Computation of Confluent Hypergeometric Functions and Application to Parabolic Boundary Control Problems

Dissertation

zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften

dem Fachbereich IV der Universität Trier
vorgelegt von

Christian Schwarz[1]

Trier 2001

[1]gefördert durch die Deutsche Forschungsgemeinschaft im Rahmen des Graduiertenkollegs "Mathematische Optimierung" an der Universität Trier

Gutachter: apl. Prof. Dr. Jürgen Müller, Universität Trier
Prof. Dr. Ekkehard Sachs, Universität Trier

Tag der mündlichen Prüfung: 5.10.2001

Danksagung

An dieser Stelle möchte ich mich bei allen bedanken, die zum Gelingen dieser Dissertation beigetragen haben.
Mein besonderer Dank geht an Herrn apl. Prof. Dr. Jürgen Müller, der durch seine stets hilfreichen Anregungen und Diskussionen sowie seine Betreuung und Unterstützung maßgeblich zum Gelingen dieser Dissertation beigetragen hat.
Ebenso danke ich Herrn Prof. Dr. Ekkehard Sachs für die Betreuung und Unterstützung dieser Arbeit sowie für seine wertvollen Hinweise, wodurch er zum Erfolg der Dissertation einen wesentlichen Beitrag geleistet hat.
Herrn Prof. Dr. Rainer Tichatschke danke ich für die Übernahme des Vorsitzes des Prüfungsausschusses.
Zudem geht mein Dank an die Deutsche Forschungsgemeinschaft und die Träger des Graduiertenkollegs "Mathematische Optimierung", die mir die Gelegenheit zur Promotion gegeben haben.
Außerdem bedanke ich mich bei meinen Kollegen und Freunden Lars Abbe, Marco Fahl, Tim Voetmann und Christoph Becker für die zahlreichen fachlichen und persönlichen Diskussionen und das Korrekturlesen des Manuskripts.
Schließlich danke ich besonders meiner Susanne für ihr Verständnis und ihre fortwährende Unterstützung, sowie meiner Familie, insbesondere meinen Eltern für ihre Unterstützung während meiner gesamten Schul- und Studienzeit.

Contents

1 Introduction **7**

2 The confluent hypergeometric functions **15**
2.1 Definition and basic properties 15
2.2 Special cases of the confluent hypergeometric function 18

3 Computation of confluent hypergeometric functions **23**
3.1 Computation on compact intervals 23
3.1.1 Chebyshev expansions 24
3.1.2 Series expansions by convolution 24
3.1.3 Construction of the series expansion 31
3.1.4 Convergence theory 33
3.1.5 Numerical results 36
3.2 Computation with recurrence relations 39
3.2.1 Basics of linear difference equations 39
3.2.2 The Miller algorithm 45
3.2.3 Applications and numerical results 50
3.3 Asymptotic expansions 53
3.3.1 Principal results on asymptotic expansions 54
3.3.2 Asymptotic expansions for confluent hypergeometric functions 54
3.3.3 Numerical results 58
3.4 Conclusion 59

4 Application to parabolic boundary value problems **61**
4.1 Conductive heat transfer 61
4.1.1 The model of heat transfer in cylindrical domains 62
4.1.2 Fourier series approach 63
4.1.3 Convergence analysis 66
4.1.4 Homogeneous boundary conditions 68
4.1.5 Different geometries of the domain 70
4.2 Thermal convection loop 74
4.2.1 The model 74

4.2.2 Fourier series approach 77
4.2.3 The Lorenz equations 80
4.2.4 Numerical results 82

5 Modelling the heat transfer in food processing 85
5.1 An optimal control problem 85
5.2 Modelling the heat transfer 88
5.2.1 The Ball-formula 88
5.2.2 The Fourier series approach 89
5.2.3 Reduced order modelling 90
5.2.4 Numerical results 91
5.3 Thermal convection loop 93

Conclusion 97

Bibliography 99

Chapter 1

Introduction

A rigorous mathematical description of physical processes often leads to (partial) differential or integral equations that, in general, can have a very complex structure. Often it is desirable to have a simplified mathematical model for the physical process in order to recognize the characteristics (and influence coefficients) of the process. In certain cases an explicit form or a series representation of the desired solutions can be obtained by using special functions. Such a representation can be obtained when the problem is of some special structure, i.e. the domain under consideration is of special geometry, e.g. a cylindrical domain. Then, the solution can be represented by using special functions, so that a suitable approximation of these function is required. However, if the structure of the problem is more complicated, especially the domain under consideration is not of such special type, we have to use discretization methods like finite differences or finite elements in order to get a numerical solution of the problem. We remark that improvements of the approximation then require a new discretization of the problem. Whereas in the case of a series representation a complete system of functions is at hand, such that we can improve the approximation or the truncated series by adding more terms of the series. Hence, in view of the numerical simulation of such mathematical models it is necessary to have efficient methods for computing these special functions.

In [22] Lozier investigated software available for the computation of special functions and stated the software needs in scientific computing. This analysis exhibits a deficiency in the availability of software packages for certain special functions. In particular, software packages are needed for the computation (in fixed precision) of confluent hypergeometric functions with complex arguments and special cases like Coulomb wave functions and Bessel functions (of complex order). Since the class of confluent hypergeometric functions covers several important special functions, it is therefore desirable to have an efficient algorithm for computing these functions (with complex parameters and argument). Moreover, every solution of linear ordinary differential equations of second order can be represented by confluent hypergeometric functions (see Erdélyi [8], p. 249), that is using a suitable transformation of the linear ordinary differential equations we obtain a differential equation of confluent hypergeometric type.

In this thesis we focus on the computation of confluent hypergeometric functions and point out the relations between these functions and parabolic boundary value problems with applications to heat transfer and fluid dynamics.
In Chapter 2 we provide the characteristics of confluent hypergeometric functions. We investigate the confluent hypergeometric function of the first kind denoted by $M(a;c;z)$ with two, in general complex, parameters a and c as well as argument z, whose power series representation is given by

$$M(a;c;z) = \sum_{\nu=0}^{\infty} \frac{(a)_\nu z^\nu}{(c)_\nu \nu!}.$$

This function, which is also called the Kummer function or M-function, is an entire function, so that convergence of the series for all $z \in \mathbb{C}$ is ensured. We give alternative definitions and representations of the function. Of particular importance are the integral representation and the connections to the hypergeometric functions. Furthermore, we address the connections to many other special functions which are included in the class of confluent hypergeometric functions as special cases.
In Chapter 3 the advantages and disadvantages of several methods for the (numerical) computation of confluent hypergeometric functions are discussed as no single method can be expected to be equally successful for all parameters and arguments. If we consider e.g. the partial sums of the above power series representation

$$s_n(a,c,z) = \sum_{\nu=0}^{n} \frac{(a)_\nu z^\nu}{(c)_\nu \nu!}$$

for $n \in \mathbb{N}$, we find that they provide a reasonable approximation to the function $M(a;c;z)$ in a small neighbourhood of the origin only. The main disadvantage of these partial sums are the cancellation errors which occur when computing in fixed precision arithmetic outside this neighbourhood. The computation of small function values in modulus causes problems, because the summation of the terms of the partial sums $s_n(a,c,z)$, which are inherently large in modulus, leads to a loss or cancellation of several decimal digits. Therefore we develop and investigate different methods for the computation of confluent hypergeometric functions regarding stability of these methods with respect to cancellation errors.
We start with the computation of the Kummer function on (real or complex) bounded intervals $[0,\beta]$ for some $\beta \neq 0$, so that the partial sums $s_n(a,c,z)$ do not provide an appropriate approximation. Instead, it is possible to use the shifted Chebyshev expansion of the function $M(a;c;z)$, given by

$$M(a;c;z) = B_0(a,c) + 2\sum_{k=1}^{\infty} B_k(a,c) T_k(z/\beta)$$

with Chebyshev coefficients $B_k(a,c)$ and shifted Chebyshev polynomials T_k. However, the Chebyshev coefficients are functions of generalized hypergeometric type whose numerical evaluation is difficult. Moreover, the coefficients depend on the parameters a and c, so that the

computation of the function $M(a;c;z)$ for different parameters requires a new computation for each change of parameters yielding a high computational effort.
In order to overcome this disadvantage of the Chebyshev expansion, we develop a method based on the Hadamard product (also called convolution product) of two power series $\varphi(z) = \sum_{\nu=0}^{\infty} \varphi_\nu z^\nu$ and $\psi(z) = \sum_{\nu=0}^{\infty} \psi_\nu z^\nu$, defined by

$$(\varphi * \psi)(z) = \sum_{\nu=0}^{\infty} \varphi_\nu \psi_\nu z^\nu.$$

This product is also called convolution product due to the following integral representation. For a holomorphic function φ, an entire function ψ and a suitable integration curve γ we have for $z \in \mathbb{C}$

$$(\varphi * \psi)(z) = \frac{1}{2\pi i} \int_\gamma \varphi(\zeta)\psi(z/\zeta)\frac{d\zeta}{\zeta}.$$

With the help of this representation we can show a continuity property of the Hadamard product: For entire functions ψ and functions φ holomorphic in $\mathbb{C} \setminus [1, \infty)$, we obtain in the sup-norm $\|\cdot\|$

$$\|\varphi * \psi\|_K \leq M_L \cdot \|\psi\|_L$$

with a certain constant M_L and for certain compact sets K and L (dependent on K) in the complex plane.
After providing these main properties of the Hadamard product we establish certain estimates which are of particular importance for the following convergence analysis.
Starting from the power series representation of the confluent hypergeometric function $M(a;c;z)$ we then construct a series expansion with the help of a suitable convolution, that is

$$M(a;c;\cdot) = F(a,1;c;\cdot) * \exp(\cdot)$$

of a hypergeometric function $F(a,1;c;\cdot)$ and the exponential function. Replacing the exponential function $\exp(\cdot)$ by its shifted Chebyshev expansion we obtain the following series expansion of the confluent hypergeometric function

$$M(a;c;z) = e^{\beta/2} I_0(\beta/2) + 2e^{\beta/2} \sum_{k=1}^{\infty} (-1)^k I_k(\beta/2) {}_3F_2(-k,k,a;c,1/2;z/\beta)$$

denoting by $I_k(\beta/2)$ the modified Bessel functions of order k and by ${}_3F_2(-k,k,a;c,1/2;\cdot)$ the polynomials resulting from the convolution of the hypergeometric function $F(a,1;c;\cdot)$ and the shifted Chebyshev polynomials T_k. We see that the coefficients depend only on the considered interval and not on the parameters a and c. Hence, for a fixed chosen interval these coefficients do not change and can be stored. Moreover, the polynomials ${}_3F_2(-k,k,a;c,1/2;\cdot)$ can be computed by a four-term recurrence formula, so the partial sums of this series expansion can be used for an efficient approximation of the function $M(a;c;z)$. In Section 3.1.4 a detailed error analysis is given for the partial sums of the series expansion. In addition, results are proved for the asymptotic behaviour of the absolute error. In Section 3.1.5 on

numerical results we compute relative errors of the partial sums of the series expansion and of the Taylor sections of $M(a;c;z)$, where the parameters a and c are real or complex. It can be seen that the problem of cancellation errors can be reduced considerably by using the partial sums of this expansion.
A similar method was developed by Müller [27] for the computation of Bessel functions of variable order on compact intervals.
Another important tool for the computation of special functions are recurrence formulae. We give a short review of the basic theory of linear difference equations (of second order), i.e. we compute solutions (y_n) satisfying the equation

$$y_{n+1} + a_n y_n + b_n y_{n-1} = 0,$$

with given coefficients a_n and $b_n \neq 0$. After considering some examples we state recurrence relations for the confluent hypergeometric functions. Although easy to implement, such recurrence relations are often numerically unstable e.g. due to rounding errors. In particular, forward recurrences cause these problems. In order to circumvent these problems we apply a method for computing recurrence relations in backward direction: the Miller algorithm. Since the application of the Miller algorithm does not require exact starting values for the backward recurrence we need a so-called "normalizing series" instead, i.e. for the computation of the functions $f_n(z)$ with a real or complex variable z we need a convergent series expansion of the form

$$S = \sum_{n=0}^{\infty} c_n f_n(z)$$

with known $S \neq 0$. However, the determination of the starting index is a critical point. Gautschi [11] presented various versions of Miller algorithms particularly considering estimates of the starting values for the backward recursion and providing an asymptotic theory of three-term recurrence relations.
We apply the Miller algorithm to recurrence formulae for the computation of the confluent hypergeometric function $M(a;c;z)$ and we prove results concerning the asymptotic behaviour of the relative errors for these computations. Furthermore, numerical results show the efficiency of this method.
Finally, another method for computing confluent hypergeometric functions is discussed. If we want to compute the function $M(a;c;z)$ on unbounded intervals we have to consider methods like asymptotic expansions for large arguments z in modulus. We use for the computation of a function f in some sector in the complex plane an expansion of the form

$$f(z) \sim \sum_{\nu=0}^{\infty} \frac{a_\nu}{z^\nu} \qquad (z \to \infty).$$

After providing the basic characteristics of asymptotic expansions, we consider the asymptotic expansion of the confluent hypergeometric function $M(a;c;z)$. For numerical computation, the determination of the number of terms used for the approximation is a crucial point.

We present a method to determine the number of terms with regard to numerical stability. We will see that from the numerical point of view the monotonicity behaviour of the terms of the series yields a suitable criterion for the truncation of the series.
Further contributions on computing the confluent hypergeometric function M are given by Nardin et al. [28], where results on numerical computation of $M(a;c;z)$, based on the power series representation, for large $|z|$ are presented. In [5], Barnett used a method based on continued fractions in order to compute the Coulomb wave functions.
As an application of the discussed methods we consider initial-boundary value problems with partial differential equations of parabolic type in Chapter 4. In order to determine solutions of the initial-boundary value problems we apply the method of eigenfunction expansion or (generalized) Fourier series representation, where the arising eigenfunctions of the associated eigenvalue problem depend directly on the geometry of the considered domain. For certain domains with some special geometry like cylinders (circular, elliptic, parabolic), rectangles or spheres the corresponding eigenfunctions can be obtained in an explicit form. In this case the eigenfunctions are often of confluent hypergeometric type, so that the above methods for computing these functions can be applied.
First, we consider (conductive) heat transfer in a cylinder with boundary control. To be more precise, we investigate an initial-boundary value problem including the two-dimensional conductive heat equation in the domain Ω, i.e. we consider the linear parabolic differential equation

$$\frac{\partial \theta}{\partial t} = \chi \Delta \theta \qquad \text{in } \Omega \times (0,T)$$

with the initial temperature distribution

$$\theta(\cdot,0) = \theta_0(\cdot) \qquad \text{in } \Omega$$

and the so-called Robin boundary condition

$$\frac{\partial \theta}{\partial n} = \alpha(u-\theta) \qquad \text{on } \partial\Omega \times (0,T)$$

modelling the heat transfer from the surrounding medium. By θ we denote the temperature and by u the boundary control or the temperature of the surrounding medium. Furthermore, we exploit the geometry of the domain and reduce the problem above to a problem of dimension one by introducing cylindrical coordinates.
We show how to construct an eigenfunction expansion or (generalized) Fourier series representation of the solution θ of this initial-boundary value problem. In general such a representation has the form

$$\theta = \sum_n \alpha_n \psi_n$$

with time-dependent (generalized) Fourier coefficients α_n. These are determined by solving an initial value problem. The eigenfunctions ψ_n come from an associated eigenvalue problem and form a complete orthonormal system.

Moreover, the uniform convergence of the series is proved. After separating the (generalized) Fourier series representation into a general solution and a particular solution, we find convergent majorants for these series representations. Furthermore, we simplify the heat transfer model for constant boundary controls, such that the resulting series representation is of simple structure being of particular interest in view of the application to sterilization processes in Chapter 5. Also from the numerical point of view this representation provides some advantages.

Finally, we explicitly show how the confluent hypergeometric function $M(a;c;z)$ arises, if different special geometries of the underlying domain, like circular, elliptic or parabolic cylinders, are considered.

We close this chapter by presenting an application in fluid dynamics. It will be discussed how the dynamics of a fluid filled loop which is heated on one side and cooled on the other can be simulated. Since we obtain a fluid flow only driven by temperature, termed natural convection, the loop is often called thermal convection loop. The adequate mathematical model is based on the Boussinesq equations including the Navier-Stokes equations for incompressible Newtonian fluids and the (convective) heat equation

$$\begin{aligned}\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} &= (1+\beta(\theta_0-\theta))\mathbf{g} - \frac{1}{\rho}\nabla p + \nu\Delta\mathbf{v}\\ \operatorname{div}\mathbf{v} &= 0\\ \frac{\partial\theta}{\partial t} + \mathbf{v}\cdot\nabla\theta &= \chi\Delta\theta\end{aligned}$$

in the domain $\Omega\times(0,T)$ with appropriate initial and boundary conditions denoting by $\mathbf{v}$ the velocity field and by p the pressure. After introducing several simplifications we can solve the resulting initial-boundary value problem again with the method of (generalized) Fourier series. Aside from the formulation of the mathematical model we perform a detailed derivation of the (generalized) Fourier series representation of the velocity and the temperature. A crucial point is the determination of the (generalized) Fourier coefficients, since these coefficients satisfy an infinite system of coupled ordinary differential equations. In order to compute approximations of the velocity and the temperature of the fluid, we truncate the (generalized) Fourier series and obtain a finite system of coupled ordinary differential equations. If we truncate to one mode, the simplest approximation, the system of coupled ordinary differential equations can be formulated by the Lorenz system

$$\begin{aligned}\frac{dX}{d\tau} &= -\sigma X + \sigma Y\\ \frac{dY}{d\tau} &= -XZ - Y + rX\\ \frac{dZ}{d\tau} &= XY - bZ\end{aligned}$$

(see Yorke et al. [48]). After providing the essential characteristics of this dynamical system, specifically focussing on the asymptotic stability of the system, we give numerical results

illustrating how this system (and enhanced systems) can be applied to the approximation of the flow behaviour.
In Chapter 5 we apply the heat transfer models given in Chapter 4 to sterilization processes in food industry, where the food is filled into containers and heated in an autoclave by steam or hot water in order to destroy harmful microorganisms. Unfortunately, heat sensitive nutrients and vitamins are also affected during the heating process. Since, in practice, the temperature of the autoclave is often empirically determined, it may happen that during the sterilization process not only the harmful microorganisms but also the nutrients and vitamins are destroyed. Hence, we need appropriate mathematical models for the sterilization process and the heat transfer, in order to optimize the nutritional quality of the product. Moreover, common sterility measures are very sensitive to the evolution of temperature, so that it is very important to have a reasonable heat transfer model for simulating the heating process. We present and discuss several models of heat transfer, a very simple and empirically determined model (the Ball-formula) which is still in practical use as well as models based on partial differential equations of different complexity.
We end this thesis by giving some concluding remarks on the combination of computing confluent hypergeometric functions and solving (parabolic) partial differential equations via eigenfunction expansions.

Chapter 2

The confluent hypergeometric functions

The confluent hypergeometric functions which are also called Kummer functions generate an important class of functions including many other special functions, e.g. Coulomb wave functions, parabolic cylinder functions, Bessel functions, incomplete gamma functions, exponential integrals, Fresnel integrals and error functions.

2.1 Definition and basic properties

In this section we define the confluent hypergeometric functions and state the basic properties, cf. Temme [40], pp. 171-177 or Andrews [3], pp. 385-390.
In order to define the confluent hypergeometric or Kummer functions, we start with the Gaussian hypergeometric equation

$$z(1-z)w'' + (c-(a+b+1)z)w' - abw = 0, \tag{2.1}$$

with its singularities $z = 0$, $z = 1$ and $z = \infty$. It is solved by the Gaussian hypergeometric function $w = F(a, b; c; z)$, defined by

$$F(a,b;c;z) = \sum_{\nu=0}^{\infty} \frac{(a)_\nu (b)_\nu z^\nu}{(c)_\nu \nu!} \qquad (|z| < 1) \tag{2.2}$$

with complex parameters a, b and c, where $c \notin (-\mathbb{N}_0)$. For $\zeta \in \mathbb{C}$ and $\nu \in \mathbb{N}$ the Pochhammer symbol or shifted factorial is defined by

$$(\zeta)_\nu = \zeta(\zeta+1)(\zeta+2)\cdots(\zeta+\nu-1), \qquad (\zeta)_0 = 1.$$

We obtain the confluent hypergeometric functions when two of the singularities of (2.1) merge into one singularity ("confluence" of two singularities). In order to describe this formal process we consider the hypergeometric function $F(a, b; c; z/b)$ which has a singularity at $z = b$. We define

$$M(a;c;z) = \lim_{b\to\infty} F(a,b;c;z/b). \tag{2.3}$$

Since we have

$$\lim_{b\to\infty} \frac{(b)_n}{b^n} = 1,$$

the computation of the limits of the terms of the power series (2.2) yields finally the power series representation of the *confluent hypergeometric function of the first kind*

$$M(a;c;z) = \sum_{\nu=0}^{\infty} \frac{(a)_\nu z^\nu}{(c)_\nu \nu!} \qquad (z \in \mathbb{C}) \tag{2.4}$$

with complex parameters a and c, where $c \notin (-\mathbb{N}_0)$. From (2.4) follows that the confluent hypergeometric function can be written as

$$M(a;c;z) = {}_1F_1(a;c;z),$$

a member of the class of generalized hypergeometric functions ${}_pF_q$. The same limit process is applicable to known results of the hypergeometric function. In this way we obtain the Kummer differential equation

$$zw'' + (c-z)w' - aw = 0.$$

Linearly independent solutions of the Kummer equation are the function $M(a;c;z)$ and the *confluent hypergeometric function of the second kind* $U(a;c;z)$, defined by

$$U(a;c;z) = \frac{\pi}{\sin(\pi c)} \left(\frac{M(a;c;z)}{\Gamma(1+a-c)\Gamma(c)} - z^{1-c} \frac{M(1+a-c;2-c;z)}{\Gamma(a)\Gamma(2-c)} \right).$$

The important functional relation

$$M(a;c;z) = e^z M(c-a;c;-z)$$

is called *Kummer transformation*. If we again apply the limit process above to the Euler integral representation of the hypergeometric function, given by

$$F(a,b;c;z) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-a)} \int_0^1 t^{a-1}(1-t)^{c-a-1}(1-tz)^{-b}\,dt \tag{2.5}$$

for $\mathrm{Re}(c) > \mathrm{Re}(a) > 0$ and $|\arg(1-z)| < \pi$, we obtain the following representation for the confluent hypergeometric function

$$M(a;c;z) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-a)} \int_0^1 e^{zt} t^{a-1}(1-t)^{c-a-1}\,dt$$

for $\mathrm{Re}(c) > \mathrm{Re}(a) > 0$.

Order and type of the confluent hypergeometric functions

We denote by

$$M(r,f)=\max_{|z|\le r}|f(z)| \qquad (r>0)$$

the maximum modulus of an entire function $f:\mathbb{C}\to\mathbb{C}$. Then the order $\rho\in[0,\infty]$ of f is defined by

$$\rho=\rho_f=\limsup_{r\to\infty}\frac{\log\log M(r,f)}{\log r}$$

and for $\rho\in(0,\infty)$ the type $\tau\in[0,\infty]$ of f is defined by

$$\tau=\tau_f=\limsup_{r\to\infty}\frac{\log M(r,f)}{r^\rho}.$$

Further it is possible to characterize the order and type of entire functions by means of the magnitude of the Taylor coefficients. For the Taylor coefficients a_n of an entire function $f(z)=\sum_{\nu=0}^{\infty}a_\nu z^\nu$ it is well-known that

$$\lim_{n\to\infty}|a_n|^{1/n}=0.$$

For the following results on the characterization of order and type we refer to Boas [6], p. 9.

Theorem 2.1 *Let $f(z)=\sum_{\nu=0}^{\infty}a_\nu z^\nu$ be an entire function of order ρ. Then we have*

$$\rho=\limsup_{n\to\infty}\frac{\log n}{-\log|a_n|^{1/n}}. \tag{2.6}$$

Theorem 2.2 *Let $f(z)=\sum_{\nu=0}^{\infty}a_\nu z^\nu$ be an entire function of order $\rho\in(0,\infty)$ and type τ. Then we have*

$$\rho\tau e=\limsup_{n\to\infty}n|a_n|^{\rho/n}. \tag{2.7}$$

With the help of these theorems we obtain immediately that the confluent hypergeometric function $M(a;c;\cdot)$ is an entire function of order $\rho=1$ and type $\tau=1$.
Indeed: With the use of the asymptotic representation of the Pochhammer symbol (for arbitrary $\zeta\in\mathbb{C}$)

$$(\zeta)_{n+1}\sim\frac{n^\zeta n!}{\Gamma(\zeta)} \qquad (n\to\infty)$$

and a simplified version of the Stirling formula

$$(n!)^{1/n}\sim\frac{n}{e} \qquad (n\to\infty),$$

we obtain

$$\rho=\limsup_{n\to\infty}\frac{\log n}{-\log|(a)_n/((c)_n n!)|^{1/n}}=1$$

and

$$\tau=\limsup_{n\to\infty}\frac{n}{e}\left|\frac{(a)_n}{(c)_n n!}\right|^{1/n}=1.$$

2.2 Special cases of the confluent hypergeometric function

As mentioned above the class of confluent hypergeometric functions includes many important special functions. We give a summary of these special functions with their connections to the confluent hypergeometric functions.

Whittaker functions

In the literature the confluent hypergeometric functions are often characterized as *Whittaker functions*, defined by

$$M_{\kappa,\mu}(z) = e^{-z/2} z^{1/2+\mu} M\left(\tfrac{1}{2}+\mu-\kappa, 1+2\mu; z\right)$$
$$W_{\kappa,\mu}(z) = e^{-z/2} z^{1/2+\mu} U\left(\tfrac{1}{2}+\mu-\kappa, 1+2\mu; z\right),$$

which are solutions of the *Whittaker equation*

$$w'' + \left(-\frac{1}{4} + \frac{\kappa}{z} + \frac{\frac{1}{4}-\mu^2}{z^2}\right) w = 0.$$

These functions occur as solutions of transformated differential equations of second order (see Temme [40], p. 178).

Parabolic cylinder functions

The *parabolic cylinder functions* are solutions of the differential equation

$$y'' - \left(a + \frac{z^2}{4}\right) y = 0,$$

given by

$$y_1 = e^{-z^2/4} M\left(\tfrac{1}{2}\,a + \tfrac{1}{4}; \tfrac{1}{2}; \tfrac{1}{2}\,z^2\right)$$
$$y_2 = z e^{-z^2/4} M\left(\tfrac{1}{2}\,a + \tfrac{3}{4}; \tfrac{3}{2}; \tfrac{1}{2}\,z^2\right)$$

(see Temme [40], p. 179).

Error functions

The error functions play an important role in statistics and probability theory. The *error function* erf and *complementary error function* erfc are defined through

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$$

$$\operatorname{erfc}(z) = 1 - \operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_z^\infty e^{-t^2} dt,$$

and the relations to the Kummer functions are

$$\operatorname{erf}(z) = z\, M\left(\tfrac{1}{2}; \tfrac{3}{2}; -z^2\right)$$
$$\operatorname{erfc}(z) = e^{-z^2} U\left(\tfrac{1}{2}; \tfrac{1}{2}; z^2\right)$$

(see Temme [40], p. 180).

Exponential integrals

For $n = 1, 2, \ldots$ the *exponential integrals* are defined by

$$E_n(z) = \int_1^\infty \frac{e^{-zt}}{t^n} dt \qquad (\operatorname{Re}(z) > 0).$$

The connection to the Kummer function of the second kind U is given by

$$E_n(z) = e^{-z} U(1; 2-n; z) = z^{n-1} e^{-z} U(n; n; z).$$

For $n = 1$ we write

$$-Ei(-z) = E_1(z) = \int_z^\infty \frac{e^{-t}}{t}\, dt \qquad (|\arg z| < \pi).$$

If $z = x$ is real, we define the exponential integral

$$Ei(x) = \int_{-\infty}^x \frac{e^t}{t}\, dt$$

(see Temme [40], p. 180).

Incomplete gamma functions

The *incomplete gamma function* is defined by

$$\gamma(a, z) = \int_0^z t^{a-1} e^{-t}\, dt \qquad (\operatorname{Re}(a) > 0,\ |\arg z| < \pi)$$

and the *complementary incomplete gamma function* is defined by

$$\Gamma(a, z) = \int_z^\infty t^{a-1} e^{-t}\, dt \qquad (a \in \mathbb{C},\ |\arg z| < \pi).$$

These functions are called incomplete because of their incomplete interval of integration compared to the (Euler) gamma function. If in particular $\operatorname{Re}(a) > 0$, we have

$$\Gamma(a) = \Gamma(a, z) + \gamma(a, z).$$

The connections to the Kummer functions are

$$\gamma(a,z) = \frac{z^a}{a} e^{-z} M(1; a+1; z),$$
$$\Gamma(a,z) = z^a e^{-z} U(1; a+1; z)$$

(see Temme [40], pp. 185-186, p. 277).
For computation of $\gamma(a,x)$ and $\Gamma(a,x)$ for $x \geq 0$ and $a \in \mathbb{R}$ by methods based on Taylor sections and continued fractions see Gautschi [13].

Orthogonal polynomials

The *Laguerre polynomials* which are defined by

$$L_n^\alpha(z) = \frac{1}{n!} e^z z^{-\alpha} \frac{d^n}{dz^n} (z^{n+\alpha} e^{-z}),$$

where α is in general a complex parameter, are related to the Kummer function through

$$L_n^\alpha(z) = \frac{\Gamma(n+\alpha+1)}{n!\,\Gamma(\alpha+1)} M(-n; \alpha+1; z) = \frac{(-1)^n}{n!} U(-n; \alpha+1; z).$$

(see Nikiforov [29], p. 23, p. 284).
The definition of the *Hermite polynomials* is

$$H_n(z) = (-1)^n e^{z^2} \frac{d^n}{dz^n} (e^{-z^2}).$$

The relations to the Kummer function are

$$H_{2n}(z) = \frac{2^{2n}\sqrt{\pi}}{\Gamma(1/2-n)} M\left(-n; \tfrac{1}{2}; z^2\right),$$
$$H_{2n+1}(z) = -\frac{2^{2n+2}\sqrt{\pi}}{\Gamma(-1/2-n)} M\left(-n; \tfrac{3}{2}; z^2\right)$$

(see Nikiforov [29], p. 23, p. 284).

Bessel functions

The *Bessel functions* $J_\lambda(x)$ of the order λ of the first kind, defined by

$$J_\lambda(x) = \frac{(x/2)^\lambda}{\Gamma(\lambda+1)} \sum_{\nu=0}^{\infty} \frac{(-x^2/4)^\nu}{(\lambda+1)_\nu \nu!}$$

with a real argument x are connected to the confluent hypergeometric function in the following way:

$$J_\lambda(x) = \frac{(x/2)^\lambda}{\Gamma(\lambda+1)} e^{-ix} M\left(\lambda+\tfrac{1}{2}; 2\lambda+1; 2ix\right) \tag{2.8}$$

(see Andrews [3], p. 400, Temme [40], p. 227).

