-z/2}$, respectively.

There are several reasons for preferring the diagonal Padé approximants ($m = n$). For these

$$p_j = \frac{(2m-j)!\, m!}{(2m)!\, (m-j)!j!}, \qquad q_j = (-1)^j p_j, \quad j = 0 : m. \qquad (3.5.28)$$

The coefficients satisfy the recursion

$$p_0 = 1, \quad p_{j+1} = \frac{(m-j)p_j}{(2m-j)(j+1)}, \quad j = 0 : m-1. \qquad (3.5.29)$$

For the diagonal Padé approximants the error $R_{m,n}(z)$ satisfy $|R_{m,n}(z)| < 1$, for $\Re z < 0$. This is an important property in applications to solving differential equations.[79] To evaluate a diagonal Padé approximant of even degree we write

$$\begin{aligned} P_{2m,2m}(z) &= p_{2m}z^{2m} + \cdots + p_2z^2 + p_0 \\ &\quad + z(p_{2m-1}z^{2m-2} + \cdots + p_3z^2 + p_1) = u(z) + v(z). \end{aligned}$$

and evaluate $u(z)$ and $v(z)$ separately. Then $Q_{2m}(z) = u(z) - v(z)$. A similar splitting can be used for an odd degree.

It was remarked in Sec. 2.2.4 that in order to compute the exponential function a range reduction should first be performed. If an integer k is determined such that

$$z^* = z - k\ln 2, \quad |z^*| \in [0, \ln 2] \tag{3.5.30}$$

then $\exp(z) = \exp(z^*) \cdot 2^k$. Hence only an approximation of $\exp(z)$ for $|z| \in [0, \ln 2]$ is needed; see Problem 5.

We now consider how to determine the Padé approximants in the general case.

Theorem 3.5.4.

Let $f(z)$ be a function defined by the power series (3.5.22). *The coefficients q_j $j = 1 : n$, of the denominator of the Padé approximant $f_{m,n}(z)$ are determined by the linear system,*

$$\sum_{j=1}^{n} c_{i-j}q_j + c_i = 0, \quad i = m+1 : m+n, \tag{3.5.31}$$

where we set $c_i = 0$ for $i < 0$, provided that this linear system has a unique solution. Further, the coefficients of the numerator are

$$\sum_{j=0}^{k} c_{i-j}q_j = p_i, \quad i = 0 : m, \quad k = \min(i, n), \tag{3.5.32}$$

and the error constant R in (3.5.24) *reads*

$$R = \sum_{j=0}^{k} c_{i-j}q_j, \quad i = m+n+1.$$

Proof. Insert (3.5.22) and (3.5.24) into (3.5.23) and multiply both sides by the denominator:

$$\Big(\sum_{l=0}^{\infty} c_l z^l + Rz^{m+n+1} + O(z^{m+n+2})\Big) \sum_{j=0}^{n} q_j z^j = \sum_{i=0}^{m} p_i z^i.$$

[79] Diagonal Padé approximants are used also for the evaluation of the matrix exponential e^A, $A \in \mathbf{R}^{n\times n}$; see Chapter 9.

Match the coefficients of z^i, $i = 0 : m + n + 1$, and remember that $q_0 = 1$:

$$\sum_{j=0}^{n} c_{i-j} q_j = \begin{cases} p_i, & \text{if } 0 \leq i \leq m; \\ 0, & \text{if } m+1 \leq i \leq m+n; \\ R, & \text{if } i = m+n+1. \end{cases}$$

The statements follow from this. □

Note that $f_{m,n}$ uses c_l for $l = 0 : m + n$ only; R uses c_{m+n+1} also. So, if c_l is given for $l = 0 : r$ then $f_{m,n}$ is defined for $m + n \leq r$, $m \geq 0$, $n \geq 0$.

There is an "if" in the theorem. There are in fact simple exceptional situations, where the linear system (3.5.31) is singular. The system can be written in more detail as

$$\begin{pmatrix} c_{m-n+1} & c_{m-n+2} & \cdots & c_m \\ c_{m-n+2} & c_{m-n+3} & \cdots & c_{m+1} \\ \vdots & \vdots & \cdots & \vdots \\ c_m & c_{m+1} & \cdots & c_{m+n-1} \end{pmatrix} \begin{pmatrix} q_n \\ q_{n-1} \\ \vdots \\ q_1 \end{pmatrix} = - \begin{pmatrix} c_{m+1} \\ c_{m+2} \\ \vdots \\ c_{m+n} \end{pmatrix}$$

where $c_i = 0$, $i < 0$. Note the system matrix has constant elements along the anti-diagonals. Such matrices are called **Hankel matrices**. It can be shown that singular cases occur in square blocks of the Padé table, where all the approximants are equal. This property, investigated by Padé, is known as the *block structure of the Padé table*. A Padé table where all the approximants are different is called **normal**. Otherwise it is called **non-normal**.

We shall indicate, how such singular situations can often be avoided by a more reasonable formulation of the request. These matters are discussed more thoroughly, e.g., in Cheney [12, Chap. 5].

Example 3.5.5.

Let for $f(z) = \cos z = 1 - \frac{1}{2}z^2$ and try to find

$$f_{1,1}(z) = (p_0 + p_1 z)/(q_0 + q_1 z), \quad q_0 = 1.$$

The coefficient matching according to the theorem, yields the equations,

$$p_0 = q_0 = 1, \qquad p_1 = q_1, \qquad -\frac{1}{2}q_0 = 0.$$

The last equation contradicts the condition that $q_0 = 1$. This single contradictory equation is in this case the "system" (3.5.31).

