wi3097: Numerieke methoden voor differentiaalvergelijkingen

Supplement of the book

Numerical Analysis Burden and Faires

Dr.ir. C. Vuik

2002

Delft University of Technology
Faculty of Information Technology and Systems
Department of Applied Mathematical Analysis

This is a supplement of the book:

Numerical Analysis (7^{e} edition)
R.L. Burden, J.D. Faires
Brodes/Cole Publishing Company, Pacific Grove, 2001

Contents

Chapter 1

Supplement of Chapter 5

In Chapter 5.10 and 5.11 of the book *Numerical Analysis of Burden and Faires*, some results are given concerning stability of numerical methods. In this supplement these results are summarized and additional material concerning stability is given.

Stability
A general definition of stability is: small changes or perturbations in the initial conditions produce correspondingly small changes in the subsequent approximations. Phenomena which have unstable behavior are: buckling of a column under compression and resonance of a bridge due to wind forces. In this supplement we only consider stable applications.

1.1 Stability of initial-value problems

Suppose that the initial condition y_0 is perturbed with ε_0. The perturbed solution $\tilde{y}$ satisfies:

$$\tilde{y}' = f(t, \tilde{y}) \text{ with } \tilde{y}(0) = y_0 + \varepsilon_0.$$

The difference of the exact and perturbed solution is defined as ε: $\varepsilon(t) = \tilde{y}(t) - y(t)$. An initial-value problem is *stable* if

$$|\varepsilon(t)| \text{ is bounded for all } t > 0.$$

If $|\varepsilon(t)|$ is not bounded for all t, we call the initial-value problem *unstable*. An initial-value problem is *absolutely stable* if

$$\lim_{t \to \infty} |\varepsilon(t)| = 0.$$

Stability of a linear initial-value problem
Consider the linear initial-value problem

$$y' = \lambda y + g \text{ with } y(0) = y_0. \tag{1.1}$$

It is easily seen that ε satisfies the *test equation* :

$$\varepsilon' = \lambda \varepsilon \text{ with } \varepsilon(0) = \varepsilon_0. \tag{1.2}$$

The solution ε of (1.2) is given by $\varepsilon(t) = \varepsilon_0 e^{\lambda t}$. This implies that a linear initial-value problem is stable if and only if $\lambda \leq 0$.

Stability of a one-step difference method
Consider two numerical solutions of (1.1): w_j with initial condition $w_0 = y_0$ and v_j with initial condition $v_0 = y_0 + \varepsilon_0$. The difference method is stable if

$$|\varepsilon_j| \text{ is bounded for all } j$$

and absolutely stable if

$$\lim_{j\to\infty} |\varepsilon_j| = 0 \, .$$

It appears that $\varepsilon_j = v_j - w_j$ is the numerical solution of the test equation. It is easy to see that every one-step method applied to the test equation gives

$$\varepsilon_{j+1} = Q(h\lambda)\varepsilon_j, \tag{1.3}$$

where $Q(h\lambda)$, the *amplification factor*, depends on the numerical method: for Forward Euler $Q(h) = 1 + h\lambda$ and for the Modified Euler method $Q(h) = 1 + h\lambda + \frac{1}{2}(h\lambda)^2$. By induction it follows that $\varepsilon_j = [Q(h\lambda)]^j \varepsilon_0$. So a numerical method is stable if and only if

$$|Q(h\lambda)| \leq 1 \, . \tag{1.4}$$

For the Forward Euler method we have $Q(h) = 1 + h\lambda$. Suppose that $\lambda \leq 0$ then the differential equation is stable. Inequality (1.4) can be written as:

$$-1 \leq 1 + h\lambda \leq 1 \, ,$$

which is equivalent to

$$-2 \leq h\lambda \leq 0 \, .$$

Since $h > 0$ and $\lambda \leq 0$ it follows that $h \leq \frac{2}{|\lambda|}$.

For Backward Euler the amplification factor is given by $Q(h) = \frac{1}{1-h\lambda}$. So for stability h should be such that:

$$-1 \leq \frac{1}{1 - h\lambda} \leq 1 \, .$$

It is easy to see that these inequalities hold for all $h \geq 0$ because $\lambda \leq 0$. This implies that the Backward Euler method is always stable as long as $\lambda \leq 0$.

Example (linear problem)
Consider the initial-value problem:

$$\begin{aligned} y' &= -10y \, , \qquad t \in [0,1] \, , \\ y(0) &= 1 \, . \end{aligned}$$

The exact solution is given by $y(t) = e^{-10t}$. From the theory it follows that Forward Euler is stable if $h \leq 0.2$. In Figure 1.1 the perturbations are plotted for the step sizes $h = \frac{1}{3}, \frac{1}{6}$ and $\frac{1}{12}$ with $\varepsilon_0 = 10^{-4}$. The method is indeed unstable if $h = \frac{1}{3}$ because $|\varepsilon_j| > |\varepsilon_{j-1}|$. For the other values of h we see that the perturbations decrease.

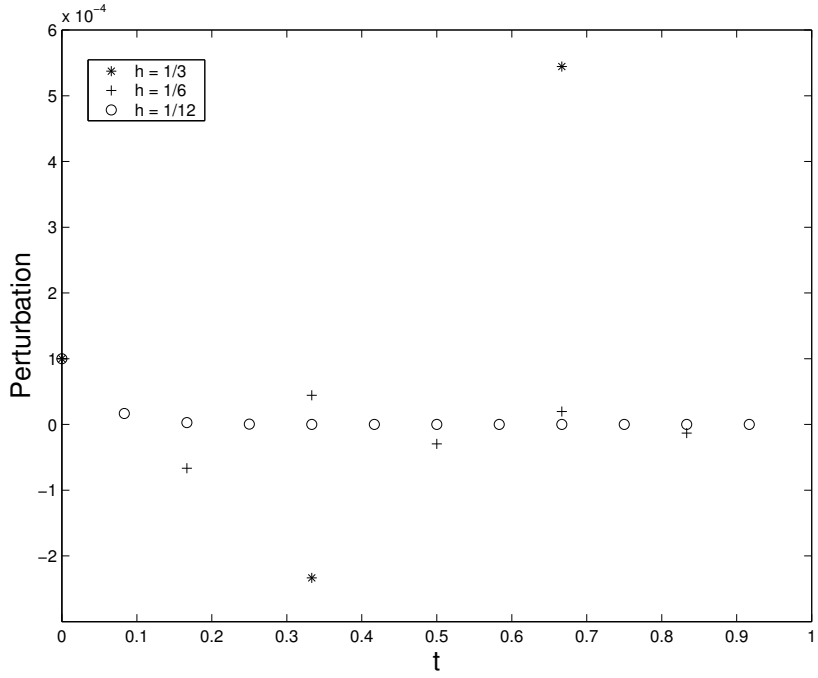

Figure 1.1: Forward Euler applied to $y^{'} = -10y$

Stability of a nonlinear initial-value problem
For a general initial-value problem we investigate the stability properties for the linearized problem. So the initial-value problem

$$y^{'} = f(t, y) \quad \text{and} \quad y(0) = y_0 \ , \tag{1.5}$$

is linearized by using the linear approximation around (t_0, y_0):

$$f(t, y) \approx f(t_0, y_0) + (y - y_0)\frac{\partial f}{\partial y}(t_0, y_0) + (t - t_0)\frac{\partial f}{\partial t}(t_0, y_0) \ .$$

If (t, y) is close to (t_0, y_0) equation (1.5) can be approximated by

$$y^{'} = f(t_0, y_0) + (y - y_0)\frac{\partial f}{\partial y}(t_0, y_0) + (t - t_0)\frac{\partial f}{\partial t}(t_0, y_0) \ .$$

Comparison with the linear equation shows that in this case $\lambda = \frac{\partial f}{\partial y}(t_0, y_0)$. Note that the stability condition depends on the values of y_0 and t_0.