Fresnel integrals

The *Fresnel integrals* $C(x)$ and $S(x)$, defined by

$$C(x) = \int_0^x \cos\left(\frac{\pi t^2}{2}\right) dt \quad \text{and} \quad S(x) = \int_0^x \sin\left(\frac{\pi t^2}{2}\right) dt$$

with a real argument x are also connected to the confluent hypergeometric function. The relations are

$$C(x) = \frac{x}{2}\left(M\left(\tfrac{1}{2};\tfrac{3}{2};\tfrac{1}{2}\,i\pi x^2\right) + M\left(\tfrac{1}{2};\tfrac{3}{2};-\tfrac{1}{2}\,i\pi x^2\right)\right)$$
$$S(x) = \frac{x}{2i}\left(M\left(\tfrac{1}{2};\tfrac{3}{2};\tfrac{1}{2}\,i\pi x^2\right)) - M\left(\tfrac{1}{2};\tfrac{3}{2};-\tfrac{1}{2}\,i\pi x^2\right)\right)$$

(see Andrews [3], p. 113, p. 400).

Coulomb wave functions

The Coulomb wave functions are solutions of the nonrelativistic Coulomb wave equation

$$\frac{d^2w}{d\rho^2} + \left(1 - \frac{2\eta}{\rho} - \frac{\lambda(\lambda+1)}{\rho^2}\right) w = 0.$$

This equation is of particular interest in physics, more exact in quantum mechanics as a form of the Schrödinger equation in a central Coulomb field. Two linearly independent solutions of the equation are the *regular* and *irregular Coulomb wave functions* denoted by $F_\lambda(\eta,\rho)$ and $G_\lambda(\eta,\rho)$. The relations to the Kummer functions are

$$F_\lambda(\eta,\rho) = C_\lambda(\eta)\rho^{\lambda+1}e^{-i\rho}M(\lambda+1-i\eta;2\lambda+2;2i\rho),$$

with $C_\lambda(\eta) = 2^\lambda e^{-\pi\eta/2}|\Gamma(\lambda+1+i\eta)|/\Gamma(2\lambda+2)$ and

$$G_\lambda(\eta,\rho) = iF_\lambda(\eta,\rho) + iD_\lambda(\eta)\rho^{\lambda+1}e^{-i\rho}U(\lambda+1-i\eta;2\lambda+2;2i\rho),$$

with $D_\lambda(\eta) = 2^{\lambda+1}e^{\pi\eta/2+\lambda\pi i-i\sigma_\lambda}$.
In nuclear and atomic physics the angular momentum number λ is integer and often denoted by L, the parameter η is real and the argument ρ is positive. Moreover, the functions F_λ and G_λ are real valued for real values of η, $\rho > 0$ and $\lambda \geq 0$ (cf. Temme [40], p. 171, p. 178).

We will see in the next chapter how to compute functions of confluent hypergeometric type. Since we also consider complex parameters, especially the Coulomb wave functions are of particular interest.

Chapter 3

Computation of confluent hypergeometric functions

In this chapter we consider several methods for computing confluent hypergeometric functions of first kind. We develop a method for computing confluent hypergeometric functions on bounded real or complex intervals using the partial sums of certain series expansions. These partial sums are easily computable and provide a better rate of convergence in comparison to the Taylor sections. Furthermore, the problem of cancellation errors is considerably reduced compared to the corresponding Taylor sections. Also the use of recurrence formulae is discussed in this chapter. Thereby we have to take the numerical stability of recurrence formulae into account. Finally, we consider the asymptotic expansion of the confluent hypergeometric function in order to compute the function $M(a;c;z)$ for large arguments z in modulus.

3.1 Computation on compact intervals

We consider the confluent hypergeometric function $M(a;c;\cdot)$ with complex parameters a and c where $c \notin (-\mathbb{N}_0)$ as defined in (2.4). Because M is an entire function, we have convergence of the series for all $z \in \mathbb{C}$. If the parameter a is a negative integer, the function M reduces to a polynomial.
Given $\beta \neq 0$ we would like to compute the function $M(a;c;z)$ for $z \in K_\beta$, where K_β denotes the compact interval $[0,\beta]$. A simple possibility to evaluate the confluent hypergeometric function is the use of the Taylor sections, given by

$$s_n(a,c,z) = \sum_{\nu=0}^{n} \frac{(a)_\nu z^\nu}{(c)_\nu \nu!} \tag{3.1}$$

for $n \in \mathbb{N}$. These partial sums might be taken as approximations in a certain neighbourhood of the origin. The size of the neighbourhood, however, is seriously restricted due to cancellation errors which arise when computing in fixed precision arithmetic. In particular, the computation of function values of small modulus causes problems, because the occurring

terms of the partial sums of larger modulus generate a loss or a cancellation of some decimal digits. For this reason we look for series expansions, which are more stable than Taylor sections with respect to cancellation errors.

3.1.1 Chebyshev expansions

In order to compute the confluent hypergeometric function on the compact interval K_β we may use the shifted Chebyshev expansion of the Kummer function M, given by

$$M(a;c;z) = B_0(a,c) + 2\sum_{k=1}^{\infty} B_k(a,c)T_k(z/\beta),$$

with the Chebyshev coefficients

$$B_k(a,c) = \frac{(a)_k(\beta/4)^k}{(c)_k k!}\,{}_2F_2(\tfrac{1}{2}+k, a+k; 1+2k, c+k; \beta)$$

and the shifted Chebyshev polynomials T_k, defined by

$$T_k(z) = t_k(1-2z),$$

using the k-th Chebyshev polynomial t_k (cf. Luke [23], p. 300 and [24], p. 30). The problem, however, is the computation of the coefficients B_k which are of generalized hypergeometric type. The efficient numerical evaluation of these functions is difficult. The use of the power series representation of B_k is, as mentioned above, reliable only in a small neighbourhood of the origin. Additionally the coefficients B_k are dependent on the parameters a and c of the function M. It is, however, desirable to have methods for the computation of confluent hypergeometric functions for continuously varying parameters a and c at hand. But it is not possible to store these Chebyshev coefficients since each change of parameters requires a new computation of the coefficients B_k. Because of this high effort for the computation of these coefficients it is very expensive to compute the partial sums of the expansion above.
In contrast to this, we will use series expansions, where the coefficients are independent of the varying parameters and thus can be precomputed.

3.1.2 Series expansions by convolution

In this section we will give theoretical results on the Hadamard product or convolution product of power series, which is the essential tool for the construction of the underlying series expansion (cf. Müller [26]).

Definition 3.1 For two formal power series $\varphi(z) = \sum_{\nu=0}^{\infty} \varphi_\nu z^\nu$ and $\psi(z) = \sum_{\nu=0}^{\infty} \psi_\nu z^\nu$ we denote their Hadamard product by

$$(\varphi * \psi)(z) = \sum_{\nu=0}^{\infty} \varphi_\nu \psi_\nu z^\nu.$$

By a cycle γ in $\mathbb{C}$ we understand a union of closed piecewise smooth curves as in Rudin [37], p. 217, and from there we also adopt the notation of $\int_\gamma f(z)dz$ and of Ind_γ. The Hadamard product is also often called convolution product because of its integral representation. For the following result and its proof we refer to Müller [26].

Lemma 3.2 *Let $\Omega \subset \mathbb{C}$ be a region which contains the origin and let φ be holomorphic in Ω. Suppose further that γ is a cycle in Ω, such that*

$$\mathrm{Ind}_\gamma(0) = 1 \quad \textit{and} \quad \mathrm{Ind}_\gamma(\alpha) = 0 \quad \textit{for all} \quad \alpha \notin \Omega.$$

If ψ is an entire function, then we have for every $z \in \mathbb{C}$

$$(\varphi * \psi)(z) = \frac{1}{2\pi i} \int_\gamma \varphi(\zeta)\psi(z/\zeta)\frac{d\zeta}{\zeta}. \tag{3.2}$$

Proof: We denote with

$$\varphi(z) = \sum_{\nu=0}^{\infty} \varphi_\nu z^\nu \qquad \text{and} \qquad \psi(z) = \sum_{\nu=0}^{\infty} \psi_\nu z^\nu$$

the Taylor expansions of φ and ψ around the origin. If we apply the general Cauchy Theorem (see Rudin [37], Theorem 10.35), we get

$$\varphi_\nu = \frac{1}{2\pi i} \int_\gamma \frac{\varphi(\zeta)}{\zeta^{\nu+1}}\, d\zeta \qquad (\nu \in \mathbb{N}_0).$$

Since $\varphi * \psi$ is entire, we have by interchanging the order of summation and integration for every $z \in \mathbb{C}$

$$\begin{aligned} \sum_{\nu=0}^{\infty} \varphi_\nu \psi_\nu z^\nu &= \frac{1}{2\pi i} \int_\gamma \sum_{\nu=0}^{\infty} \psi_\nu (z/\zeta)^\nu \frac{\varphi(\zeta)}{\zeta}\, d\zeta \\ &= \frac{1}{2\pi i} \int_\gamma \varphi(\zeta)\psi(z/\zeta)\frac{d\zeta}{\zeta}. \end{aligned}$$

□

For a compact set K and a continuous function φ we define

$$\|\varphi\|_K = \sup_{z \in K} |\varphi(z)|.$$

We only consider regions Ω which are cut planes, more precisely we assume without loss of generalization $\Omega = \mathbb{C} \setminus [1, \infty)$. Hence, the integral representation of Lemma 3.2 leads to the following continuity property of the Hadamard product (cf. Müller [26]).

Theorem 3.3 *Let φ be holomorphic in the cut plane $\Omega = \mathbb{C} \setminus [1, \infty)$. If K is a compact set in $\mathbb{C}$, then for every compact set L with $K^* \subset L^0$, where $K^* := K \cdot [0,1]$, a constant $M_L = M_L(\varphi, K) > 0$ exists, such that for every entire function ψ we have*

$$\|\varphi * \psi\|_K \leq M_L \cdot \|\psi\|_L. \tag{3.3}$$

Proof: We use the notation

$$K \cdot \gamma^{-1} := \{z/\zeta : z \in K, \zeta \in \gamma\}$$

with a cycle γ as in Lemma 3.2. Since K^* is starlike with respect to the origin, for every open set $U \supset K^*$ a simple closed piecewise smooth curve γ in $\mathbb{C} \setminus [1, \infty)$ of the form as in Figure 3.1 containing the origin in its interior exists, such that $K \cdot \gamma^{-1} \subset U$. Hence, we may choose $\gamma = \gamma(L)$ such that $K \cdot \gamma^{-1} \subset L$.
With the given parameterization $\gamma = \{\zeta \in \mathbb{C} : \zeta = \zeta(t), t \in [\underline{t}, \overline{t}]\}$ we get for every $z \in K$

$$|(\varphi * \psi)(z)| = \frac{1}{2\pi} \left| \int_{\underline{t}}^{\overline{t}} \varphi(\zeta(t)) \psi(z/\zeta(t)) \frac{d\zeta(t)}{\zeta(t)} \right| \leq \frac{1}{2\pi} \int_{\underline{t}}^{\overline{t}} |\varphi(\zeta(t))| \cdot \frac{|\zeta'(t)|}{|\zeta(t)|}\, dt \cdot \|\psi\|_{K \cdot \gamma^{-1}}.$$

Defining

$$M_L := \frac{1}{2\pi} \int_{\underline{t}}^{\overline{t}} |\varphi(\zeta(t))| \cdot \frac{|\zeta'(t)|}{|\zeta(t)|}\, dt$$

we get the asserted estimate. □

Remark 3.4 In particular, the result of Theorem 3.3 can be formulated for compact sets K which are starlike with respect to the origin (see Müller [26]). Then, for every compact set L with $K \subset L^0$, a constant $M_L = M_L(\varphi, K) > 0$ exists, such that for every entire function ψ we have

$$\|\varphi * \psi\|_K \leq M_L \cdot \|\psi\|_L.$$

We introduce for some fixed entire function g the family of entire functions

$$\mathcal{F} = \{f : \exists\ \varphi = \varphi_f \text{ holomorphic in } \mathbb{C} \setminus [1, \infty) : f = \varphi * g\}.$$

Moreover, let $K \subset \mathbb{C}$ be a given compact set which is starlike with respect to the origin. This is the set on which the function f ought to be approximated. Denoting by (P_n) a sequence of polynomials of degree $\leq n$ converging uniformly to g on L (with L as in Theorem 3.3), application of Theorem 3.3 yields

$$\|f - \varphi_f * P_n\|_K \leq M_{f,L} \cdot \|g - P_n\|_L \qquad (n \in \mathbb{N}), \tag{3.4}$$

i.e. the error of approximating f by the polynomials $\varphi_f * P_n$ can be estimated by the error of approximating g by the polynomials P_n and the factor $M_{f,L} = M_{f,L}(g, K)$. Hence, we have to choose polynomials P_n which provide a fast rate of convergence on L to g, at least essentially faster than the rate of convergence of the Taylor sections $(s_n(g))$ of g.
The idea is the use of some series expansion of the function g, i.e.

$$g = \sum_{k=0}^{\infty} \alpha_k T_k$$

with coefficients α_k and polynomials T_k of $\deg(T_k) \leq k$ yielding the formal series representation of the function f

$$f = \varphi_f * g = \sum_{k=0}^{\infty} \alpha_k (\varphi_f * T_k).$$

Note that the coefficients α_k are independent of the functions f of the family $\mathcal{F}$ and $\varphi_f * T_k$ are polynomials of $\deg(\varphi_f * T_k) \leq k$. Indentifying the polynomials P_n with the partial sums of the expansion of the function g we have the approximation of the function f

$$\varphi_f * P_n = \sum_{k=0}^{n} \alpha_k (\varphi_f * T_k).$$

Furthermore, we are interested in an estimate in terms of $\|\cdot\|_K$ instead of $\|\cdot\|_L$. This can be achieved by the Bernstein Lemma (see e.g. Walsh [43], p. 77). If $\delta > 1$ is given, a compact set $L = L_\delta$ exists as in Theorem 3.3 such that for all $m \in \mathbb{N}$

$$\|P_m - P_{m-1}\|_L \leq \delta^m \|P_m - P_{m-1}\|_K,$$

with $P_0 := 0$. The uniform convergence of $P_n = \sum_{m=0}^{n} (P_m - P_{m-1})$ to g on L yields

$$\|g - P_n\|_L \leq \sum_{m>n} \|P_m - P_{m-1}\|_L \leq \sum_{m>n} \delta^m \|P_m - P_{m-1}\|_K \qquad (n \in \mathbb{N}),$$

and if we use (3.4) with $M_f(\delta) := M_{f,L_\delta}$ we get

$$\|f - \varphi_f * P_n\|_K \leq M_f(\delta) \sum_{m>n} \delta^m \|P_m - P_{m-1}\|_K \qquad (n \in \mathbb{N}, f \in \mathcal{F}). \tag{3.5}$$

Henceforth, we would like to apply the results above to the confluent hypergeometric function. Setting

$$\mathcal{F} := \{M(a; c; \cdot) = (\varphi * g)(\cdot) : a,\, c \in \mathbb{C},\, c \notin -\mathbb{N}_0\}$$

we have the family of entire functions to be approximated, where the confluent hypergeometric function can be written as

$$M(a; c; z) = \sum_{\nu=0}^{\infty} \frac{(a)_\nu z^\nu}{(c)_\nu \nu!} = \sum_{\nu=0}^{\infty} \frac{(a)_\nu}{(c)_\nu} z^\nu * \exp(z),$$

so that we have $g(z) = \exp(z)$ and the Gaussian hypergeometric function $\varphi(z) = F(a, 1; c; z)$. Since the hypergeometric function $F(a, b; c; \cdot)$ is holomorphic in the cut plane $\mathbb{C} \setminus [1, \infty)$, we are allowed to apply Theorem 3.3 and can prove the following result.

Theorem 3.5 *Let ψ be an arbitrary entire function. If $\varphi = F(a,1;c;\cdot)$ is a hypergeometric function with $0 < \mathrm{Re}(a) < \mathrm{Re}(c)$, then we have*

$$\|\varphi * \psi\|_K \leq \delta_{a,c} \|\psi\|_K \tag{3.6}$$

with

$$\delta_{a,c} = \left| \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-a)} \right| \cdot \frac{\Gamma(\mathrm{Re}(a))\Gamma(\mathrm{Re}(c-a))}{\Gamma(\mathrm{Re}(c))}$$

for all compact sets $K \subset \mathbb{C}$ which are starlike with respect to the origin.

In order to prove Theorem 3.5, we need some information on the boundary behaviour of the hypergeometric functions $h(a,c;\cdot) := F(a,1;c;\cdot)$.

Lemma 3.6 *For $t > 1$ the limits*

$$h^+(a,c;t) := \lim_{\substack{z \to t \\ \mathrm{Im}(z)>0}} h(a,c;z) \quad \text{and} \quad h^-(a,c;t) := \lim_{\substack{z \to t \\ \mathrm{Im}(z)<0}} h(a,c;z)$$

both exist, and we have for $0 < \mathrm{Re}(a) < \mathrm{Re}(c)$

$$\frac{1}{2\pi i}\left[h^+(a,c;t) - h^-(a,c;t)\right] = \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-a)}\, t^{1-c}(t-1)^{c-a-1} \qquad (t>1).$$

Proof: Since the hypergeometric function $h(a,c;\cdot)$ is analytically continuable across the cut $[1,\infty)$ except of the (possible) singularity $z = 1$, for arbitrary parameters a, c with $c \notin -\mathbb{N}_0$, we know the existence of the limits $h^+(a,c;t)$ and $h^-(a,c;t)$ for $t > 1$.
For $0 < \mathrm{Re}(a) < \mathrm{Re}(c)$ we consider the Euler integral representation for the hypergeometric function (2.5)

$$h(a,c;z) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-a)} \int_0^1 t^{a-1}(1-t)^{c-a-1} \frac{dt}{1-zt}.$$

The substitution $t = 1/s$ gives the Cauchy integral

$$h(a,c;z) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-a)} \int_1^\infty s^{1-c}(s-1)^{c-a-1} \frac{ds}{s-z}.$$

In order to characterize the boundary behaviour of this integral, we can apply the formulae of Sokhotskyi (see Henrici [17], p. 94) and obtain

$$\frac{1}{2\pi i}\left[h^+(a,c;t) - h^-(a,c;t)\right] = \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-a)}\, t^{1-c}(t-1)^{c-a-1}$$

for all $t > 1$. □

As a first consequence of this result the substitution $t = 1/s$ yields

$$\int_1^\infty t^{1-c}(t-1)^{c-a-1}\frac{dt}{t} = \int_0^1 s^{a-1}(1-s)^{c-a-1}ds = \mathcal{B}(a, c-a)$$

and therefore

$$\frac{1}{2\pi i}\int_1^\infty \left[h^+(a,c;t) - h^-(a,c;t)\right]\frac{dt}{t} = \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-a)}\,\mathcal{B}(a,c-a) = 1,$$

where $\mathcal{B}$ denotes the Beta function. If we consider the modulus of the difference of the limits, we obtain as a further consequence of Lemma 3.6

$$\frac{1}{2\pi}\left|h^+(a,c;t) - h^-(a,c;t)\right| = \left|\frac{\Gamma(c)}{\Gamma(a)\Gamma(c-a)}\right| \cdot t^{\mathrm{Re}(1-c)}(t-1)^{\mathrm{Re}(c-a-1)}.$$

An analogous consideration as above yields

$$\int_1^\infty t^{\mathrm{Re}(1-c)}(t-1)^{\mathrm{Re}(c-a-1)}\frac{dt}{t} = \int_0^1 s^{\mathrm{Re}(a)-1}(1-s)^{\mathrm{Re}(c-a)-1}ds = \mathcal{B}(\mathrm{Re}(a), \mathrm{Re}(c-a))$$

and finally

$$\frac{1}{2\pi}\int_1^\infty \left|h^+(a,c;t) - h^-(a,c;t)\right|\frac{dt}{t} = \delta_{a,c}. \tag{3.7}$$

Proof of Theorem 3.5: It is sufficient to show $|(\varphi * \psi)(z)| \le \delta_{a,c}\|\psi\|_{[0,z]}$ with the straight line $[0,z] = \{tz : t \in [0,1]\}$ for every $z \in \mathbb{C}$.
For a given $\varepsilon \in (0, \frac{1}{2})$ a simple closed and piecewise smooth curve γ_ε is chosen as in Figure 3.1. It consists of two circular arcs $\gamma_\varepsilon^{(1)} = \{\zeta : |\zeta - 1| = \varepsilon\}$, $\gamma_\varepsilon^{(2)} = \{\zeta : |\zeta| = 1/\varepsilon\}$ and of two arcs $\gamma_\varepsilon^{(3)}$ and $\gamma_\varepsilon^{(4)}$ parallel to the real axis connecting $\gamma_\varepsilon^{(1)}$ and $\gamma_\varepsilon^{(2)}$. In order to obtain the required estimate, we will consider the integral representation of the Hadamard product with the curve of integration γ_ε. We are interested in limit relations for ε tending to zero. Therefore we need information on the asymptotic behaviour of hypergeometric functions at the singularities $\zeta = \infty$ and $\zeta = 1$. The use of formulae which describe this behaviour (see Olver [32], pp. 165-168 for $a \notin \mathbb{N}$, $c - a \notin \mathbb{N}$ and Abramowitz, Stegun [1], pp. 559-560 for $a \in \mathbb{N}$, $c - a \in \mathbb{N}$) leads for arbitrary $z \in \mathbb{C}$ to

$$\int_{\gamma_\varepsilon^{(j)}} h(a,c;\zeta)\psi(z/\zeta)\frac{d\zeta}{\zeta} \to 0 \qquad (\varepsilon \to 0^+,\ j = 1, 2),$$

$$\int_{\gamma_\varepsilon^{(3)}} h(a,c;\zeta)\psi(z/\zeta)\frac{d\zeta}{\zeta} \to \int_1^\infty h^+(a,c;t)\psi(z/t)\frac{dt}{t} \qquad (\varepsilon \to 0^+),$$

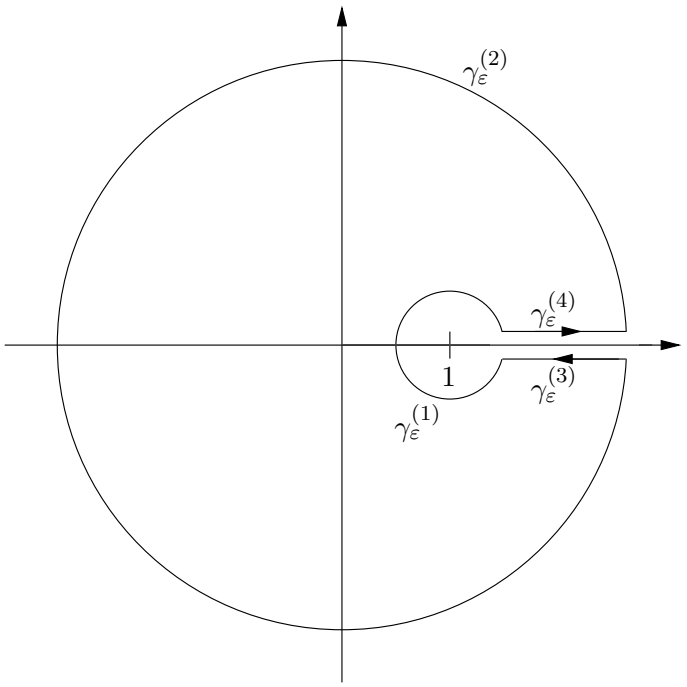

Figure 3.1: integration curve γ_ε

$$\int\limits_{\gamma_\varepsilon^{(4)}} h(a,c;\zeta)\psi(z/\zeta)\frac{d\zeta}{\zeta} \to -\int\limits_1^\infty h^-(a,c;t)\psi(z/t)\frac{dt}{t} \qquad (\varepsilon \to 0^+).$$

Altogether we have for $\varepsilon \to 0^+$

$$\int\limits_{\gamma_\varepsilon} h(a,c;\zeta)\psi(z/\zeta)\frac{d\zeta}{\zeta} \to \int\limits_1^\infty \left[h^+(a,c;t) - h^-(a,c;t)\right]\psi(z/t)\frac{dt}{t}.$$

Application of the integral representation (3.2) leads to

$$(\varphi * \psi)(z) = (h(a,c;\cdot) * \psi)(z) = \frac{1}{2\pi i}\int\limits_1^\infty \left[h^+(a,c;t) - h^-(a,c;t)\right]\psi(z/t)\frac{dt}{t}.$$

If we consider the modulus of the product and apply (3.7), we obtain

$$|(\varphi * \psi)(z)| \le \frac{1}{2\pi}\int\limits_1^\infty \left|h^+(a,c;t) - h^-(a,c;t)\right|\frac{dt}{t} \cdot \|\psi\|_{[0,z]} = \delta_{a,c}\,\|\psi\|_{[0,z]},$$

since $z \cdot [1,\infty)^{-1} = [0,z]$. □

Theorem 3.5 is a generalization of Theorem 2 in Müller [26] for the case of complex parameters a and c:

Corollary 3.7 *Let ψ be an arbitrary entire function and let $\varphi = F(a,1;c;\cdot)$ be a hypergeometric function with $0 < a \leq c$. Then we have*

$$\|\varphi * \psi\|_K \leq \|\psi\|_K \tag{3.8}$$

for all compact sets $K \subset \mathbb{C}$ which are starlike with respect to the origin.

Proof: Application of Theorem 3.5 yields $\delta_{a,c} = 1$ for $0 < a < c$. For parameters $a = c$ we have

$$\varphi(z) = h(a,a;z) = \frac{1}{1-z} = \sum_{\nu=0}^{\infty} z^\nu$$

and therefore $\varphi * \psi = \psi$. □

3.1.3 Construction of the series expansion

In this section we see how to construct the desired series expansion of the Kummer function (cf. Müller [26] and [38]). As mentioned above, the essential tool in this context is the Hadamard product of power series. In the previous section we have already seen that the confluent hypergeometric function can be written as the convolution product

$$M(a;c;\cdot) = F(a,1;c;\cdot) * \exp(\cdot). \tag{3.9}$$

The next step is the choice of an approximation of the exponential function. For this purpose we consider the shifted Chebyshev expansion on the compact interval K_β which is given by

$$e^z = e^{\beta/2} I_0(\beta/2) + 2e^{\beta/2} \sum_{k=1}^{\infty} (-1)^k I_k(\beta/2) T_k(z/\beta) \tag{3.10}$$

(cf. e.g. Luke [24], p. 32), where I_λ denote the modified Bessel functions of the first kind of order λ, defined by

$$I_\lambda(z) = \frac{(z/2)^\lambda}{\Gamma(\lambda+1)} \sum_{\nu=0}^{\infty} \frac{(z^2/4)^\nu}{(\lambda+1)_\nu \nu!},$$

for $\lambda \in \mathbb{C} \setminus (-\mathbb{N})$. In order to construct the series expansion for the confluent hypergeometric function we need some information on the convolution product of the Chebyshev polynomials T_k and the hypergeometric function $F(a,1;b;\cdot)$. Since the Chebyshev polynomials T_k can be represented as

$$T_k(z/\beta) = F(-k,k;1/2;z/\beta),$$

(cf. e.g. Luke [23], p. 301), we obtain the convolution product

$$F(a,1;c;z) * T_k(z/\beta) = {}_3F_2(-k,k,a;c,1/2;z/\beta) \tag{3.11}$$

with the generalized hypergeometric function

$${}_3F_2(-k,k,a;c,1/2;z) = \sum_{j=0}^{k} \frac{(-k)_j (k)_j (a)_j z^j}{(c)_j (1/2)_j j!}, \tag{3.12}$$

which is reduced to a polynomial. With these relations we are able to construct finally the series expansion.

Lemma 3.8 *For the confluent hypergeometric function $M(a,c;\cdot)$ we have the following series expansion*

$$M(a,c,z) = e^{\beta/2} I_0(\beta/2) + 2e^{\beta/2} \sum_{k=1}^{\infty} (-1)^k I_k(\beta/2)\, {}_3F_2(-k,k,a;c,1/2;z/\beta) \tag{3.13}$$

where the convergence is uniform on every compact set in $\mathbb{C}$.

Proof: Based on the representation (3.9) of the Kummer function using the relations (3.11) for the product $F(a,1;c;z) * T_k(z/\beta)$ and (3.10) for the Chebyshev expansion of the exponential function we obtain the asserted expansion. According to Theorem 3.3 a similar argument as in (3.5) using the Bernstein Lemma yields uniform convergence of the series expansion (3.13) due to the uniform convergence of the Chebyshev expansion (3.10) of the exponential function. □

We notice that the coefficients depend only on the interval K_β and are independent of the parameters a and c of the Kummer function M. For the computation of the polynomials ${}_3F_2$ we can apply a four-term recurrence formula, which can be found in Luke [24], p. 147. In our case we obtain the following result.

Lemma 3.9 *For the polynomials $F_k(\cdot) := {}_3F_2(-k,k,a;c,1/2;\cdot)$ we have the following recurrence formula*

$$\begin{aligned}(k-2)(k+c-1)F_k(x) = &-\left[2(k-1)(k+c-2)(k-3/2)\ -2k(k-2)(k+c-1)\right.\\ &\left.+4x\cdot(k-2)(k+a-1)\right]F_{k-1}(x)\\ &-\left[2(k-2)(k-c-1)(k-3/2)\ -2(k-3)(k-1)(k-c-2)\right.\\ &\left.-4x\cdot(k-1)(k-a-2)\right]F_{k-2}(x)\\ &+(k-1)(k-c-2)F_{k-3}(x),\end{aligned}$$

with the starting values $F_0(x)=1$, $F_1(x)=1-2ax/c$ and $F_3(x)=1-8ax/c+8(a)_2x^2/(c)_2$.

If we take into consideration, that one of our main goals is the efficient computation and programming of the confluent hypergeometric functions, we have to take the computational effort into account. We can state, that with respect to the precomputation of the coefficients and the use of the recurrence formula for the computation of the polynomials, we are able to compute with a essentially lower effort in comparison to the Chebyshev expansion.

3.1.4 Convergence theory

Application of Theorem 3.5 and Corollary 3.7 leads to the following important result for the polynomials ${}_3F_2$, which enables us to prove results on the asymptotic behaviour of the absolute errors of the partial sums of expansion (3.13) (cf. [38]).