If this equation is ignored, we obtain $f_{1,1}(z) = (1 + q_1 z)/(1 + q_1 z) = 1$, with error $\approx \frac{1}{2}z^2$, in spite that we asked for an error that is $O(z^{m+n+1}) = O(z^3)$. If we instead allow that $q_0 = 0$, then $p_0 = 0$, and we obtain the same final result, since $f_{1,1}(z) = q_1 z/(q_1 z) = 1$.

In a sense, this singular case corresponds to a rather stupid request: we ask to approximate the even function $\cos z$ by a rational function where the numerator and the denominator end with odd powers of z. One should, of course, ask for the approximation by a rational function of z^2. What would you do, if $f(z)$ is an *odd* function?

Imagine a case where $f_{m-1,n-1}(z)$ happens to be a more accurate approximation to $f(z)$ than usual, say that $f_{m-1,n-1}(z) - f(z) = O(z^{m+n+1})$. (For instance, let $f(z)$ be the ratio of two polynomials of degree $m-1$ and $n-1$, respectively.) Let b be an arbitrary number, and choose

$$Q_{m,n}(z) = (z+b)Q_{m-1,n-1}(z),$$
$$P_{m,n}(z) = (z+b)P_{m-1,n-1}(z).$$

Then

$$\begin{aligned} f_{m,n}(z) &= P_{m,n}(z)/Q_{m,n}(z) \\ &= P_{m-1,n-1}(z)/Q_{m-1,n-1}(z) = f_{m-1,n-1}(z), \end{aligned}$$

which is an $O(z^{m+n+1})$-accurate approximation to $f(z)$. Hence our request for this accuracy is satisfied by more than one pair of polynomials, $P_{m,n}(z)$, $Q_{m,n}(z)$, since b is arbitrary. This is impossible, unless the system (3.5.31) (that determines $Q_{m,n}$) is singular.

This illustrates another type of situations where the singular case occurs. Numerically, a similar situation occurs in a natural way, when one wants to approximate $f(z)$ by $f_{m,n}(z)$, although already $f_{m-1,n-1}(z)$ would represent $f(z)$ as well as possible with the limited precision of the computer. In this case we must expect the system (3.5.31) to be very close to a singular system. A reasonable procedure for handling this is to compute the Padé approximants for a sequence of increasing values of m, n, to estimate the condition numbers and to stop when it approaches the reciprocal of the machine unit. This illustrates a fact of some generality. *Unnecessary numerical trouble can be avoided by means of a well designed termination criterion.*

For $f(z) = -\ln(1-z)$, we have $c_l = 1/l$, $l > 0$. When $m = n$ the matrix of the system (3.5.31) turns out to be the notorious Hilbert matrix (with permuted columns), for which the condition number grows exponentially; see Example 2.4.7. (The elements of the usual Hilbert matrix are $a_{ij} = 1/(i+j-1)$.)

Example 3.5.6.

The Padé approximations $f_{m,n}(z)$ and the corresponding error terms were computed by a program using the formulas of Theorem 3.5.4 for $f(z) = e^z$, with $0 \le m \le 4$, $n = 4 - m$. In the Padé table these will be on the fourth diagonal, perpendicularly to the main diagonal. The input was the Malaurin coefficients of $f_{4,0}$. The results were first obtained in floating point arithmetic, but they were then converted into rational form by the algorithm described in Example 3.5.2. The coefficients p_i of the numerators $P_{m,4-m}$ and q_j of the denominators $Q_{m,4-m}$ are given in Table 3.5.1. For $m = n = 4$ the program found that the error term is $-4 \cdot 10^{-8} z^9$, while the error term of the Maclaurin expansion $f_{8,0}$ is $3 \cdot 10^{-6} z^9$.

When $m = n = 10$ the program gave warnings about divisions by zero, and it estimated the condition number of the linear system (3.5.31) to be 10^{22}. The

Table 3.5.1. *The Coefficients p_i and q_j of Padé approximations for e^z.*

$m\backslash i$	0	1	2	3	4
0	1	0	0	0	0
1	1	1/4	0	0	0
2	1	1/2	1/12	0	0
3	1	3/4	1/4	1/24	0
4	1	1	1/2	1/6	1/24

$m\backslash j$	0	1	2	3	4
0	1	−1	1/2	−1/6	1/24
1	1	−3/4	1/4	−1/24	0
2	1	−1/2	1/12	0	0
3	1	−1/4	0	0	0
4	1	0	0	0	0

reciprocal of this number is a measure of how close the matrix of the system is to a singular matrix, (see Theorem 7.5.3). The computed coefficients of the Padé approximant had large errors. Nevertheless e was computed with full machine accuracy (for $z = 1$), and the error term was estimated to be less than $10^{-25}z^{21}$.

Example 3.5.7.