Example (nonlinear problem)
A simple model of a transient channel flow is given by

$$y' = -ay^2 + p, \ \ y(0) = 0.$$

In this example $f(t, y) = -ay^2 + p$. After linearization one obtains

$$\lambda = \frac{\partial f}{\partial t}(t_0, y_0) = -2ay_0.$$

This implies that the initial-value problem is stable for all positive a and y_0. The Forward Euler method is stable if

$$h \leq \frac{1}{ay_0}.$$

Note that the bound on the step size decreases if y_0 increases.

Theorem 1.1.1 *If the numerical method is stable and consistent then the numerical solution converges to the exact solution for $h \to 0$. Furthermore the global error $y_i - w_i$ and the local truncation error $\tau_i(h)$ have the same rate of convergence.*

Proof
We only prove the theorem for the test equation $y' = \lambda y$. The following recurrence holds for the global error $y_j - w_j$:

$$\begin{aligned} y_j - w_j &= y_j - Q(h\lambda)w_{j-1} \\ &= y_j - Q(h\lambda)y_{j-1} + Q(h\lambda)(y_{j-1} - w_{j-1}) \\ &= h\tau_j + Q(h\lambda)(y_{j-1} - w_{j-1}) \, . \end{aligned} \tag{1.6}$$

Repeating this argument one obtains

$$y_j - w_j = \sum_{l=0}^{j-1} [Q(h\lambda)]^l h\tau_{j-l} \, .$$

From the stability follows

$$|y_j - w_j| \leq \sum_{l=0}^{j-1} |Q(h\lambda)|^l h|\tau_{j-l}| \leq \sum_{l=0}^{j-1} h|\tau_{j-l}| \leq \max_{1 \leq l \leq j-1} |\tau_l| \, ,$$

where we used that $jh \leq 1$. This implies that the rate of convergence of the global error is identical to that of the local truncation error. Furthermore the global error goes to zero if $h \to 0$ because the method is consistent. ⊠

1.2 Stability of a system of first order differential equations

First we consider the stability of a system of first order differential equations. Thereafter the stability of a numerical method applied to a system is investigated.

Stability of a linear system of first order differential equations
An m-th order linear system of first order differential equations has the form

$$\begin{aligned} \frac{dy_1}{dt} &= a_{11}y_1 + \ldots + a_{1m}y_m + g_1 \, , \\ \vdots \quad &\vdots \qquad\qquad\qquad \vdots \\ \frac{dy_m}{dt} &= a_{m1}y_1 + \ldots + a_{mm}y_m + g_m \, , \end{aligned} \tag{1.7}$$

where a_{ij} are real numbers and $g_1, \ldots, g_m$ are real valued functions of t. This system is also notated as

$$\boldsymbol{y}' = A\boldsymbol{y} + \boldsymbol{g} \text{ with } \boldsymbol{y}(0) = \boldsymbol{y}_0. \tag{1.8}$$

Suppose there are two solutions: $\boldsymbol{y}$ with initial condition $\boldsymbol{y}(0) = \boldsymbol{y}_0$ and $\tilde{\boldsymbol{y}}$ with the perturbed condition $\tilde{\boldsymbol{y}}(0) = \boldsymbol{y}_0 + \boldsymbol{\varepsilon}_0$. The vector function $\boldsymbol{\varepsilon} = \tilde{\boldsymbol{y}} - \boldsymbol{y}$ satisfies:

$$\boldsymbol{\varepsilon}' = A\boldsymbol{\varepsilon} \text{ with } \boldsymbol{\varepsilon}(0) = \boldsymbol{\varepsilon}_0, \tag{1.9}$$

which is a system of coupled linear differential equations. Hence the equations in the system must be solved simultaneously. In contrast, if each equation involves only a single variable then each equation can be solved independently of all the others which is much easier. This observation suggests to transform the system of equations into an equivalent uncoupled system. We assume that the matrix A is diagonalizable. This means that there is a nonsingular matrix S such that

$$AS = S\Lambda,$$

where $\Lambda = diag(\lambda_1 \dots \lambda_m)$ contains the eigenvalues of A and the columns of S are the corresponding eigenvectors of A. If A has a complex eigenpair $(\lambda, \boldsymbol{v})$ then the complex conjugate $(\bar{\lambda}, \bar{\boldsymbol{v}})$ is also an eigenpair of A.

Define a new dependent variable $\boldsymbol{\eta}$ by the relation $\boldsymbol{\varepsilon} = S\boldsymbol{\eta}$. Using this, expression (1.9) can be written as

$$\boldsymbol{\eta}' = \Lambda\boldsymbol{\eta} \text{ with } \boldsymbol{\eta}(0) = \boldsymbol{\eta}_0. \tag{1.10}$$

System (1.10) is uncoupled and easy to solve:

$$\boldsymbol{\eta}(t) = e^{\Lambda t}\boldsymbol{\eta}_0,$$

where $e^{\Lambda t}$ is a diagonal matrix with diagonal elements $e^{\lambda_1 t}, \dots e^{\lambda_m t}$. So finally the solution of (1.9) is given by

$$\boldsymbol{\varepsilon} = Se^{\Lambda t}S^{-1}\boldsymbol{\varepsilon}_0.$$

In the case that all eigenvalues are real (1.9) is a stable system if $\lambda_k \leq 0$ for $k = 1, \dots, m$. If λ_k is complex-valued the solution $\boldsymbol{\eta}_k$ can be written as an oscillating real function with amplitude equal to $e^{Re\lambda_k t}$. So in the general case ($\lambda_k \in \mathbb{C}$) system (1.9) is stable if $Re\lambda_k \leq 0$ for $k = 1, \dots, m$.

Stability of a numerical method

We consider the numerical solution of the system of linear differential equations

$$\boldsymbol{y}' = A\boldsymbol{y} \text{ with } \boldsymbol{y}(0) = \boldsymbol{y}_0. \tag{1.11}$$

The numerical solution $\boldsymbol{u}$ can be written as:

$$\boldsymbol{u}_{j+1} = G(hA)\boldsymbol{u}_j \text{ with } \boldsymbol{u}_0 = \boldsymbol{y}_0.$$

The $m \times m$ matrix $G(hA)$ is known as the *amplification matrix*. The following expressions are easy to verify.

Forward Euler	$G(hA) = I + hA$,
Backward Euler	$G(hA) = (I - hA)^{-1}$,
Implicit Trapezoidal	$G(hA) = (I - \frac{1}{2}hA)^{-1}(I + \frac{1}{2}hA)$,
Modified Euler	$G(hA) = I + hA + \frac{1}{2}h^2A^2$.