Theorem 3.10 *For parameters* $0 < \mathrm{Re}(a) < \mathrm{Re}(c)$ *we have*

$$\|{}_3F_2(-k,k,a;c,1/2;\cdot)\|_{[0,1]} \leq \delta_{a,c} \qquad (k \in \mathbb{N}_0).$$

In the case of parameters $0 < a \leq c$ *we have*

$$\|{}_3F_2(-k,k,a;c,1/2;\cdot)\|_{[0,1]} = 1 \qquad (k \in \mathbb{N}_0).$$

Proof: First, application of Theorem 3.5 with $\varphi = F(a,1;c;\cdot)$ and $\psi = T_k$ leads to the first inequality. Then, application of Corollary 3.7 with $\varphi = F(a,1;c;\cdot)$ and $\psi = T_k$ yields

$$\|{}_3F_2(-k,k,a;c,1/2;\cdot)\|_{[0,1]} = \|F(a,1;c;\cdot) * T_k\|_{[0,1]} \leq \|T_k\|_{[0,1]} = 1$$

for parameters $0 < a \leq c$. Since ${}_3F_2(-k,k,a;c,1/2;0) = 1$, we get the asserted equality. □

Let $\mathcal{S}_n(a,c,\beta,\cdot)$ denote the n-th partial sum of the expansion (3.13) and $\mathcal{T}_n(\beta,\cdot)$ the n-th partial sum of the Chebyshev expansion (3.10) of the exponential function for $n \in \mathbb{N}$. Then for the asymptotic behaviour of the absolute error we obtain

Corollary 3.11 *For parameters* $0 < a \leq c$ *and intervals* $K_\beta = [0,\beta]$ *we have*

$$\|M(a,c;\cdot) - \mathcal{S}_{n-1}(a,c,\beta,\cdot)\|_{K_\beta} \sim \frac{2e^{\mathrm{Re}(\beta)/2}}{\sqrt{2\pi n}} \left(\frac{e|\beta|}{4n}\right)^n \qquad (n \to \infty).$$

Proof: Since ${}_3F_2(-k,k,a;c,1/2;0) = 1$, the application of Theorem 3.10 leads to

$$\|M(a,c;\cdot) - \mathcal{S}_{n-1}(a,c,\beta,\cdot)\|_{K_\beta} \leq \|\exp(\cdot) - \mathcal{T}_{n-1}(\beta;\cdot)\|_{K_\beta} \tag{3.14}$$

$$\leq 2e^{\mathrm{Re}(\beta)/2} \sum_{k \geq n} |I_k(\beta/2)| \tag{3.15}$$

and

$$\|M(a,c;\cdot) - \mathcal{S}_{n-1}(a,c,\beta,\cdot)\|_{K_\beta} \geq 2e^{\mathrm{Re}(\beta)/2} \left| \sum_{k \geq n} (-1)^k I_k(\beta/2) \right|. \tag{3.16}$$

Because of the asymptotic behaviour of the modified Bessel functions I_n (see Abramowitz, Stegun [1], p. 365), we have $(n \to \infty)$

$$|I_n(z)| \sim \frac{1}{\sqrt{2\pi n}} \left(\frac{e|z|}{2n}\right)^n \qquad (z \in \mathbb{C}),$$

and therefore

$$\|M(a,c;\cdot)-\mathcal{S}_{n-1}(a,c,\beta,\cdot)\|_{K_\beta}\sim 2e^{\mathrm{Re}(\beta)/2}\,|I_n(\beta/2)|\sim\frac{2e^{\mathrm{Re}(\beta)/2}}{\sqrt{2\pi n}}\left(\frac{e|\beta|}{4n}\right)^n,$$

what is the assertion. □

In a similar way we get for $0<\mathrm{Re}(a)<\mathrm{Re}(c)$ the results

$$\|M(a,c;\cdot)-\mathcal{S}_{n-1}(a,c,\beta;\cdot)\|_{K_\beta}\le\delta_{a,c}\|\exp(\cdot)-\mathcal{T}_{n-1}(\beta;\cdot)\|_{K_\beta} \tag{3.17}$$

$$\le 2e^{\mathrm{Re}(\beta)/2}\delta_{a,c}\sum_{k\ge n}|I_k(\beta/2)| \tag{3.18}$$

and

$$\|M(a,c;\cdot)-\mathcal{S}_{n-1}(a,c,\beta;\cdot)\|_{K_\beta}\ge 2e^{\mathrm{Re}(\beta)/2}\left|\sum_{k\ge n}(-1)^k I_k(\beta/2)\right|$$

and therefore the same asymptotic behaviour for the absolute error of $\mathcal{S}_n(a,c,\beta,\cdot)$ except for the constant $\delta_{a,c}$. The comparison of this result with the asymptotic behaviour of the Taylor sections leads to

Remark 3.12 For the Taylor sections $s_n(a,c,\cdot)$ we obtain $(n\to\infty)$

$$\|M(a,c;\cdot)-s_{n-1}(a,c,\cdot)\|_{K_\beta}\sim\frac{(a)_n|\beta|^n}{(c)_n n!}\sim\frac{(a)_n}{(c)_n\sqrt{2\pi n}}\left(\frac{e|\beta|}{n}\right)^n, \tag{3.19}$$

and if we consider the quotient of the absolute errors, we find

$$\frac{\|M(a,c;\cdot)-\mathcal{S}_{n-1}(a,c,\beta,\cdot)\|_{K_\beta}}{\|M(a,c;\cdot)-s_{n-1}(a,c,\cdot)\|_{K_\beta}}\sim\frac{2e^{\mathrm{Re}(\beta)/2}(c)_n}{(a)_n 4^n}\sim\frac{2e^{\mathrm{Re}(\beta)/2}\Gamma(a)}{\Gamma(c)}\cdot\frac{n^{c-a}}{4^n},$$

with the "asymptotic acceleration factor" $\frac{2e^{\mathrm{Re}(\beta)/2}\Gamma(a)}{\Gamma(c)}\cdot\frac{n^{c-a}}{4^n}$.

Remark 3.13 In the case of arbitrary complex parameters a and c, which are no negative integer, we are not allowed to apply Theorem 3.5 and therefore the inequality (3.17) for the absolute error is no longer valid. But we have an asymptotically optimal rate of convergence of the partial sums $\mathcal{S}_n(a,c,\beta,\cdot)$ in the following sense: A sequence of polynomials P_n of degree $\le n$ is said to be *maximally convergent* on a compact set K to an entire function f of order $\rho\in(0,\infty)$, if

$$\limsup_{n\to\infty} n^{1/\rho}\|f-P_n\|_K^{1/n}=\limsup_{n\to\infty} n^{1/\rho}\,E_n(f,K)^{1/n}, \tag{3.20}$$

where $E_n(f,K)$ denotes the error of the best approximating polynomial of degree $\le n$ on the interval K. About the rate of convergence of the best approximating polynomial of an

entire functions of order $\rho \in (0,\infty)$ and type $\tau \in (0,\infty)$ we know by a result found in Rice [35] or Winiarski [47]

$$\limsup_{n\to\infty} n^{1/\rho}\, E_n(f,K)^{1/n} = c(K)(e\rho\tau)^{1/\rho}, \tag{3.21}$$

where $c(K)$ denotes the capacity of the compact set K, if it exists. In the case of $K = K_\beta$, we have $c(K_\beta) = \beta/4$.

Now we can prove the asymptotical optimality of the polynomials $\mathcal{S}_n(a,c,\beta,\cdot)$ for arbitrary complex parameters a and c. We remark that according to (3.19) we have no maximal convergence for the Taylor sections $s_n(a,c,\cdot)$.

Theorem 3.14 *For arbitrary complex parameters a and c with $c \notin (-\mathbb{N}_0)$ and intervals K_β we have*

$$\limsup_{n\to\infty} n\|M(a,c;\cdot) - S_n(a,c,\beta,\cdot)\|_{K_\beta}^{1/n} = \frac{e\beta}{4} = \limsup_{n\to\infty} n\, E_n(M,K_\beta)^{1/n}.$$

Proof: The application of Theorem 3.3 ensures for every $\delta > 1$ and k sufficiently large the inequality

$$\|{}_3F_2(-k,k,a;c,1/2;\cdot)\|_{[0,1]} \le \delta^k.$$

Therefore we get in this case inequalities for the absolute error of the form

$$\|M(a,c;\cdot) - S_{n-1}(a,c,\beta,\cdot)\|_{K_\beta} \le 2e^{\mathrm{Re}(\beta)/2} \sum_{k\ge n} \delta^k\, |I_k(\beta/2)|$$

and

$$\|M(a,c;\cdot) - S_{n-1}(a,c,\beta,\cdot)\|_{K_\beta} \ge 2e^{\mathrm{Re}(\beta)/2} \left| \sum_{k\ge n} (-1)^k I_k(\beta/2) \right|.$$

Then we obtain with the asymptotic behaviour of the modified Bessel functions for $n \to \infty$

$$2e^{\mathrm{Re}(\beta)/2} \sum_{k\ge n} \delta^k\, |I_k(\beta/2)| \sim 2e^{\mathrm{Re}(\beta)/2}\delta^n\, |I_n(\beta/2)| \sim \frac{2e^{\mathrm{Re}(\beta)/2}}{\sqrt{2\pi n}} \delta^n \left(\frac{e|\beta|}{4n}\right)^n$$

and

$$2e^{\mathrm{Re}(\beta)/2} \left| \sum_{k\ge n} (-1)^k I_k(\beta/2) \right| \sim \frac{2e^{\mathrm{Re}(\beta)/2}}{\sqrt{2\pi n}} \left(\frac{e|\beta|}{4n}\right)^n.$$

Thus we have (replacing n by $n+1$)

$$\frac{e\beta}{4} \le \limsup_{n\to\infty} n\|M(a,c;\cdot) - S_n(a,c,\beta,\cdot)\|_{K_\beta}^{1/n} \le \frac{e\beta}{4}\delta$$

for all $\delta > 1$. So we get the first asserted equality.

Since $M(a;c;z)$ has order $\rho = 1$ and type $\tau = 1$, we get now the asymptotical optimality from the equation (3.21), what proves the second asserted equation. □

In the previous chapter we have already seen that many important special functions can be represented by means of confluent hypergeometric functions. Three important examples are the Bessel functions, Fresnel integrals and the Coulomb wave functions. In [27] Müller developed a method for computing the Bessel functions J_λ with variable order on compact intervals. For computation methods based on Taylor series and asymptotic expansions see Amos et al. [2]. In order to compute these special functions with the help of the function M we have to evaluate the function $M(a;c;z)$ at pure imaginary arguments $z = ix$. Hence we consider in the next section on numerical results especially pure imaginary arguments of the confluent hypergeometric function.

3.1.5 Numerical results

As mentioned above, the major disadvantage of the Taylor sections results from cancellation errors, which occur in the evaluation of the partial sums $s_n(a,c,\cdot)$. If we consider the simple case $z = 25i$ and $a = c = 1$, we have

$$M(1;1;25i) = e^{25i} = \sum_{\nu=0}^{\infty} \frac{(25i)^\nu}{\nu!},$$

and a largest term of the magnitude

$$\max_{\nu\in\mathbb{N}} \left| \frac{(25i)^\nu}{\nu!} \right| = \frac{25^{25}}{25!} \approx 5.726042115469874 \cdot 10^9.$$

Since $|M(1,1;25i)| = |e^{25i}| = 1$, we are confronted with a loss of 9 decimal digits.
Now we present the numerical results for computing the Kummer function $M(a;c;z)$ for real and complex parameters.

Real parameters

As a first example we consider a compact interval on the imaginary axis with $\beta = i\gamma = 25i$, which means $K_\beta = [0, 25i]$. The following figures show approximations of the significant decimal digits for $M(a;c;z)$, given by

$$d_1(a,c,z) = \min\left\{15, -\log_{10}\left(\frac{|M(a;c;z) - \mathcal{S}_n(a,c,\beta;z)|}{|M(a;c;z)|}\right)\right\} \tag{3.22}$$

and

$$d_2(a,c,z) = \min\left\{15, -\log_{10}\left(\frac{|M(a;c;z) - s_n(a,c,z)|}{|M(a;c;z)|}\right)\right\}. \tag{3.23}$$

The results are computed in double precision arithmetic providing an accuracy of about 16 decimal digits. Since the usual accuracy requirement for special functions computation routines is 15 decimal digits in double precision programs, we have cut off the errors at a

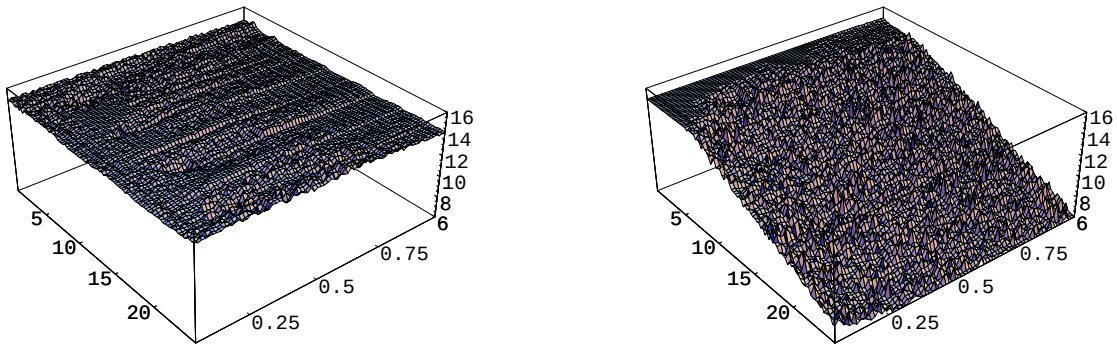

Figure 3.2: d_1 and d_2 for $z = 0.5i(0.25i)25i$, $a = 0.05(0.0125)1.0$ and $c = 0.5$ with $\deg(\mathcal{S}_n) = 45$, $\deg(s_n) = 100$

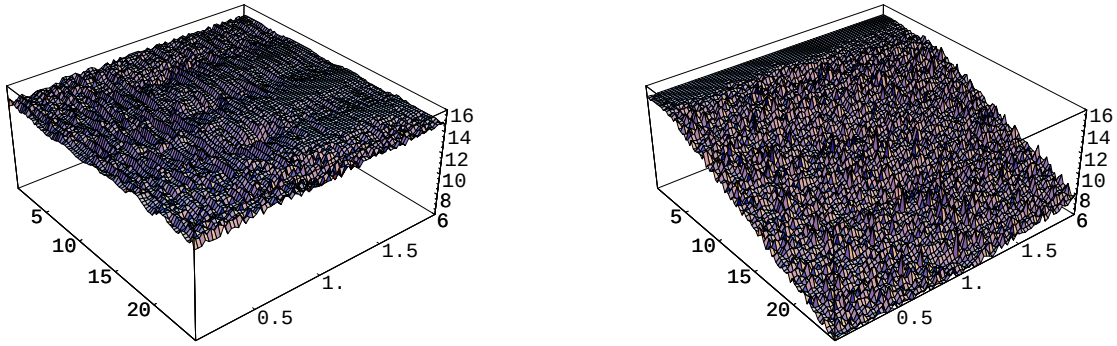

Figure 3.3: d_1 and d_2 for $z = 0.5i(0.25i)25i$, $a = 0.5$ and $c = 0.05(0.025)2.0$ with $\deg(\mathcal{S}_n) = 45$, $\deg(s_n) = 100$

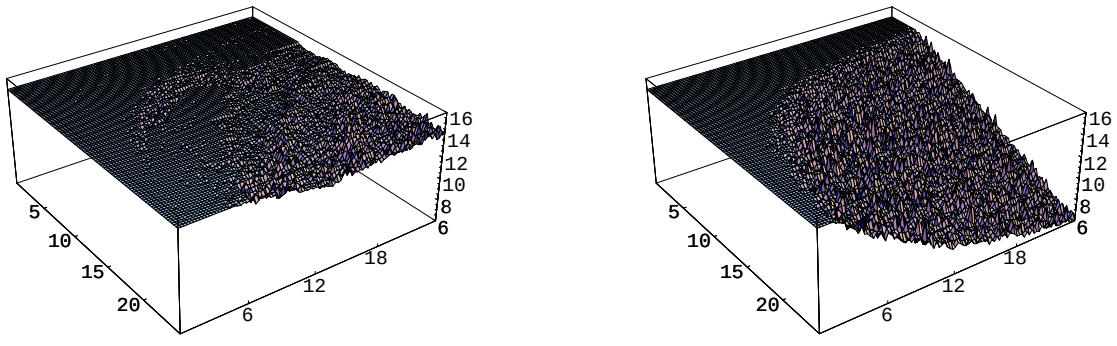

Figure 3.4: d_1 and d_2 for $z = 0.5i(0.25i)25i$, $a = 0.5(0.25)24.5$ and $c = 24.5$ with $\deg(\mathcal{S}_n) = 40$, $\deg(s_n) = 95$

level of 15 digits. Therefore all values on the 15-digits level represent approximations within the usual tolerance. The choice of degrees 100 (Figure 3.2), 100 (Figure 3.3) and 95 (Figure 3.4) for the Taylor sections and 45 (Figure 3.2), 45 (Figure 3.3) and 40 (Figure 3.4) for the sections of expansion (3.13) ensures that the errors result exclusively from cancellation, so that the choice of higher degrees does not provide a reduction of these errors. On the two horizontal axes we see the varying argument z and the varying parameter a (Figure 3.2 and Figure 3.4) or c (Figure 3.3). The figures show that the partial sums of the expansion (3.13) yield a satisfying accuracy both for parameters $a \leq c$ and in the case of $a > c$.

Complex parameters

We consider Coulomb wave functions with parameters $a = L+1-i\eta$, $c = 2L+2$ and argument $z = 2i\rho$ on compact intervals on the imaginary axis $[0, 40i]$. Because of their definition, these functions belong to the class of confluent hypergeometric functions with $0 < \mathrm{Re}(a) < \mathrm{Re}(c)$, so the theoretical convergence results from above are applicable. We also calculate the relative errors of $\mathcal{S}_n(a, c, \beta; \cdot)$ and the Taylor sections $s_n(a, c, \cdot)$ through $d_1(a, c, z)$ and $d_2(a, c, z)$. Also in this case we compute the results in double precision arithmetic, shown in Figure 3.5 and Figure 3.6, and the chosen degrees of the approximating polynomials ensure that the obtained errors result exclusively from cancellation.

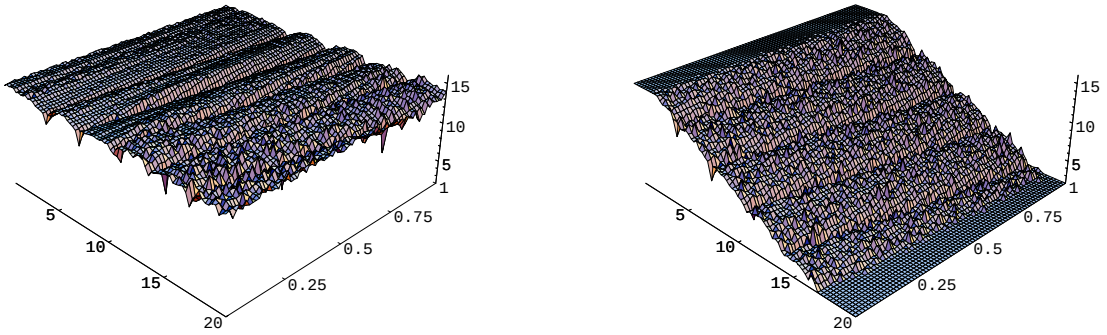

Figure 3.5: d_1 and d_2 for $\rho = 0.25(0.25)20$, $\eta = 0.0125(0.0125)1$ and $L = 1$ with $\deg(\mathcal{S}_n) = 60$, $\deg(s_n) = 120$

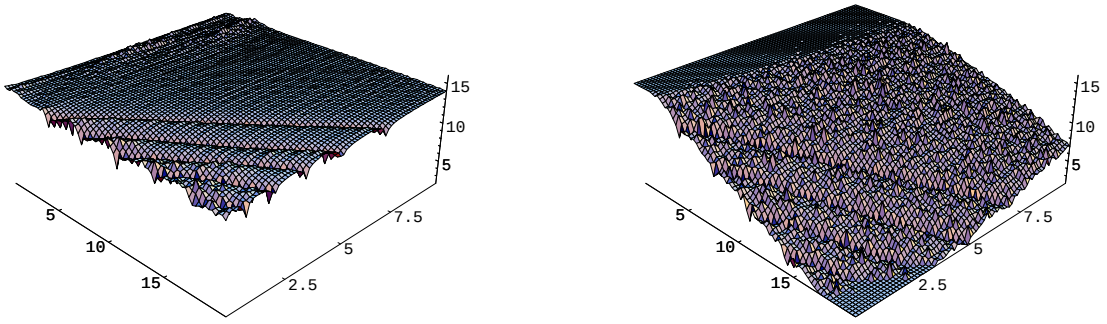

Figure 3.6: d_1 and d_2 for $\rho = 0.25(0.25)20$, $\eta = 0.125(0.125)10$ and $L = 1$ with $\deg(\mathcal{S}_n) = 60$, $\deg(s_n) = 120$

On the horizontal axes in Figure 3.5 and Figure 3.6 we see the varying argument ρ and the varying parameter η. The "waves" and the decreasing accuracy between them are due to the zeros of $M(L+1-i\eta; 2L+2; 2i\rho)$.

3.2 Computation with recurrence relations

Recurrence relations represent an important tool for computing special functions or even a sequence of special functions. The essential advantage of such methods is that in many cases they are easy to implement and the computational effort is low. But on the other hand such relations are often very susceptible to errors in numerical computation, like rounding errors. Therefore it is very important to know whether the recurrence relation is (numerically) stable.

3.2.1 Basics of linear difference equations

In this section we will give an overview on the theory of linear difference equations, especially of second order. We state the main results concerning solutions of the equations and their asymptotic behaviour. For this section we refer to Temme [40], pp. 335-342 and Gautschi [11].
We consider for $n = 1, 2, \ldots$ the following (three-term) recurrence relation

$$y_{n+1} + a_n y_n + b_n y_{n-1} = 0, \tag{3.24}$$

with given coefficients a_n and $b_n \neq 0$. An equation of the type (3.24) is called *linear homogeneous difference equation of second order*. The general solution (y_n) of equation (3.24) is given through

$$y_n = p f_n + q g_n, \tag{3.25}$$

with two linearly independent solutions (f_n) and (g_n) of the equation (3.24), and p, q are constants independent of n. Now, we are interested in those solutions (f_n), (g_n) satisfying the relation

$$\lim_{n\to\infty} \frac{f_n}{g_n} = 0. \tag{3.26}$$

This condition implies

$$\lim_{n\to\infty} \frac{f_n}{y_n} = 0$$

for any solution (y_n) not being a constant multiple of (f_n). Indeed, if (y_n) is not proportional to (f_n) we have $q \neq 0$, and therefore we get

$$\lim_{n\to\infty} \frac{f_n}{y_n} = \lim_{n\to\infty} \frac{f_n/g_n}{q + p(f_n/g_n)} = 0.$$

The set of all solutions (f_n) of equation (3.24) having the property (3.26) forms a one-dimensional subspace of the space of all solutions. It is not possible to have two linearly independent solutions (f_n), $(\tilde{f}_n)$ satisfying condition (3.26). Solutions of this subspace are called *minimal*, whereas any non-minimal solution is called *dominant*. We state that each dominant solution of equation (3.24) is asymptotically proportional to (g_n).

Labelling the initial values y_0, y_1 as well as f_0, f_1, g_0, g_1 the constants p and q can be computed as

$$p = \frac{g_1 y_0 - g_0 y_1}{f_0 g_1 - f_1 g_0}, \quad q = \frac{f_1 y_0 - f_0 y_1}{f_1 g_0 - f_0 g_1}.$$

We remark that the denominators are different from zero when we have linearly independent solutions (f_n), (g_n).

Now, we will illustrate the problem when computing minimal solutions (or any constant multiple of them). If we calculate such a sequence (f_n) by the relation (3.24) using approximate initial values $y_0 \approx f_0$, $y_1 \approx f_1$, caused by rounding errors for example, but computing with infinite precision, the resulting solution (y_n) will be in general linearly independent of the minimal solution (f_n). So, the condition (3.26) implies

$$\left| \frac{y_n - f_n}{f_n} \right| \to \infty \quad (n \to \infty).$$

That means the relative error of the computed approximation (y_n) to (f_n) diverges. The following examples will show the difficulty of forward computing of minimal solutions. For Example 3.15 and Example 3.16 cf. Gautschi [11] and [12].

Example 3.15 We compute the Bessel functions of the first kind $J_n(x)$ for fixed real x and $n = 0, 1, 2, \ldots$ using the recurrence relation

$$y_{n+1} - \frac{2n}{x} y_n + y_{n-1} = 0. \tag{3.27}$$

We compute for $x = 1$ with known initial values $J_0(1)$ and $J_1(1)$ in double precision. The results of the forward recurrence, $\hat{f}_n$, and the exact values, $f_n = J_n(1)$, are shown in Table 3.1, where the underlined digits coincide with the correct ones. We recognize the absurd results of the forward recurrence since we know that $J_n(1)$ decrease with increasing n and $J_n(1) \to 0$ for $n \to \infty$. Also the negative values for $n \geq 10$ indicate that something goes wrong. The problem can be explained by the arguments introduced before. Since $Y_n(x)$, the Bessel function of the second kind, is also a solution of equation (3.27) we have, recalling the asymptotic behaviour of the Bessel functions for $n \to \infty$ (see Abramowitz, Stegun [1], p. 365)

$$J_n(x) \sim \frac{1}{\sqrt{2\pi n}} \left(\frac{ex}{2n} \right)^n, \qquad Y_n(x) \sim -\sqrt{\frac{2}{\pi n}} \left(\frac{2n}{ex} \right)^n$$

and setting $f_n = J_n(x)$, $g_n = Y_n(x)$,

$$\frac{f_n}{g_n} \sim -\frac{1}{2} \left(\frac{ex}{2n} \right)^{2n} \qquad (n \to \infty),$$

so that the condition (3.26) is fulfilled and therefore $f_n = J_n(x)$ is the minimal solution of equation (3.27).

n	$\hat{f}_n$	$f_n = J_n(1)$
0	7.651976865579666E-01	7.651976865579666E-01
1	4.400505857449335E-01	4.400505857449335E-01
2	1.149034849319004E-01	1.149034849319005E-01
3	1.956335398266806E-02	1.956335398266841E-02
4	2.476638964107991E-03	2.476638964109955E-03
5	2.497577301958653E-04	2.497577302112344E-04
6	2.093833785066224E-05	2.093833800238927E-05
7	1.502324012081502E-06	1.502325817436808E-06
8	9.419831847878868E-08	9.422344172604501E-08
9	4.849083579117064E-09	5.249250179911875E-09
10	-6.914814054681528E-09	2.630615123687453E-10
11	-1.431453646727476E-07	1.198006746303137E-11
12	-3.142283208745766E-06	4.999718179448405E-13
13	-7.527165164522565E-05	1.925616764480173E-14
14	-1.953920659567121E-03	6.885408200044226E-16
15	-5.463450681623416E-02	2.297531532210344E-17
16	-1.637081283827458E+00	7.186396586807492E-19
17	-5.233196657566241E+01	2.115375568053261E-20
18	-1.777649782288695E+03	5.880344573595758E-22
19	-6.394306019581734E+04	1.548478441211653E-23
20	-2.428058637658770E+06	3.873503008524658E-25

n	$\hat{f}_n$	$f_n = n!(e^x - e_n(x))$
0	1.718281828459045E+00	1.718281828459045E+00
1	7.182818284590450E-01	7.182818284590452E-01
2	4.365636569180901E-01	4.365636569180904E-01
3	3.096909707542705E-01	3.096909707542714E-01
4	2.387638830170821E-01	2.387638830170856E-01
5	1.938194150854109E-01	1.938194150854282E-01
6	1.629164905124653E-01	1.629164905125694E-01
7	1.404154335872576E-01	1.404154335879862E-01
8	1.233234686980608E-01	1.233234687038897E-01
9	1.099112182825479E-01	1.099112183350075E-01
10	9.911218282547906E-02	9.911218335007541E-02
11	9.023401108026973E-02	9.023401685082952E-02
12	8.280813296323685E-02	8.280820220995427E-02
13	7.650572852207915E-02	7.650662872940557E-02
14	7.108019930910813E-02	7.109280221167809E-02
15	6.620298963662207E-02	6.639203317517139E-02
16	5.924783418595325E-02	5.863302364662063E-02
17	7.213181161205284E-03	5.539442563917151E-02
18	-8.701627390983048E-01	5.249408714425881E-02
19	-1.753309204286779E+01	4.988174288517624E-02
20	-3.516618408573558E+02	4.751660058870116E-02

Table 3.1: Approximated solution $\hat{f}_n$ and exact values f_n of the Examples 3.15 and 3.16

Example 3.16 We consider the first order recurrence

$$(n+1)y_n - y_{n+1} = x^{n+1},$$

with the solution $f_n = n!(e^x - e_n(x))$, where $e_n(x) = \sum_{k=0}^{n} \frac{x^k}{k!}$ and $x > 0$. We execute some iterations with the starting value $f_0 = e^x - 1$ and $x = 1$. The computation is carried out in double precision arithmetic, so that we have an accuracy of 16 decimal digits. The results are shown in Table 3.1, where again the underlined digits coincide with the correct ones.

For a detailed investigation on the numerical instability of forward recursions we refer to [46], where Wimp defined a so-called index of stability for the forward computation. With the use of this index Wimp characterized the stability of recurrence relations.
Often it is possible to determine dominant and minimal solutions of linear homogeneous difference equations using the asymptotic behaviour of the coefficients a_n and b_n. For an overview on the asymptotic theory of linear second order difference equations we refer to Gautschi [11].

Recurrence formulae for confluent hypergeometric functions

Now we consider recurrence formulae for confluent hypergeometric functions. The following three examples show the properties of three different recurrence relations for the confluent hypergeometric functions in a single parameter as well as in both parameters.

Example 3.17 We consider the recurrence relation with respect to parameter a (see Temme [40], p. 341)

$$(n+a+1-c)y_{n+1} + (c-z-2a-2n)y_n + (a+n-1)y_{n-1} = 0 \tag{3.28}$$

which has the solutions

$$f_n = \frac{\Gamma(a+n)}{\Gamma(a)} U(a+n;c;z), \quad g_n = \frac{\Gamma(a+n)}{\Gamma(a+n+1-c)} M(a+n;c;z).$$

The asymptotic formulae ($n \to \infty$) of the solutions (f_n) and (g_n) can be obtained (see Temme [40], p. 341) with use of the asymptotic expansions of the confluent hypergeometric functions for large parameter a (see Abramowitz, Stegun [1], p. 508)

$$f_n \sim c_1 n^{c/2-3n/4} e^{-2\sqrt{nz}}, \quad g_n \sim c_2 n^{c/2-3n/4} e^{2\sqrt{nz}},$$

where the constants c_1 and c_2 do not depend on n.