To evaluate $\ln(1 + x)$ one can use the relation

$$\ln(1+x) = \ln\left(\frac{1+z}{1-z}\right), \quad z = \frac{x/2}{1+x/2},$$

and use the continued fraction expansion given in (3.5.3). The convergents of this continued fraction are odd functions and Padé approximants. The first few are

$$s_{00} = 2z, \qquad s_{01} = \frac{3}{3-z^2}, \qquad s_{11} = 2z\frac{15+4z^2}{3(5-3z^2)},$$
$$s_{12} = \frac{105-55z^2}{105-90z^2+9z^4}, \qquad s_{22} = 2z\frac{945-735z^2+64z^4}{15(63-70z^2+15z^4)}.$$

Here the diagonal approximants s_{mm} are most interest. For example, the approximation s_{22} matches the Taylor series up to the term z^8 and the error is approximately equal to the term $z^{10}/11$. Note that the denominators are the Legendre polynomials in $1/z$,

3.5.3 The Epsilon Algorithm.

We shall here briefly introduce the important ϵ-algorithm and indicate the connections between Padé approximation, Aitken acceleration, linear difference equations and this algorithm.

If n is large, the heavy part of the computation of a Padé approximant

$$f_{m,n}(z) = P_{m,n}(z)/Q_{m,n}(z)$$

of $f(z)$ in (3.5.22) is the solution of the linear system (3.5.31). We see that if m or n is decreased by 1, most of the equations of the system will be the same. There are therefore relations between the polynomials $Q_{m,n}(z)$ for adjacent values of m, n, which have been subject to intensive research that has resulted in several interesting algorithms. See, e.g., the monographs of Brezinski [8, 9] and the literature cited there.

Here we are primarily interested in the use of Padé approximants as a convergence accelerator in the *numerical* computation of values of $f(z)$ for (say) $z = e^{i\phi}$, in particular for $z = \pm 1$. A natural question is whether it is possible to omit the calculation of the coefficients p_j, q_j, and find a recurrence relation that gives the function values directly. A very elegant solution to this problem, called the ϵ-algorithm, was found in 1956 by P. Wynn [48], after complicated calculations. We shall present the algorithm, but we refer to the original paper of Wynn for the proof.

A two-dimensional array of numbers $\epsilon_k^{(p)}$ is computed by the recurrence relation,

$$\epsilon_{k+1}^{(p)} = \epsilon_{k-1}^{(p+1)} + \frac{1}{\epsilon_k^{(p+1)} - \epsilon_k^{(p)}}, \tag{3.5.33}$$

which involves quantities in a rhombus

$$\begin{array}{ccc} & \epsilon_k^{(p)} & \\ \epsilon_{k-1}^{(p+1)} & & \epsilon_{k+1}^{(p)} \\ & \epsilon_k^{(p+1)} & \end{array}$$

If the following boundary conditions are used:

$$\begin{aligned} \epsilon_{-1}^{(p)} &= 0, \\ \epsilon_0^{(p)} &= f_{p,0}(z) = \sum_{j=0}^{p} c_j z^j, \\ \epsilon_{2n}^{(-n)} &= f_{0,n}(z) = \frac{1}{\sum_{j=0}^{n} d_j z^j}, \end{aligned} \tag{3.5.34}$$

this yields for *even* subscripts

$$\epsilon_{2n}^{(p)} = f_{p+n,n}(z), \tag{3.5.35}$$

The values of $\epsilon_{2n+1}^{(p)}$ with *odd* subscripts are auxiliary quantities only. The polynomials $f_{0,n}(z)$ are obtained from the Taylor expansion of $1/f(z)$. Several procedures for obtaining this were given in Sec. 3.1.

It seems easier to program the ϵ-algorithm it after a slight change of notation. We introduce an $r \times 2r$ matrix $A = [a_{ij}]$, $a_{ij} = \epsilon_k^{(p)}$, where $k = j - 2$, $p = i - j + 1$. Conversely, $i = k + p + 1, j = k + 2$. The ϵ-algorithm, together with the boundary conditions now takes the form:

```
for $i = 1 : r$
  $a_{i,1} = 0;\quad a_{i,2} = f_{i-1,0}(z);\quad a_{i,2i} = f_{0,i-1}(z);$
  for $j = 2 : 2 * i - 2$
    $a_{i,j+1} = a_{i-1,j-1} + 1/(a_{ij} - a_{i-1,j}).$
```

```
        end
      end
```

Results:

$$f_{m,n}(z) = a_{m+n+1,2n+2}, \quad (m, n \geq 0, \quad m+n+1 \leq r).$$

The above program sketch must be improved for practical use, e.g., something should be done about the risk for a division by zero.

An extension of the Aitken acceleration, due to Shanks [39] 1955, uses a comparison series with terms of the form

$$c_j = \sum_{\nu=1}^{p} \alpha'_\nu k_\nu^j, \quad j \geq 0, \quad k_\nu \neq 0. \tag{3.5.36}$$

Here α'_ν and k_ν are $2p$ parameters, to be determined, in principle, by means of c_j, $j = 0 : 2p-1$. The parameters may be complex. The power series becomes

$$S(z) = \sum_{j=0}^{\infty} c_j z^j = \sum_{\nu=1}^{p} \alpha'_\nu \sum_{j=0}^{\infty} k_\nu^j z^j = \sum_{\nu=1}^{p} \frac{\alpha'_\nu}{1 - k_\nu z}.$$

This is a rational function of z, and the "Ansatz" of Shanks is thus related to Padé approximation, but note that the poles at k_ν^{-1} should be simple and that $m < n$ for $S(z)$, because $S(z) \to 0$, as $z \to \infty$. Recall that the calculations for the Padé approximation determines the coefficients of $S(z)$ *without calculating the* $2n$ *parameters* α'_ν *and* k_ν. It can happen that m becomes larger than n, and if α'_ν and k_ν are afterwards determined, by the expansion of $S(z)$ into partial fractions, it can turn out that some of the k_ν are multiple poles.