Theorem 1.2.1 *If A is a diagonalizable $m \times m$ matrix and S (matrix of eigenvectors) is a nonsingular matrix such that*

$$S^{-1}AS = diag(\lambda_1, \ldots, \lambda_m),$$

the amplification matrix $G(hA)$ of a numerical method has the following properties: $G(hA)$ is diagonalizable and $S^{-1}G(hA)S = M$, where $M = diag(Q(h\lambda_1), \ldots, Q(h\lambda_m))$ and $Q(h\lambda)$ is the amplification factor of the numerical method.

Proof
Using the related definitions the results easily follows. ⊠

To investigate the stability of a numerical method applied to (1.8) we consider the numerical solution of

$$\boldsymbol{\varepsilon}' = A\boldsymbol{\varepsilon} \text{ with } \boldsymbol{\varepsilon}(0) = \boldsymbol{\varepsilon}_0, \tag{1.12}$$

Using the amplification matrix it follows that $\boldsymbol{\varepsilon}_{j+1} = G(hA)\boldsymbol{\varepsilon}_j$. The numerical method is only stable if $\|\boldsymbol{\varepsilon}_j\|$ is bounded for all j. Using the substitution $\boldsymbol{\varepsilon}_j = S\boldsymbol{\eta}_j$ we obtain

$$\boldsymbol{\eta}_{j+1} = M\boldsymbol{\eta}_j.$$

So it appears that the numerical method is stable if the inequalities

$$|Q(h\lambda_k)| \leq 1 \text{ hold for all } 1 \leq k \leq m.$$

Region of stability
In the case that A has complex eigenvalues we define the *region of stability* R for a numerical method:

$$R = \{z \in \mathbb{C} \mid |Q(z)| \leq 1\}.$$

So R is the set of all points z in the complex plane such that the modulus of $Q(z)$ is less than or equal to 1. This implies that a numerical method is stable when $h\lambda_i \in R$, for $i = 1, \ldots, m$.

The regions of stability of some explicit methods are given in Figure 1.2 and for implicit methods in Figure 1.3. A region of stability can be used in the following way. Suppose λ_k is an eigenvalue of A. Then the step size h should be chosen such that $h\lambda_k \in R$ for all k. Note that the regions of stability for the explicit methods given in Figure 1.2 do not include the imaginary axis.This implies that these methods cannot be used for systems with an imaginary eigenvalue ($\lambda_k \in \mathbb{I}$ for some k) which implies that the solution contains an undamped oscillation. In Figure 1.4 we give the region R for the fourth order Runge Kutta method. Since part of the imaginary axis is contained in R the explicit Runge Kutta method is conditionally stable for undamped oscillations.

