# TMA4212 Numerical solution of partial differential equations with finite difference methods

Brynjulf Owren[1]

January 31, 2017

[1]Translated and amended by E. Celledoni

# Preface

The preparation of this note started in the winter of 2004. The note is a teaching aid for the first half of the course TMA4210 "Numerical solution of partial differential equations with difference methods". In the winter of 2006 the note was updated and several new sections were added to adapt it to the course TMA4212. I want to thank all the students who followed the courses during these semesters. They have been a source of inspiration for the writing and they helped me in the correction of many typos and mistakes in the earlier versions of this note.

# Contents


**1 Introduction** **1**

**2 Background material** **3**
2.1 Background on matrix theory . . . 3
2.1.1 Jordan form . . . 3
2.1.2 Symmetric matrices . . . 4
2.1.3 Positive definite matrices . . . 4
2.1.4 Gershgorin's theorem . . . 4
2.1.5 Vector and matrix norms . . . 5
2.1.6 Consistent and subordinate matrix norms . . . 6
2.1.7 Matrix norms and spectral radius. . . . 7
2.2 Difference formulae . . . 8
2.2.1 Taylor expansion . . . 8
2.2.2 Big $\mathcal{O}$-notation . . . 9
2.2.3 Difference approximations to the derivatives . . . 10
2.2.4 Difference operators and other operators . . . 11
2.2.5 Differential operator. . . . 13

**3 Boundary value problems** **15**
3.1 A simple case example . . . 15
3.1.1 2-norm stability for the case example . . . 18
3.1.2 Neumann boundary conditions . . . 18
3.2 Linear boundary value problems . . . 22
3.2.1 A self-adjoint case . . . 22
3.3 A nonlinear example . . . 23

**4 Discretization of the heat equation** **27**
4.1 On the derivation of the heat equation . . . 27
4.2 Numerical solution of the initial/boundary value problem . . . 28
4.2.1 Numerical approximation on a grid. . . . 28
4.2.2 Euler, Backward Euler and Crank–Nicolson . . . 29
4.2.3 Solution of the linear systems in Backward Euler's method and Crank–Nicolson . . . 33
4.2.4 Solution of linear systems in Matlab . . . 34
4.2.5 The $\theta$-method . . . 35
4.3 Semi-discretization . . . 36
4.3.1 Semi-discretization of the heat equation . . . 36
4.3.2 Semidiscretization principle in general . . . 37
4.3.3 General approach . . . 38

4.3.4 $u_t = Lu$ with different choices of $L$ . . . 39
4.4 Boundary conditions involving derivatives . . . 41
4.4.1 Different types of boundary conditions . . . 41
4.4.2 Discretization of the boundary conditions . . . 42
4.5 Nonlinear parabolic differential equations . . . 44

**5 Stability, consistency and convergence** **47**
5.1 Properties of the continuous problem . . . 47
5.2 Convergence of a numerical method . . . 48
5.3 Domain of dependence of a numerical method . . . 49
5.4 Proof of convergence for the Euler's method on the (I/BVP) with $r \leq \frac{1}{2}$ . . . 50
5.5 Stability on unbounded time interval ($F$-stability) . . . 51
5.6 Stability on $[0, T]$ when $h \to 0$, $k \to 0$ . . . 52
5.7 Stability and roundoff error . . . 57
5.8 Consistency and Lax' equivalence theorem . . . 58
5.9 von Neumann's stability criterion . . . 59

**6 Elliptic differential equations** **63**
6.1 Elliptic equation on the plane . . . 63
6.2 Difference methods derived using Taylor series . . . 64
6.2.1 Discretization of a self-adjoint equation . . . 66
6.3 Boundary conditions of Neumann and Robin type . . . 67
6.4 Grid-like net and variable step-size . . . 69
6.5 General rectangular net . . . 70
6.6 Discretization using Taylor expansion on a completely general net . . . 71
6.7 Difference formulae derived by integration . . . 72
6.8 Net based on triangles . . . 75
6.9 Difference equations . . . 75
6.10 Convergence of the methods for elliptic equations . . . 77
6.10.1 Convergence for the 5-point formula on a Dirichlet problem . . . 77
6.10.2 Some general comments on convergence . . . 78
6.11 A discussion on the solution of large linear systems . . . 79