This suggests a generalization of the Shanks approach but how? If we consider the coefficients q_j, $j = 1 : n$, occurring in (3.5.31) as known quantities then (3.5.31) can be interpreted as a *linear difference equation*[80] . The general solution of this is given by (3.5.36), if the zeros of the polynomial

$$Q(x) := 1 + \sum_{j=1}^{n} q_j x^j$$

are simple, but if multiple roots are allowed, the general solution reads,

$$c_l = \sum_{\nu} p_\nu(l) k_\nu^n,$$

where k_ν runs through the different zeros of $Q(x)$, and p_ν is an arbitrary polynomial, the degree of which equals the multiplicity -1 of the zero k_ν.

Essentially the same mathematical relations occur in several areas of numerical analysis, such as interpolation and approximation by a sum of exponentials, and in the design of quadrature rules with free nodes (see Sec. 5.2). For an application of the ϵ-algorithm to numerical quadrature, see Sec. 5.3.3.

[80] This can also be expressed in terms of the z-transform; see § 3.2.3.

3.5.4 The QD Algorithm.

Given the continued fraction

$$c(z) = \frac{a_1}{1+}\,\frac{a_2 z}{1+}\,\frac{a_3 z}{1+}, \tag{3.5.37}$$

we denote the nth approximant by

$$w_n(z) = P_n(z)/Q_n(z), \quad n = 1, 2, \ldots. \tag{3.5.38}$$

This corresponds to the finite continued fraction obtained by setting $a_{n+1} = 0$. The sequence of numerators $\{P_n\}$ and denominators $\{Q_n\}$ in (3.5.38) satisfy the recurrence relations:

$$\begin{aligned} P_0 = 0, \quad P_1 = 1, \qquad & P_n = za_nP_{n-2} + P_{n-1}, \\ Q_0 = Q_1 = 1, \qquad & Q_n = za_nQ_{n-2} + Q_{n-1}, \quad n \geq 2, \end{aligned}$$

Hence both P_n and Q_n are polynomials in z of degree $[(n-1)/2]$ and $[n/2]$, respectively. It can be shown that the polynomials P_n and Q_n have no common zero for $n = 1, 2, \ldots,$ and for all z.

In the special case that all $a_i > 0$, the continued fraction

$$c(z) = \frac{a_1}{1+}\,\frac{a_2 z}{1+}\,\frac{a_3 z}{1+}, \tag{3.5.39}$$

is called a Stieltjes fraction.[81]

From the initial conditions and recurrence relations it follows that $Q_n(0) = 1$, $n = 0, 1, 2, \ldots$. Hence the rational function $w_n(z)$ is analytic at $z = 0$ and thus can be expanded in a Taylor series

$$\frac{P_n(z)}{Q_n(z)} = c_0^{(n)} + c_1^{(n)} z + c_2^{(n)} z^2 + \cdots \tag{3.5.40}$$

that converges for z sufficiently small. The coefficients $c_k^{(n)}$ in (3.5.40) can be shown to be independent of n for $k < n$. We denote by $c_k := c_k^{(n+1)}$ the ultimate value of $c_k^{(n)}$ for increasing values n and let

$$f(z) = c_0 + c_1 z + c_2 z^2 + \cdots, \tag{3.5.41}$$

be the formal power series formed with these coefficients. Then the power series $f(z)$ and the fraction $c(z)$ are said to **correspond** to each other. Note that the formal power series $f(z)$ corresponding to a given fraction $c(z)$ converges for any $z \neq 0$.

We now consider the converse problem: Given a (formal) power series $f(z)$, find a continued fraction $c(z)$ of the form (3.5.50) corresponding to it. Note that we do not require that the formal power series corresponding to the continued fraction converges, merely that the nth approximant w_n of the continued fraction satisfies

$$f(z) - w_n(z) = O(z^n).$$

[81] The theory of such fractions was first expounded by Stieltjes in a famous memoir, which appeared in 1894, the year of his death.

Example 3.5.8.

For $|z| < 1$,

$$\arctan z = z - \tfrac{1}{3}z^3 + \tfrac{1}{5}z^5 - \tfrac{1}{7}z^7 + \cdots.$$

The corresponding partial numerators and the corresponding continued fractions are

$$a_{2k} = \frac{(2k-1)^2}{(4k-3)(4k-1)}, \qquad a_{2k+1} = \frac{(2k)^2}{(4k-1)(4k+1)}.$$

The fraction converges for all z such that z^2 is not real and $z^2 \leq 1$. After an equivalence transformation we obtain

$$\arctan z = \frac{z}{1+}\,\frac{z^2}{3+}\,\frac{4z^2}{5+}\,\frac{9z^2}{7+}\,\frac{16z^2}{9+} \tag{3.5.42}$$

The convergents of the corresponding continued fractions are equal to Padé approximants.

The **qd algorithm**[82], can be used to compute such a continued fraction, if it exists.

For arbitrary integers n and $k \geq 0$, we define the Hankel matrices

$$H_k^{(n)} = \begin{pmatrix} c_n & c_{n+1} & \cdots & c_{n+k-1} \\ c_{n+1} & c_{n+2} & \cdots & c_{n+k} \\ \vdots & \cdots & \cdots & \vdots \\ c_{n+k-1} & c_{n+k-2} & \cdots & c_{n+2k-2} \end{pmatrix} \in \mathbf{R}^{k\times k}, \tag{3.5.43}$$

where we set $c_k = 0$ for $k < 0$. Further, we define the Hankel determinants

$$\mathbf{H}_k^{(n)} = \det\left(H_k^{(n)}\right), \quad k = 1, 2, \ldots. \tag{3.5.44}$$

associated with the formal power series (3.5.41).