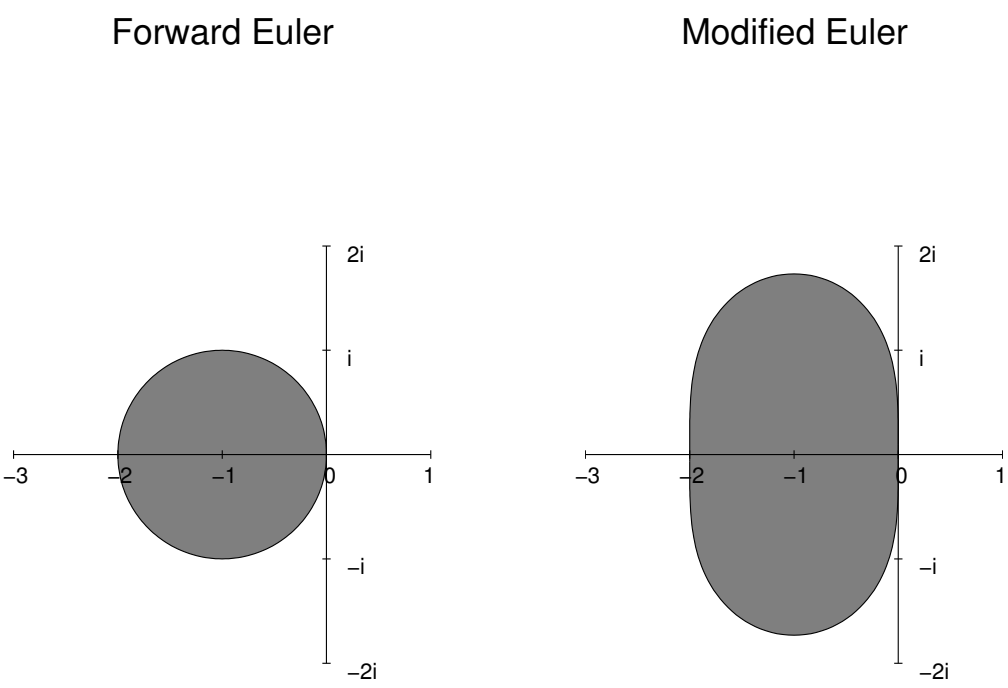

Figure 1.2: Regions of stability of explicit methods

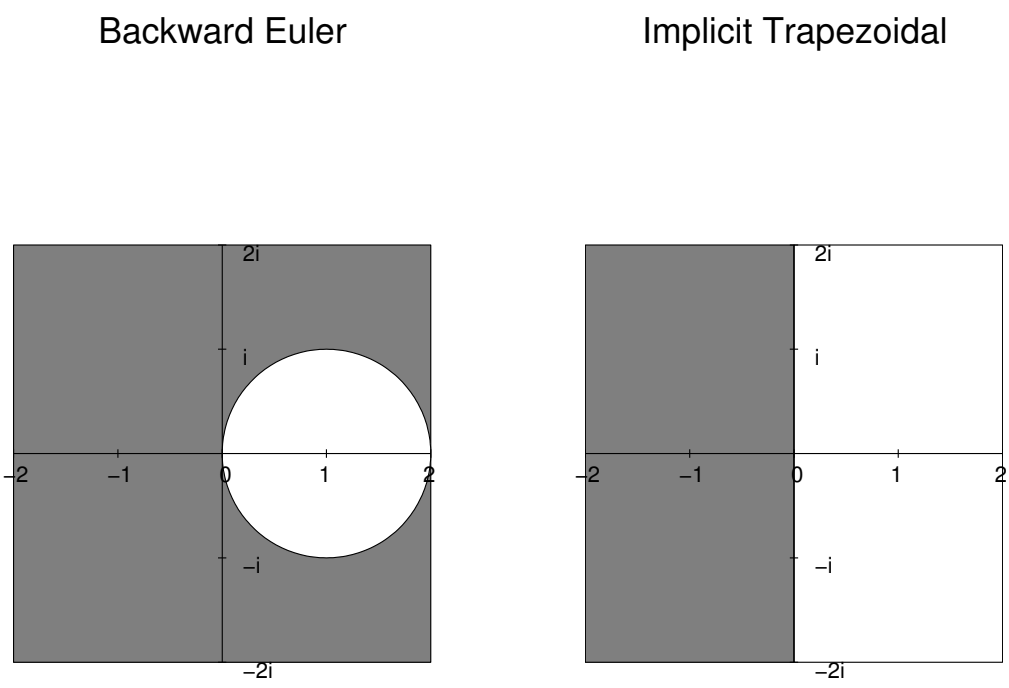

Figure 1.3: Regions of stability of implicit methods

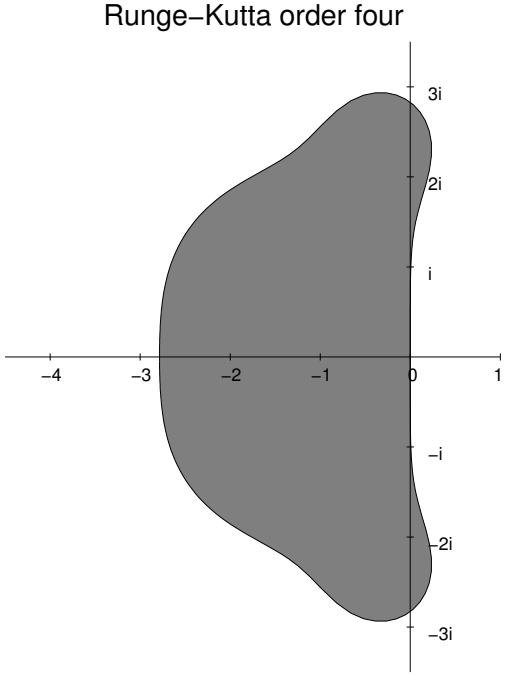

Figure 1.4: Regions of stability of the Runge Kutta method

Stability of a nonlinear system
A nonlinear system

$$\begin{aligned} \frac{dy_1}{dt} &= f_1(t, y_1, \ldots, y_m) \,, \\ \vdots \quad &\vdots \qquad\qquad \vdots \\ \frac{dy_m}{dt} &= f_m(t, y_1, \ldots, y_m) \,, \end{aligned} \tag{1.13}$$

has locally the same properties as a linear system (1.8), where the matrix A is replaced by the Jacobian matrix

$$\begin{pmatrix} \frac{\partial f_1}{\partial y_1} & \cdots & \frac{\partial f_1}{\partial y_m} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial y_1} & \cdots & \frac{\partial f_m}{\partial y_m} \end{pmatrix} .$$

Example
We consider the stability of the undamped oscillating pendulum. The angle ψ satisfies the nonlinear differential equation:

$$\psi'' + \sin\psi = 0 \,, \quad \psi(0) = \psi_0 \quad \text{and} \quad \psi'(0) = 0 \,.$$

This equation can be transformed to a system of first order equations:

$$\begin{aligned} y_1' &= \quad y_2 \quad = f_1(t, y_1, y_2) \,, \\ y_2' &= -\sin y_1 = f_2(t, y_1, y_2) \,. \end{aligned}$$

The Jacobian matrix is given by

$$\begin{bmatrix} 0 & 1 \\ -\cos y_1 & 0 \end{bmatrix} .$$

Assuming that $-\frac{\pi}{2} < \psi_0 < \frac{\pi}{2}$ the eigenvalues are

$$\lambda_{1,2} = \pm i\sqrt{\cos y_1} .$$

The differential equation is stable since $Re\lambda_{1,2} = 0$. Note that Forward Euler and Modified Euler are unstable for all step sizes h. Backward Euler, Implicit Trapezoidal and the fourth order Runge Kutta method are stable. It appears that the solution obtained with the Backward Euler method shows unphysical damping.

1.3 Stiff systems of differential equations

In this section we only consider systems of the form:

$$\boldsymbol{y}' = A\boldsymbol{y} , \tag{1.14}$$

where the eigenvalues of A are real and negative. The eigenvalues are ordered in the following way

$$\lambda_m \leq \lambda_{m-1} .... \leq \lambda_1 < 0 .$$

Stiff systems of differential equations
System (1.14) is called *stiff* if

$$\frac{|\lambda_m|}{|\lambda_1|} \gg 1 .$$

The solution of (1.14) is given by $\boldsymbol{y} = \sum_{k=1}^{m} c_k e^{\lambda_k t}\boldsymbol{v}_k$, where $\boldsymbol{v}_k$ is the eigenvector corresponding to λ_k. Note that the components of $\boldsymbol{y}$ with respect to the basis $\{\boldsymbol{v}_1, \ldots \boldsymbol{v}_m\}$, corresponding to strong negative eigenvalues rapidly decay to zero. These components are only important for small times. In many applications one is not interested in the transient but in the long time behavior of the solution.

Explicit methods
Explicit methods are conditionally stable, so the step size h should be chosen sufficiently small. We consider the Forward Euler method. This method is only stable if

$$h \leq \frac{2}{|\lambda_m|} .$$

However the accuracy for the long time behavior is strongly influenced by the eigenvalues close to zero, which implies that h can be chosen relatively large. So for stiff systems of differential equations the stability requirements for explicit methods leads to unpractical small values of the step size h.

Implicit methods
Most implicit methods are unconditionally stable if $Re\lambda_k \leq 0$. Using such a method the step size h can be controlled by the required accuracy. This implies that for moderate values of h there may be a large error for small times, whereas the accuracy is high at medium and large

times, due to the damping of the initial errors.

We compare the Backward Euler and the Implicit Trapezoidal method with respect to the transient components. The amplification factors of both methods are given in Figure 1.5. There is an important difference in the value of the amplification factors for $h\lambda \to -\infty$.

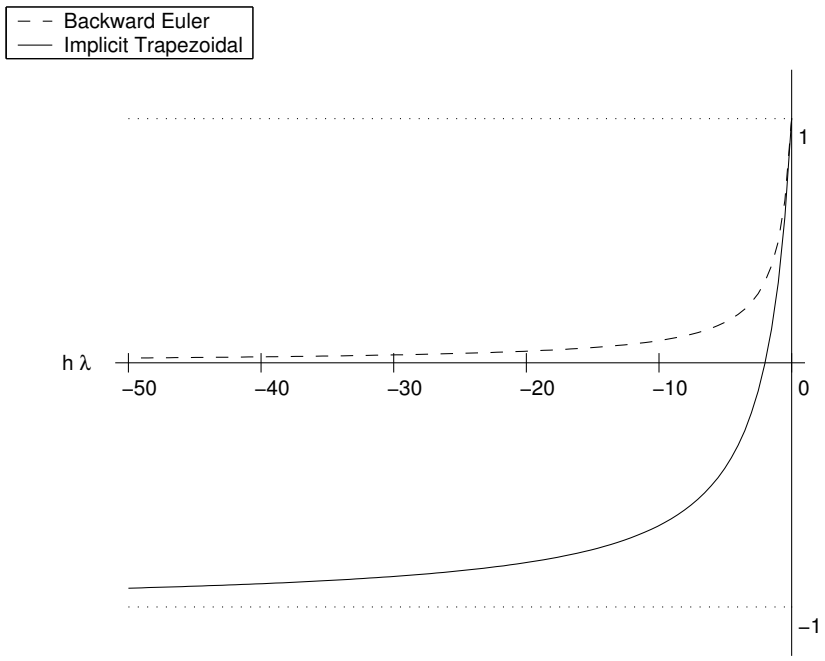

Figure 1.5: Amplification factors of the Backward Euler and the Implicit Trapezoidal method

Super-stable
A numerical method is called super-stable if

$$\lim_{h\lambda \to -\infty} |Q(h\lambda)| < 1 \ .$$

From Figure 1.5 we conclude that Backward Euler is super-stable, whereas the Implicit Trapezoidal method is not super-stable because

$$\lim_{h\lambda \to -\infty} |Q(h\lambda)| = 1 \ .$$

This means that perturbations in the initial conditions of the transient components are very slowly damped if the Implicit Trapezoidal method is used.