**7 Hyperbolic equations** **81**
7.1 Examples . . . 81
7.2 Characteristics . . . 82
7.3 Explicit difference formulae for $u_t + au_x = 0$ . . . 84
7.4 Stability . . . 85
7.5 Implicit methods for $u_t + au_x = 0$ . . . 88
7.6 Hyperbolic systems of first order equations . . . 90
7.7 Dissipation and dispersion . . . 96

# Chapter 1

# Introduction

The numerical approximation of partial differential equations is an important component in the simulation of natural processes. Examples where simulation techniques are useful are chemical processes, fluid mechanics, structural dynamics, quantum physical processes, electromagnetism, finance, etc.

When we talk about partial differential equations (PDEs) we mean equations where the solution is a function (or a vector of functions) of at least two variables, which are called independent variables. The equation describes a relation involving the solution and its partial derivatives. But the specification of a mathematical model in applications involves much more than just this relation. First of all the model is associated to a *geometry*. This means that we specify a domain in the space of the independent variables where the differential equation should be satisfied.

This domain can be finite or infinite. In two dimensions such domain can be a subset of the plane, but also the surface of a cylinder or a sphere.

Typically a PDE in itself has infinitely many solutions. The specification of a PDE on a domain must be supplemented with the specification of boundary conditions. There are many different types of boundary conditions, but in general they specify the solution or its derivatives on the boundary of the domain. If one of the independent variables is physical time, then the boundary conditions at the initial time are called initial conditions, and this is just a particular type of boundary conditions.

Most PDEs are such that it is not possible to write their solution in a closed form. To understand the behavior of the solutions, it is therefore necessary to use approximation methods and computer simulations. There are several numerical methods which can be used for this purpose, among them we will focus mostly on difference methods. Other techniques are spectral methods and finite element methods. These can be studied in other courses (TMA4220).

Partial differential equations can be linear or nonlinear. Since this is a introductory course we will mostly consider the linear case. Linear PDEs are divided in three subclasses, parabolic, elliptic and hyperbolic differential equations. The PhD course MA8103 considers nonlinear PDEs with particular focus on hyperbolic PDEs.

A typical prototype of parabolic partial differential equation is the heat equation, this is the main subject of chapter 3. In this chapter we work most of the time with 2 independent variables, time and one space dimension. For this reason the discussion about the geometry of the problems is somewhat limited in this chapter.

In chapter 4 we consider general topics regarding difference approximation of PDEs which are of interest for all the three types of PDEs. This chapter is about convergence of the numerical approximations to the exact solution of the PDEs. The concepts of stability

and consistency will be also introduced as important tools for showing convergence of a numerical scheme.

In chapter 5 we discuss elliptic equations and typical examples in this case are the Laplace's equation and the Poisson's equation. We work with 2-dimensional domains, where both the independent variables are space variables. In this case it is interesting to look at more complex geometries compared to the parabolic case, and the domain is not necessarily rectangular like in chapter 3. As a consequence most of the discussion is about how to approximate different boundary conditions for domains which are not rectangular.

The first chapter of this note reviews some background material on matrix theory and Taylor expansion, in this chapter we will also set the notation on differential operators.

# Chapter 2

# Background material

## 2.1 Background on matrix theory

Let $A$ be a $n \times n$-matrix with real (or complex) entries, we write $A \in \mathbf{R}^{n\times n}$ (or $A \in \mathbf{C}^{n\times n}$). We say that $A$ *is diagonalizable* if it exist a matrix $X \in \mathbf{C}^{n\times n}$ such that

$$\Lambda = X^{-1}AX = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}.$$

Every $\lambda_i \in \mathbf{C}$ is called *egenvalue* of $A$. The matrix $X$ has $n$ columns denoted by $X = [x_1, \ldots, x_n]$, and every $x_i \in \mathbf{C}^n$ is called *egenvector* (*associated to the egenvalue* $\lambda_i$). We write a diagonal matrix as $\Lambda$ above, in the following form