Theorem 3.5.5. Henrici [26, Theorem 12.4c]

Given a formal power series (3.5.41)*, there exists at most one corresponding continued fraction. It exists precisely one such fraction if and and only if the Hankel determinants* (3.5.44) *satisfy $H_k^{(n)} \neq 0$ for $n = 0, 1$ and $k = 1, 2, \ldots$.*

The Hankel determinants satisfy the following important identity called **Jacobi's identity**:

For all integers n and $k \geq 1$

$$(H_k^{(n)})^2 - H_k^{(n-1)}H_k^{(n+1)} + H_{k+1}^{(n-1)}H_{k-1}^{(n+1)} = 0. \tag{3.5.45}$$

[82] The qd algorithm was originally given by the Swiss mathematician Heinz Rutishauser [38]

If the determinants $H_k^{(n)}$ are arranged in a triangular array

$$\begin{array}{lllll}
1 & & & & \\
1 & H_1^{(0)} = c_0 & & & \\
1 & H_1^{(1)} = c_1 & H_2^{(0)} & & \\
1 & H_1^{(2)} = c_2 & H_2^{(1)} & H_3^{(0)} & \\
1 & H_1^{(3)} = c_3 & H_2^{(2)} & H_3^{(1)} & H_4^{(0)}
\end{array}$$

then Jacobi's identity links together the entries in a star like configuration. Since the two first columns are trivial (3.5.45) may be used to calculate the Hankel determinants recursively from left to right.

The **quotient-difference scheme**, or qd scheme is a scheme

$$\begin{array}{cccccc}
 & q_1^{(0)} & & & & \\
0 & & e_1^{(0)} & & & \\
 & q_1^{(1)} & & q_2^{(0)} & & \\
0 & & e_1^{(1)} & & e_2^{(0)} & \\
 & q_1^{(2)} & & q_2^{(1)} & & q_3^{(0)} \\
0 & & e_1^{(2)} & & e_2^{(1)} & \\
 & q_1^{(3)} & & q_2^{(2)} & & q_3^{(1)} \\
0 & & e_1^{(3)} & & e_2^{(2)} & \\
 & \vdots & & q_2^{(3)} & & \vdots \\
 & & \vdots & & \vdots &
\end{array},$$

where the quantities are connected by the two **rhombus rules**

$$e_m^{(n)} = q_m^{(n+1)} - q_m^{(n)} + e_{m-1}^{(n+1)}; \quad m = 1, 2, \dots, \; n = 0, 1, 2, \dots, \tag{3.5.46}$$

$$q_{m+1}^{(n)} = \frac{e_m^{(n+1)}}{e_m^{(n)}} q_m^{(n+1)}; \quad m = 1, 2, \dots, \; n = 0, 1, 2, \dots, \tag{3.5.47}$$

The qd scheme associated with the formal power series (3.5.41) is obtained by taking the entries in the second column to be

$$q_1^{(n)} = c_{n+1}/c_n, \quad n = 0, 1, 2, \dots, \tag{3.5.48}$$

The remaining elements in the qd scheme can then be generated column by column using the rhombus rules. If the columns $q_{m+1}^{(n)}$, $m = 1, 2, \dots$ exist, then the continued fraction corresponding to f is given by

$$c = \frac{c_0}{1-}\,\frac{q_1^{(0)}z}{1-}\,\frac{e_1^{(0)}z}{1-}\,\frac{q_2^{(0)}z}{1-}\,\frac{e_2^{(0)}z}{1-} - \cdots, \tag{3.5.49}$$

Example 3.5.9.

For the power series $c(z) = 0! + 1!z + 2!z^2 + 3!z^3 + \cdots$, the following qd scheme is obtained:

$$\begin{array}{ccccccc}
 & 1 & & & & & \\
0 & & 1 & & & & \\
 & 2 & & 2 & & & \\
0 & & 1 & & 2 & & \\
 & 3 & & 3 & & 3 & \\
0 & & 1 & & 2 & & 3 \\
 & 4 & & 4 & & 4 & \\
0 & & 1 & & 2 & & 3
\end{array}$$

Hence the corresponding continued fraction is

$$c(z) = \frac{1}{1+}\,\frac{z}{1+}\,\frac{z}{1+}\,\frac{2z}{1+}\,\frac{2z}{1+}\,\frac{3z}{1+}.$$

It is sometimes convenient to consider continued fractions in z^{-1}

$$c(z) = \frac{a_1}{1+}\,\frac{a_2 z^{-1}}{1+}\,\frac{a_3 z^{-1}}{1+}, \tag{3.5.50}$$

that corresponds to the formal series

$$p = c_0 + c_1 z^{-1} + c_2 z^{-2} + \cdots.$$

Review Questions

1. Define a continued fraction. Show how the convergents can be evaluated either backwards or forwards.
2. Show how any positive number can be expanded into a continued fraction with integer elements. In what sense are the convergents the best approximations? How accurate are they?
3. The denominators or numerators of the approximants of a continued fraction can be evaluated by Clenshaw's algorithm. Why is that?
4. What is the Padé table? Describe how the Padé approximants can be computed, if they exist. Tell something about singular and almost singular situations that can be encountered, and how to avoid them.
5. Describe the ϵ-algorithm, and tell something about its background and its efficiency.