Chapter 2

Supplement of Chapter 11

In Chapter 11.3 of the book *Numerical Analysis of Burden and Faires*, the finite difference method for linear boundary value problems is presented. In this supplement additional material concerning conditioning, convergence and other boundary conditions are given.

2.1 Summary of relevant linear algebra subjects

In this section we summarize a number of linear algebra subjects relevant for the solution of boundary value problems by the finite difference method. Most of these subjects are given in the book of *Burden and Faires*, but they are somewhat scattered in the text.

Condition of a linear system
The scaled Eulerian norm of the vector $\boldsymbol{x}$ is defined as (compare Definition 7.2, p. 419):

$$\|\boldsymbol{x}\| = \sqrt{\frac{1}{N}\sum_{j=1}^{N} x_j^2}\ .$$

The *natural*, or *induced*, matrix norm associated with the vectornorm $\|.\|$ is defined as (Theorem 7.9, p. 425):

$$\|A\| = \max_{\|\boldsymbol{x}\|=1} \|A\boldsymbol{x}\|.$$

The following inequality is often used (Corollary 7.10, p. 425):

$$\|A\boldsymbol{x}\| \leq \|A\|\ \|\boldsymbol{x}\|.$$

Suppose vector $\boldsymbol{x}$ is the solution of the linear system $A\boldsymbol{x} = \boldsymbol{b}$. Furthermore the right-hand side vector $\boldsymbol{b}$ is perturbed with the vector $\Delta\boldsymbol{b}$. As a result of this the solution vector also contains an error $\Delta\boldsymbol{x}$ since we are solving the following perturbed system:

$$A(\boldsymbol{x} + \Delta\boldsymbol{x}) = \boldsymbol{b} + \Delta\boldsymbol{b}\ .$$

The relative error satisfies the following inequality (Theorem 7.27, p. 455, 456):

$$\frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \leq \|A\|\ \|A^{-1}\|\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}\ .$$

The condition number of the nonsingular matrix A relative to the norm $\|.\|$ is (Definition 7.28, p. 456)

$$K(A) = \|A\| \ \|A^{-1}\|.$$

For a nonsingular symmetric matrix the equalities:

$$\|A\| = \lambda_{\max} = \max_{1\le j\le N} |\lambda_j| \quad \text{and} \quad \|A^{-1}\| = \frac{1}{\lambda_{\min}} = \frac{1}{\min\limits_{1\le j\le N} |\lambda_j|} \ ,$$

imply that $K(A) = \frac{\lambda_{\max}}{\lambda_{\min}}$. In order to estimate the condition number it is important to know bounds for the largest and the smallest eigenvalue of A. In many application the Gerschgorin Circle Theorem can be used to estimate these eigenvalues (Theorem 9.13, p. 556).

Gerschgorin Circle Theorem
Let A be an $N \times N$ matrix and R_j denote the circle in the complex plane with center a_{jj} and radius $\sum\limits_{\substack{k=1\\k\neq j}}^{N} |a_{jk}|$; that is

$$R_j = \{z \in \mathbb{C} \ \big| \ \ |z - a_{jj}| \le \sum_{\substack{k=1\\k\neq j}}^{N} |a_{jk}|\},$$

where $\mathbb{C}$ denotes the complex plane. The eigenvalues of A are contained within $R = \cup_{j=1}^{N} R_j$.

2.2 Consistency, stability and convergence

In this section we prove that under certain conditions the difference between the numerical and exact solution goes to zero if the step size h goes to zero. To show this we shall give some definitions.

The second order boundary value problem is given by

$$-y'' + py' + qy = -r \ , \quad 0 < x < 1 \ ,$$

with boundary conditions

$$y(0) = \alpha \quad \text{and} \quad y(1) = \beta \ .$$

The interval $[0, 1]$ is divided into $N + 1$ equal subintervals whose endpoints are the mesh points $x_i = ih$, for $i = 0, ...N + 1$, where $h = \frac{1}{N+1}$.

Truncation error
The truncation error $\boldsymbol{\varepsilon}$ of the numerical scheme

$$F\boldsymbol{w} = \boldsymbol{f} \tag{2.1}$$

is defined as

$$\varepsilon_j = (F\boldsymbol{y} - \boldsymbol{f})_j \ , \quad j = 1, ..., N \ ,$$

where the components of $\boldsymbol{y}$, the exact solution, are given by $y_j = y(x_j)$. The matrix F and vector $\boldsymbol{f}$ are given by: $F = A/h^2$ and $\boldsymbol{f} = \boldsymbol{b}/h^2$ where A and $\boldsymbol{b}$ are defined in formula (11.19)

on page 661 of Burden and Faires.

As an example we consider the differential equation $-y'' + qy = -r$, which is discretized by:

$$\frac{-w_{j-1} + 2w_j - w_{j+1}}{h^2} + q_j w_j = -r_j.$$

The truncation error follows from the Taylor expansion:

$$\varepsilon_j = -y_j'' + q_j y_j + r_j + O(h^2) \; .$$

Combined with equation $-y_j'' + q_j y_j = -r_j$ one obtains:

$$\varepsilon_j = O(h^2). \tag{2.2}$$

The order of the truncation error is equal to 2 for this method.

Consistency
A finite difference method is called consistent if

$$\lim_{h \to 0} \|\boldsymbol{\varepsilon}\| = 0 \; .$$

From (2.2) it follows that $\|\boldsymbol{\varepsilon}\| = O(h^2)$ so this numerical scheme is consistent.

Stability
A finite difference scheme is stable, if there is a constant M independent of h, such that

$$\|F^{-1}\| \leq M \; , \quad \text{for} \quad h \to 0 \; .$$

If a method is stable then the resulting system has a unique solution. The matrix F for our method is symmetric. This implies that the eigenvalues are real numbers and

$$\|F^{-1}\| = \frac{1}{\lambda_{\min}} \; .$$

If the function q satisfies the following inequalities $0 < q_{\min} \leq q(x) \leq q_{\max}$ for $0 \leq x \leq 1$, then Gerschgorin's theorem implies

$$q_{\min} \leq \lambda_j \leq q_{\max} + \frac{4}{h^2} \quad \text{for} \quad j = 1, ..., N \; .$$

From this it follows that $\|F^{-1}\| \leq \frac{1}{q_{\min}}$ so the numerical scheme is stable.

If $q \equiv 0$ then the Gerschgorin's theorem does not give usefull information for the smallest eigenvalue. However for this special case the eigenvalues of F are known:

$$\lambda_j = (2 - 2\cos(N + 1 - j)h\pi)/h^2 \; , \quad j = 1, ..., N \; .$$

From this it follows that

$$\lambda_{\min} = (2 - 2\cos h\pi)/h^2 = (4\sin^2 \frac{h\pi}{2})/h^2 \approx \pi^2 \; .$$

So also for the case $q \equiv 0$ the numerical scheme is stable.

Convergence
Suppose $\boldsymbol{w}$ is the solution of (2.1). A numerical scheme is called convergent if the global error $\boldsymbol{y} - \boldsymbol{w}$ has the property

$$\lim_{h \to 0} \|\boldsymbol{y} - \boldsymbol{w}\| = 0 \, .$$

In the next theorem we give a relation between the notions: consistency, stability, and convergence.