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

### 2.1.1 Jordan form

For any $A \in \mathbf{R}^{n\times n}$ (or $A \in \mathbf{C}^{n\times n}$) it exists a matrix $M \in \mathbf{C}^{n\times n}$ such that

$$M^{-1}AM = J = \begin{bmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_k \end{bmatrix}, \qquad \text{(block-diagonal)}. \tag{2.1}$$

Here $J_i$ is a $m_i \times m_i$-matrix, and $\sum_{i=1}^{k} m_i = n$. The *Jordan-blocks* $J_i$ have the form

$$J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix}, \quad \text{if } m_i \geq 2$$

and $J_i = [\lambda_i]$ if $m_i = 1$. If all $m_i = 1$, then $k = n$ and the matrix is diagonalizable. If $A$ has $n$ distinct egenvalues, it is always diagonalizable. The converse is not true, that is a matrix can be diagonalizable even if it has multiple eigenvalues.

### 2.1.2 Symmetric matrices

When we talk about symmetric matrices, we mean normally *real* symmetric matrices. The *transpose* $A^T$ of a $m \times n$-matrix $A$, is a $n \times m$-matrix with $a_{ji}$ as the $(ij)$-element (a matrix whose columns are the rows of $A$). A $n \times n$ matrix is symmetric if $A^T = A$.

A symmetric $n \times n$ matrix has real eigenvalues $\lambda_1, \ldots, \lambda_n$ and a set of real orthonormal eigenvectors $x_1, \ldots, x_n$. Let $\langle \cdot, \cdot \rangle$ denote the standard inner-product on $\mathbf{C}^n$, then $\langle x_i, x_j \rangle = \delta_{ij}$ (Kronecker-delta).

A consequence of this is that the matrix of eigenvectors $X = [x_1, \ldots, x_n]$ is real and orthogonal and its inverse is therefore the transpose

$$X^{-1} = X^T.$$

The diagonalization of $A$ is given by

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \quad X = [x_1, \ldots, x_n], \quad X^T X = I, \quad X^T A X = \Lambda \Leftrightarrow A = X \Lambda X^T$$

### 2.1.3 Positive definite matrices

If $A$ is symmetric and $\langle x, Ax \rangle = x^T Ax > 0$ for all $0 \neq x \in \mathbf{R}^n$ $A$ is called *positive definite*.

$A$ (symmetric) is positive semi-definite if $\langle x, Ax \rangle \geq 0$ for all $x \in \mathbf{R}^n$ and $\langle x, Ax \rangle = 0$ for at least a $x \neq 0$.

$A$ positive definite $\Leftrightarrow$ $A$ has only positive eigenvalues.

$A$ positive semi-definite $\Leftrightarrow$ $A$ has only non-negative eigenvalues, and at least a 0-eigenvalue.

### 2.1.4 Gershgorin's theorem

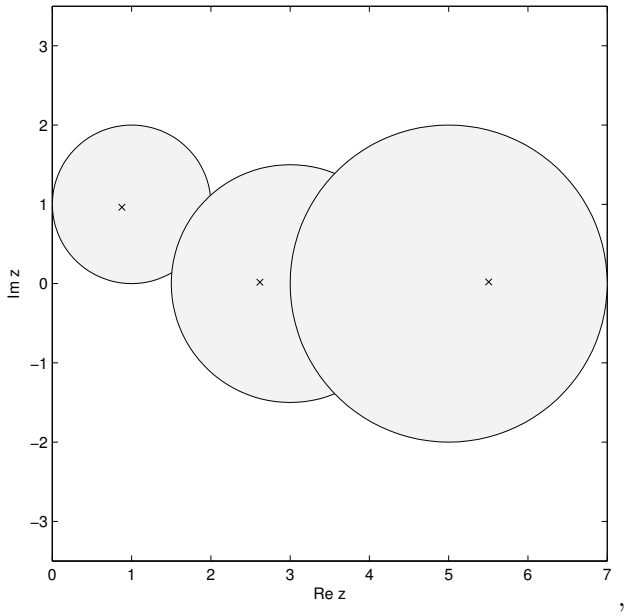
**Gershgorin's theorem.** Is given $A = (a_{ik}) \in \mathbf{C}^{n \times n}$. Define $n$ disks $S_j$ in the complex plane by

$$S_j = \left\{ z \in \mathbf{C} : |z - a_{jj}| \leq \sum_{k \neq j} |a_{jk}| \right\}.$$