6. Describe the qd algorithm. What can it be used for?

Problems and Computer Exercises

1. (a) Write a program for the algorithm of Example 3.5.2. Apply it to find a few coefficients of the continued fractions for

$$\tfrac{1}{2}(\sqrt{5}+1), \quad \sqrt{2}, \quad e, \quad \pi, \quad \log 2/\log 3, \quad 2^{j/12}$$

for a few integers j, $1 \le j \le 11$.

(b) Check the accuracy of the convergents. What happens when you apply your program to a rational number, e.g., 729/768 ?

(c) The metonic cycle used for calendrical purposes by the Greeks consists of 235 lunar month, which nearly equal 19 solar years. Show, using the algorithm in Example 3.5.2, that 235/19 is the sixth convergent of the ratio 365.2495/29.53059 of the Lunar phase synodic) period and solar period

2. A matrix formalism for continued fractions.

(a) We use the same notations as in Sec. 3.4.1, but we set, with no loss of generality, $b_0 = 0$. Set

$$P(n) = \begin{pmatrix} p_{n-1} & p_n \\ q_{n-1} & q_n \end{pmatrix}, \qquad A(n) = \begin{pmatrix} 0 & a_n \\ 1 & b_n \end{pmatrix}.$$

Show that $P(0) = I$,

$$P(n) = P(n-1)A(n), \qquad P(n) = A(1)A(2)\cdots A(n-1)A(n), \quad n \ge 1.$$

Comment: This does not minimize the number of arithmetic operations but, in a matrix-oriented programming language, it often gives very simple programs.

(b) Write a program for this with some termination criterion, and test it on a few cases, e.g.,

$$1 + \frac{1}{1+}\,\frac{1}{1+}\,\frac{1}{1+}\ldots; \quad 2 + \frac{1}{3+}\,\frac{1}{2+}\,\frac{1}{3+}\,\frac{1}{2+}\,\frac{1}{3+}\ldots; \quad 2 + \frac{2}{2+}\,\frac{3}{3+}\,\frac{4}{4+}\ldots.$$

As a post-processing, apply in the first two cases, e.g., Aitken acceleration in order to obtain a very high accuracy. Does the result look familiar in the last case? See Problem 3 concerning the exact results in the two other cases.

(c) Write a version of the program with some strategy for scaling $P(n)$ in order to eliminate the risk of overflow and underflow.

Hint: Note that the convergents $x_n = p_n/q_n$ are unchanged if you multiply the $P(n)$ by arbitrary scalars.

(d) Use this matrix form for working out a short proof of (3.5.6).

Hint: What is the determinant of a matrix product?

3. (a) Explain that $x = 1 + 1/x$ for the continued fraction in (3.5.11)?

(b) Compute the periodic continued fraction

$$2 + \frac{1}{3+}\,\frac{1}{2+}\,\frac{1}{3+}\,\frac{1}{2+}\,\frac{1}{3+}\cdots$$

exactly (by paper and pencil). (The convergence is assured by Seidel's Theorem 3.5.2.)

(c) Suggest a generalization of (a) and (b), where you can always obtain a quadratic equation with a positive root.

(d) Show that

$$\frac{1}{\sqrt{x^2-1}} = \frac{1}{x-}\,\frac{\frac{1}{2}}{x-y} \quad \text{where} \quad y = \frac{\frac{1}{4}}{x-}\,\frac{\frac{1}{4}}{x-}\,\frac{\frac{1}{4}}{x-}\cdots.$$

4. (a) Prove the equivalence transformation (3.5.7). Show that the errors of the convergents have alternating signs, if the elements of the continued fraction are positive.

(b) Show how to bring a general continued fraction to the special form of equation (3.5.10).

5. Let $P_{m,m}(z)/Q_{m,m}(z)$ be the diagonal Padé approximants of the exponential function. Show that the coefficients for $P_{m,m}(z)$ satisfy the recursion

$$p_0 = 1, \quad p_{j+1} = \frac{m-j}{(2m-j)(j+1)}p_j, \quad j = 0 : m-1. \tag{3.5.51}$$

(b) Show that for $m = 6$ we have

$$P_{6,6}(z) = 1 + \frac{1}{2}z + \frac{5}{44}z^2 + \frac{1}{66}z^3 + \frac{1}{792}z^4 + \frac{1}{15840}z^5 + \frac{1}{665280}z^6.$$

and $Q_{6,6}(z) = P_{6,6}(-z)$. How many operations are needed to evaluate this approximation for a given z?

(c) Use the error estimate in (3.5.27), neglecting higher order terms, to compute a bound for the relative error of the approximation in (b) when $|z| \in [0, \ln 2]$. What degree of the diagonal Padé approximant is needed for the relative error is required to be of the oder of the unit roundoff $2^{-53} = 1.11 \cdot 10^{-16}$ in IEEE double precision?

6. (a) Write a program for computing a Padé approximant and its error term. Apply it (perhaps after a transformation), for various values of m, n to, e.g., e^z, $\arctan z$, $\tan z$. (Note that two of these examples are odd functions.) Use the algorithm of Example 3.5.2 for expressing the coefficients as rational numbers. For how large m, n can you (in these examples) use your program without severe trouble with rounding errors.

(b) (b) Try to determine for which other functions the Padé table has a similar symmetry as shown in the text for the exponential function e^z.