Example 3.18 We consider the recurrence relation with respect to parameter c (see Temme [40], p. 342)

$$zy_{n+1} + (1-c-n-z)y_n + (c+n-a-1)y_{n-1} = 0 \tag{3.29}$$

which has the solutions

$$f_n = \frac{\Gamma(c+n-a)}{\Gamma(c+n)} M(a; c+n; z), \quad g_n = U(a; c+n; z).$$

Here, with the help of the asymptotic relation $M(a; c+n; z) \sim 1$ for $n \to \infty$, we have the asymptotic relations $(n \to \infty)$

$$f_n \sim n^{-a}, \quad g_n \sim z^{1-c-n} \frac{\Gamma(c+n-1)}{\Gamma(a)}.$$

Example 3.19 We consider the recurrence relation with respect to both parameters a and c (see Wimp [46], p. 61) for $n = 1, 2, \ldots$

$$(2n+c-2)(n+c-a)y_{n+1} + (2n+c-1)\left((2a-c) + \frac{(2n+c-2)(2n+c)}{z}\right) y_n - (2n+c)(n+a-1)y_{n-1} = 0 \quad (3.30)$$

which has the solutions

$$f_n = \frac{z^n (a)_n}{(c)_{2n}} M(n+a; 2n+c; z), \quad (3.31)$$

$$g_n = \frac{z^{-n}\Gamma(2n+c-1)}{\Gamma(n+c-a)} M(a+1-c-n; 2-c-2n; z). \quad (3.32)$$

We consider the asymptotic behaviour for $n \to \infty$ of f_n and g_n. First, we get the asymptotic relation

$$M(n+a; 2n+c; z) \sim e^{z/2} \qquad (n \to \infty)$$

and for the quotient of Pochhammer symbols, with use of the representation

$$(z)_{n+1} \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \frac{n^z}{\Gamma(z)}$$

for $n \to \infty$ and $z \in \mathbb{C} \setminus (-\mathbb{N}_0)$ (see Henrici [16], p. 28 and Stirling's formula)

$$\frac{(a)_n}{(c)_{2n}} \sim \frac{\Gamma(c)}{\Gamma(a)} \frac{(n-1)^{n+a-1/2}}{(2n-1)^{2n+c-1/2}} e^n \sim \frac{\Gamma(c)}{\Gamma(a)} \frac{1}{2^{2n+c-1/2}} \frac{1}{n^{n-a+c}} e^n.$$

Finally, we obtain for the minimal solution

$$f_n \sim \frac{\Gamma(c)}{\Gamma(a)} \frac{1}{2^{c-1/2}} \left(\frac{ez}{4}\right)^n e^{z/2} \frac{1}{n^{n-a+c}}. \quad (3.33)$$

A similar argument shows, with use of the Stirling formula

$$\Gamma(z) \sim \sqrt{2\pi} e^{-z} z^{z-1/2}$$

for $|z| \to \infty$ and $|\arg(z)| \leq \pi - \varepsilon$, $\varepsilon > 0$ (see Luke [24], p. 32),

$$\begin{aligned}\frac{\Gamma(2n+c-1)}{\Gamma(n+c-a)} &\sim \frac{(2n+c-1)^{2n+c-3/2}}{(n+c-a)^{n+c-a-1/2}}\, e^{-n-a+1} \\ &\sim \frac{(2n+c-1)^{n+a-3/2}}{(n+c-a)^{-1/2}} \frac{(n+c-a+a-1)^{n+c-a}}{(n+c-a)^{n+c-a}}\, 2^{n+c-a} e^{-n-a+1} \\ &\sim 2^{2n+c-3/2} n^{n+a-1} e^{-n}\end{aligned}$$

and we get for the dominant solution

$$g_n \sim 2^{c-3/2} \left(\frac{4}{ez}\right)^n e^{z/2} n^{n+a-1}. \tag{3.34}$$

Remark 3.20 The recurrence relation (3.30) for the confluent hypergeometric function (see Example 3.19) can be understood as a generalization of the Bessel recurrence relation (3.27) introduced in Example 3.15. We choose the parameters $a = 1/2$, $c = 1$ and the argument $z = 2ix$, for $x \in \mathbb{R}$. Then the recurrence relation (3.30) reads as

$$(2n-1)(n+1/2)y_{n+1} + (2n-1)(n+1/2)\frac{2n}{ix}\, y_n - (2n+1)(n-1/2)y_{n-1} = 0$$

and simplifying this equation leads to

$$y_{n+1} + \frac{2n}{ix}\, y_n - y_{n-1} = 0.$$

On the other hand the minimal solution f_n (3.31) can be written as

$$f_n = \frac{(ix/2)^n}{n!}\, M(n+1/2; 2n+1; 2ix).$$

Now we understand the solution f_n as a transformation is the following way

$$f_n = i^n e^{ix} \tilde{f}_n,$$

where, recalling the representation of the confluent hypergeometric function in terms of Bessel functions (2.8),

$$\tilde{f}_n = \frac{(x/2)^n}{n!}\, e^{-ix} M(n+1/2; 2n+1; 2ix) = J_n(x).$$

Hence, this leads to the recurrence relation

$$i^{n+1} e^{ix} \tilde{y}_{n+1} + \frac{2n}{ix}\, i^n e^{ix} \tilde{y}_n - i^{n-1} e^{ix} \tilde{y}_{n-1} = 0,$$

which can be simplified to

$$\tilde{y}_{n+1} - \frac{2n}{x}\, \tilde{y}_n + \tilde{y}_{n-1} = 0$$

what is exactly the Bessel recurrence relation (3.27) with the solution $\tilde{f}_n = J_n(x)$. In this way we have shown that the recurrence relation (3.30) for the confluent hypergeometric function is a generalization of the Bessel recurrence relation.

In the next section we consider the (numerically stable) computation using such recurrence formulae, especially for computing minimal solutions.

3.2.2 The Miller algorithm

As we have seen in the previous section the numerical computation of minimal solutions might be very problematic. The idea is now to apply (3.24) in backward direction in order to compute the values $f_1, \ldots, f_N$ for fixed integer N. Then, (f_n) becomes a dominant solution and (g_n) a minimal solution. But we need the starting values f_N and f_{N-1} then. For the Miller algorithm these starting values are not required as we will see soon. For the following formulation of the algorithm and the statement of the main known properties we refer to Gautschi [11], Temme [40], pp. 343-346 and Wimp [46], pp. 29-34.

The algorithm

We assume to have a convergent series of the form

$$S := \sum_{k=0}^{\infty} c_k f_k, \tag{3.35}$$

with known $S \neq 0$ and known coefficients c_k. This series is also called a "normalizing series". Now we choose a nonnegative integer, usually large, starting value $\nu > N$ and compute a solution of (3.24) for $n = \nu - 1, \nu - 2, \ldots, 0$ through

$$y_{n+1}^{(\nu)} + a_n y_n^{(\nu)} + b_n y_{n-1}^{(\nu)} = 0$$

with the incorrect initial values

$$y_{\nu+1}^{(\nu)} = 0, \quad y_{\nu}^{(\nu)} = 1,$$

where these initial values can be replaced by other ones, but at least one of these values must be different from zero. (The choice of these initial values, however, may affect the rate of convergence of the algorithm.) Then the computed solution is a linear combination of the solutions (f_n) and (g_n) of the form

$$y_n^{(\nu)} = p_\nu f_n + q_\nu g_n, \tag{3.36}$$

for $n = 0, 1, \ldots, \nu + 1$, where the coefficients are

$$p_\nu = \frac{g_{\nu+1}}{g_{\nu+1} f_\nu - g_\nu f_{\nu+1}}, \quad q_\nu = \frac{f_{\nu+1}}{g_{\nu+1} f_\nu - g_\nu f_{\nu+1}}.$$

Since we made the assumption (3.26) on the solutions (f_n) and (g_n) we can conclude

$$\lim_{\nu\to\infty} \frac{y_n^{(\nu)}}{p_\nu} = f_n - \lim_{\nu\to\infty} \frac{f_{\nu+1}}{g_{\nu+1}} g_n = f_n \tag{3.37}$$

for $n = 0, \ldots, N$. Obviously, we have an approximation to f_n, if ν is large enough, given by the quantities $y_n^{(\nu)}$ and p_ν, whereas indeed p_ν is unknown, in general. But, by the use of the normalizing series we can compute the approximation to f_n, if ν is large enough, defining

$$f_n^{(\nu)} := y_n^{(\nu)} \frac{S}{S^{(\nu)}} \tag{3.38}$$

where $S^{(\nu)}$ is defined by

$$S^{(\nu)} := \sum_{k=0}^{\nu} c_k y_k^{(\nu)}.$$

Now, if we replace $y_k^{(\nu)}$ with $p_\nu f_k$ in $S^{(\nu)}$ having (3.37) in mind we get the asymptotic relation $p_\nu \sim S^{(\nu)}/S$. Therefore we are allowed to consider $f_n^{(\nu)}$ as an approximation to f_n for large enough integer ν. Consequently, we call the algorithm convergent if for $n = 0, 1, \ldots, N$

$$\lim_{\nu\to\infty} f_n^{(\nu)} = f_n.$$

The following result, for which we refer to Gautschi [11], yields a characterization of the convergence of the algorithm. (Here, the presented proof is slightly more detailed.)

Theorem 3.21 *Let (f_n) with $f_n \neq 0$ be a minimal solution of the recurrence relation (3.24) with the normalizing series (3.35) and let (g_n) be any other solution of the relation (3.24) (satisfying the condition (3.26)). Then the Miller algorithm converges, i.e.*

$$\lim_{\nu\to\infty} f_n^{(\nu)} = f_n$$

if and only if the condition

$$\lim_{\nu\to\infty} \frac{f_{\nu+1}}{g_{\nu+1}} \sum_{k=0}^{\nu} c_k g_k = 0 \tag{3.39}$$

is fulfilled.

Proof: Let (f_n) with $f_n \neq 0$ be a minimal solution of (3.24) and let (g_n) be any other solution of (3.24) satisfying (3.26). By using the initial values $y_\nu^{(\nu)} = 1$ and $y_{\nu+1}^{(\nu)} = 0$ we obtain

$$\begin{aligned} p_\nu f_{\nu+1} + q_\nu g_{\nu+1} &= 0 \\ p_\nu f_\nu + q_\nu g_\nu &= 1 \end{aligned}$$

and can determine $q_\nu = -(f_{\nu+1}/g_{\nu+1})p_\nu$ by the first equation. Putting this in the representation of the approximate solution $y_n^{(\nu)}$ we get

$$y_n^{(\nu)} = p_\nu \left(f_n - \frac{f_{\nu+1}}{g_{\nu+1}} g_n \right).$$

Setting

$$\sigma_\nu = \sum_{k=\nu+1}^{\infty} c_k f_k, \quad \rho_n = \frac{f_n}{g_n}, \quad \tau_\nu = \rho_{\nu+1} \sum_{k=0}^{\nu} c_k g_k$$

and using the normalization (3.38) we obtain

$$\begin{aligned} f_n^{(\nu)} &= \frac{S f_n \left(1 - \rho_{\nu+1}/\rho_n\right)}{\sum_{k=0}^{\nu} c_k f_k - \rho_{\nu+1} \sum_{k=0}^{\nu} c_k g_k} \\ &= \frac{S f_n \left(1 - \rho_{\nu+1}/\rho_n\right)}{S - \sum_{\nu+1}^{\infty} c_k f_k - \rho_{\nu+1} \sum_{k=0}^{\nu} c_k g_k} \\ &= \frac{f_n \left(1 - \rho_{\nu+1}/\rho_n\right)}{1 - \sigma_\nu/S - \tau_\nu/S}. \end{aligned}$$

We conclude that $\sigma_\nu \to 0$ for $\nu \to \infty$ since we have convergence of the normalizing series and $\rho_{\nu+1} \to 0$ for $\nu \to \infty$ because of the condition (3.26). So, we have convergence of the Miller algorithm if and only if $\tau_\nu \to 0$. □

We note that the condition (3.39) is fulfilled for bounded coefficients c_k and if we have

$$\lim_{\nu\to\infty} \frac{g_{\nu+1}}{g_\nu} = t_1, \quad \lim_{\nu\to\infty} \frac{f_{\nu+1}}{f_\nu} = t_2$$

for $|t_1| > |t_2|$ and $|t_2| < 1$. Indeed: This implies the existence of numbers r_1 and r_2 satisfying $|t_1| > r_1 > r_2 > |t_2|$ such that

$$\left|\frac{g_{\nu+1}}{g_\nu}\right| \geq r_1, \quad \left|\frac{f_{\nu+1}}{f_\nu}\right| \leq r_2$$

for ν large enough. Then, we obtain the following estimate

$$|\tau_{\nu+1}| \leq \frac{r_2}{r_1}|\tau_\nu| + \frac{r_2}{r_1}|c_{\nu+1} f_{\nu+1}|$$

which already implies that $\tau_\nu \to 0$ for $\nu \to \infty$ since $r_2/r_1 < 1$ and $c_\nu f_\nu \to 0$ (see Polyak [34], p. 45, Lemma 3).
For numerical purposes and to obtain results concerning the (asymptotic) rate of convergence, we have to consider the relative error of $f_n^{(\nu)}$. We define for $f_n \neq 0$ the relative error

$$\varepsilon_n^{(\nu)} = \frac{f_n^{(\nu)} - f_n}{f_n}$$

and obtain the following representation of the relative error (cf. Temme [40], p. 345 and Gautschi [11]).

Lemma 3.22 *For the relative error $\varepsilon_n^{(\nu)}$ of the approximation $f_n^{(\nu)}$ we have*

$$\varepsilon_n^{(\nu)} = \frac{\sigma_\nu/S - \rho_{\nu+1}/\rho_n + \tau_\nu/S}{1 - \sigma_\nu/S - \tau_\nu/S}, \tag{3.40}$$

where ρ_n, σ_ν and τ_ν are defined as before.

Proof: First, using the definition of the approximation $f_n^{(\nu)}$, we can write

$$\varepsilon_n^{(\nu)} = \frac{Sy_n^{(\nu)}/S^{(\nu)} - f_n}{f_n} = \frac{S(p_\nu + q_\nu g_n/f_n) - S^{(\nu)}}{S^{(\nu)}}.$$

Putting in the normalizing series and the partial sums $S^{(\nu)}$ yields

$$\varepsilon_n^{(\nu)} = \frac{p_\nu \sum_{k=\nu+1}^{\infty} c_k f_k + q_\nu/\rho_n \sum_{k=0}^{\infty} c_k f_k - q_\nu \sum_{k=0}^{\nu} c_k g_k}{p_\nu \sum_{k=0}^{\nu} c_k f_k + q_\nu \sum_{k=0}^{\nu} c_k g_k}.$$

Adding zero to the denominator of the form $0 = Sp_\nu\sigma_\nu - Sp_\nu\sigma_\nu$ and using the definitions of σ_ν and τ_ν we get

$$\begin{aligned}\varepsilon_n^{(\nu)} &= \frac{p_\nu\sigma_\nu + Sq_\nu/\rho_n - q_\nu\tau_\nu/\rho_{\nu+1}}{Sp_\nu - p_\nu\sigma_\nu + q_\nu\tau_\nu/\rho_{\nu+1}} \\ &= \frac{\sigma_\nu/S + q_\nu/(p_\nu\rho_n) - q_\nu\tau_\nu/(Sp_\nu\rho_{\nu+1})}{1 - \sigma_\nu/S + q_\nu\tau_\nu/(Sp_\nu\rho_{\nu+1})}.\end{aligned}$$

But this expression can still be simplified. We consider the initial value $y_{\nu+1}^{(\nu)}$ and obtain

$$y_{\nu+1}^{(\nu)} = p_\nu f_{\nu+1} + q_\nu g_{\nu+1} = 0,$$

what leads to

$$\frac{q_\nu}{p_\nu} = -\frac{f_{\nu+1}}{g_{\nu+1}} = -\rho_{\nu+1}.$$

Putting this in the last equality for the relative error we can rewrite it as

$$\varepsilon_n^{(\nu)} = \frac{\sigma_\nu/S - \rho_{\nu+1}/\rho_n + \tau_\nu/S}{1 - \sigma_\nu/S - \tau_\nu/S}$$

what is the asserted representation. □

By the use of this representation of the relative error we can state again the convergence condition for the algorithm: The relative error $\varepsilon_n^{(\nu)}$ converges to zero for $n = 0, \ldots, N$ and $\nu \to \infty$ if and only if $\tau_\nu \to 0$.

For numerical purposes it is important to have information on the asymptotic behaviour of the relative error $\varepsilon_n^{(\nu)}$. Since it is difficult to get strict estimates we approximate the representations of τ_ν and σ_ν through the dominant terms of the series. This leads to the approximations

$$\sigma_\nu \approx c_{\nu+1} f_{\nu+1}, \quad \tau_\nu \approx \rho_{\nu+1} c_\nu g_\nu$$

and with it to the approximate relative error

$$\varepsilon_n^{(\nu)} \approx \frac{c_{\nu+1} f_{\nu+1}}{S} + \frac{f_{\nu+1}}{g_{\nu+1}} \frac{c_\nu g_\nu}{S} - \frac{f_{\nu+1}}{g_{\nu+1}} \frac{g_n}{f_n} \tag{3.41}$$

There are two aspects influencing the convergence behaviour of the algorithm. First, the terms $c_{\nu+1} f_{\nu+1}$ or $c_\nu f_{\nu+1}$ on the right-hand side of (3.41) indicate that the normalizing series (3.35) has to converge fast. For the second fraction on the right-hand side we have $g_\nu/g_{\nu+1} \to 1/t_1$, as we stated earlier in order to achieve convergence of the algorithm. Second, the last fraction on the right-hand side of (3.41) indicates that the extent of dominance of the solution (g_n) in comparison to the minimal solution (f_n) is very important. In other words, the rate of convergence of (3.26) has to be taken into account.

Example 3.23 We compute again the Bessel functions as in Example 3.15, but now in backward direction using the Miller algorithm. To get an impression of the fast convergence of the Miller algorithm in this case we consider the relative error and its asymptotic behaviour. Using the asymptotic representation of the relative error of the Miller algorithm above and the asymptotic formulae for the minimal solution (f_n) and the dominant solution (g_n) of the Bessel differential equation (see Example 3.15) we obtain the following asymptotic

n	$f_n^{(\nu)}$	f_n
25	1.902219281208027E-33	1.902951751891382E-33
24	9.511096406040136E-32	9.511097932712494E-32
23	4.563424055618057E-30	4.563424055950106E-30
22	2.098223955943702E-28	2.098223955943777E-28
21	9.227621982096672E-27	9.227621982096670E-27
20	3.873503008524659E-25	3.873503008524658E-25
19	1.548478441211654E-23	1.548478441211653E-23
18	5.880344573595760E-22	5.880344573595758E-22
17	2.115375568053262E-20	2.115375568053261E-20
16	7.186396586807494E-19	7.186396586807493E-19
15	2.297531532210345E-17	2.297531532210344E-17
14	6.885408200044227E-16	6.885408200044226E-16
13	1.925616764480173E-14	1.925616764480173E-14
12	4.999718179448405E-13	4.999718179448405E-13
11	1.198006746303137E-11	1.198006746303137E-11
10	2.630615123687453E-10	2.630615123687453E-10
9	5.249250179911875E-09	5.249250179911875E-09
8	9.422344172604501E-08	9.422344172604501E-08
7	1.502325817436808E-06	1.502325817436808E-06
6	2.093833800238927E-05	2.093833800238927E-05
5	2.497577302112344E-04	2.497577302112344E-04
4	2.476638964109955E-03	2.476638964109955E-03
3	1.956335398266841E-02	1.956335398266841E-02
2	1.149034849319005E-01	1.149034849319005E-01
1	4.400505857449336E-01	4.400505857449335E-01
0	7.651976865579666E-01	7.651976865579666E-01

Table 3.2: Approximated solution $f_n^{(\nu)}$ and exact values f_n of Example 3.23

representation of the relative error for $n = 25$ and $\nu \to \infty$

$$\varepsilon_n^{(\nu)} \approx \left(\frac{ez}{2(\nu+1)}\right)^{2(\nu+1)} \left[\frac{1}{2\pi(\nu+1)} - \frac{1}{4\pi\nu}\left(\frac{ez}{2\nu}\right)^{2\nu} + \frac{1}{2}\frac{Y_{25}(z)}{J_{25}(z)}\right]. \tag{3.42}$$

We recognize the dominant factors $(ez/(2\nu+2))^{2\nu+2}$ which are responsible for the fast convergence. For illustration we present a numerical example where we chose $\nu = 25$ as starting value. The results are presented in Table 3.2. Again, the underlined digits show the exact decimal digits.

Another task is the determination or estimation of a starting value ν, where some information on asymptotic estimates of the dominant and minimal solution is needed. A further detailed investigation of the Miller algorithm can be found in Gautschi [11]. Here, Gautschi used asymptotic estimates of the underlying special functions in order to obtain estimates of the starting value of the backward recursion. There are estimated and empirical ratios ν/N given. Also the influence of the choice of the normalizing series is investigated. We note that serious problems may occur if there is cancellation in the series (3.35) itself. A further aspect is the sensitivity of the Miller algorithm to rounding errors, although our numerical analysis does not give such indication. An analysis of the error accumulation of the Miller algorithm is due to Olver and can be found in [30]. An approach for numerical computation of starting values was performed by Olver [31], where also inhomogeneous difference equations were considered.

3.2.3 Applications and numerical results

In this section we apply the previous results on the asymptotic behaviour of the relative error of the Miller algorithm to recurrence formulae for confluent hypergeometric functions and give some numerical results. For this purpose we compute the recurrence relation, introduced in Example 3.19, in backward direction using the Miller algorithm in double precision arithmetic. Computing the minimal solution (f_n) of the recurrence relation (3.30) we choose the normalizing series (see Wimp [46], p. 62)

$$S = c - 1 = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!}(c-1)_k(c+2k-1)f_k.$$

In order to investigate the asymptotic rate of convergence of the Miller algorithm we consider the relative error, with fixed n and starting value ν,

$$\varepsilon_n^{(\nu)} = \frac{y_n^{(\nu)} - f_n}{f_n} \tag{3.43}$$

approximated asymptotically by $(\nu \to \infty)$

$$\varepsilon_n^{(\nu)} \approx \frac{c_{\nu+1}f_{\nu+1}}{S} + \frac{f_{\nu+1}}{g_{\nu+1}}\frac{c_\nu g_\nu}{S} - \frac{f_{\nu+1}}{g_{\nu+1}}\frac{g_n}{f_n}.$$

Using the asymptotic formulae of the solutions (f_n) (3.33) and (g_n) (3.34) the relative error can be written as

$$\varepsilon_n^{(\nu)} \approx \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-1)(c-1)} \frac{1}{2^{c-1/2}} \left(\frac{ez}{4(\nu+1)}\right)^{\nu+1} \cdot \left[(-1)^{\nu+1}(\nu+1)^{a-1}(c+2\nu+1)e^{z/2} + (-1)^{\nu}\nu^{a-2}(c+2\nu-1)e^{z/2}\frac{z}{4} - \Gamma(c-1)(c-1)\frac{1}{2^{c-1/2}} \left(\frac{ez}{4(\nu+1)}\right)^{\nu+1} \frac{1}{(\nu+1)^{c-1}} \frac{g_n}{f_n}\right].$$

Since the influence of the last term to the rate of convergence is neglectable it can be omitted. Hence, the relative error reads as

$$\varepsilon^{(\nu)} \approx \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-1)(c-1)} \frac{1}{2^{c-1/2}} \left(\frac{ez}{4(\nu+1)}\right)^{\nu+1} \cdot \left[(-1)^{\nu+1}(\nu+1)^{a-1}(c+2\nu+1)e^{z/2} + (-1)^{\nu}\nu^{a-2}(c+2\nu-1)e^{z/2}\frac{z}{4}\right], \quad (3.44)$$

where we omit the index n of $\varepsilon_n^{(\nu)}$. Hence, we investigate the numerical behaviour of the relative error $\varepsilon^{(\nu)}$, defined in (3.43), in comparison to its asymptotic approximation (3.44) for fixed parameters and argument, in order to check the quality of the asymptotic approximation. For this purpose we compute the number of significant decimal digits by

$$d^{(\nu)} = d^{(\nu)}(a,c,z) = \min\left\{15, -\log_{10}\left(\frac{\left|M(a;c;z) - f_0^{(\nu)}(z)\right|}{|M(a;c;z)|}\right)\right\}$$

and its asymptotic approximation by

$$\hat{d}^{(\nu)} = \min\left\{15, -\log_{10}\left|\varepsilon^{(\nu)}\right|\right\}.$$

The results are presented in Figure 3.7 and Figure 3.8, where the dashed lines show the asymptotic approximation $\hat{d}^{(\nu)}$ of the relative error $d^{(\nu)}$, which is shown by the dotted lines.

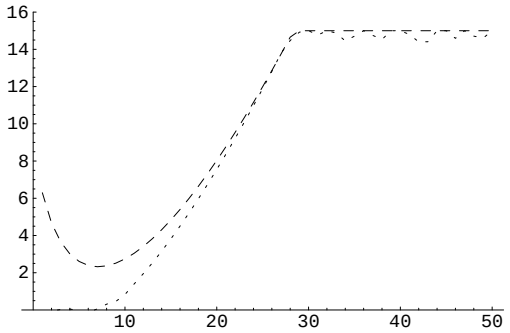

Figure 3.7: $d^{(\nu)}(10, 20, 10i)$, $\hat{d}^{(\nu)}$

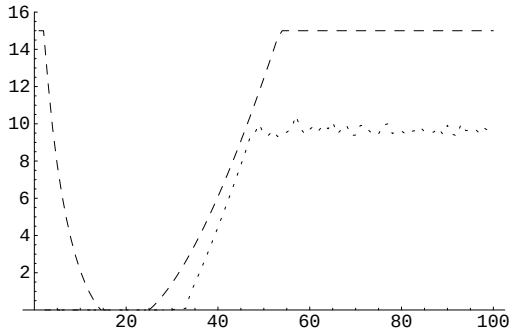

Figure 3.8: $d^{(\nu)}(24.5, 24.5, 25i)$, $\hat{d}^{(\nu)}$

Both figures show the number of significant decimal digits where (the starting value) ν is varied ($\nu = 1, \ldots, 50$ in Figure 3.7 and $\nu = 1, \ldots, 100$ in Figure 3.8). We recognize in Figure 3.8 cancellation of nearly 5 decimal digits. This phenomenon can be explained by cancellation occurring during the computation of the normalizing series. So, the choice of the normalizing series plays an important role. We can conclude that in particular increasing modulus of the parameter a and the argument z lead to worse results, in the sense of slower convergence and with respect to cancellation errors.
In order to see into the numerical behaviour of the relative error (3.43) of the Miller algorithm for varying parameters and argument, we consider again real and complex parameters. We compute again the (logarithmic) relative errors, or more precisely the number of significant decimal digits $d^{(\nu)}$ in the case of real parameters a and c. The results are shown in Figure 3.9 for small positive a and c (left figure) and for larger positive parameters a and c (right figure). We recognize a decrease of the number of significant decimal digits especially for the larger parameter case. Analogously, we compute in the complex case, the case of Coulomb wave functions, the number of significant decimal digits $d^{(\nu)}$ for parameters $a = L + 1 - i\eta$, $c = 2L + 2$ and argument $z = 2i\rho$. A crucial point is the determination of the starting value ν. As can be seen in Figure 3.10 there is no significant difference between $d^{(80)}$ and $d^{(110)}$.

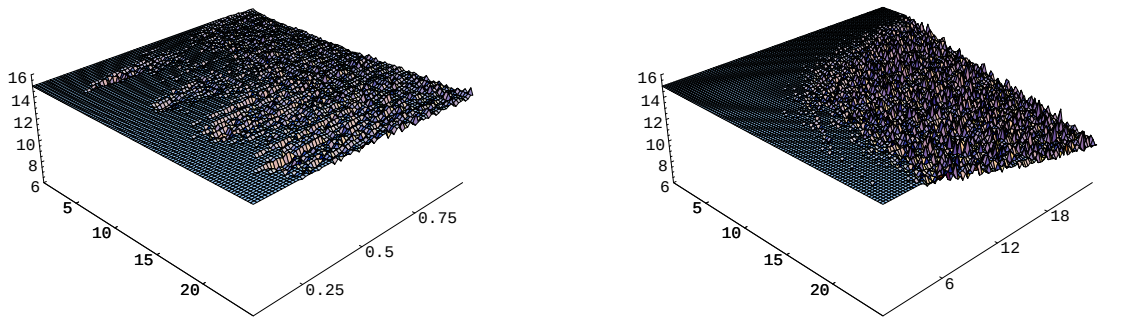

Figure 3.9: $d^{(110)}$ for $a = 0.05(0.0125)1.0$, $c = 0.5$ and $z = 0.5i(0.25i)25i$ and $d^{(100)}$ for $a = 0.5(0.25)24.5$, $c = 24.5$ and $z = 0.5i(0.25i)25i$

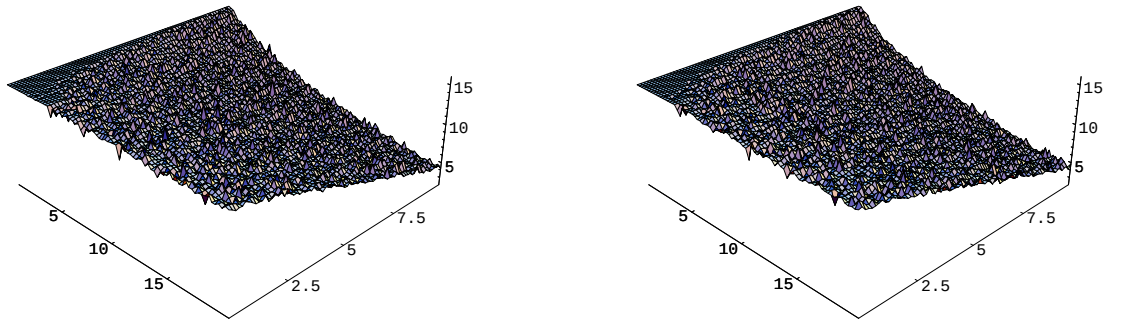

Figure 3.10: $d^{(80)}$ and $d^{(110)}$for $\rho = 0.25(0.25)20$, $\eta = 0.125(0.125)10$ and $L = 1$

Remark 3.24 We point out an important application of the recurrence formula introduced in Example 3.17, a recurrence relation for the confluent hypergeometric functions with respect to the first parameter a. We remember the results in the previous Section 3.1, where we have shown favourable convergence behaviour of the partial sums of the constructed series expansion (3.13) in the case $0 < \mathrm{Re}(a) < \mathrm{Re}(c)$, or $0 < a \leq c$ in the case of real parameters. Now it is possible to apply the recurrence relation (3.28) in forward direction using starting values computed with the partial sums of the series expansion (3.13), in order to compute the function $M(a; c; z)$ for parameter $a > c$. From the numerical point of view this procedure

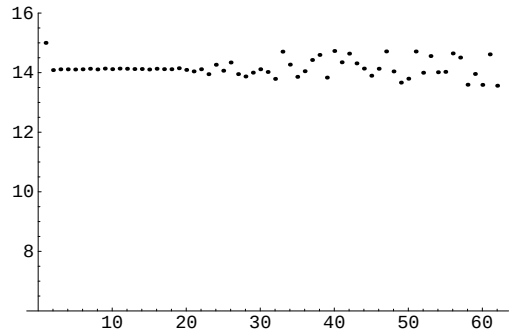

Figure 3.11: $d_{62}(2, 1.5, 25i)$

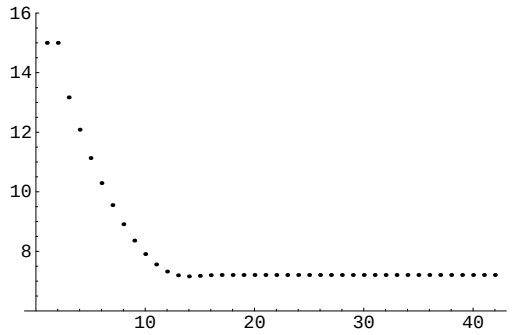

Figure 3.12: $d_{42}(1.5, 25, 25i)$

is interesting since the computed solution (g_n) of the recurrence relation (3.28) is dominant. We compute the number of significant decimal digits for different n and fixed parameters and argument, i.e. we have to evaluate an expression

$$d_n = d_n(a, c, z) = \min\left\{15, -\log_{10}\left(\frac{|\hat{g}_n - g_n|}{|g_n|}\right)\right\},$$

where d_n denotes the number of significant decimal digits, $\hat{g}_n$ is the computed solution of the recurrence relation and g_n the exact solution, given by (see Example 3.17)

$$g_n = \frac{\Gamma(a+n)}{\Gamma(a+n+1-c)}\, M(a+n; c; z).$$

In fact, Figure 3.11 shows that even for $n = 1, \ldots, 62$ the recurrence nearly remains numerically stable in the sense of forward stability for $a > c$. On the other hand we recognize some problems in the case of $a < c$. The result can be seen in Figure 3.12, where we at least observe a stabilization at a level of 7 decimal digits, for $n = 1, \ldots, 42$.