The union $S = \bigcup_{j=1}^{n} S_j$ contains all the eigenvalues of $A$. For every eigenvalue $\lambda$ of $A$ there is a $j$ such that $\lambda \in S_j$.

**Example.**

$$A = \begin{bmatrix} 1+i & 1 & 0 \\ 0.5 & 3 & 1 \\ 1 & 1 & 5 \end{bmatrix}.$$




,

□

*Proof of Gershgorin's theorem:* Let $\lambda$ be a eigenvalue with associate eigenvector $x = [\xi_1, \ldots, \xi_n]^T \neq 0$. Choose $\ell$ among the indexes $1, \ldots, n$ such that $|\xi_\ell| \geq |\xi_k|,\ k = 1, \ldots, n$, and so $|\xi_\ell| > 0$. The equation $Ax = \lambda x$ has component $\ell$:

$$\sum_{k=1}^{n} a_{\ell k}\xi_k = \lambda\,\xi_\ell \ \Rightarrow\ (\lambda - a_{\ell\ell})\xi_\ell = \sum_{k\neq\ell} a_{\ell k}\xi_k$$

Divide by $|\xi_\ell|$ on each side and take the absolute value

$$|\lambda - a_{\ell\ell}| \quad = \quad \left|\sum_{k\neq\ell} a_{\ell k}\frac{\xi_k}{\xi_\ell}\right| \quad \leq \quad \sum_{k\neq\ell}|a_{\ell k}|\frac{|\xi_k|}{|\xi_\ell|} \quad \leq \quad \sum_{k\neq\ell}|a_{\ell k}|$$

Then we get $\lambda \in S_\ell$.

**Example.** Diagonally dominant matrices with positive diagonal elements are positive definite. Why?

### 2.1.5 Vector and matrix norms

Consider a vector space $X$ (real or complex). A norm $\|\cdot\| : X \to \mathbf{R}$ satisfies the following axioms

1. $\|x\| \geq 0$ for all $x$, $\|x\| = 0 \Leftrightarrow x = 0$.
2. $\|\alpha x\| = |\alpha|\|x\| \quad (\alpha \in \mathbf{R}\ (\mathbf{C}))$
3. $\|x + y\| \leq \|x\| + \|y\|$

**Examples.** $x = (\xi_k)$, $X = \mathbf{R}^n$.

$$\|x\|_1 = \sum_{k=1}^{n}|\xi_k|, \qquad \|x\|_2 = \left(\sum_{k=1}^{n}|\xi_k|^2\right)^{1/2}, \qquad \|x\|_\infty = \max_{1\leq k\leq n}|\xi_k|.$$

The matrix spaces $\mathbf{R}^{n\times n}$ and $\mathbf{C}^{n\times n}$ are also vector spaces over $\mathbf{R}$ ($\mathbf{C}$). We say that $\|\cdot\|$ is a matrix norm if for all $A, B \in \mathbf{R}^{n\times n}$ ($\mathbf{C}^{n\times n}$)

1. $\|A\| > 0$ for all $A$, $\|A\| = 0 \Leftrightarrow A = 0$,
2. $\|\alpha A\| = |\alpha|\|A\|, \quad (\alpha \in \mathbf{R}\ (\mathbf{C}))$,
3. $\|A + B\| \leq \|A\| + \|B\|$,
4. $\|A \cdot B\| \leq \|A\| \cdot \|B\|$.

*Remark.* The last point requires that a matrix-matrix product is defined (this operation is not defined in a general vector space). In abstract terms the axioms 1–4 give an example of *Banach-algebra*.