Theorem 2.2.1 *If a numerical scheme is stable and consistent, then it is convergent.*

Proof
The global error $\boldsymbol{y} - \boldsymbol{w}$ is a solution of the system

$$F(\boldsymbol{y} - \boldsymbol{w}) = F\boldsymbol{y} - F\boldsymbol{w} = \boldsymbol{f} + \boldsymbol{\varepsilon} - \boldsymbol{f} = \boldsymbol{\varepsilon} \, ,$$

which implies: $\boldsymbol{y} - \boldsymbol{w} = F^{-1}\boldsymbol{\varepsilon}$. Taking the norm of both sides yields

$$\|\boldsymbol{y} - \boldsymbol{w}\| \leq \|F^{-1}\| \, \|\boldsymbol{\varepsilon}\| \, .$$

Using the stability and the consistency it follows that

$$\lim_{h \to 0} \|\boldsymbol{y} - \boldsymbol{w}\| = 0 \, .$$

⊠

Note that consistency alone is not sufficient to guarantee convergence.

Example (convergence)
Suppose that heat transfer in a rod is described by the boundary value problem:

$$\begin{aligned} -y'' &= 25e^{5x} \, , \qquad 0 < x < 1 \, , \\ y(0) &= y(1) = 0 \, . \end{aligned}$$

Due to the heat source ($25e^{5x}$) we expect that the temperature y is positive. In Figure 2.1 the graphs of y and the numerical solutions obtained with the step sizes: $h = \frac{1}{4}, \frac{1}{8}$ and $\frac{1}{16}$ are given. Note that there is a rapid convergence from the numerical solutions to the exact solution. The accuracy using $h = \frac{1}{16}$ is sufficient from engineering point of view.

As a variant of this application we consider

$$\begin{aligned} -y'' + 9y &= 25e^{5x} \, , \qquad 0 < x < 1 \, , \\ y(0) &= y(1) = 0 \, . \end{aligned}$$

The additional term $9y$ describes heat loss due to the temperature difference of the rod and its surroundings. The exact and numerical solutions are given in Figure 2.2. Due to the extra heat loss the temperature is indeed less than in the first example (Figure 2.1). On the other hand the convergence is similar for both problems.

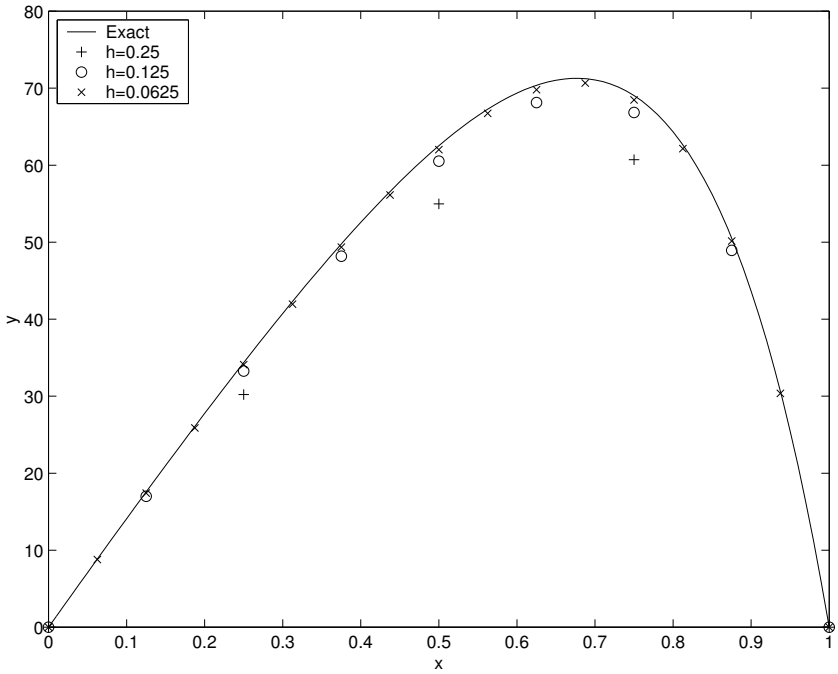

Figure 2.1: The exact and numerical solutions of the stationary heat problem

2.3 The condition number of the discretization matrix

For $q \equiv 0$ we have seen that

$$\lambda_{\min} \simeq \pi^2 \quad \text{and} \quad \lambda_{\max} \simeq \frac{4}{h^2} \, .$$

This means that $K(F) = \frac{4}{\pi^2 h^2}$. If h goes to zero the condition number of F goes to infinity. So it seems that we are unable to solve this problem numerically in a stable way. In practice the error bounds appears to be too pessimistic. For a better estimate we use the following analysis. The numerical solution $\boldsymbol{w}$ satisfies

$$F\boldsymbol{w} = \boldsymbol{f} \, .$$

With a perturbation $\Delta\boldsymbol{f}$ of $\boldsymbol{f}$ the computed solution satisfies

$$F(\boldsymbol{w} + \Delta\boldsymbol{w}) = \boldsymbol{f} + \Delta\boldsymbol{f} \, ,$$

so $\Delta\boldsymbol{w}$ is a solution of $F\Delta\boldsymbol{w} = \Delta\boldsymbol{f}$. From this we deduce

$$\|\Delta\boldsymbol{w}\| = \|F^{-1}\Delta\boldsymbol{f}\| \leq \frac{1}{\lambda_{\min}} \|\Delta\boldsymbol{f}\| \, ,$$

and for the relative error we obtain the bound

$$\frac{\|\Delta\boldsymbol{w}\|}{\|\boldsymbol{w}\|} \leq \frac{1}{\lambda_{\min}} \frac{\|\boldsymbol{f}\|}{\|\boldsymbol{w}\|} \cdot \frac{\|\Delta\boldsymbol{f}\|}{\|\boldsymbol{f}\|} \, .$$

In this bound the condition number $K(F)$ is replaced by the "effective" condition number $\frac{1}{\lambda_{\min}} \frac{\|\boldsymbol{f}\|}{\|\boldsymbol{w}\|}$. Since $\lambda_{\min} \simeq \pi^2$ and in many applications $\frac{\|\boldsymbol{f}\|}{\|\boldsymbol{w}\|}$ is bounded if h goes to zero we have that the "effective" condition number is bounded independent of h.

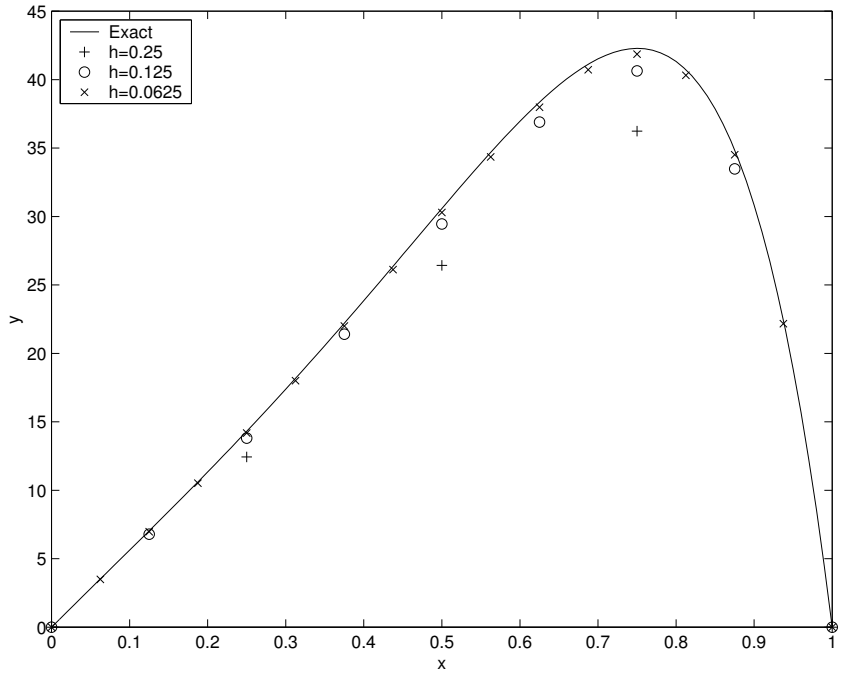

Figure 2.2: The exact and numerical solutions of the stationary heat problem with additional heat loss