3.3 Asymptotic expansions

In this section we will consider another method for the computation of confluent hypergeometric functions. The series expansions developed in Section 3.1 are only applicable in some bounded domains. Also recurrence relations do not seem to be suitable for the computation of $M(a; c; z)$ at large arguments z in modulus. If we want to compute functions on a sector

$S = \{z \in \mathbb{C} : |\arg(z) - \alpha| \leq \delta\}$, we have to consider methods like asymptotic expansions. These expansions play an important role if we want to compute the Kummer functions on an unbounded interval.

3.3.1 Principal results on asymptotic expansions

First, we give the definition and fundamental properties of asymptotic expansions (see Olver [32], p. 16).

Definition 3.25 Let $\sum_{\nu=0}^{\infty} a_\nu z^{-\nu}$ be a formal power series, and let $f(z)$ be a function, such that

$$f(z) = \sum_{\nu=0}^{n-1} \frac{a_\nu}{z^\nu} + R_n(z), \tag{3.45}$$

where we have for each fixed $n \in \mathbb{N}$

$$R_n(z) = \mathcal{O}\left(\frac{1}{|z|^n}\right) \qquad (z \to \infty \text{ in } S). \tag{3.46}$$

Then we call the series $\sum_{\nu=0}^{\infty} a_\nu z^{-\nu}$ *an asymptotic expansion of* $f(z)$ *in* S, and write

$$f(z) \sim \sum_{\nu=0}^{\infty} \frac{a_\nu}{z^\nu} \qquad (z \to \infty \text{ in } S). \tag{3.47}$$

If the series $\sum_{\nu=0}^{\infty} a_\nu z^{-\nu}$ converges for all sufficiently large $|z|$, then it is the asymptotic expansion of its sum without restriction to $\arg(z)$.
We give a necessary and sufficient condition for the existence of asymptotic expansions (see Olver [32], p. 17).

Remark 3.26 The function $f(z)$ has an asymptotic expansion of the form (3.47) if and only if we have for all $n \in \mathbb{N}_0$

$$z^n \left(f(z) - \sum_{\nu=0}^{n-1} \frac{a_\nu}{z^\nu} \right) \to a_n \qquad (z \to \infty \text{ in } S)$$

uniformly with respect to $\arg(z)$. This implies that the sequence (a_n) is uniquely determined. Therefore every function f can have at most one asymptotic expansion in S (see Olver [32], p. 17).

3.3.2 Asymptotic expansions for confluent hypergeometric functions

The asymptotic expansions for the Kummer function can be inferred from integral representations and connections to the confluent hypergeometric function of the second kind (see Abramowitz, Stegun [1], p. 508).

Theorem 3.27 *For fixed parameters a and c the function $M(a;c;z)$ has the asymptotic expansion*

$$M(a,c;z) = \frac{\Gamma(c)}{\Gamma(c-a)}\frac{e^{i\pi a}}{z^a}\sum_{n=0}^{N_1}\frac{(a)_n(a-c+1)_n}{n!\,(-z)^n} + \mathcal{O}\left(\frac{1}{|z|^{N_1+1}}\right)$$

$$+ \frac{\Gamma(c)}{\Gamma(a)}e^z z^{a-c}\sum_{n=0}^{N_2}\frac{(c-a)_n(1-a)_n}{n!\,z^n} + \mathcal{O}\left(\frac{1}{|z|^{N_2+1}}\right) \qquad (z\to\infty \text{ in } S)$$

with $S = \{z : -\frac{1}{2}\pi < \arg(z) < \frac{3}{2}\pi\}$.

An important fact especially for numerical evaluation of the partial sums of the asymptotic expansion is the choice of numbers N_1 and N_2. Therefore we investigate the monotonicity behaviour of the terms of the partial sums. Note that the asymptotic expansion truncates if $a \in \mathbb{Z}$ and $c \in \mathbb{Z}$. We define the sequences

$$f_n = f_n(a,c,z) = \frac{(a)_n(a-c+1)_n}{n!\,(-z)^n} \qquad \text{and} \qquad g_n = g_n(a,c,z) = \frac{(c-a)_n(1-a)_n}{n!\,z^n}$$

and formulate the following result concerning the determination of degrees N_1 and N_2.

Lemma 3.28 *For real parameters a and c, numbers $N_f \in \mathbb{N}$ and $N_g \in \mathbb{N}$ exist, such that we have for $n \in \mathbb{N}$*

$$|f_{n+1}| \geq |f_n| \quad (n > N_f), \qquad |g_{n+1}| \geq |g_n| \quad (n > N_g).$$

Moreover, the numbers N_f and N_g are given by

$$N_f = \begin{cases} \max\left(0, \left\lceil \frac{1}{2}\left(-(a+\gamma-|z|) + \sqrt{(a+\gamma-|z|)^2 - 4(a\gamma-|z|)}\right)\right\rceil\right) & \text{if } a \leq \hat{a}_f, \\ 0 & \text{if } a > \hat{a}_f \end{cases}$$

and

$$N_g = \begin{cases} \max\left(0, \left\lceil \frac{1}{2}\left((a+\tilde{\gamma}+|z|) + \sqrt{(a+\tilde{\gamma}+|z|)^2 - 4(a\tilde{\gamma}+c-|z|)}\right)\right\rceil\right) & \text{if } a \geq \hat{a}_g, \\ 0 & \text{if } a < \hat{a}_g \end{cases}$$

denoting by $\lceil x \rceil$ the largest integer less than or equal to the real number x as well as setting $\gamma = a - c + 1$, $\tilde{\gamma} = a - c - 1$,

$$\hat{a}_f = \frac{1}{4|z|}\,(1 - 2c + c^2 + 2|z| + 2c|z| + |z|^2),$$

and

$$\hat{a}_g = -\frac{1}{4|z|}\,(1 - 2c + c^2 + 2|z| - 2c|z| + |z|^2).$$

Proof: We first prove the assertion for the sequence (f_n). The assertion for (g_n) can be shown analogously. We consider the quotient

$$\left|\frac{f_{n+1}}{f_n}\right| = \left|\frac{(a)_{n+1}(a-c+1)_{n+1}n!(-z)^n}{(a)_n(a-c+1)_n(n+1)!(-z)^{n+1}}\right|$$

which simplifies to

$$\left|\frac{f_{n+1}}{f_n}\right| = \left|\frac{(a+n)(a-c+n+1)}{(n+1)(-z)}\right|. \tag{3.48}$$

Since we have to find a number N_f, such that

$$\left|\frac{f_{n+1}}{f_n}\right| \geq 1$$

for $n > N_f$ we consider for the moment the function $h_f : \mathbb{R}_+ \to \mathbb{R}$, defined by

$$h_f(x) = \frac{(a+x)(a-c+x+1)}{(x+1)|z|}. \tag{3.49}$$

In order to find a number x_f with $h_f(x_f) = 1$ we have to solve the quadratic equation

$$x^2 + (2a - c + 1 - |z|)x + a(a-c+1) - |z| = 0$$

which leads to the possible solutions ($\gamma = a - c + 1$)

$$x_f^{1,2} = \frac{1}{2}\left(-(a+\gamma-|z|) \pm \sqrt{(a+\gamma-|z|)^2 - 4(a\gamma-|z|)}\right).$$

Then, we have one solution x_f if the parameter a satisfies

$$a = \hat{a}_f = \frac{1}{4|z|}\left(1 - 2c + c^2 + 2|z| + 2c|z| + |z|^2\right),$$

two solutions $x_f^{1,2}$ if $a < \hat{a}$ and no solution if $a > \hat{a}$. In the case of two positive solutions we choose the largest one. Since we recognize from (3.49) that $h_f \to \infty$ for $x \to \infty$ the choice of the largest solution x_f and setting $N_f = \lceil x_f \rceil$ yield $|f_{n+1}| \geq |f_n|$ for $n > N_f$. If we have no solution or negative solutions the same argument as before yields $|f_{n+1}| \geq |f_n|$ for $n > N_f$ with $N_f = 0$. Then, the assertion follows by setting $N_f = \max(0, \lceil x_f \rceil)$ for $a < \hat{a}$ and since $N_f = 0$ for $a > \hat{a}$.
Analogously, we consider the quotient of the sequence (g_n)

$$\left|\frac{g_{n+1}}{g_n}\right| = \left|\frac{(c-a+n)(1-a+n)}{(n+1)z}\right|. \tag{3.50}$$

In order to find a number N_g with

$$\left|\frac{g_{n+1}}{g_n}\right| \geq 1$$

for $n > N_g$ we consider the corresponding real function $h_g : \mathbb{R}_+ \to \mathbb{R}$, defined by

$$h_g(x) = \frac{(c-a+x)(1-a+x)}{(x+1)|z|}. \tag{3.51}$$

Here, the determination of a number x_g satisfying $h_g(x_g) = 1$ leads to the quadratic equation

$$x^2 + (c - 2a + 1 - |z|)x + a(a - c - 1) + c - |z| = 0$$

yielding the possible solutions (setting $\tilde{\gamma} = a - c - 1$)

$$x_g^{1,2} = \frac{1}{2}\left((a + \tilde{\gamma} + |z|) \pm \sqrt{(a + \tilde{\gamma} + |z|)^2 - 4(a\tilde{\gamma} + c - |z|)}\right).$$

In this case we have one solution x_g if

$$a = \hat{a} = -\frac{1}{4|z|}\,(1 - 2c + c^2 + 2|z| - 2c|z| + |z|^2),$$

two solutions $x_g^{1,2}$ if $a > \hat{a}$ and no solution if $a < \hat{a}$. Again we choose the largest one if there are two positive solutions and since we recognize from (3.51) that $h_g \to \infty$ for $x \to \infty$ the choice of the largest solution x_g and setting $N_g = \lceil x_g \rceil$ yield $|g_{n+1}| \geq |g_n|$ for $n > N_g$. If we have no solution or negative solutions the same argument as before yields $|g_{n+1}| \geq |g_n|$ for $n > N_g$ with $N_g = 0$. Then, the assertion follows by setting $N_g = \max(0, \lceil x_g \rceil)$ for $a > \hat{a}$ and since $N_g = 0$ for $a < \hat{a}$. □

In order to illustrate the result of Lemma 3.28 we consider the following numerical example.

Example 3.29 We choose the parameters $a = -4.5$, $c = 70.1$ and the argument $z = 45i$. Then, the application of Lemma 3.28 leads to the equation

$$x^2 - 123.1x + 286.2 = 0$$

which yields the solutions $N_f^1 = 2$ and $N_f^2 = 120$. We choose N_f^2 for truncating the expansion. Moreover, we obtain the equation

$$x^2 + 35.1x + 365.3 = 0$$

which has no real solution and we therefore set $N_g = 0$. So, the sequence (f_n) increases monotonely for $n > 120$ and the sequence (g_n) increases monotonely for $n > 0$. This behaviour can be seen in Figure 3.13, where we computed $\log_{10}|f_n|$ and $\log_{10}|g_n|$ for $n = 1, \ldots, 130$.

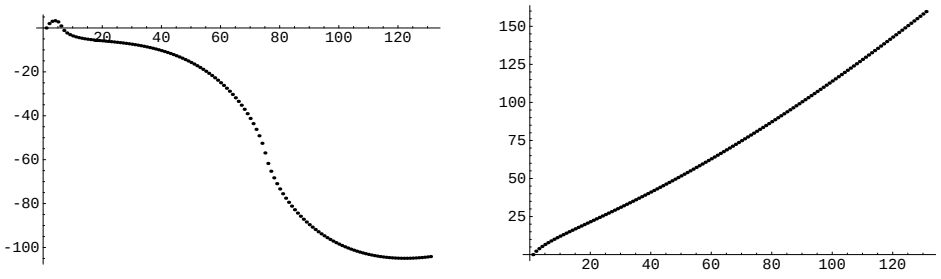

Figure 3.13: Monotonicity of the sequences (f_n) and (g_n)

3.3.3 Numerical results

We consider the approximations of the asymptotic expansion

$$\sigma_{N_1,N_2}(a,c,z) = \frac{\Gamma(c)}{\Gamma(c-a)}\frac{e^{i\pi a}}{z^a}\sum_{n=0}^{N_1}\frac{(a)_n(a-c+1)_n}{n!\,(-z)^n} + \frac{\Gamma(c)}{\Gamma(a)}e^z z^{a-c}\sum_{n=0}^{N_2}\frac{(c-a)_n(1-a)_n}{n!\,z^n}$$

and compute the relative error or the number of significant decimal digits by

$$d_{N_1,N_2}(a,c,z) = \min\left\{15, -\log_{10}\left(\frac{|M(a;c;z)-\sigma_{N_1,N_2}(a,c,z)|}{|M(a;c;z)|}\right)\right\}.$$

The results are computed in double precision arithmetic and are shown in Figure 3.14 and Figure 3.15, where we see the results obtained by using the abort criterion, presented in Lemma 3.28, on the left-hand side of Figures 3.14 and 3.15, whereas on the right-hand side in Figures 3.14 and 3.15 we see the results obtained by computing without abort criterion and a degree of $N_1 = N_2 = 45$ terms. We recognize that we achieve an improvement of the relative error d_{N_1,N_2} especially for arguments $|z| \leq 45$. But we also have to observe a not so excellent behaviour for increasing a. Nevertheless we can conclude that we have an efficient tool for computing $M(a;c;z)$ for large values of $|z|$ by using the partial sums σ_{N_1,N_2} of the asymptotic expansion of the confluent hypergeometric function $M(a;c;z)$ combined with the criterion for truncating the asymptotic expansion.

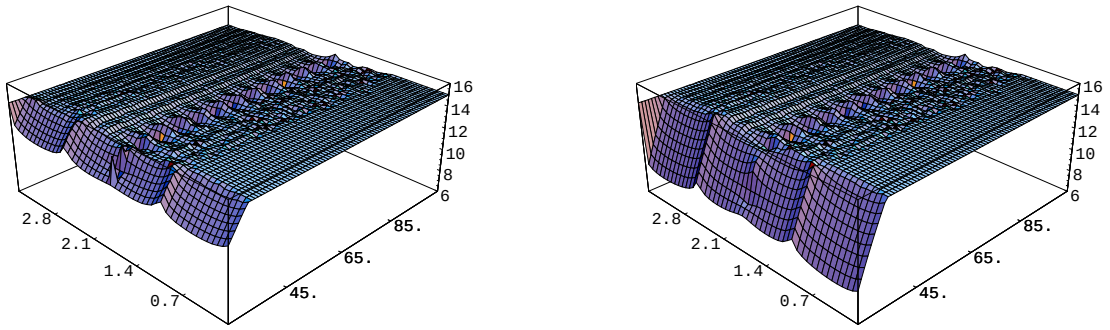

Figure 3.14: d_{N_1,N_2} with (left) and without (right) use of abort criterion for $z = 25i(1.0i)105i$, $a = 0.07(0.07)3.5$ and $c = 3.5$

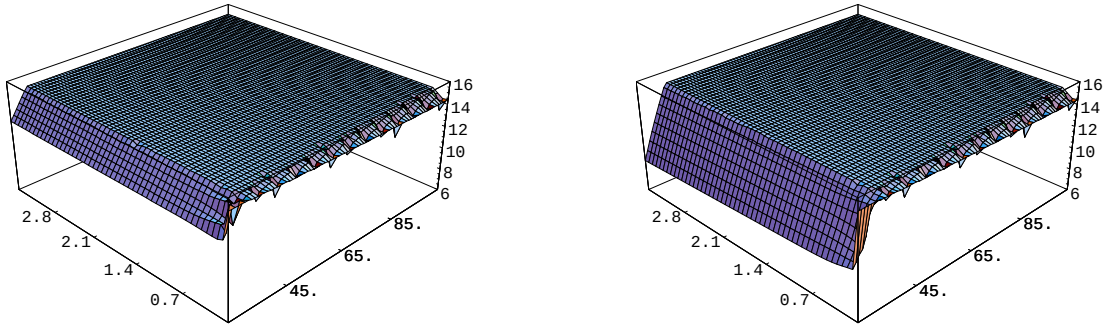

Figure 3.15: d_{N_1,N_2} with (left) and without (right) use of abort criterion for $z = 25i(1.0i)105i$, $a = 0.07$ and $c = 0.07(0.07)3.5$

3.4 Conclusion

We give some concluding remarks about the previous chapter in order to make suggestions to combine the presented methods for computing confluent hypergeometric functions.
At first, the computation method on compact intervals, investigated in Section 3.1, leads to reliable results in the case of the parameter combination $0 < \mathrm{Re}(a) < \mathrm{Re}(c)$ and not too large intervals, i.e. we are able to compute on intervals from zero to $30i$ or even $40i$. But this restriction may be dropped if we combine this method with the asymptotic expansion of the confluent hypergeometric function (see Section 3.3). Here, the numerical results show that we may use the partial sums of the asymptotic expansion in general for arguments exceeding the modulus of $30i$ or $40i$. We could even improve the results by applying a suitable abort criterion for the determination of the maximal degree of the partial sums. So, it should be possible to close the gap between the reliability of these two methods. Furthermore, numerical results show that especially large (real) parameters a may cause problems when computing the confluent hypergeometric functions. Indeed, the convergence theory developed in Section 3.1.4 does not cover the case $\mathrm{Re}(a) > \mathrm{Re}(c)$. Therefore we investigated in Section 3.2 the computation with recurrence relations. In particular, the simple forward recurrence relation for the confluent hypergeometric functions with respect to parameter a, introduced in Example 3.17, may be applied to overcome the described problem above. We compute two starting values for small (real) parameters a and $a+1$ using the method based on the series expansion for the computation on compact intervals (see Section 3.1) and apply the forward recurrence relation with respect to a until we obtain the desired function value. We also investigated another method for the computation of confluent hypergeometric functions with the help of recurrence relations, the Miller algorithm, which especially for small (real) parameters a and c leads to very good results. This algorithm is easy to implement and we do not need any (exact) starting value providing essential advantages. But also in this case large (real) parameters a cause problems (see Section 3.2.3). The reason is the cancellation when computing the normalizing series, which is responsible for the quality of the results.

Chapter 4

Application to parabolic boundary value problems

In this chapter we consider initial-boundary value problems governed by parabolic partial differential equations occurring in heating processes.
First, we describe a conductive heat transfer model in cylindrical domains. In order to simplify the model by introducing a cylindrical coordinate system the discussion of the underlying geometry is essential. Based on the associated eigenvalue problem of the initial-boundary value problems we then use the method of eigenfunction expansion to obtain a generalized solution of the problem. The required eigenfunctions are, in general, of confluent hypergeometric type, so that the computational methods of the previous chapter can be used to obtain the desired results.
In addition to the classical cylindrical domains we make suggestions to generalize the given model to domains with different geometry.
Besides of the conductive heat transfer we present a heat transfer model including convection. We consider a fluid filled loop where the flow behaviour is described by the Navier-Stokes equations. Using the aforementioned method of eigenfunction expansions we obtain solutions of the underlying initial-boundary value problem.
Finally, we will illustrate how the Lorenz system can be used to approximate the flow behaviour of the fluid.

4.1 Conductive heat transfer

We consider initial-boundary value problems describing the heat transfer in cylindrical domains. After transforming the initial-boundary value problem into cylindrical coordinates, the construction of a (generalized) Fourier series representation is presented and convergence of the resulting Fourier series is proved. Furthermore, it will be shown that in some cases the transformation of the initial-boundary value problem into a problem with homogeneous boundary conditions is appropriate. The resulting simplified model provides advantageous

characteristics for later numerical computations. Finally, we illustrate that different geometries of the domain cause the appearance of different eigenfunctions in the Fourier series expansions establishing dependence of the confluent hypergeometric functions on the chosen geometry.

4.1.1 The model of heat transfer in cylindrical domains

Considering a cross-section of an infinite cylinder, we start with an initial-boundary value problem including the two-dimensional conductive heat equation in the domain Ω, where $\Omega = \{\xi \in \mathbb{R}^2 : |\xi| < 1\}$, i.e. we consider the linear parabolic differential equation

$$\frac{\partial \theta}{\partial t}(\xi, t) = \chi \Delta \theta(\xi, t)$$

for $(\xi, t) \in \Omega \times (0, T)$ with the process time $T > 0$. We denote by Δ the Laplace operator which is for $\xi = (\xi_1, \xi_2) \in \Omega$ defined by $\Delta = \partial^2/\partial\xi_1^2 + \partial^2/\partial\xi_2^2$. The parameter χ is called the heat diffusivity (of the product to be heated) and is defined by $\chi = k/(\rho c)$, where ρ is the density, c the heat capacity and k the thermal conductivity (of the product). In general the parameters k and c depend on the temperature θ, but in the following simplified model we assume a constant heat diffusivity.
The initial condition is given by

$$\theta(\xi, 0) = \theta_0(\xi)$$

with the initial temperature distribution $\theta_0 \in C^2(\Omega)$.
Finally, the boundary condition is described by the so-called mixed Neumann boundary condition (or Robin boundary condition)

$$\frac{\partial \theta}{\partial n}(\xi, t) = \alpha(u(t) - \theta(\xi, t))$$

for $(\xi, t) \in \partial\Omega \times (0, T)$ modelling the heat transfer from the surrounding medium. We denote by n the outward unit normal vector and the function u represents the boundary control, the temperature of the surrounding medium. The parameter α is called the heat transfer coefficient which is also assumed to be constant.
Due to the geometry of the cylinder we introduce a different coordinate system. Since we consider a cross-section of a cylinder we define for a given point $\xi = (\xi_1, \xi_2) \in \Omega$ the polar coordinates by

$$\xi_1 = r\cos\varphi, \quad \xi_2 = r\sin\varphi$$

with radius $r \in [0, 1)$ and angle $\varphi \in [0, 2\pi]$. By doing this, we can consider the solution θ dependent only on the distance r from the origin due to the symmetry of the problem (cf. Tröltzsch [42], p. 150). This leads to the following one-dimensional problem,

$$\begin{aligned} r\,\frac{\partial \theta}{\partial t}(r, t) &= \chi\left(r\,\frac{\partial^2 \theta}{\partial r^2}(r, t) + \frac{\partial \theta}{\partial r}(r, t)\right) \\ \theta(r, 0) &= \theta_0(r) \\ \frac{\partial \theta}{\partial r}(1, t) &= \alpha(u(t) - \theta(1, t)) \end{aligned} \tag{4.1}$$

for $(r,t) \in Q$, where $Q = (0,1) \times (0,T)$. We call the function θ a *classical solution* of the initial-boundary value problem (4.1), if θ is twice continuously differentiable with respect to the spatial variable r on $(0,1)$ and once continuously differentiable with respect to the time variable t on $(0,T)$. It can be shown that a unique classical solution of the problem (4.1) exists (cf. McOwen [25], pp. 134-137 and p. 312).

4.1.2 Fourier series approach

In this section we derive the (generalized) Fourier series representation of the solution of the given initial-boundary value problem.
In order to apply the method of eigenfunction expansion we have to consider the associated homogeneous eigenvalue problem with respect to the initial-boundary value problem (4.1). For $\psi \in C^2(0,1)$ the eigenvalue problem

$$\chi r\psi''(r) + \chi\psi'(r) + r\lambda\psi(r) = 0, \tag{4.2}$$

$$\psi'(1) + \alpha\psi(1) = 0 \tag{4.3}$$

is called a *singular Sturm-Liouville eigenvalue problem* with the singularity $r = 0$ (cf. Pinsky [33], pp. 23-26 and González-Velasco [14], pp. 370-378). Setting $k_n = \sqrt{\lambda_n/\chi}$ we obtain the (bounded) solutions of this eigenvalue problem

$$\psi_n(r) = \sqrt{c_n} J_0(k_n r)$$

with constants c_n. The eigenvalues λ_n can be computed solving the boundary condition (4.3)

$$\alpha J_0(k_n) - k_n J_1(k_n) = 0. \tag{4.4}$$

The system $\{\psi_n\}$ forms a complete orthonormal system in $L^2_\omega(0,1)$ with the weight function $\omega(r) = r$ (cf. e.g. Triebel [41], pp. 362-365), i.e.

$$\int_0^1 r\,\psi_n(r)\psi_m(r)\,dr = \begin{cases} 0 & \text{for } \lambda_n \neq \lambda_m, \\ 1 & \text{for } \lambda_n = \lambda_m, \end{cases}$$

and we can now specify the normalizing coefficients to be

$$c_n = 2\left(J_0^2(k_n) + J_1^2(k_n)\right)^{-1}.$$

We can thus expand the function $\theta(r,t)$ for each $t \in (0,T)$ into a (generalized) Fourier series

$$\theta(r,t) = \sum_{n=1}^{\infty} \alpha_n(t)\psi_n(r), \tag{4.5}$$

with the (generalized) Fourier coefficients

$$\alpha_n(t) = \int_0^1 r\,\theta(r,t)\psi_n(r)\,dr.$$

In order to determine the coefficients α_n we consider the derivative

$$\frac{\partial\theta}{\partial t}(r,t) = \sum_{n=0}^{\infty} \frac{d}{dt}\,\alpha_n(t)\psi_n(r)$$

assuming for the moment that the coefficients are differentiable and the order of summation and differentiation can be interchanged (cf. Haberman [15], p. 258). Then the equation (4.1) reads

$$\sum_{n=0}^{\infty} \frac{d}{dt}\,\alpha_n(t)\psi_n(r) = \chi\left(\frac{\partial^2\theta}{\partial r^2}(r,t) + \frac{1}{r}\,\frac{\partial\theta}{\partial r}(r,t)\right)$$

and thus $\frac{d}{dt}\,\alpha_n$ is given by

$$\frac{d}{dt}\,\alpha_n(t) = \chi \int_0^1 r\,\theta_{rr}(r,t)\psi_n(r)\,dr + \chi \int_0^1 \theta_r(r,t)\psi_n(r)\,dr$$

denoting the partial derivatives of θ with respect to r by θ_r and θ_{rr}. Integrating by parts twice we get

$$\begin{aligned}
\int_0^1 r\,\theta_{rr}\psi_n\,dr &= \left[r\,\theta_r\psi_n\right]_0^1 - \int_0^1 \theta_r\left(r\psi_n' + \psi_n\right)\,dr \\
&= \theta_r(1,t)\psi_n(1) - \int_0^1 r\,\theta_r\psi_n'\,dr - \int_0^1 \theta_r\psi_n\,dr \\
&= \alpha(u(t)-\theta(1,t))\psi_n(1) - \left[r\,\theta\psi_n'\right]_0^1 + \int_0^1 \theta\left(r\,\psi_n'' + \psi_n'\right)\,dr - \int_0^1 \theta_r\psi_n\,dr \\
&= \alpha(u(t)-\theta(1,t))\psi_n(1) - \theta(1,t)\psi_n'(1) + \int_0^1 r\,\theta\psi_n''\,dr + \int_0^1 \theta\psi_n'\,dr - \int_0^1 \theta_r\psi_n\,dr
\end{aligned}$$

and since the eigenfunctions ψ_n satisfy the differential equation (4.2) and the homogeneous boundary condition (4.3) we obtain the equation

$$\begin{aligned}
\frac{d}{dt}\,\alpha_n(t) - \chi\alpha u(t)\psi_n(1) &= \int_0^1 \theta(r,t)\left(\chi r\psi_n''(r) + \chi\psi_n'(r)\right)\,dr \\
&= \int_0^1 \theta(r,t)\left(-r\lambda_n\psi_n(r)\right)\,dr \\
&= -\lambda_n\alpha_n(t).
\end{aligned}$$

Using the initial condition $\theta(r,0) = \theta_0(r)$ the functions $\alpha_n(t)$ have to satisfy the initial value

problem

$$\frac{d}{dt}\,\alpha_n(t) + \lambda_n \alpha_n(t) = \chi\alpha u(t)\psi_n(1)$$
$$\alpha_n(0) = \int_0^1 \eta\,\theta_0(\eta)\psi_n(\eta)\,d\eta$$

which has the solution

$$\alpha_n(t) = e^{-\lambda_n t}\int_0^1 \eta\,\theta_0(\eta)\psi_n(\eta)\,d\eta + \chi\alpha\int_0^t e^{-\lambda_n(t-s)}u(s)\psi_n(1)\,ds.$$

Summing up, we finally have the Fourier series representation

$$\theta(r,t) = \sum_{n=0}^{\infty}\psi_n(r)\left(e^{-\lambda_n t}\int_0^1 \eta\theta_0(\eta)\psi_n(\eta)d\eta + \psi_n(1)\chi\alpha\int_0^t e^{-\lambda_n(t-s)}u(s)ds\right). \tag{4.6}$$

Introducing the Green function

$$G(r,\eta;t-s) = \sum_{n=0}^{\infty}\psi_n(r)\eta\psi_n(\eta)e^{-\lambda_n(t-s)} \tag{4.7}$$

we may write the solution as

$$\theta(r,t) = \int_0^1 G(r,\eta;t)\theta_0(\eta)d\eta + \chi\alpha\int_0^t G(r,1;t-s)u(s)ds \tag{4.8}$$

(see Tröltzsch [42], p. 113 and p. 151).
It can be shown that for a continuous control $u(t)$ exactly one bounded solution of (4.1) exists (see Tröltzsch [42], p. 151). If $u(t)$ is not continuous, a classical solution $\theta(r,t)$ of (4.1) need not exist. Hence, for $u \in L^\infty(0,T)$ a continuous function $\theta(r,t)$ that satisfies the integral equation (4.8) is called a *generalized solution of* (4.1).
We split this representation in two series in order to investigate the convergence behaviour. The function

$$\theta_{\text{free}}(r,t) = \sum_{n=0}^{\infty}\psi_n(r)e^{-\lambda_n t}\int_0^1 \eta\theta_0(\eta)\psi_n(\eta)d\eta$$

is called the *free solution*, whereas the function

$$\theta_{\text{part}}(r,t) = \sum_{n=0}^{\infty}\psi_n(r)\psi_n(1)\chi\alpha\int_0^t e^{-\lambda_n(t-s)}u(s)ds$$

is called the *particular solution* of the initial-boundary value problem (cf. e.g. Logan [21], p. 216).