**Example.** The Frobenius-norm of a matrix is defined as

$$\|A\|_F = \left( \sum_{j=1}^{n} \sum_{k=1}^{n} |a_{jk}|^2 \right)^{1/2}.$$

### 2.1.6 Consistent and subordinate matrix norms

A given matrix norm is *consistent* with a given vector norm on $\mathbf{R}^n$ if

$$\|Ax\| \leq \|A\| \cdot \|x\| \quad \text{for all} \quad A \in \mathbf{R}^{n\times n},\ x \in \mathbf{R}^n.$$

A given matrix norm is *subordinate* to a given vector norm on $\mathbf{R}^n$ if

$$\|A\| = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|}.$$

**Examples.** We give here as examples some of the most common subordinate matrix norms. We look for matrix norms subordinate to the three vector norms $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$.

1. Let $\|\cdot\|_1$ be the matrix norm subordinate to the vector norm $\|\cdot\|_1$. One can show that $A \in \mathbf{R}^{n\times n}$ ($\mathbf{C}^{n\times n}$) is
$$\|A\|_1 = \max_{1\leq k\leq n} \sum_{i=1}^{n} |a_{ik}|.$$
In other words we can say that $\|A\|_1$ is the "maximal column-sum in $A$".

2. To find the matrix norm subordinate to the vector norm $\|\cdot\|_2$ we must define the *spectral radius* of a matrix $M \in \mathbf{R}^{n\times n}$ ($\mathbf{C}^{n\times n}$). If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $M$, we denote the spectral radius of $M$ by $\rho(M)$, and it is defined as
$$\rho(M) = \max_{1\leq k\leq n} |\lambda_k|. \tag{2.2}$$
If we plot the eigenvalues of $M$ in the complex plane, the spectral radius is the minimal radius of a circle centered in the origin and containing all eigenvalues of $M$.

   We define now the 2-norm of a matrix $A$ as
$$\|A\|_2 = \sqrt{\rho(A^T A)}.$$

Note that $A^TA$ is positive (semi)definite, so all the eigenvalues are real and positive. Taking the square root of the biggest eigenvalue, we obtain $\|A\|_2$. Note also that the spectral radius of $A$ can be very different from (the square root of ) the spectral radius of $A^TA$. On the other hand if $A$ is symmetric then $\|A\|_2 = \rho(A)$.

3. Let $\|\cdot\|_\infty$ be the matrix norm subordinate to the vector norm $\|\cdot\|_\infty$. We have

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{k=1}^{n} |a_{ik}|.$$

That is $\|A\|_\infty$ is the "maximal row-sum in $A$". Observe also that $\|A\|_1 = \|A^T\|_\infty$.

### 2.1.7 Matrix norms and spectral radius.

For any matrix norm $\|\cdot\|$ it is true that

$$\|A\| \ge \rho(A). \tag{2.3}$$

*Proof:* Let $x$ be a eigenvector of $A$ associated to an eigenvalue $\lambda$ such that

$$Ax = \lambda x.$$

Let $y \in \mathbf{C}^n$ be arbitrary. Then we have

$$A(xy^T) = (Ax)y^T = \lambda(xy^T),$$

such that

$$\|A(xy^T)\| \le \|A\|\,\|xy^T\|.$$

As a consequence

$$|\lambda|\,\|xy^T\| = \|\lambda(xy^T)\| = \|A(xy^T)\| \le \|A\|\,\|xy^T\|.$$

Therefore $|\lambda| \le \|A\|$, and since this is true for every eigenvalue of $A$, it must be $\rho(A) \le \|A\|$.