7. (a) Show that there is at most one rational function $R(z)$, where the degrees of the numerator and denominator do not exceed, respectively, m and n, such that

$$f(z) - R(z) = O(z^{m+n+1}), \quad \text{as} \quad z \to 0,$$

even if the system (3.5.31) is singular. (Note, however, that P_m and Q_n are not uniquely determined, if the system is singular; they have common factors.)

(b) Is it true that if $f(z)$ is a rational function of degrees m', n', then

$$f_{m,n}(z) = f(z), \quad \forall\ m \geq m', \quad n \geq n'?$$

8. Write a program for evaluation the incomplete gamma function. Use the continued fraction (3.5.21) for x greater than about $a + 1$. For x less than about $a + 1$ use the power series for $\gamma(a, x)$.

9. Check that the program sketch for the ϵ-algorithm is equivalent with the scheme with the quantities $\epsilon_k^{(p)}$ given earlier in the text. How do you obtain the boundary values?

Notes and References

Much work on approximations to special functions, e.g., Gauss hypergeometric function and the Kummer function, was done around the end of World War II. This work culminated with the publication by the US Department of Commerce of the classical *Handbook of Mathematical Functions* edited by Milton Abramowitz and Irene A. Stegun [1, 1965]. Chapters 13 and 15 in this handbook contain many useful formulas and tables. Sections 15.1, 15.4, and Table 13.6, show how *many other important functions, elementary as well as advanced special functions, can be expressed in terms of these functions.* Tables and formulas in this handbook can be useful in preliminary surveys before turning to computer programs.

The basic properties of these functions are derived in Lebedev's monograph on Special Functions [30]. Lebedev's compact book will often be referred to, because it provides a good background to the applications of advanced Analysis, that lacks complete proofs in our book. For example, the chapter on the gamma function contains numerous instances of the use of series expansions and analytic continuation that are efficient as well as instructive, important and beautiful. Codes and other interesting information concerning the evaluation of special functions are also found in a modern classic, Numerical Recipes [36, Chapter 5–6].

The idea of using Cauchy's formula and FFT for numerical differentiation seems to have been first suggested by Lyness and Moler [32].

A rigorous theory of semi-convergent series was developed by Stieltjes and Poincaré in 1886.

More information about the classical methods for polynomial interpolation of equidistant data is found in, e.g., Fröberg [20] and Steffensen [42], in particular § 18 about "the calculus of symbols". For the history of these matters see, e.g., Goldstine [21].

More complete presentation of extrapolation methods is given in the monograph by Claude and Redivo-Zaglia [11], and more recently Sidi [40]. The historical development of the field is nicely surveyed by Brezinski [10].

Convergence acceleraton methods (due to Lindelöf, Plana and others) that transform an infinite series to an integral in the complex plane, can, with appropriate numerical procedures for computing the integral, compete with the methods mentioned in Sec. 3.4. They have the additional property to be applicable to some *ill-conditioned series*; see Dahlquist [16].

The theory of continued fractions started to develop in the 17th century. The main contributors were Euler, Lambert and Lagrange; see Brezinski [9]. The basic algorithmic aspects of what we today call Padé approximants were established by Frobenius [18]. Padé [34] gave a systematic study of these approximants and introduced the table named after him. The analytic theory of continued fractions has earlier origins and contributors include Chebyshev, A. A. Markov and Stieltjes. Modern related developments are the epsilon algorithm of P. Wynn and the quotient-difference algorithm of Rutishauser.

An easy to read introduction to continued fractions and Padé approximations is Baker [2]. Their use in numerical computations is surveyed in Blanche [4]. More recent developments of Padé approximations is found in Gragg [22]. Continued fractions of special functions are found in Abramowitz and Stegun [1]. Codes and further references are given in Numerical Recipes, Press et al. [36, Chapters 5 and 6]. The following example contains a different type of continued fraction. More information about arithmetic continued fractions, from a computational point of view is found in Riesel [37].

Bibliography

[1] Milton Abramowitz and Irene A. Stegun (eds.). *Handbook of Mathematical Functions.* Dover, New York, NY, 1965.

[2] G. A. Baker, Jr. *Essentials of Padé Approximants.* Academic Press, New York, 1975.

[3] Petter Bjørstad, Germund Dahlquist, and Eric Grosse. Extrapolations of asymptotic expansions by a modified Aitken δ^2-formula. *BIT*, 21:56–65, 1981.

[4] G. Blanche. Numerical evaluation of continued fractions. *SIAM Review*, 6:4:383–421, 1964.

[5] G. A. Bliss. *Algebraic Functions.* Amer. Math. Soc., New York, 1933.

[6] Jon M. Borwein, Peter B. Borwein, and K. Dilcher. Pi, Euler numbers and asymptotic expansions. *Amer. Math. Monthly*, 96:681–687, 1989.

[7] RIchard P. Brent and H. T. Kung. Fast algorithms for manipulating formal power series. *J. ACM*, 25:581–595, 1978.

[8] Claude Brezinski. *Padé-Type Approximations and General Orthogonal Polynomials.* Birkhäuser Verlag, Basel, 1980.

[9] Claude Brezinski. *History of Continued Fractions and Padé Approximants.* Springer-Verlag, Berlin, 1991.

[10] Claude Brezinski. Convergence acceleration during the 20th century. *J. Comput. Appl. Math.*, 122:1–21, 2000.

[11] Claude Brezinski and Michela Redivo-Zaglia. *Extrapolation Methods.* North-Holland, Amsterdam, 1991.