4.1.3 Convergence analysis

We prove the following result regarding the convergence behaviour of the series representations of the free and particular solutions of the initial-boundary value problem.

Theorem 4.1 *Let $\theta(r,t) = \theta_{\text{free}}(r,t) + \theta_{\text{part}}(r,t)$ be the (generalized) Fourier series representation of the solution of the initial-boundary value problem (4.1). If $u \in L^p(0,T)$ for some $p > 2$, then the series*

$$\theta_{\text{free}}(r,t) = \sum_{n=1}^{\infty} c_n J_0(k_n r) e^{-k_n^2 \chi t} \int_0^1 \eta \theta_0(\eta) J_0(k_n \eta) d\eta$$

and

$$\theta_{\text{part}}(r,t) = \sum_{n=1}^{\infty} c_n J_0(k_n r) J_0(k_n) \chi \alpha \int_0^t e^{-k_n^2 \chi (t-s)} u(s) ds$$

converge absolutely and uniformly on $Q_\delta = (\delta, 1) \times (0, T)$ for every $\delta > 0$ and are continuous on Q_δ.

Proof: We first consider the terms of the series expansion θ_{part}. Since we are interested in estimates of these terms for large n we use the asymptotic expansion for the Bessel function J_ν (see Nikiforov [29], p. 210) with integer order ν,

$$J_\nu(z) = \sqrt{\frac{2}{\pi z}} \left\{ \cos\left(z - \frac{\pi}{2}\left(\nu + \frac{1}{2}\right)\right) + \mathcal{O}\left(\frac{1}{z}\right)\right\} \qquad (z \to \infty).$$

Then we can estimate the terms of the series

$$\begin{aligned} |c_n J_0(k_n r) J_0(k_n)| &= \frac{2|J_0(k_n r) J_0(k_n)|}{|J_0^2(k_n) + J_1^2(k_n)|} \\ &= \frac{2r^{-1/2} |\left[\cos(k_n r - \pi/4) + \mathcal{O}(1/(k_n r))\right] \cdot \left[\cos(k_n - \pi/4) + \mathcal{O}(1/k_n)\right]|}{\left[\cos(k_n - \pi/4) + \mathcal{O}(1/k_n)\right]^2 + \left[\cos(k_n - 3\pi/4) + \mathcal{O}(1/k_n)\right]^2} \\ &\le \frac{2r^{-1/2}\left[1 + \mathcal{O}(1/(k_n r))\mathcal{O}(1/k_n) + \mathcal{O}(1/(k_n r)) + \mathcal{O}(1/k_n)\right]}{1 + 4\mathcal{O}(1/k_n) + 2\mathcal{O}(1/k_n^2)}, \end{aligned}$$

since $\cos^2(k_n - \pi/4) + \cos^2(k_n - 3\pi/4) = 1$. Then, for $\varepsilon > 0$ and $\delta > 0$ a number $N^* = N(\delta, \varepsilon) \in \mathbb{N}$ exists, such that

$$|c_n J_0(k_n r) J_0(k_n)| \le \frac{2}{\sqrt{\delta}} (1 + \varepsilon)$$

for all $r \ge \delta$ and all $n \ge N^*$. Now we have to estimate the integral. The use of Hölder's inequality in its general version (see Werner [45], p. 20) yields

$$\left| \int_0^t e^{-k_n^2 \chi (t-s)} u(s) ds \right| \le \left[\int_0^t e^{-q k_n^2 \chi (t-s)} ds \right]^{1/q} \cdot \|u\|_p,$$

where $1/p + 1/q = 1$ and $\|\cdot\|_p$ denotes the norm in the space $L^p(0,T)$. The integral on the right hand side can be computed as

$$\int_0^t e^{-qk_n^2\chi(t-s)}ds = \frac{1}{qk_n^2\chi}\left(1 - e^{-qk_n^2\chi t}\right).$$

Since the eigenvalues $\lambda_n = k_n^2\chi$ fulfill $\sqrt{\lambda_n} = n\pi + c(\alpha)$, where $c(\alpha)$ is a positive constant, we finally have

$$\left| c_n J_0(k_n r) J_0(k_n) \int_0^t e^{-k_n^2\chi(t-s)} u(s) ds \right| \leq \text{const} \cdot \frac{1}{\sqrt{\delta}} \cdot \frac{1}{n^{2/q}}$$

for all $r \geq \delta$ and all $n \geq N^*$. Hence, we have a convergent majorant of the series θ_{part} because $q < 2$. The assertion follows by the Weierstrass majorant criterion.
Now we consider the series expansion θ_{free}. For the terms $c_n J_0(k_n r)$ we can apply the same estimate as above. With given $\delta > 0$ and $\varepsilon > 0$ a number $N^* = N(\delta, \varepsilon) \in \mathbb{N}$ exists, such that

$$|c_n J_0(k_n r)| \leq \sqrt{\frac{2\pi k_n}{\delta}}\,(1 + \varepsilon)$$

for all $r \geq \delta$ and all $n \geq N^*$. The integral can be estimated with a positive constant C by

$$\left| \int_0^1 \eta\theta_0(\eta) J_0(k_n\eta) d\eta \right| \leq C \int_0^1 |J_0(k_n\eta)| d\eta.$$

If we substitute $k_n\eta = \tau$, we can further estimate

$$\int_0^1 |J_0(k_n\eta)| d\eta = \frac{1}{k_n}\int_0^{k_n} |J_0(\tau)| d\tau \leq \frac{1}{k_n}\int_0^{k_n} \frac{1}{\sqrt{\tau}} d\tau = \frac{2}{\sqrt{k_n}},$$

where we used the relation $J_0(\tau) \leq 1/\sqrt{\tau}$. Finally, we have

$$\left| c_n J_0(k_n r) e^{-k_n^2\chi t} \int_0^1 \eta\theta_0(\eta) J_0(k_n\eta) d\eta \right| \leq \text{const} \cdot \frac{1}{\sqrt{\delta}} \cdot \left(e^{-\chi t}\right)^{k_n^2},$$

a convergent majorant of the series. The assertion follows also by the Weierstrass majorant criterion. The continuity of θ_{free} and θ_{free} on Q_δ follows immediately by the continuity of the terms of the series as well as the already shown uniform convergence. □

Since the boundary control $u(t)$, which is in general in $L^\infty(0,T)$ (see Tröltzsch [42], pp. 150-151), is often of simple structure, the representation of the solution can be simplified, which is shown in the following section.

4.1.4 Homogeneous boundary conditions

In order to evaluate the series expansion numerically some parameters have to be taken into account. A crucial point for numerical evaluation is the choice of the heat transfer coefficient α. In practical applications this parameter is often of the magnitude 10^4. Numerical tests exhibit certain problems for such values of α. We recognize that in these cases the eigenvalues of the problem (4.2), which are determined by the equation (4.4)

$$\alpha J_0(k_n) - k_n J_1(k_n) = 0,$$

are close to the zeros of J_0. This fact is responsible for requiring a larger number of terms of the partial sums of the Fourier series expansion. This is illustrated in Figure 4.1. The dotted lines represent the function $J_0(x)$ and the solid lines represent the function $J_0(x) - xJ_1(x)/\alpha$ for $x \in [0, 20]$. We observe that the latter function approaches the former for increasing values of α. In particular, the first zeros of $J_0(x)$ nearly coincide with the computed eigenvalues, so that the first terms of the partial sums of the series representation (4.6) achieve very small contribution to the approximation, since $\psi_n(1) = \sqrt{c_n} J_0(k_n)$ has to be evaluated for the particular solution θ_{part}.

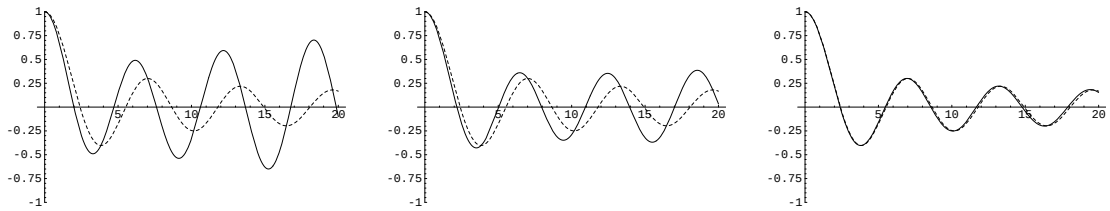

Figure 4.1: Functions $J_0(x)$ and $J_0(x) - xJ_1(x)/\alpha$ for $\alpha = 5$, 10 and 100

Under certain conditions we can overcome the occurring problems when computing the temperature distribution via the Fourier series. When we have a constant boundary control and transform the initial-boundary value problem into one with homogeneous boundary conditions, the resulting Fourier series is of the type θ_{free}, which should not cause problems in numerical evaluation. This is favourable for the numerical computations in the practical applications treated in Chapter 5.

We consider the initial-boundary value problem including the one-dimensional heat equation in cylindrical coordinates

$$\begin{aligned} r\,\frac{\partial \theta}{\partial t}(r,t) &= \chi\left(r\,\frac{\partial^2 \theta}{\partial r^2}(r,t) + \frac{\partial \theta}{\partial r}(r,t)\right) \\ \theta(r,0) &= \theta_0(r) \\ \frac{\partial \theta}{\partial r}(1,t) &= \alpha(u(t) - \theta(1,t)) \end{aligned}$$

for $(r,t) \in Q$, control u and state θ. In order to perform a transformation to homogeneous boundary conditions we assume a constant boundary control $u(t) = u_c$ for all $t \in (0,T)$ and

set $\phi(r,t) = \theta(r,t) - u_c$ for $(r,t) \in Q$. Then, we obtain the following initial-boundary value problem

$$\begin{aligned} r\,\frac{\partial \phi}{\partial t}(r,t) &= \chi\left(r\,\frac{\partial^2 \phi}{\partial r^2}(r,t) + \frac{\partial \phi}{\partial r}(r,t)\right) \\ \phi(r,0) &= \theta_0(r) - u_c \\ \frac{\partial \phi}{\partial r}(1,t) &= -\alpha\phi(1,t) \end{aligned} \tag{4.9}$$

for $(r,t) \in Q$. Considering the already known associated Sturm-Liouville problem (4.2) the eigenfunctions in this case are again $\psi_n(r) = \sqrt{c_n} J_0(k_n r)$ and the eigenvalues $k_n = \sqrt{\lambda_n/\chi}$. Hence, we expand for each $t \in (0,T)$ the function $\phi(r,t)$ into a (generalized) Fourier series in eigenfunctions

$$\phi(r,t) = \sum_{n=1}^{\infty} \alpha_n(t)\psi_n(r),$$

with Fourier coefficients

$$\alpha_n(t) = \int_0^1 r\,\phi(r,t)\psi_n(r)\,dr.$$

Now the Fourier coefficients are given by the equation

$$\frac{d}{dt}\,\alpha_n(t) = \chi \int_0^1 r\,\phi(r,t)\psi_n''(r)\,dr + \chi \int_0^1 \phi(r,t)\psi_n'(r)\,dr$$

allowing determination by the initial value problem

$$\begin{aligned} \frac{d}{dt}\,\alpha_n(t) &= -\lambda_n \alpha_n(t) \\ \alpha_n(0) &= \int_0^1 r\,(\theta_0 - u_c)\psi_n(r)\,dr \end{aligned}$$

yielding the solution

$$\alpha_n(t) = e^{-k_n^2 \chi t} \int_0^1 r\,(\theta_0 - u_c)\psi_n(r)\,dr.$$

Summing up, in case of a constant boundary control we get the following representation of the solution of (4.9)

$$\phi(r,t) = \sum_{n=1}^{\infty} c_n J_0(k_n r) e^{-k_n^2 \chi t} \int_0^1 \eta\,(\theta_0(\eta) - u_c) J_0(k_n \eta)\,d\eta. \tag{4.10}$$

If additionally the initial temperature is constant, $\theta_0(\eta) = \theta_0$, we are able to compute the integral explicitly. The substitution $\rho = k_n \eta$ leads to

$$\int_0^1 \eta(\theta_0 - u_c) J_0(k_n\eta)d\eta = \frac{\theta_0 - u_c}{k_n^2} \int_0^{k_n} \rho J_0(\rho)d\rho = \frac{\theta_0 - u_c}{k_n} J_1(k_n).$$

Defining

$$d_n = \frac{\theta_0 - u_c}{k_n} J_1(k_n) c_n = \frac{2(\theta_0 - u_c) J_1(k_n)}{k_n (J_0^2(k_n) + J_1^2(k_n))}$$

the solution ϕ has the simple form

$$\phi(r,t) = \sum_{n=1}^{\infty} d_n J_0(k_n r) e^{-k_n^2 \chi t}.$$

In Chapter 5 we present numerical examples and applications of this representation.

4.1.5 Different geometries of the domain

In this section we suggest extensions of the previous heat transfer model to different geometries of the domain. Up to now we considered the (classical) case of cylindrical domains. Henceforth, we introduce geometries of e.g. elliptical cylinder and parabolic cylinder type. The idea is to exploit the special geometry of the domain by introducing different coordinate systems.
Starting from the initial-boundary value problem

$$\begin{aligned} \frac{\partial \phi}{\partial t} &= \chi \Delta \phi && \text{on } \Omega \times (0,T) \\ \phi(\cdot, 0) &= \theta_0(\cdot) - u_c && \text{on } \Omega \\ \frac{\partial \phi}{\partial n} + \alpha \phi &= 0 && \text{on } \partial\Omega \times (0,T) \end{aligned}$$

in the bounded domain $\Omega \subset \mathbb{R}^3$ with boundary $\partial\Omega$ and time interval $(0,T)$ we have to determine generalized solutions of this problem. In doing so, functions $\psi \in C^2(\Omega)$ have to solve the multidimensional eigenvalue problem

$$\begin{aligned} \Delta \psi + \lambda \psi &= 0 && \text{on } \Omega \\ \frac{\partial \psi}{\partial n} + \alpha \psi &= 0 && \text{on } \partial\Omega \end{aligned} \tag{4.11}$$

with eigenvalue parameter λ (the so-called Helmholtz equation).
Then, the first step is the transformation of this equation into the different coordinate systems, and second, find the solutions with the method of separation of variables.
We give three examples for possible coordinate transformations and show how to achieve solutions of the Helmholtz equation using separation of variables. For the given examples see Temme [40], pp. 257-266 and Nikiforov [29], pp. 297-299. Except for the third example we consider only cross-sections of cylindrical domains Ω and since we neglect the third independent variable the problem is reduced to the cross-section $\Omega' \subset \mathbb{R}^2$.

Elliptical cylinder coordinates

We transform the point $(x,y) \in \Omega'$ in the Cartesian coordinate system into the point (ξ, η) in the elliptical cylinder coordinate system, defined by

$$x = \cosh\xi \cos\eta, \quad y = \sinh\xi \sin\eta$$

with $\xi \in [0, \infty)$, $\eta \in [0, 2\pi]$. Figure 4.2 shows the cross-section of the elliptical cylinder and the orthogonal trajectories for fixed $\xi = \xi_0$ and $\eta = \eta_0$. The transformation of the Laplace

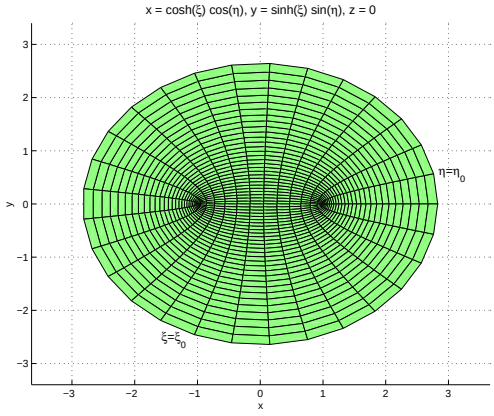

Figure 4.2: Elliptical cylinder coordinates

operator Δ (with respect to the Cartesian coordinate system) into the new coordinate system yields

$$\Delta\psi = \frac{1}{\sinh^2 \xi + \sin \eta} \left(\frac{\partial^2 \psi}{\partial \xi^2} + \frac{\partial^2 \psi}{\partial \eta^2} \right).$$

If we set $\psi(\xi, \eta) = f_1(\xi) f_2(\eta)$, the separation of $\Delta\psi + \lambda\psi = 0$ leads to

$$f_1'' + \left(-\nu + \tfrac{1}{2} \left(\lambda - \mu^2 \right) \cosh 2\xi \right) f_1 = 0, \tag{4.12}$$

$$f_2'' + \left(\nu - \tfrac{1}{2} \left(\lambda - \mu^2 \right) \cos 2\eta \right) f_2 = 0 \tag{4.13}$$

with separation constants ν and μ. The solutions of the equations (4.12) and (4.13) are Mathieu functions, which are related to Bessel functions (see Abramowitz, Stegun, p. 730).

Parabolic cylinder coordinates

We transform the point $(x, y) \in \Omega'$ in the Cartesian coordinate system into the point (ξ, η) in the parabolic cylinder coordinate system, defined by

$$x = \frac{1}{2}(\xi^2 - \eta^2), \quad y = \xi\eta$$

with ξ, $\eta \in \mathbb{R}$. Figure 4.3 shows the cross-section of the parabolic cylinder and the orthogonal trajectories for fixed $\xi = \xi_0$ and $\eta = \eta_0$. The transformation of the Laplace operator Δ into the new coordinate system yields

$$\Delta\psi = \frac{1}{\xi^2 + \eta^2} \left(\frac{\partial^2 \psi}{\partial \xi^2} + \frac{\partial^2 \psi}{\partial \eta^2} \right).$$

If we set $\psi(\xi, \eta) = f_1(\xi) f_2(\eta)$, the separation of $\Delta\psi + \lambda\psi = 0$ leads to

$$f_1'' + \left(\nu + \left(\lambda - \mu^2 \right) \xi^2 \right) f_1 = 0, \tag{4.14}$$

$$f_2'' + \left(-\nu + \left(\lambda - \mu^2 \right) \eta^2 \right) f_2 = 0 \tag{4.15}$$

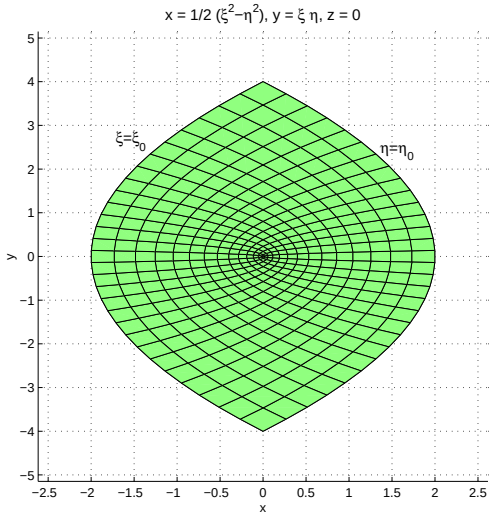

Figure 4.3: Parabolic cylinder coordinates

with separation constants ν and μ. The solutions of the equations (4.14) and (4.15) are parabolic cylinder functions, which can be represented in terms of confluent hypergeometric functions.

Rotational paraboloidal coordinates

We transform the differential equation (4.11) into rotational paraboloidal coordinates in the following way (see Nikiforov [29], pp. 297-299). For points $(x, y, z) \in \Omega$ we introduce

$$x = \xi\eta\cos\varphi, \quad y = \xi\eta\sin\varphi, \quad z = \frac{1}{2}(\xi^2 - \eta^2).$$

In this case the Helmholtz equation becomes

$$\frac{1}{\xi^2+\eta^2}\left[\frac{1}{\xi}\frac{\partial}{\partial\xi}\left(\xi\frac{\partial\psi}{\partial\xi}\right) + \frac{1}{\eta}\frac{\partial}{\partial\eta}\left(\eta\frac{\partial\psi}{\partial\eta}\right)\right] + \frac{1}{(\xi\eta)^2}\frac{\partial^2\psi}{\partial\varphi^2} + \lambda\psi = 0. \tag{4.16}$$

Using separation of variables

$$\psi(\xi, \eta, \varphi) = f_1(\xi)f_2(\eta)f_3(\varphi)$$

we obtain the system of ordinary differential equations

$$f_1'' + \frac{1}{\xi}f_1' + (\lambda\xi^2 - \mu\xi^{-2} + \nu)f_1 = 0 \tag{4.17}$$

$$f_2'' + \frac{1}{\eta}f_2' + (\lambda\eta^2 - \mu\eta^{-2} - \nu)f_2 = 0 \tag{4.18}$$

$$f_3'' + \mu f_3 = 0. \tag{4.19}$$

with separation constants ν and μ. It is only necessary to solve equation (4.17), since we can solve equation (4.18) by the substitution $-\nu$ for ν, and equation (4.19) is solved by trigonometric functions. The substitution $\xi^2 = s$ transforms equation (4.17) into a generalized

hypergeometric equation

$$f_1'' + \frac{1}{s} f_1' + (\lambda s^2 + \nu s - \mu) f_1 = 0. \tag{4.20}$$

The next step is the transformation into an equation of hypergeometric type. This was done by the method described in Nikiforov [29], pp. 1-3. An equation of the form

$$f_1'' + \frac{\tilde{\tau}(s)}{\sigma(s)} f_1' + \frac{\tilde{\sigma}(s)}{\sigma^2(s)} f_1 = 0,$$

where $\sigma(s)$ and $\tilde{\sigma}(s)$ are polynomials of degree at most 2, and $\tilde{\tau}(s)$ is a polynomial of degree at most 1, can be reduced to an equation of hypergeometric type

$$\sigma(s) y'' + \tau(s) y' + \kappa y = 0,$$

where $\tau(s)$ is a polynomial of degree at most 1 and κ is constant. This was done by the substitution $f_1 = g(s)y$ with a suitable chosen $g(s)$. Here, the transformation of equation (4.20) leads to

$$sy'' + (1 + l(i\sqrt{\mu} + s))y' + (l(i\sqrt{\mu} + 1)/2 - \nu/4)y = 0. \tag{4.21}$$

Equation (4.21) can be transformed into a canonical form by a linear change of the independent variable (see Nikiforov [29], p. 253). There are three possibilities, dependent on the degree of the polynomial $\sigma(s)$. In the case of equation (4.21) the degree of $\sigma(s)$ is $\deg(\sigma(s)) = 1$. The linear transformation $s = \beta w$ leads to

$$wy'' + \tau(\beta w)y' + \kappa\beta y = 0,$$

and with $\beta = -1/\tau'(s) = -1/l$ we get the confluent hypergeometric equation

$$wy'' + (c - w)y' - ay = 0, \tag{4.22}$$

where the parameters are $a = -\kappa\beta = (1 + i\sqrt{\mu})/2 - \nu/(4k)$ and $c = \tau(0) = 1 + ki\sqrt{\mu}$ with $k = \sqrt{\lambda}$. So, we obtain the solutions

$$y_1 = M(a; c; -ks)$$

and, by the Kummer identity,

$$y_2 = e^{-ks} M(c - a; c; ks).$$

The transformation function $g(s)$ is determined by the expression (see Nikiforov [29], p. 2)

$$\frac{g'(s)}{g(s)} = \frac{ls/2 + ki\sqrt{\mu}/2}{s},$$

and we get

$$g(s) = s^{ki\sqrt{\mu}/2} e^{ks/2},$$

so that we have the solution of equation (4.20)

$$f_1(s) = g(s)y(s) = s^{ki\sqrt{\mu}/2}e^{ks/2}M(a;c;-ks) = s^{ki\sqrt{\mu}/2}e^{-ks/2}M(c-a;c;ks).$$

In this way we recognize how confluent hypergeometric functions appear in the context of solving (parabolic) partial differential equations, where the crucial point is the geometry of the domain. We remark that it is necessary to prove orthogonality and completeness of the respective system of eigenfunctions. For known results on multidimensional eigenvalue problems including the Helmholtz equation we refer to e.g. Haberman [15], pp. 203-210.

4.2 Thermal convection loop

We consider a system of partial differential equations describing the flow behaviour as well as the heat transfer in a fluid filled loop. The underlying initial-boundary value problem can be solved by the use of the (generalized) Fourier series approach. Also the connection to the Lorenz system is presented. Finally, some numerical results are given in order to show how to approximate the dynamics of the system originated by convection.

4.2.1 The model

We consider a thermal convection loop consisting of a circular pipe standing in a vertical plane, as shown in Figure 4.4. Such devices are often called thermosyphons (see Yorke et al. [48] and Bau et al. [44]) and are of particular interest since a circulation of fluid can be achieved without use of pumps. Usually such loops are heated from below and cooled from above. This induces a fluid flow only driven by temperature called natural convection. Our goal is to give a reasonable model for the simulation of the flow behaviour of the fluid (cf. Yorke et al. [48], Bau et al. [44] and Rubio [36]).
We denote the radius of the pipe by r_p and the radius of the torus by R. If we denote the velocity vector by $\mathbf{v}$, the pressure by p, the density by ρ and the external body force by $\mathbf{f}$, the behaviour of incompressible Newtonian fluids in a loop $\Omega \subset \mathbb{R}^2$ for time $t \in [0,T]$ can be described by the Navier-Stokes equations

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} = \frac{1}{\rho}\,\mathbf{f} - \frac{1}{\rho}\,\nabla p + \nu\Delta\mathbf{v} \tag{4.23}$$

$$\operatorname{div}\mathbf{v} = 0, \tag{4.24}$$

where the density ρ is constant since the considered fluid is incompressible. Furthermore, ν denotes the constant kinematic viscosity. As mentioned above we consider an enhanced model including the heat transfer. That means thermodynamic characteristics of the fluid as well as the vertical buoyancy force have to be taken into account. If we denote the temperature by θ, the heat equation (convection diffusion equation) is given by

$$\frac{\partial \theta}{\partial t} + \mathbf{v}\cdot\nabla\theta = \chi\Delta\theta, \tag{4.25}$$

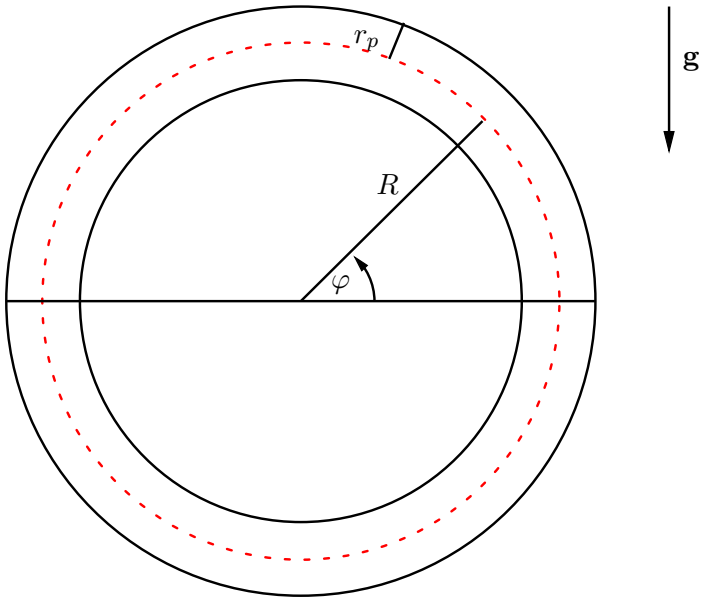

Figure 4.4: Thermal convection loop

with the coefficient of thermal diffusivity χ. Heat transfer by convection is described by the term $\mathbf{v} \cdot \nabla\theta$ and heat transfer through diffusion by means of the term $\chi\Delta\theta$. We make the following assumptions: Variations of the fluid's characteristics due to changes in density are neglected with the exception of the vertical buoyancy force, which is due to changes in density induced by temperature changes. We assume the density ρ depending linearly on the temperature θ,

$$\rho = \rho_0(1 + \beta(\theta_0 - \theta))$$

with the thermal expansion coefficient β, the reference density ρ_0 of the fluid and the reference temperature θ_0 of the fluid. These simplifications are called Boussinesq approximation. In the case of the thermal convection loop the external force is induced by the gravity acceleration $\mathbf{g}$ and the vertical buoyancy force, given by $\mathbf{f_b} = (\rho_0 - \rho)(-\mathbf{g})$. Hence the external force becomes $\mathbf{f} = \rho\mathbf{g} + \mathbf{f_b} = \mathbf{g}(\rho + \rho_0\beta(\theta_0 - \theta))$ (see Rubio [36]) and we obtain the Boussinesq equations describing the flow behaviour

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = (1 + \beta(\theta_0 - \theta))\mathbf{g} - \frac{1}{\rho}\nabla p + \nu\Delta\mathbf{v} \tag{4.26}$$

$$\operatorname{div} \mathbf{v} = 0 \tag{4.27}$$

$$\frac{\partial \theta}{\partial t} + \mathbf{v} \cdot \nabla\theta = \chi\Delta\theta. \tag{4.28}$$

Based on the loop's geometry it is convenient to consider polar coordinates (r, φ) with radius $r \in [0, r_p)$ and angle $\varphi \in [0, 2\pi)$. Additionally, we make the assumption of a thin loop, that

means the radius of the pipe r_p is much smaller than the radius of the loop R, so that the velocity depends only on the radial coordinate. So, the fluid flow obeys circular streamlines, i.e. the flow of the fluid particles takes place at a fixed distance of the origin. Making these assumptions the velocity vector $\mathbf{v}$ can be represented by

$$\mathbf{v}(r, \varphi, t) = v(r, t)\, \mathbf{e}_\varphi, \tag{4.29}$$

where $\mathbf{e}_\varphi$ denotes a unit vector in direction of increasing angle φ. It can be shown that all flows of the form (4.29) satisfy the divergence free condition (see Rubio [36]). So, the Boussinesq equations can be written as

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = (1 + \beta(\theta_0 - \theta))\mathbf{g} - \frac{1}{\rho_0} \nabla p + \nu \Delta \mathbf{v} \tag{4.30}$$

$$\frac{\partial \theta}{\partial t} + \mathbf{v} \cdot \nabla \theta = \chi \Delta \theta. \tag{4.31}$$

But these equations can still be simplified by integrating both sides of equation (4.30) along a circular path at fixed radius $\bar{r}$. For a detailed derivation of the equations see Rubio [36]. We make the additional assumption that there is no diffusion along the pipe, i.e. in the convection diffusion equation we consider

$$\Delta_r \theta = \frac{\partial^2 \theta}{\partial r^2} + \frac{1}{r} \frac{\partial \theta}{\partial r}.$$

This is justified as in thin pipes the rate of change in temperature across the pipe is (much) larger than along the pipe.
Hence, the initial-boundary value problem describing the fluid flow in a thermal convection loop can be written as (cf. Rubio [36] and Yorke et al. [48])

$$\frac{\partial v}{\partial t}(r, t) = \nu \Delta_r v(r, t) + \frac{\beta g}{2\pi} \int_0^{2\pi} \theta(r, \varphi, t) \cos \varphi \, d\varphi \tag{4.32}$$

$$\frac{\partial \theta}{\partial t}(r, \phi, t) = -\frac{v(r, t)}{R} \frac{\partial \theta}{\partial \varphi}(r, \varphi, t) + \chi \Delta_r \theta(r, \varphi, t), \tag{4.33}$$

for $r \in (0, r_p)$, $\varphi \in [0, 2\pi)$ and $t \in (0, T)$. The magnitude of the gravity acceleration is denoted by g.
The boundary conditions are assumed to be "no-slip" conditions, i.e. at the wall the velocity of the flow is zero,

$$v(r_p, t) = 0, \qquad t \geq 0.$$

Further we assume equality of the wall temperature, at the inner and outer wall, and the fluid temperature at the wall, i.e. we have the Dirichlet boundary condition

$$\theta(r_p, \varphi, t) = u_w(\varphi)$$

for $t \in (0, T)$ and $\varphi \in (0, 2\pi)$, where the wall temperature is assumed to satisfy

$$u_w(\varphi) = \theta_c - W \sin(\varphi)$$

with a constant W and constant temperature $\theta_c = \theta(r_p, 0, t) = \theta(r_p, \pi, t)$ for $t \in (0, T)$. This boundary condition describes the difference of the wall temperature between the bottom $\varphi = \frac{3}{2}\pi$ and the top $\varphi = \frac{1}{2}\pi$ of the torus.
The initial conditions are given by

$$v(r, 0) = v_0(r), \qquad r \in (0, r_p)$$

for the velocity and

$$\theta(r, \varphi, 0) = \theta_0(r, \varphi), \qquad r \in (0, r_p),\ \varphi \in (0, 2\pi)$$

for the temperature distribution.