2.4 Neumann boundary condition

In all our examples we have used Dirichlet boundary conditions. What about other boundary conditions? We only consider the Neumann boundary condition

$$\frac{dy}{dx}(1) = 0 \; . \tag{2.3}$$

Discretization
We introduce a virtual point $x_{N+2} = (N+2)h = 1 + h$. For $j = N + 1$ the equation

$$\frac{-w_N + 2w_{N+1} - w_{N+2}}{h^2} + q_{N+1}w_{N+1} = -r_{N+1} \tag{2.4}$$

is valid. Discretization of (2.3) gives

$$\frac{w_{N+2} - w_N}{2h} = 0 \; ,$$

so $w_{N+2} = w_N$. Substituting this into (2.4) and division by 2 leads to

$$\frac{-w_N + w_{N+1}}{h^2} + \frac{1}{2}q_{N+1}w_{N+1} = -\frac{1}{2}r_{N+1} \; . \tag{2.5}$$

The motivation to divide by 2 is to obtain a symmetric matrix F. Note that now the vector $\boldsymbol{w}$ has length $N + 1$.

Convergence
Using a Taylor expansion it is easy to see that

$$y_{N+2} = y_N + O(h^3) \; ,$$

so the truncation error in (2.5) is equal to $O(h)$ and in all the other equations it is equal to $O(h^2)$. In the Dirichlet case the truncation error is $O(h^2)$ in all equations. Does this mean that the global error for the Neumann boundary condition is also $O(h)$? The answer is no, it is possible to show that the global error remains $O(h^2)$.

To show that the global error is $O(h^2)$ we consider the case $q \equiv 0$. It appears that the resulting numerical scheme is stable, so $\|F^{-1}\|$ is bounded independent of h. The truncation error is splitted into two parts, one proportional to h^2 and the other proportional to h:

$$F\boldsymbol{y} = \boldsymbol{f} + h^2\boldsymbol{u} + h\boldsymbol{v} \ , \quad \text{with} \quad \boldsymbol{v} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ v_{N+1} \end{pmatrix} \ .$$

The global error $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{w}$ is also splitted into two parts: $\boldsymbol{e} = \boldsymbol{e}^{(1)} + \boldsymbol{e}^{(2)}$, where

$$F\boldsymbol{e}^{(1)} = h^2\boldsymbol{u} \quad \text{and} \quad F\boldsymbol{e}^{(2)} = h\boldsymbol{v} \ .$$

Due to stability we obtain $\|\boldsymbol{e}^{(1)}\| = O(h^2)$, so it remains to bound $\|\boldsymbol{e}^{(2)}\|$. It is possible to give an expression for $e_j^{(2)}$:

$$e_j^{(2)} = jh^3 v_{N+1} \ , \quad j = 1, ..., N+1 \ .$$

The inequality $jh \leq 1$ can be used to show that $\|\boldsymbol{e}^{(2)}\| = O(h^2)$. So the norm of the global error is $O(h^2)$.

2.5 A general boundary value problem

A general second order boundary value problem is given by

$$-(sy')' + py' + qy = -r \ , \quad a < x < b \ ,$$

with boundary conditions

$$y(a) = \alpha \quad \text{and} \quad y'(b) = \beta \ .$$

Discretization in point x_j yields

$$\frac{-s_{j+\frac{1}{2}}(w_{j+1} - w_j) + s_{j-\frac{1}{2}}(w_j - w_{j-1})}{h^2} + p_j \frac{w_{j+1} - w_{j-1}}{2h} + q_j w_j = -r_j \ , \quad j = 1, ..., N+1 \ . \tag{2.6}$$

Note that if $p_j \neq 0$ then the discretization matrix is nonsymmetric. The Dirichlet boundary condition at $x = a$ leads to $w_0 = \alpha$ which can be directly substituted into (2.6) for $j = 1$. Introducing a virtual point, the boundary condition at $x = b$ is discretized as follows:

$$\frac{1}{2}(s_{N+1\frac{1}{2}} \frac{w_{N+2} - w_{N+1}}{h} + s_{N+\frac{1}{2}} \frac{w_{N+1} - w_N}{h}) = s(b)\beta \ .$$

In this expression the left-hand side is an average of

$$s_{N+1\frac{1}{2}} \frac{dy}{dx}(b + \frac{1}{2}h) \quad \text{and} \quad s_{N-\frac{1}{2}} \frac{dy}{dx}(b - \frac{1}{2}h) \ .$$

This average is an approximation of $s_{N+1}\frac{dy}{dx}(b) = s(b)\beta$. We choose for this formulation in order to replace expression $s_{N+1\frac{1}{2}} \frac{w_{N+2} - w_{N+1}}{h}$ in equation (2.6) for $j = N+1$, by the expression $-s_{N+\frac{1}{2}} \frac{w_{N+1} - w_N}{h} + 2hs(b)\beta$.

2.6 The convection-diffusion equation

In the previous section we have discretized the first order derivative in the differential equation by a central difference quotient. In some applications this leads to unphysical wiggles in the numerical solution. To illustrate this we use the following boundary value problem:

$$\begin{aligned} -y'' + py' &= 0\,, \qquad 0 < x < 1\,, \\ y(0) = 1\,, \quad y(1) &= \alpha\,, \end{aligned}$$

with $p > 0$. This equation can be seen as a model for a heat transport problem, where there is transport by conduction $-y''$ (diffusion) and by convection py'. This equation is known as the convection-diffusion equation.

Central difference quotient
Application of central difference quotients to all the terms leads to the following linear system of equations:

$$\frac{1}{h^2}\begin{pmatrix} 2 & -1+\frac{ph}{2} & & \\ -1-\frac{ph}{2} & 2 & \ddots & \varnothing \\ \varnothing & \ddots & \ddots & \ddots \end{pmatrix}\begin{pmatrix} w_1 \\ \vdots \\ \vdots \\ w_N \end{pmatrix} = \frac{1}{h^2}\begin{pmatrix} 1+\frac{ph}{2} \\ 0 \\ \vdots \\ 0 \\ (1-\frac{ph}{2})\alpha \end{pmatrix}.$$

Note that the discretization matrix becomes lower triangular if $ph = 2$. This implies that the numerical solution is independent of the value of α. This is incorrect. It appears indeed that the numerical solution is only acceptable if h is chosen such that $ph < 2$.