[12] E. W. Cheney. *Introduction to Approximation Theory.* McGraw-Hill, New York, NY, 1966.

[13] C. W. Clenshaw. A note on summation of Chebyshev series. *Math. Tables Aids Comput.*, 9:118–120, 1955.

[14] R. M. Corless, Gaston H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and Donald H. Knuth. On the Lambert W function. *Adv. Comput. Math.*, 5:329–359, 1996.

[15] R. Courant. *Differential and Integral Calculus*, volume I. Blackie & Son, London, 1934. Reprinted 1988 in Classics Library, John Wiley.

[16] Germund Dahlquist. On summation formulas due to Plana, Lindelöf and Abel, and related Gauss–Christoffel rules II. *BIT*, 37:4:804–832, 1997.

[17] Peter Deuflhard and Andreas Hohmann. *Numerical Analysis in Modern Scientific Computing.* Springer, Berlin, second edition, 2003.

[18] G. Frobenius. Über Relationen zwischen den Näherungsbrüchen von Potenzreihen. *J. für Math.*, 90:1–17, 1881.

[19] Carl-Erik Fröberg. *Lärobok i Numerisk Analys.* Svenska Bokförlaget/Bonniers, Stockholm, 1962. In Swedish.

[20] Carl-Erik Fröberg. *Numerical Mathematics.* Benjamin/Cummings, Menlo Park, CA, 1980.

[21] H. H. Goldstine. *The Computer from Pascal to von Neumann.* Princeton University Press, Princeton, NJ, 1972.

[22] W. B. Gragg. The Padé table and its relation to certain algorithms of numerical analysis. *SIAM Review*, 14:1–62, 1972.

[23] Peter Henrici. *Discrete Variable Methods in Ordinary Differential Equations.* John Wiley, New York, 1962.

[24] Peter Henrici. *Elements of Numerical Analysis.* John Wiley, New York, 1964.

[25] Peter Henrici. *Applied and Computational Complex Analysis. Volume 1 Power Series, Integration, Conformal Mapping, Location of Zeros.* Wiley Classics Library, New York, 1988. Reprint of the 1974 original.

[26] Peter Henrici. *Applied and Computational Complex Analysis. Volume 2 Special Functions, Integral Transforms, Asymptotics, Continued Fractions.* Wiley Classics Library, New York, 1991. Reprint of the 1977 original.

[27] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms.* SIAM, Philadelphia, second edition, 2002.

[28] Felix Klein. *Elementary Mathematics from an Advanced Standpoint.* Dover, New York, 1945. Translation from German original, 1924.

[29] Donald E. Knuth. *The Art of Computer Programming, Vol. 2. Seminumerical Algorithms.* Addison-Wesley, Reading, MA, second edition, 1997.

[30] N. N. Lebedev. *Special Functions and Their Applications.* Dover, New York, 1972. Translation from Russian original.

[31] C. C. Lin and L. A. Segel. *Mathematics Applied to Deterministic Problems in the Natural Sciences.* Macmillan, New York, 1974.

[32] James N. Lyness and Cleve B. Moler. Numerical differentiation of analytic functions. *SIAM J. Numer. Anal.*, 4:202–210, 1967.

[33] N. E. Nörlund. *Vorlesungen über Differenzenrechnung.* Springer-Verlag, Berlin, 1924. Reprinted by Chelsea, New York, 1954.

[34] H. Padé. Sur la représentation approcheée d'un fonction par des fraction rationelles. *Thesis Anal. Ecole Norm. Sup..*, 3:1–93, supplement, 1892.

[35] O. Perron. *Die Lehre von den Kettenbrüchen. Vol. II.* Teubner, Stuttgart, third edition, 1957.

[36] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical Recipes in Fortran; The Art of Scientific Computing.* Cambridge University Press, Cambridge, GB, second edition, 1992.

[37] Hans Riesel. *Prime Numbers and Computer Methods for Factorization.* Progr. Math. 126, Birkhäuser, Boston, MA, second edition, 1994.

[38] Heinz Rutishauser. Der Quotienten-Differenzen-Algoritmus. *Z. Angew. Math. Phys.*, 5:233–251, 1954.

[39] D. Shanks. Nonlinear transformations of divergent and slowly convergent sequences. *J. Math. Phys.*, 34:1–42, 1955.

[40] Avram Sidi. *Practical Extrapolation Methods. Theory and Applications.* Cambridge University Press, Cambridge, UK, 2003.

[41] A. Smoktunowicz. Backward stability of Clenshaw's algorithm. *BIT*, 42:3:600–610, 2002.

[42] J. F. Steffensen. *Interpolation.* Chelsea, New York, second edition, 1950.

[43] Frank Stenger. *Numerical Methods Based on Sinc and Analytic Functions.* Springer-Verlag, Berlin, 1993.

[44] Gilbert Strang. *Introduction to Applied Mathematics.* Wellesley-Cambridge Press, Wellesley, MA, 1986.

[45] E. C. Titchmarsh. *The Theory of Functions.* Oxford University Press, London, second edition, 1939.

[46] John Todd. Notes on numerical analysis i, solution of differential equations by recurrence relations. *Math. Tables Aids Comput.*, 4:39–44, 1950.

[47] H. S. Wall. *Analytic Theory of Continued Fractions.* Van Nostrand, Princeton, NJ, 1948.

[48] Peter Wynn. On a device for computing the $e_m(s_n)$ transformation. *Math. Tables Aids Comput.*, 10:91–96, 1956.

Index