4.2.2 Fourier series approach

We formally expand the difference between the temperature θ and the wall temperature u_w into a Fourier series in angle φ with coefficients c_n and s_n depending on the radius r and time t, i.e.

$$\theta(r, \varphi, t) = u_w(\varphi) + \sum_{n=1}^{\infty} \left(c_n(r, t)\cos(n\varphi) + s_n(r, t)\sin(n\varphi)\right). \tag{4.34}$$

Hence, we have to determine the coefficients s_n and c_n. Putting the representation of the temperature (4.34) in equation (4.32) we compute the integral

$$\int_0^{2\pi} \theta(r, \varphi, t)\cos\varphi\, d\varphi = c_1(r, t)\pi,$$

with the help of the trigonometric integrals ($k \in \mathbb{N} \setminus \{1\}$)

$$\int_0^{2\pi} \cos^2(\varphi)\, d\varphi = \pi, \quad \int_0^{2\pi} \cos(\varphi)\cos(k\varphi)\, d\varphi = 0, \quad \int_0^{2\pi} \cos(\varphi)\sin(k\varphi)\, d\varphi = 0.$$

Then the differential equation for the velocity (4.32) can be rewritten as

$$\frac{\partial v}{\partial t}(r, t) = \nu\Delta_r v(r, t) + \frac{\beta g}{2}\, c_1(r, t).$$

In order to determine the Fourier coefficients of θ we put the representation (4.34) in the convection diffusion equation (4.33). In doing so, we have to compute the partial derivatives

$$\begin{aligned}
\frac{\partial \theta}{\partial t}(r, \varphi, t) &= \sum_k \left(\frac{\partial c_k}{\partial t}(r, t)\cos(k\varphi) + \frac{\partial s_k}{\partial t}(r, t)\sin(k\varphi)\right) \\
\frac{\partial \theta}{\partial \varphi}(r, \varphi, t) &= -W\cos\varphi + \sum_k \left(-kc_k(r, t)\sin(k\varphi) + ks_k(r, t)\cos(k\varphi)\right) \\
\Delta_r \theta &= \sum_k \left(\Delta_r c_k\cos(k\varphi) + \Delta_r s_k\sin(k\varphi)\right),
\end{aligned}$$

then we multiply both sides of equation (4.33) once with $\cos(k\varphi)$, once with $\sin(k\varphi)$ and integrate each time along the circular path from 0 to 2π. In a similar way as before we obtain the differential equations

$$\frac{\partial c_1}{\partial t}(r,t) = \chi\Delta_r c_1(r,t) - \frac{v(r,t)}{R}\, s_1(r,t) + \frac{W}{R}\, v(r,t)$$
$$\frac{\partial s_1}{\partial t}(r,t) = \chi\Delta_r s_1(r,t) + \frac{v(r,t)}{R}\, c_1(r,t)$$

for the case $k = 1$ and for each $k > 1$ we obtain

$$\frac{\partial c_k}{\partial t}(r,t) = \chi\Delta_r c_k(r,t) - \frac{v(r,t)}{R}\, s_k(r,t)$$
$$\frac{\partial s_k}{\partial t}(r,t) = \chi\Delta_r s_k(r,t) + \frac{v(r,t)}{R}\, c_k(r,t).$$

For abbreviation we drop the index 1 of c_1 and s_1. We recognize that the partial differential equations for $v(r,t)$, $c(r,t)$ and $s(r,t)$ decouple from the coefficients c_k and s_k for $k > 1$. Moreover, it can be shown that

$$\int_0^{r_p} (c_k^2(r,t) + s_k^2(r,t))r\, dr \to 0 \qquad (t \to \infty),$$

i.e. for increasing time $t \to \infty$ the higher frequency terms of the series can be neglected (see Yorke et al. [48]). Summarizing, we can determine the coefficients c, s and the velocity v by solving the three (parabolic) partial differential equations

$$\frac{\partial v}{\partial t}(r,t) = \nu\Delta_r v(r,t) + \frac{\beta g}{2}\, c(r,t) \tag{4.35}$$
$$\frac{\partial c}{\partial t}(r,t) = \chi\Delta_r c(r,t) - \frac{v(r,t)}{R}\, s(r,t) + \frac{W}{R}\, v(r,t) \tag{4.36}$$
$$\frac{\partial s}{\partial t}(r,t) = \chi\Delta_r s(r,t) + \frac{v(r,t)}{R}\, c(r,t). \tag{4.37}$$

We determine generalized solutions of the partial differential equations above by constructing the corresponding eigenfunction expansions of the functions $c(r,t)$, $s(r,t)$ and $v(r,t)$. Without loss of generality we consider a radius of the pipe $r_p = 1$ and obtain for $\psi \in C^2(0,1)$ the associated Sturm-Liouville eigenvalue problem

$$r\psi'' + \psi' + \lambda r\psi = 0, \qquad \psi(1) = 0.$$

Since we recognize the equation of Bessel type we can determine the system of eigenfunctions $\psi_n(r) = \sqrt{c_n}J_0(z_n r)$, which form a complete orthonormal system in $L^2_\omega(0,1)$ with $\omega(r) = r$ as we already know. The eigenvalues λ_n are given by $J_0(\sqrt{\lambda_n}) = 0$, i.e. the eigenvalues are the zeros of J_0, denoted by $\lambda_n = z_n^2$. Here, the constants for normalization are $c_n = 2/J_1^2(z_n)$.

Now we can expand v, c and s for each $t \in (0,T)$ into (generalized) Fourier series in eigenfunctions

$$v(r,t) = \sum_{n=1}^{\infty} \alpha_n^v(t)\psi_n(r), \qquad \alpha_n^v(t) = \int_0^1 r\, v(r,t)\psi_n(r)dr,$$
$$c(r,t) = \sum_{n=1}^{\infty} \alpha_n^c(t)\psi_n(r), \qquad \alpha_n^c(t) = \int_0^1 r\, c(r,t)\psi_n(r)dr,$$
$$s(r,t) = \sum_{n=1}^{\infty} \alpha_n^s(t)\psi_n(r), \qquad \alpha_n^s(t) = \int_0^1 r\, s(r,t)\psi_n(r)dr,$$

where the Fourier coefficients $\alpha_n^v(t)$, $\alpha_n^c(t)$ and $\alpha_n^s(t)$ are given by

$$\frac{d}{dt}\,\alpha_n^v(t) = \nu \int_0^1 r\, v(r,t)\psi''(r)\,dr + \nu \int_0^1 v(r,t)\psi'(r)\,dr + \frac{\beta g}{2}\int_0^1 r\, c(r,t)\psi(r)\,dr$$

for the velocity $v(r,t)$,

$$\frac{d}{dt}\,\alpha_n^c(t) = \chi \int_0^1 r\, c(r,t)\psi''(r)\,dr + \chi \int_0^1 c(r,t)\psi'(r)\,dr$$
$$- \frac{1}{R}\int_0^1 r\, v(r,t)s(r,t)\psi(r)\,dr + \frac{W}{R}\int_0^1 r\, v(r,t)\psi(r)\,dr$$

for the function $c(r,t)$ and

$$\frac{d}{dt}\,\alpha_n^s(t) = \chi \int_0^1 r\, s(r,t)\psi''(r)\,dr + \chi \int_0^1 s(r,t)\psi'(r)\,dr + \frac{1}{R}\int_0^1 r\, v(r,t)c(r,t)\psi(r)\,dr$$

for the function $s(r,t)$. So, the Fourier coefficients can be determined by solving the ordinary differential equations (cf. Yorke et al. [48])

$$\begin{aligned}
\frac{d}{dt}\,\alpha_n^v &= -\nu\left(\frac{z_n}{r_p}\right)^2 \alpha_n^v + \frac{\beta g}{2}\,\alpha_n^c \\
\frac{d}{dt}\,\alpha_n^c &= -\chi\left(\frac{z_n}{r_p}\right)^2 \alpha_n^c - \frac{1}{R}\sum_{j,k}\gamma_{jk}^{(n)}\alpha_j^v\alpha_k^s + \frac{W}{R}\alpha_n^v \\
\frac{d}{dt}\,\alpha_n^s &= -\chi\left(\frac{z_n}{r_p}\right)^2 \alpha_n^s + \frac{1}{R}\sum_{j,k}\gamma_{jk}^{(n)}\alpha_j^v\alpha_k^c,
\end{aligned} \tag{4.38}$$

with

$$\gamma_{jk}^{(n)} = \frac{2}{J_1^2(z_n)}\int_0^1 x\, J_0(z_n x)J_0(z_j x)J_0(z_k x)\,dx.$$

Hence, we have to solve a system of coupled ordinary differential equations in order to compute the desired functions θ and v for simulating the thermal convection loop.
The initial conditions $\alpha_n^v(0)$, $\alpha_n^c(0)$ and $\alpha_n^s(0)$ are determined by the initial conditions for the velocity $v(r,0) = v_0(r)$ and the temperature $\theta(r,\varphi,0) = \theta_0(r,\varphi)$.

4.2.3 The Lorenz equations

The connection between the system (4.38) of ordinary differential equations and the Lorenz equations can be seen by the use of dimensionless quantities. We introduce the dimensionless variable τ for measuring time by

$$\tau = \frac{\chi z_1^2}{r_p^2}\, t$$

and the quantities

$$\xi = \frac{\beta g}{2\chi\nu}\left(\frac{r_p}{z_1}\right)^4 \frac{W}{R}, \qquad \sigma = \frac{\nu}{\chi} = \mathrm{Pr}$$

where σ is the Prandtl number of the fluid. The corresponding quantities with respect to the coefficients are given by

$$X_n = h_n^v \alpha_n^v, \qquad Y_n = h_n^c \alpha_n^c, \qquad Z_n = h_n^s \alpha_n^s,$$

with suitable quantities h_n^v, h_n^c and h_n^s. For an exact definition of the quantities and a more detailed derivation of the Lorenz equations see Yorke et al. [48].
In order to show the connection to the Lorenz equations we restrict ourselves to the case $n = 1$. The differentiation of the (new) coefficients X_1, Y_1 and Z_1 with respect to τ yields

$$\begin{aligned}
\frac{dX_1}{d\tau} &= -\sigma X_1 + \sigma Y_1 \\
\frac{dY_1}{d\tau} &= -X_1 Z_1 - Y_1 + \xi X_1 \\
\frac{dZ_1}{d\tau} &= X_1 Y_1 - Z_1,
\end{aligned}$$

and we recognize the well-known Lorenz equations as a truncation of the system above to one mode. The usual formulation of the Lorenz equations, which can be found e.g. in Doering [7], p. 75, is

$$\begin{aligned}
\frac{dX}{d\tau} &= -\sigma X + \sigma Y \\
\frac{dY}{d\tau} &= -XZ - Y + rX \\
\frac{dZ}{d\tau} &= XY - bZ
\end{aligned}$$

with positive parameters σ, b and r. In our case the parameter σ is determined by the fluids Prandtl number, the (geometric) parameter is $b = 1$ and the parameter r, or ξ, corresponds to the Rayleigh number Ra, which expresses the proportion of thermal expansion to viscosity. If the number Ra exceeds a certain critical value $\mathrm{Ra_c}$ we obtain a chaotic flow.

Now it is possible to characterize the flow behaviour using the known characteristics of the Lorenz system. A detailed description of the characteristics of the dynamics of the system can be found in Doering [7], (pp. 75-87) and in Sparrow [39]. We give a short summary of the basic properties of the system and implications to the flow behaviour (cf. Doering [7], pp. 75-80).
We start with the determination of the system's stationary points and the stability analysis of these points. Writing the nonlinear Lorenz system with the help of the function $F : \mathbb{R}^3 \to \mathbb{R}^3$, defined by

$$F(X, Y, Z) = (-\sigma X + \sigma Y, -XZ - Y + \xi X, XY - bZ)^T,$$

the Lorenz system can be written as

$$\frac{d}{d\tau}(X, Y, Z)^T = F(X, Y, Z)$$

and we can determine the stationary points of the system. We obtain

$$p_1 = (0, 0, 0)$$

and the two points

$$p_2 = \left(\sqrt{\xi - 1}, \sqrt{\xi - 1}, \xi - 1\right), \quad p_3 = \left(-\sqrt{\xi - 1}, -\sqrt{\xi - 1}, \xi - 1\right),$$

existing for $\xi > 1$, such that $F(p_i) = 0$ for $i = 1, 2, 3$. The origin p_1 corresponds to the case of no convection in our flow problem. First, we investigate the stability of the origin. For this purpose we consider the linearized system given by the Jacobian of the function F, defined by

$$J_F(X, Y, Z) = \begin{pmatrix} -\sigma & \sigma & 0 \\ \xi - Z & -1 & -X \\ Y & X & -b \end{pmatrix}.$$

The next step is the computation of the eigenvalues of the matrix J_F at the three fixed points. At the origin the Jacobian reads as

$$J_F(0, 0, 0) = \begin{pmatrix} -\sigma & \sigma & 0 \\ \xi & -1 & 0 \\ 0 & 0 & -b \end{pmatrix}.$$

The solutions of the equation

$$(\lambda + b)(\lambda^2 + (\sigma + 1)\lambda - \sigma(\xi - 1)) = 0$$

are the eigenvalues $\lambda_1 = -b$ and

$$\lambda_{2,3} = -\tfrac{1}{2}(\sigma + 1) \pm \tfrac{1}{2}\sqrt{(\sigma + 1)^2 + 4\sigma(\xi - 1)}.$$

We see that the three eigenvalues are always real (assuming $\sigma > 0$) and that λ_2 has negative sign for $\xi < 1$ and positive sign for $\xi > 1$. Hence, we have for $\xi < 1$ (asymptotic) stability

of the origin $p_1 = (0, 0, 0)$ since the real parts of all eigenvalues are negative. Whereas for $\xi > 1$ the origin p_1 is non-stable.
At the fixed points p_2 and p_3 the Jacobian reads

$$J_F(\delta, \delta, r-1) = \begin{pmatrix} -\sigma & \sigma & 0 \\ 1 & -1 & -\delta \\ \delta & \delta & -b, \end{pmatrix}$$

setting $\delta = \pm\sqrt{\xi - 1}$ for abbreviation leading to the characteristic equation

$$\lambda^3 + (\sigma + b + 1)\lambda^2 + b(\sigma + \xi)\lambda + 2\sigma b(\xi - 1) = 0.$$

Setting the critical value

$$\xi_c = \frac{\sigma(\sigma + 4)}{\sigma - 2}$$

it can be shown that for $\xi < \xi_c$ all eigenvalues have negative real part and therefore the two stationary points p_2 and p_3 are stable, whereas for $\xi > \xi_c$ the points p_2 and p_3 are non-stable (see Doering [7], p. 79).

Example 4.2 We choose the parameter $\sigma = 6$ implying the critical value $\xi_c = 15$. So, we can determine the stationary points for

$$\xi = 10 : p_2 = (3, 3, 9), \quad p_3 = (-3, -3, 9),$$
$$\xi = 26 : p_2 = (5, 5, 25), \quad p_3 = (-5, -5, 25).$$

We recognize the stable case shown in the illustration in Figures 4.5 and 4.6 on the left-hand side and the non-stable case on the right-hand side.

4.2.4 Numerical results

In this section we compute approximations of the velocity and the temperature of a fluid flow in a thermal convection loop.

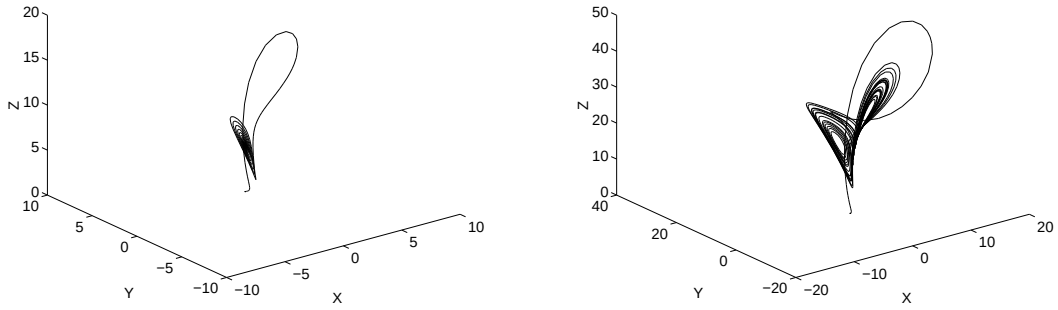

Figure 4.5: Lorenz system for $\xi = 10$ and $\xi = 26$

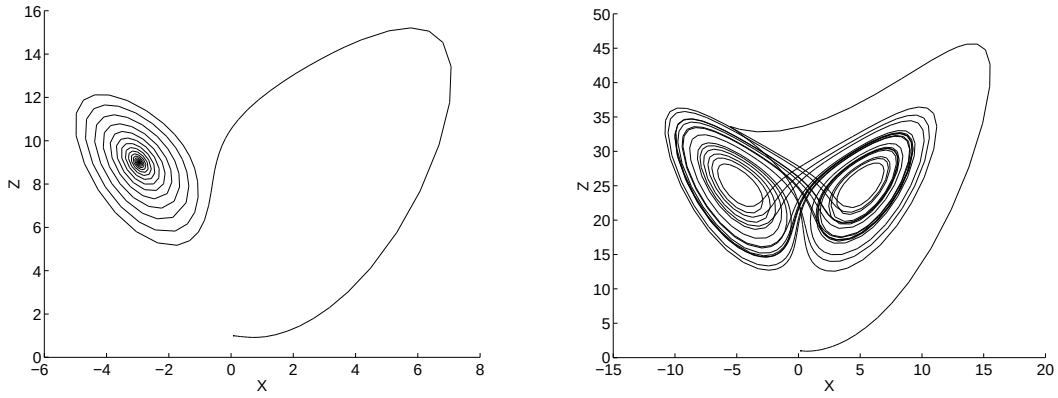

Figure 4.6: $(X, Z)-$projection of Lorenz system for $\xi = 10$ and $\xi = 26$

The following data of the fluid are given: the thermal diffusivity is $\chi = 1.514 \cdot 10^{-6}$ ft^2/s, the kinematic viscosity is $\nu = 1.22 \cdot 10^{-5}$ ft^2/s, the thermal expansion coefficient is $\beta = 8.0 \cdot 10^{-4}$ /K. The radius of the loop is given by $R = 1.2467$ ft and the radius of the pipe is $r_p = 4.921 \cdot 10^{-2}$ ft. The wall temperature is $u_w(\varphi) = \theta_c - W \sin\varphi$ with $\theta_c = 314.15$ K and variation $W = 1$. The time interval is $[0, 1000]$.
We compute approximations of the velocity

$$v(r,t) \approx \sum_{n=1}^{N} \alpha_n^v(t)\psi_n(r)$$

and of the temperature

$$\theta(r,\varphi,t) \approx u_w(\varphi) + \cos\varphi \sum_{n=1}^{N} \alpha_n^c(t)\psi_n(r) + \sin\varphi \sum_{n=1}^{N} \alpha_n^s(t)\psi_n(r).$$

The initial values are given by $v_0 = 0$, $\theta_0 = \theta_c$.
In Figure 4.9 the velocity profile (left-hand side) and the temperature profile (right-hand side) for various terms of the partial sums are shown for $\varphi = \frac{3}{2}\pi$. The upper pair of figures shows the one-mode model $N = 1$, the pair of figures in the middle are the results for $N = 2$ and the lower pair of figures represents the profiles for degree $N = 5$. It can be seen that the addition of further terms of the partial sums improves the quality of the approximations.
In Chapter 5 we present possible applications of the model and further numerical results.

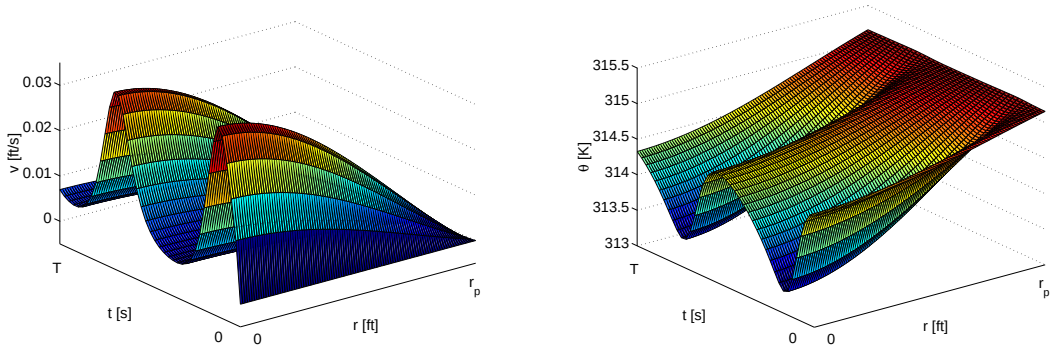

Figure 4.7: $v(r,t)$, $\theta(r,\frac{3}{2}\pi,t)$, $N=1$

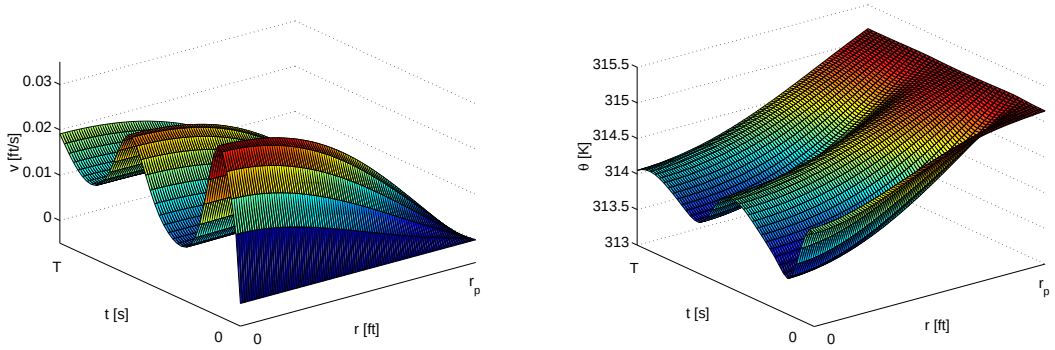

Figure 4.8: $v(r,t)$, $\theta(r,\frac{3}{2}\pi,t)$, $N=2$

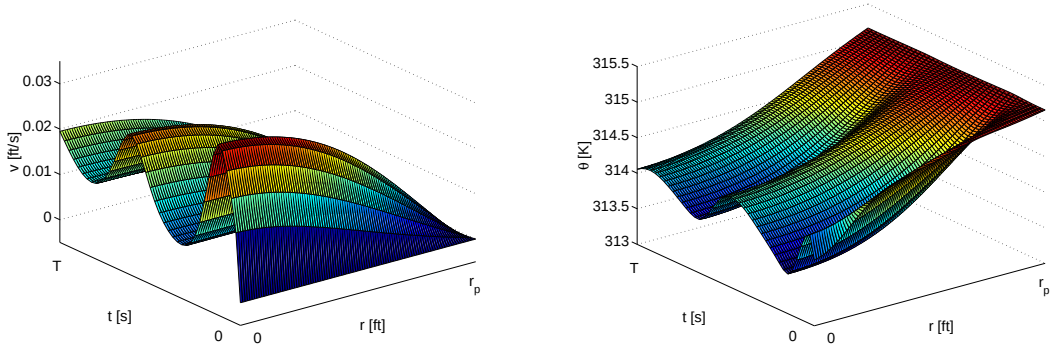

Figure 4.9: $v(r,t)$, $\theta(r,\frac{3}{2}\pi,t)$, $N=5$

Chapter 5

Modelling the heat transfer in food processing

In food processing the sterilization of food products in so-called autoclaves plays an important role. Thereby the product is filled into containers which are heated in the autoclave by steam or hot water. During the heating processes harmful microorganisms are destroyed. Unfortunately heat sensitive nutrients and vitamins are also affected during the heating process. Since in practice the temperature of the autoclave is often empirically determined, it may happen that during the sterilization process not only the harmful microorganisms but also the nutrients and vitamins are destroyed. Hence, it is necessary to have appropriate mathematical models for the sterilization process and the heat transfer in order to optimize the nutritional quality of the product.
In this chapter we first present a sterilization model which enables us to measure the microorganism destruction during heating processes. In this way we can formulate an optimal control problem with the goal of optimizing the nutritional quality of the product during the heating process subject to a sterility constraint. We will then recognize the importance of having a reasonable model of the heat transfer presenting different models and discussing their advantages and disadvantages.

5.1 An optimal control problem

In this section we specify an optimal control problem describing the maximization of the nutritional quality of the product subject to the sterility constraint. We consider sterilization processes where prepackaged food is heated in an autoclave in order to eliminate existing microorganisms, i.e. the thermal effect on microorganisms and nutrients has to be taken into account. Consequentially, it is necessary to provide a sterility requirement in the model in order to formulate a suitable control problem. For the derivation and investigation of the model for sterilization of prepackaged food we refer to Kleis/Sachs [20] and Kleis [19].
Let Ω denote the domain of the container of the product under consideration and $\partial\Omega$ its

boundary. The concentration of microorganisms, $C(x,t)$, at the point (x,t) can be described by an initial value problem consisting of a linear differential equation

$$\begin{aligned} \frac{\partial C}{\partial t}(x,t) &= -K(\theta(x,t))C(x,t), \\ C(x,0) &= C_0(x) \end{aligned} \tag{5.1}$$

where $\theta(x,t)$ is the absolute temperature at the point $x \in \Omega$ at time t and C_0 denotes the initial concentration of microorganisms at $x \in \Omega$. The function $K(\theta)$ is given by the Arrhenius equation

$$K(\theta) = K_{\text{ref}} \exp\left(-\frac{E}{R}\left(\frac{1}{\theta} - \frac{1}{\theta_{\text{ref}}}\right)\right),$$

with the reference temperature θ_{ref}, the activation energy E and the universal gas constant R. The data θ_{ref} and E are dependent on the considered microorganism. In order to obtain a sterility condition we have to solve the initial value problem (5.1). For a continuous function $\theta(x,\cdot)$ in $[0,T]$ and fixed $x \in \Omega$ the continuity of the function $K(\theta(x,\cdot))$ yields the unique solution of the initial value problem (5.1)

$$C(x,t) = C_0(x) \exp\left(-\int_0^t K(\theta(x,\tau))d\tau\right).$$

The concentration of the microorganisms is affected by the temperature at the corresponding location. We recognize that for fixed location the concentration decreases with increasing temperature. Therefore, in practice, one has to consider only that location where the product is the coldest during the heating period. We call the point $x_c \in \Omega$ the coldest point which is often assumed to be located in the geometrical center of the container.

In order to describe the sterility condition for the product we consider the concentration of microorganisms at the point x_c for the processing time T with respect to the initial concentration

$$\frac{C(x_c,T)}{C_0(x_c)} \leq 10^{-\beta}$$

with the given reduction rate β corresponding to the considered microorganisms. The concentration $C(x_c,T)$ can be rewritten as

$$\begin{aligned} C(x_c,T) &= C_0(x_c) \exp\left(-\int_0^T K(\theta(x_c,\tau))d\tau\right) \\ &= C_0(x_c) \exp\left(-\int_0^T K_{\text{ref}} 10^{-\frac{E}{R\ln 10}\left(\frac{1}{\theta(x_c,\tau)} - \frac{1}{\theta_{\text{ref}}}\right)} d\tau\right) \\ &= C_0(x_c) \exp\left(-\int_0^T K_{\text{ref}} 10^{\frac{\theta(x_c,\tau)-\theta_{\text{ref}}}{z(\theta(x_c,\tau))}} d\tau\right) \end{aligned}$$

with the z-value

$$z(\theta(x_c,\tau)) = \frac{R}{E}\,\theta_{\text{ref}}\theta(x_c,\tau)\ln 10.$$

If we neglect the dependence of the z-value on the temperature and use the approximation $\theta_{\text{ref}}\theta(x_c,\tau) \approx \theta_{\text{ref}}^2$ we obtain the sterility condition

$$\exp\left(-\int_0^T K_{\text{ref}} 10^{\frac{\theta(x_c,\tau)-\theta_{\text{ref}}}{z_{\text{ref}}}} d\tau\right) \leq 10^{-\beta}$$

with $z_{\text{ref}} = (R/E)\theta_{\text{ref}}^2$. Using logarithms we can reformulate this condition as

$$\int_0^T 10^{\frac{\theta(x_c,\tau)-\theta_{\text{ref}}}{z_{\text{ref}}}} d\tau \geq \frac{\beta \ln 10}{K_{\text{ref}}}.$$

The expression on the left hand side is characteristic for the measurement of microorganism destruction and common in food industry. For the following definition cf. Kessler [18], p. 172.