Upwind difference quotient
In the case that the condition $ph < 2$ leads to a very small step size h, it is better to use an upwind difference quotient. This means that the term py'_j is approximated by

$$\begin{aligned} p\frac{w_j - w_{j-1}}{h} \qquad &\text{if} \quad v \geq 0 \quad \text{and} \\ p\frac{w_{j+1} - w_j}{h} \qquad &\text{if} \quad v < 0\,. \end{aligned}$$

Example (convection-diffusion)
To illustrate the difference between a central and upwind discretization, we consider the following boundary value problem:

$$\begin{aligned} -y'' + py' &= 1\,, \qquad 0 < x < 1\,, \\ y(0) = y(1) &= 0\,. \end{aligned}$$

The numerical solutions are computed for $h = 0.1$ and $p = 10, 20$ and 100 (see Figure Figuur 2.3, 2.4 and 2.5). For $p = 10$ the inequality $ph < 2$ holds and the results obtained with a central difference quotient are better than the ones obtained with an upwind difference quotient. For $p = 20$ we note that $ph = 2$ so the numerical solution obtained from the central difference quotient is independent of the boundary condition in $x = 1$. For $p = 100$ we observe that the solution obtained by upwind differences leads to an acceptable numerical solution, whereas that obtained by central differences is not acceptable due to severe wiggles. Furthermore the numerical solution with central differences has large errors.

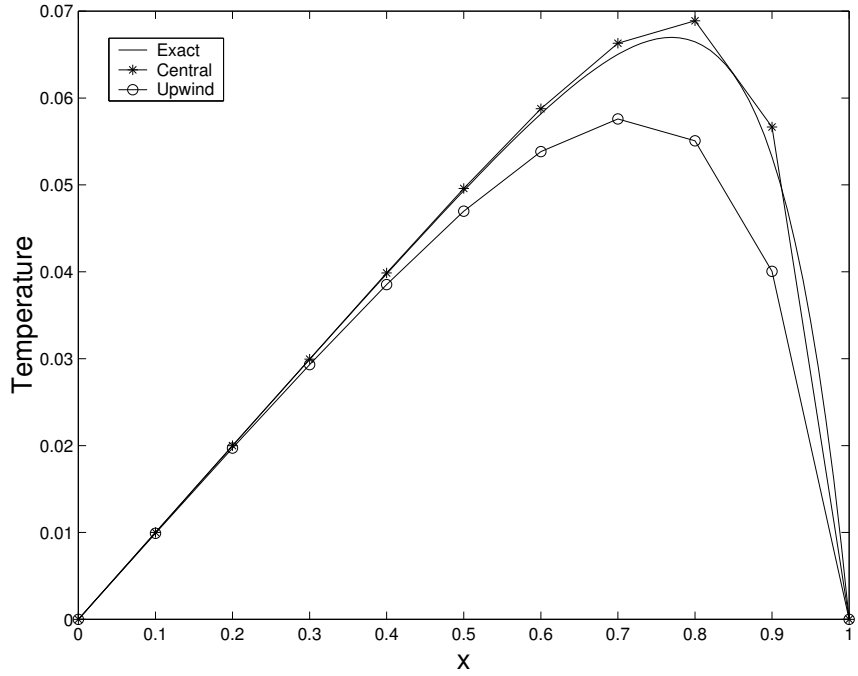

Figure 2.3: The solution of the convection-diffusion problem for $p = 10$

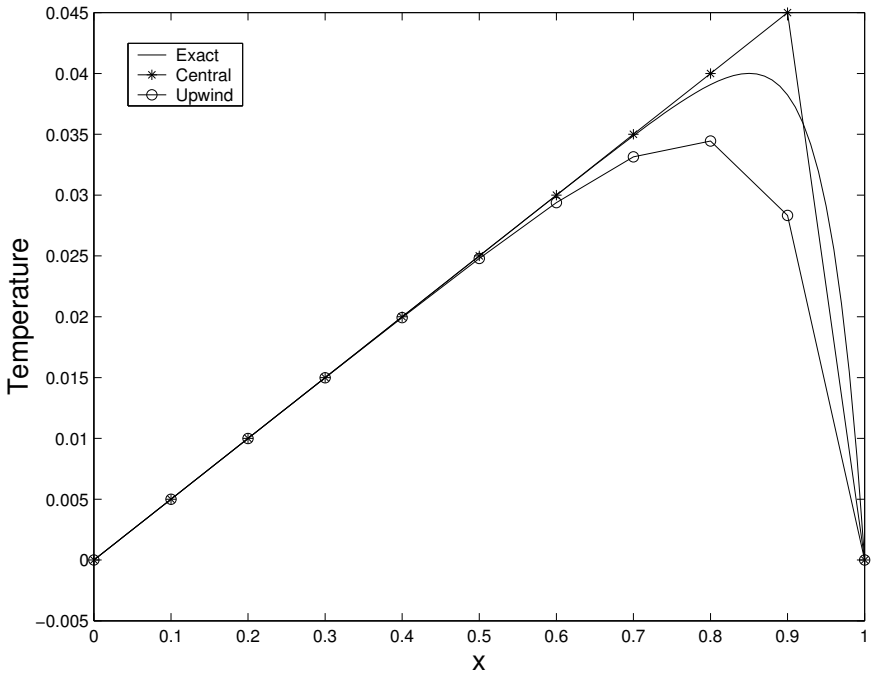

Figure 2.4: The solution of the convection-diffusion problem for $p = 20$

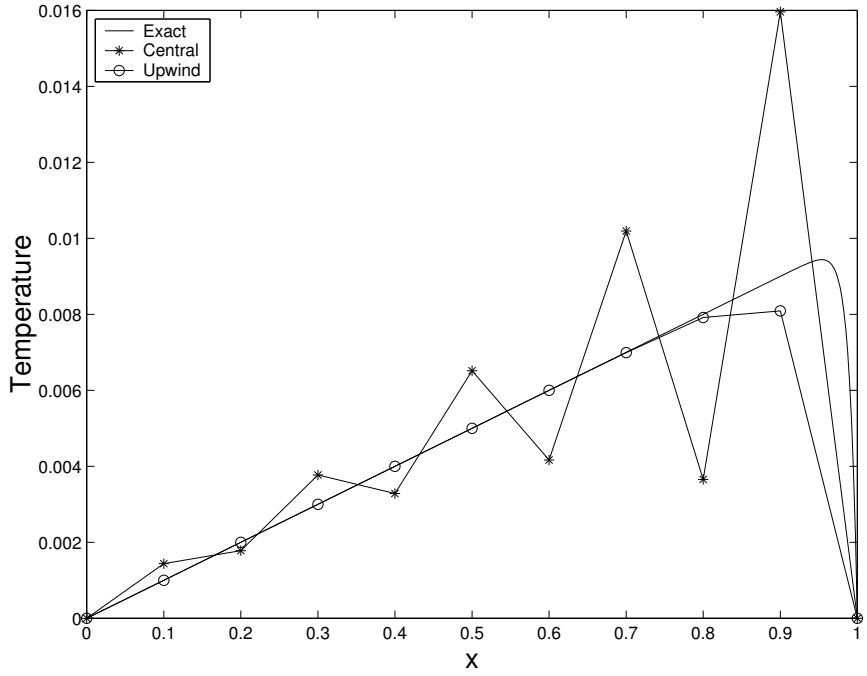

Figure 2.5: The solution of the convection-diffusion problem for $p = 100$

Index