Definition 5.1. The function

$$F(\theta,\theta_{\text{ref}})(x) = \int_0^T 10^{(\theta(x,\tau)-\theta_{\text{ref}})/z_{\text{ref}}} d\tau$$

is called the *F-value* at the point x to the reference temperature θ_{ref} and z_{ref}.

Using the definition of the F-value we can rewrite the sterility condition as

$$F(\theta,\theta_{\text{ref}})(x_c) \geq \frac{\beta \ln 10}{K_{\text{ref}}}.$$

This requirement can be interpreted as follows: The F-value is the needed time to achieve a given F_0-value, where the F_0-value is defined as

$$F_0 = \beta\,\frac{\ln(10)}{K_{\text{ref}}}.$$

In engineering practice this value is often related to the microorganism *Clostridium botulinum* with the value $\beta = 12$. The reference value K_{ref} depends on the heated product and the reference temperature is usually $\theta_{\text{ref}} = 394.25$ K (see Kessler [18], p. 172).
We can interpret the F-value as a measure for the reduction of microorganisms for a given temperature profile $\theta(x,t)$.
The differential equation (5.1) can also be used to describe the concentration of nutrients. Determining the related reference data θ_q and z_q we can define analogously to the definition of the F-value

$$J(\theta,\theta_q)(x) = \int_0^T 10^{(\theta(x,\tau)-\theta_q)/z_q} d\tau$$

to reflect the desired value. Since it is desired to maximize the concentration of nutrients we consider the function J at the boundary of the container x_b because this region is the most critical, in terms of destruction of nutrients, during the heating period.
Now we are able to formulate the control problem. Introducing the boundary control $u(t)$ for $t \in (0,T)$, the temperature of the autoclave, the control problem can be formulated by the minimization of the objective function

$$J(\theta, \theta_q)(x_b) + \frac{\mu}{2} \int_0^T |u(t)|^2 dt$$

subject to the sterility condition

$$F(\theta, \theta_{\text{ref}})(x_c) \geq F_0$$

and the constraint

$$u_{\text{low}} \leq u(t) \leq u_{\text{up}} \qquad \text{for } t \in (0,T)$$

due to technical restrictions on the control for $t \in (0,T)$, i.e. there are upper and lower bounds for the temperature of the autoclave. The objective function includes a regularizing energy term weighted with a parameter $\mu > 0$. For investigation of such optimal control problems we refer to Kleis/Sachs [20] and Kleis [19].
We notice that $\theta = \theta(x,t;u)$ is the temperature distribution corresponding to the boundary control $u(t)$. Hence, the determination of a reasonable heat transfer model is very important. In particular, the sterility constraint given by the F-value requires a reasonable approximation of the temperature profile since the F-value is very sensitive to changes in temperature particularly if the temperature is near the reference value.

5.2 Modelling the heat transfer

The choice of the heat transfer model for representing the temperature evolution inside the food container (often of cylindrical shape) is very important. We will give a survey of some of the available models and their approximations. The available input parameters are the temperature of the autoclave, the initial temperature of the product, the size of the container and the heat diffusivity.

5.2.1 The Ball-formula

A very simple approach for modelling heat transfer in food processing is the Ball-formula. Although presented in 1957 in Ball/Olson [4], p. 228, the Ball-formula is still very popular in engineering practice.
If we assume constant and uniform temperature of the autoclave, $u(t) = u_c$, and constant initial temperature θ_0, the formula for the scaled difference

$$\vartheta = \frac{\theta - u_c}{\theta_0 - u_c}$$

is given by

$$\vartheta = je^{-(t/f)\ln(10)}. \tag{5.2}$$

Considering the absolute temperature θ we can rewrite the formula as

$$\theta(x_c, t) = j(\theta_c - u_c)e^{-(t/f)\ln(10)} + u_c.$$

The Ball-formula is determined by two parameters j and f dependent on the geometry of the problem and the heat transfer coefficient. The parameter f is also dependent on the heat diffusivity. For heating processes the two parameters are positive. A detailed description of the parameters is given in (5.3). Usually the parameters are determined empirically using measured data of the heated product.
On the other hand the Ball-formula can be interpreted as a solution of a time-dependent ordinary differential equation

$$\frac{d}{dt}\vartheta = -\frac{\ln(10)}{f}\ \vartheta, \quad \vartheta(0) = j$$

or

$$\frac{d}{dt}\theta = -\frac{\ln(10)}{f}\ (\theta - u_c), \quad \theta(x_c, 0) = j(\theta_c - u_c) + u_c.$$

Since this differential equation only depends on time the conductivity can not be represented by this equation. The spatial effect is neglected in this equation which is important for a heat conduction process. Moreover, the Ball-formula leads to reliable results only for sufficiently large time values t (see Ball/Olson [4], p. 288). Hence, in this form it cannot be used for cooling processes, or in general, changes in autoclave temperature. The reason is the initial lag of the cooling curve, since the conductivity of the process cannot be modelled by (5.2). This problem is illustrated in the section on numerical results. In order to compensate this lag portion, this part of the curve is instead approximated by a hyperbola (see Ball/Olson [4], p. 321).
Despite these disadvantages this formula is still in practical use. The main reasons may be the simplicity of the formula as well as the often required large time values in heating processes.

5.2.2 The Fourier series approach

Another approach imbeding the Ball-formula (5.2) as a special case, is based on the (conductive) heat equation and the representation of its solution via Fourier series. Based on a multi-dimensional heat equation we show how the Ball-formula can be derived in this case. For a detailed derivation of the following model we refer to Section 4.1.4. Starting from a cylindrical domain, a cross-section of a cylinder, considered in a cylindrical coordinate system we can reduce the two-dimensional problem to a one-dimensional problem due to symmetry. Then the initial-boundary value problem for the one-dimensional (conductive) heat equation

reads as

$$\begin{aligned} r\,\frac{\partial\theta}{\partial t}(r,t) &= \chi\left(r\,\frac{\partial^2\theta}{\partial r^2}(r,t) + \frac{\partial\theta}{\partial r}(r,t)\right) \\ \theta(r,0) &= \theta_0(r) \\ \frac{\partial\theta}{\partial r}(1,t) &= \alpha(u(t)-\theta(1,t)) \end{aligned}$$

for $(r,t) \in Q$, where $Q = (0,1)\times(0,T)$ with radius r and time t, control u and state θ. We denote again with χ the heat diffusivity and with θ_0 the initial temperature distribution. Assuming constant autoclave temperature $u(t) = u_c$ and constant initial temperature $\theta_0(r) = \theta_c$ the temperature inside the container can be represented by the series expansion

$$\theta(r,t) = \sum_{n=1}^{\infty} d_n J_0(k_n r) e^{-k_n^2 \chi t} + u_c$$

with constant coefficients

$$d_n = c_n(\theta_0 - u_c)J_1(k_n)/k_n.$$

Hence, the computed approximations are the partial sums of the series $\theta(r,t)$, defined by $\theta_N(r,t)$ for $N \in \mathbb{N}$. If we consider only the first term and let $r \to 0^+$,

$$\theta_1(0^+,t) = d_1 e^{-k_1^2\chi t},$$

we (formally) obtain the Ball-formula (5.2) with parameters

$$j = c_1 J_1(k_1)/k_1, \quad f = \ln(10)/(k_1^2\chi). \tag{5.3}$$

The dependencies of the parameters of the Ball-formula are easy to see from these representations. By appearance of the first eigenvalue k_1 the parameter j depends on the geometry of the product and the heat transfer coefficient α. The parameter f is also dependent on the geometry and the heat transfer coefficient but furthermore inversely correlated to the heat diffusivity χ.

5.2.3 Reduced order modelling

If the given problem is of more complex structure a different model is required, e.g. a nonlinear heat equation of the form

$$\begin{aligned} c(\theta)\frac{\partial}{\partial t}\theta - \nabla\cdot(k(\theta)\nabla(\theta)) &= 0 && \text{in } \Omega\times(0,T) \\ \theta(\cdot,0) &= \theta_0(\cdot) && \text{in } \Omega \\ k(\theta)\frac{\partial}{\partial n}\theta &= \tilde{\alpha}(u-\theta) && \text{on } \partial\Omega\times(0,T) \end{aligned}$$

with density ρ, heat capacity c and heat conductivity k of the product might be solved approximately by finite elements methods. But after discretizing the partial differential equation often the resulting problem to be solved is of large-scale type. Therefore reduced order models like proper orthogonal decomposition are suitable to achieve a reduction of dimension (see Fahl [9] and [10]).

5.2.4 Numerical results

In this section we compare the different approaches for the computation of the temperature evolution inside the container for a simple model problem. The given data are the domain $\Omega = (0, 0.03)$, the time $T = 3000$ s, a constant heat diffusivity χ, so that we have to consider the following initial-boundary value problem

$$\begin{aligned} r\,\frac{\partial\theta}{\partial t} &= \chi\left(r\frac{\partial^2\theta}{\partial r^2} + \frac{\partial\theta}{\partial r}\right) \\ \theta(r,0) &= \theta_0(r) \\ \frac{\partial\theta}{\partial r}(0.03,t) &= \alpha(u(t) - \theta(0.03,t)) \end{aligned}$$

with the constants $\chi(= k/c\rho) = 10^{-6}$, $\alpha = \tilde{\alpha}/k = 1.25 \cdot 10^4$ and constant initial temperature $\theta_0 = 314.15$ K. The control representing the heating and cooling process is given by

$$u(t) = \begin{cases} 403.15 \text{ K}, & 0 < t < 1500 \text{ s} \\ 303.15 \text{ K}, & 1500 \text{ s} \le t \le 3000 \text{ s} \end{cases}.$$

As derived before we use the following approximation to the temperature distribution

$$\theta_N(r,t) = \sum_{n=0}^{N} d_n J_0(k_n r) e^{-k_n^2 \chi t} + u(t).$$

The numerical results are shown in Figure 5.1. The replacement of the area with the peak by an appropriate hyperbola is shown in Figure 5.2, for which we refer to Ball/Olson [4], p. 321. We remark that the construction of this hyperbola is based on empirical data being one of the main weaknesses of this method. Figure 5.1 shows the comparison of the unmodified Ball-formula (dashed line) with the Fourier series solution. We see the heating up and cooling down process with the specified temperatures. We notice that the peaks in the solution are due to the model. In particular the evident peak of the unmodified Ball-formula was the reason why this portion is replaced by an appropriate hyperbola in the Ball-formula. The

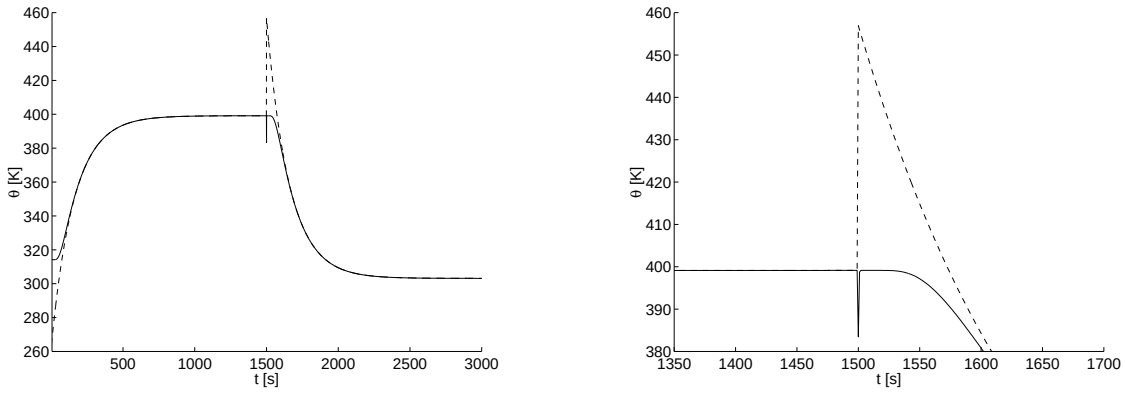

Figure 5.1: Ball-formula and Fourier Approximation

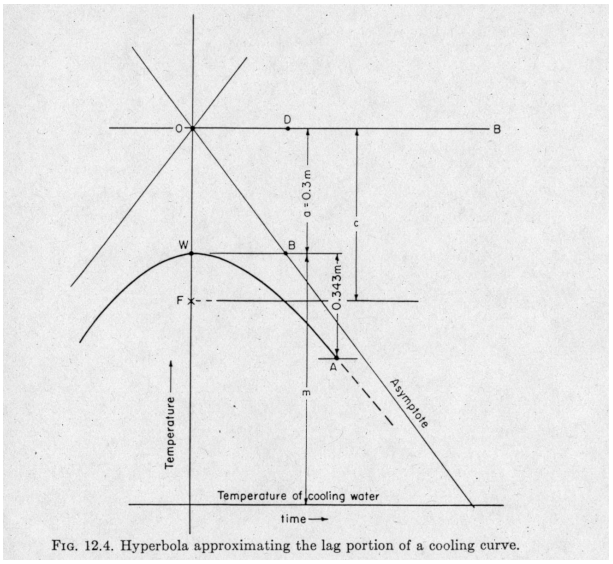

Figure 5.2: From Ball-Olson [4], Sterilization in food technology, 1957

plot for the Fourier series was obtained by using 18 terms in the Fourier series expansion. The additional use of terms of the series further improves the behaviour at the switching point, the peaks get smaller and the gap can be closed.
In order to see the effect of the different models on the F-value for a typical industrial sterilization process we use the following data: $\chi = 10^{-6}$, $\alpha = 1.25 \cdot 10^4$. The domain is still chosen as $\Omega = (0, 0.03)$ as well as the initial temperature $\theta_0 = 314.15$ K. For the control (autoclave temperature) we use

$$u(t) = \begin{cases} 399.15 & 0 \leq t \leq 1500\ s \\ 303.15 & 1500 \leq t \leq 3000\ s \end{cases} .$$

The computation of the F-value using the unmodified Ball-formula (dashed dotted line) as well as the Fourier partial sums of degree $N = 18$ (solid line) is shown in Figure 5.3.

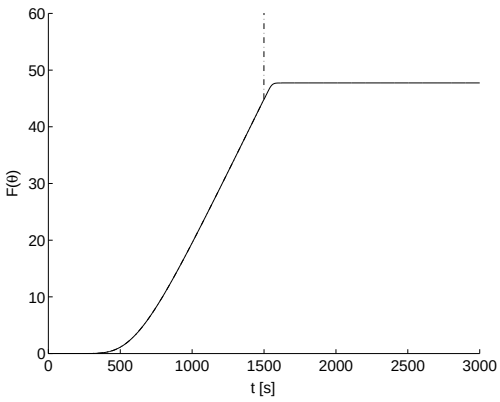

Figure 5.3: F-value computation using different models

We recognize the computed F-value of 47.7 (min.) using the Fourier partial sums, whereas the use of the unmodified Ball-formula delivers a completely wrong result, (the computed F-value is $1.5 \cdot 10^5$).
The presented results show how simple methods like the Ball-formula can significantly improved with the help of the Fourier series approach in order to achieve evidently better results. The improvements are primarily based on the explicit consideration of the underlying structure, i.e. geometry and symmetry. Of course more general problems of nonlinear type require more general and more complex models suited for simulation and optimal control of such problems (see [10]).

5.3 Thermal convection loop

We consider the model derived in Section 4.2 in the light of the sterilization processes presented in this chapter. That means we have a fluid filled loop, e.g. a cross-section of a hollow cylinder, and we would like to model the process of heating up. The radius of the loop is R, the radius of the pipe r_p. The heating time is given by T. Then, the initial-boundary value problem is given by (see (4.32) and (4.33))

$$\frac{\partial v}{\partial t}(r,t) = \nu \Delta_r v(r,t) + \frac{\beta g}{2\pi} \int_0^{2\pi} \theta(r,\varphi,t) \cos\varphi \, d\varphi$$
$$\frac{\partial \theta}{\partial t}(r,\phi,t) = -\frac{v(r,t)}{R} \frac{\partial \theta}{\partial \varphi}(r,\varphi,t) + \chi \Delta_r \theta(r,\varphi,t),$$

for $r \in (0, r_p)$, $\varphi \in [0, 2\pi)$ and $t \in (0, T)$. The initial conditions are given by

$$v(r,0) = v_0(r), \qquad r \in (0, r_p)$$

for the velocity and

$$\theta(r,\varphi,0) = \theta_0(r,\varphi), \qquad r \in (0, r_p),\ \varphi \in (0, 2\pi)$$

for the temperature distribution. The boundary conditions are

$$v(r_p,t) = 0, \qquad t \geq 0$$

and

$$\theta(r_p,\varphi,t) = u_w(\varphi)$$

with the wall temperature (boundary control) $u_w(\varphi) = \bar{u}_c - W \sin(\varphi)$. Then, the Fourier series representations of the velocity and the temperature are

$$v(r,t) = \sum_{n=1}^{\infty} \alpha_n^v \psi_n(r)$$

and

$$\theta(r,\varphi,t) = u_w(\varphi) + \cos(\varphi)\sum_{n=1}^{\infty}\alpha_n^c(t)\psi_n(r) + \sin(\varphi)\sum_{n=1}^{\infty}\alpha_n^s(t)\psi_n(r),$$

where the system of ordinary differential equations for the determination of the Fourier coefficients of the series expansions reads as $(n = 1, 2, \ldots)$

$$\begin{aligned}
\frac{d}{dt}\alpha_n^v(t) &= -\nu\left(\frac{z_n}{r_p}\right)^2\alpha_n^v(t) + \frac{\beta g}{2}\alpha_n^c(t)\\
\frac{d}{dt}\alpha_n^c(t) &= -\chi\left(\frac{z_n}{r_p}\right)^2\alpha_n^c(t) - \frac{1}{R}\sum_{j,k}\gamma_{jk}^{(n)}\alpha_j^v(t)\alpha_k^s(t) + \frac{W}{R}\alpha_n^v(t)\\
\frac{d}{dt}\alpha_n^s(t) &= -\chi\left(\frac{z_n}{r_p}\right)^2\alpha_n^s(t) + \frac{1}{R}\sum_{j,k}\gamma_{jk}^{(n)}\alpha_j^v(t)\alpha_k^c(t).
\end{aligned}$$

Numerical results

We give a comparison of the approximations of the temperature and the F-value for a different number of terms of the partial sums. The following data are given: the thermal diffusivity is $\chi = 1.514\cdot 10^{-6}$ ft^2/s, the kinematic viscosity is $\nu = 1.22\cdot 10^{-5}$ ft^2/s, the thermal expansion coefficient is $\beta = 8.0\cdot 10^{-4}$ /K. The radius of the loop is given by $R = 1.2467$ ft and the radius of the pipe is $r_p = 4.921\cdot 10^{-2}$ ft. The heating time is $T = 1000$ s and $\theta_0 = 314.15$ K. We compute the approximations

$$\theta(r,\varphi,t) \approx u_w(\varphi) + \cos\varphi\sum_{n=1}^{N}\alpha_n^c(t)\psi_n(r) + \sin\varphi\sum_{n=1}^{N}\alpha_n^s(t)\psi_n(r)$$

for the temperature and

$$F(\theta) = \int_0^T 10^{(\theta(r,\varphi,s)-\theta_{\mathrm{ref}})/z_{\mathrm{ref}}}\,ds$$

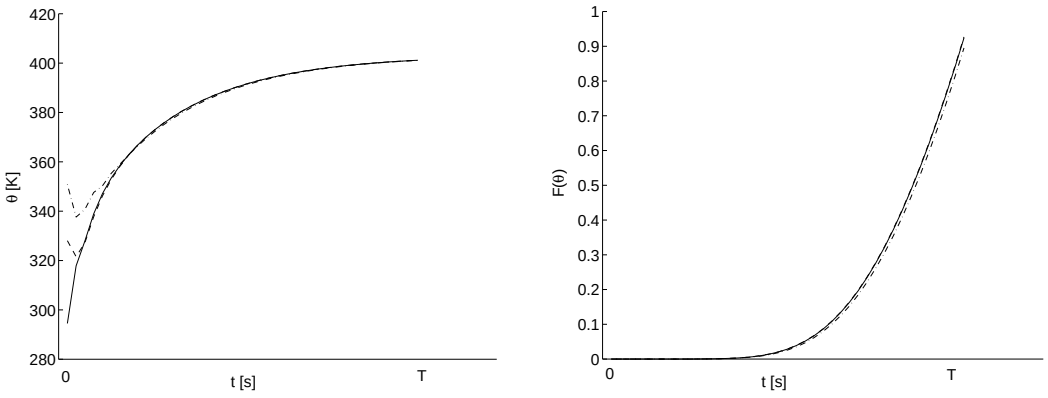

Figure 5.4: Temperature $\theta(r, \frac{3}{2}\pi, t)$ and F-value $F(\theta)$

for the F-value at the center of the pipe ($r \to 0^+$). The reference data are $\theta_{\text{ref}} = 394.25$ K, $z_{\text{ref}} = 10$ K and the wall temperature (boundary control) is $u_w(\varphi) = 399.15 - 4\sin(\varphi)$ K. So, the wall temperature at the top and the bottom of the loop is

$$u_w(\varphi) = \begin{cases} 395.15 \text{ K} & \text{for } \varphi = \frac{1}{2}\pi, \\ 403.15 \text{ K} & \text{for } \varphi = \frac{3}{2}\pi. \end{cases}$$

The results are shown in Figure 5.4. It can be seen that the increase of temperature, and also of F-value, is faster if the number of terms of the partial sums increases. The curve of the five-mode approximation (solid line) with $N = 5$ is steeper than the curves of the one-mode (dashed-dotted line) with $N = 1$ or two-mode approximation (dotted line) with $N = 2$. Moreover, we recognize that for small values of time t the approximations might be improved by the use of further terms of the series expansion of θ. But the addition of one term of the series requires the solution of a system of ordinary differential equations with three additional equations.

Conclusion

The first subject of this thesis was the development and analysis of efficient methods for the computation of confluent hypergeometric functions. Based on series expansions, recurrence relations and asymptotic expansions we presented and analyzed approaches for the computation of these functions yielding promising results in a great variety of cases. In particular we have seen that the computation with recurrence relations using the Miller algorithm leads to very good results for small (real) parameters a and c. Using the series expansion and the asymptotic expansion of the confluent hypergeometric function we are able to compute on compact intervals and on unbounded domains.
The second subject was the solution of certain boundary value problems using the (generalized) Fourier series approach for solving (parabolic) partial differential equations. We have shown how the computation of solutions of different heat transfer processes via the computation of confluent hypergeometric functions can be carried out. The advantage of this proceeding is the easy improvement of the approximation by adding more terms of the series, since a complete system of eigenfunctions is at hand. As a result we have seen that if the underlying geometry is of a suitable type the computation is practicable and efficient achieving a low computational effort compared to other methods for the numerical solution of partial differential equations based on discretization, e.g. finite differences. In these methods the improvement of the approximation requires a new discretization of the partial differential equations resulting in large-scale problems. On the other hand they are more flexible with respect to the problem geometry. Furthermore, we have seen how the Fourier series approach can be applied to modelling heat transfer processes in food processing. We have shown that these representations are generalizations of standard methods in food processing which are still in practical use, e.g. the Ball-formula.
As a final result we have seen again that also in the field of partial differential equations no method can be termed "standard" as the choice of the method and the efficiency of the implementation always depends on the specific structure of the problem. We hope we could contribute to this theory by pointing out the areas and circumstances where the analyzed methods are specifically advantageous.

Bibliography

[1] M. Abramowitz and I. Stegun. *Handbook of mathematical functions*. Dover Publ., New York, 1968.

[2] D. E. Amos, S. L. Daniel, and M. K. Weston. CDC 6600 Subroutines IBESS and JBESS for Bessel Functions $I_\nu(x)$ and $J_\nu(x)$, $x \geq 0$, $\nu \geq 0$. *ACM Trans. Math. Software*, **3**(1):76–92, 1977.

[3] L. C. Andrews. *Special functions of mathematics for engineers*. McGraw Hill, New York, 2nd edition, 1992.

[4] C. O. Ball and F. C. Olson. *Sterilization in food technology*. McGraw-Hill, New York, 1957.

[5] A. R. Barnett. High-precision evaluation of the regular and irregular coulomb wavefunction. *J. Comput. Appl. Math.*, **8**:29–33, 1982.

[6] R. P. Boas. *Entire functions*. Academic Press, New York, 1954.

[7] C. Doering and J. D. Gibbon. *Applied analysis of the Navier-Stokes equations*. Cambridge Univ. Press, Cambridge, 1995.

[8] A. Erdélyi. *Higher transcendental functions*, volume I of *Bateman Manuscript Project*. McGraw Hill, New York, 1953.

[9] M. Fahl. *Trust-region Methods for Flow Control based on Reduced Order Modelling*. PhD thesis, Universität Trier, 2000.

[10] M. Fahl, E. W. Sachs, and C. Schwarz. Modeling Heat Transfer for Optimal Control Problems in Food Processing. In *Proceedings of the 2000 IEEE CCA/CACSD, Sept. 25-27, 2000, Anchorage, AK, to appear*.

[11] W. Gautschi. Computational aspects of three-term recurrence relations. *SIAM Review*, **9**:24–82, 1967.

[12] W. Gautschi. Zur Numerik rekurrenter Relationen. *Computing*, **9**:107–126, 1972.

[13] W. Gautschi. A Computational Procedure for Incomplete Gamma Functions. *ACM Trans. Math. Software*, **5**(4):466–481, 1979.

[14] E. A. Gonzáles-Velasco. *Fourier analysis and boundary value problems*. Academic Press, San Diego, 1995.

[15] R. Haberman. *Elementary applied partial differential equations*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 1987.

[16] P. Henrici. *Applied and computational complex analysis*, volume 2. Wiley, New York, 1991.

[17] P. Henrici. *Applied and computational complex analysis*, volume 3. Wiley, New York, 1993.

[18] H. G. Kessler. *Lebensmittel- und Bioverfahrenstechnik - Molkereitechnologie*. Verlag A. Kessler, München, 1996.

[19] D. Kleis. *Augmented Lagrange SQP Methods and Application to the Sterilization of Prepackaged Food*. PhD thesis, Universität Trier, 1997.

[20] D. Kleis and E. W. Sachs. Optimal control of the sterilization of prepackaged food. *SIAM Journal on Optimization*, **10**(4):1180–1195, 2000.

[21] J. D. Logan. *Applied mathematics*. Wiley, New York, 2nd edition, 1997.

[22] D. W. Lozier. Software needs in special functions. *J. Comp. and Appl. Math.*, **66**:345–358, 1996.

[23] Y. L. Luke. *The special functions and their approximations*, volume I. Academic Press, New York, 1969.

[24] Y. L. Luke. *The special functions and their approximations*, volume II. Academic Press, New York, 1969.

[25] R. C. McOwen. *Partial differential equations: methods and applications*. Prentice Hall, Upper Saddle River, NJ, 1996.

[26] J. Müller. Convergence Acceleration of Taylor Sections by Convolution. *Constr. Approx.*, **15**:523–536, 1999.

[27] J. Müller. Series expansions for computing Bessel functions of variable order on compact intervals. *Numerical Algorithms*, **24**:299–308, 2000.

[28] M. Nardin, W. F. Perger, and A. Bhalla. Numerical evaluation of the confluent hypergeometric function for complex arguments of large magnitudes. *J. Comput. Appl. Math.*, **39**:193–200, 1992.

[29] A. F. Nikiforov. *Special functions of mathematical physics*. Birkhäuser, Basel, 1988.

[30] F. W. J. Olver. Error analysis of Miller's recurrence algorithm. *Math. Comput.*, **18**:65–74, 1964.

[31] F. W. J. Olver. Numerical solutions of second order linear difference equations. *J. Res. Nat. Bur. Standards*, **71B**:11–29, 1967.

[32] F. W. J. Olver. *Introduction to asymptotics and special functions*. Academic Press, New York, 1974.

[33] M. A. Pinsky. *Partial differential equations and boundary value problems with applications*. McGraw Hill, New York, 2nd edition, 1991.

[34] B. T. Polyak. *Introduction to Optimization*. Optimization Software, Inc., Publications Division, New York, 1987.

[35] J. R. Rice. The degree of convergence for entire functions. *Duke Math. J.*, **38**:429–440, 1971.

[36] D. Rubio. *Distributed Parameter Control of Thermal Fluids*. PhD thesis, Virginia Polytechnic Institute and State University, Blacksburg, 1997.

[37] W. Rudin. *Real and complex analysis*. McGraw-Hill, New York, 3rd edition, 1987.

[38] C. Schwarz. Computation of Confluent Hypergeometric Functions on Compact Intervals. In *Proceedings of the Alexits Memorial Conference: Functions, Series, Operators, Aug. 9-14, 1999, Budapest, to appear*.

[39] C. Sparrow. *The Lorenz equations: bifurcations, chaos, and strange attractors*. Springer, New York, 1982.

[40] N. M. Temme. *Special functions: An introduction to the classical functions of mathematical physics*. Wiley, New York, 1996.

[41] H. Triebel. *Höhere Analysis*. Deutscher Verl. d. Wiss., Berlin, 1972.

[42] F. Tröltzsch. *Optimality conditions for parabolic control problems and applications*. Number 62 in Teubner-Texte zur Mathematik. Teubner, Leipzig, 1984.

[43] J. L. Walsh. *Interpolation and approximation by rational functions in the complex domain*. AMS, Providence, RI, 5th edition, 1969.

[44] Y. Wang, J. Singer, and H. H. Bau. Controlling chaos in a thermal convection loop. *J. Fluid Mech.*, **237**:479–498, 1992.

[45] D. Werner. *Funktionalanalysis*. Springer, Berlin, 2nd edition, 1997.

[46] J. Wimp. *Computation with recurrence relations*. Pitman, Boston, 1984.

[47] T. Winiarski. Approximation and interpolation of entire functions. *Ann. Polon. Math.*, **23**:259–273, 1970.

[48] J. A. Yorke, E. D. Yorke, and J. Mallet-Paret. Lorenz-like chaos in a partial differential equation for a heated fluid loop. *Physica D*, **24**:279–291, 1987.

Tabellarische Zusammenfassung des Bildungsweges

Name	Christian Schwarz
Geburtstag	21. März 1974
Geburtsort	Saarlouis
08/1980 – 07/1984	Grundschule Bergstrasse, Völklingen
08/1984 – 06/1993	privates, staatlich anerkanntes Gymnasium Marienschule Saarbrücken
10/1993 – 04/1999	Studium der Angewandten Mathematik an der Universität Trier mit Abschluss Diplom
seit 05/1999	Stipendiat im Graduiertenkolleg "Mathematische Optimierung" der Deutschen Forschungsgemeinschaft an der Universität Trier