*Question to the reader:* What is wrong with the following line of reasoning

$$|\lambda|\,\|x\| = \|\lambda\, x\| = \|A\, x\| \le \|A\|\,\|x\| \quad \text{etc.}$$

**Convergent matrix.** A matrix $A$ is said to be *convergent* (to zero) if

$$A^k \to 0 \quad \text{when} \quad k \to \infty.$$

| **A sufficient criterion.** If $\|A\| < 1$ for a particular matrix norm, $A$ is convergent. |
|---|

*Proof.*

$$\|A^k\| = \|A \cdot A^{k-1}\| \le \|A\| \cdot \|A^{k-1}\| \le \cdots \le \|A\|^k \ \to 0 \qquad \text{if } \|A\| < 1$$

| **Necessary and sufficient criterion.** $A$ is convergent if and only if the spectral radius $\rho(A)$, defined by (2.2), satisfies $\rho(A) < 1$. |
|---|

*Proof:* We use Jordan form, and let $A = MJM^{-1}$ where $M \in \mathbf{C}^{n\times n}$ and $J$ is like in (2.1). Then we have $A^2 = MJM^{-1}MJM^{-1} = MJ^2M^{-1}$, and by induction we get $A^k = MJ^kM^{-1}$. Now $A^k \to 0$ if and only if $J^k \to 0$. And $J^k \to 0$ if and only if every Jordan block $J_i^k \to 0$. Assume such a Jordan block has diagonal element $\lambda$ and the $m_i \times m_i$-matrix $F$ has its $(j, j+1)$ elements, for $j = 1, \ldots, m_i - 1$, equal to 1, and the other elements equal to zero. Then $J_i = \lambda I + F$ where $I$ is the identity matrix. The matrix $F$ is nihilpotent, i.e. $F^m = 0,\ m \geq n$. We assume that $k \geq n-1$ and compute

$$J_i^k = (\lambda I + F)^k = \sum_{m=0}^{k} \binom{k}{m} \lambda^{k-m} F^m = \sum_{m=0}^{n-1} \binom{k}{m} \lambda^{k-m} F^m = \sum_{m=0}^{n-1} \varphi_k^{(m)}(\lambda) F^m$$

where $\varphi_k(\lambda) = \lambda^k/k!$. When $k \to \infty$ then $\varphi_k^{(m)}(\lambda) \to 0$ for $0 \leq m \leq n-1$ if and only if $|\lambda| < 1$. This must be true for all Jordan blocks (i.e. eigenvalues of $A$) and this concludes the proof. □

## 2.2 Difference formulae

### 2.2.1 Taylor expansion

**1 free variable.** Let $u \in C^{n+1}(I)$ where $I \subset \mathbf{R}$ is a interval of the real line. This means that the $n+1$-th derivative of $u$ exists and is continuous on the interval $I$. Then the following formula is valid.

**Taylor's formula with reminder.** With $x \in I$, $x + h \in I$ is

$$u(x+h) = \sum_{m=0}^{n} \frac{h^m}{m!} u^{(m)}(x) + r_n$$

where

$$r_n = \frac{h^{n+1}}{(n+1)!} u^{(n+1)}(x + \theta h), \quad 0 < \theta < 1.$$

**2 free variables.** Assume now that $u \in C^{n+1}(\Omega)$ where $\Omega \subset \mathbf{R}^2$. It is convenient to use a operatornotation for the partial derivatives. We write $\mathbf{h} = [h, k]$, and let

$$\mathbf{h} \cdot \nabla := h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y} \quad \text{i.e.} \quad \mathbf{h} \cdot \nabla u = h\frac{\partial u}{\partial x} + k\frac{\partial u}{\partial y}$$

The operator produces the derivative of a function in the direction $\mathbf{h} = [h, k]$, and we find the *directional derivative*.

We can also define powers of the operator by for example

$$(\mathbf{h} \cdot \nabla)^2 = \left(h\frac{\partial}{\partial x} + k\frac{\partial}{\partial y}\right)^2 = h^2\frac{\partial^2}{\partial x^2} + 2hk\frac{\partial^2}{\partial x \partial y} + k^2\frac{\partial^2}{\partial y^2}$$

The extension to the $m$-th power is obvious. then we can write

**Taylor's formula with reminder for functions of two variables.**

$$u(x+h,y+k)=\sum_{m=0}^{n}\frac{1}{m!}\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)^{m}u(x,y)+r_n \tag{2.4}$$

where

$$r_n=\frac{1}{(n+1)!}\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial y}\right)^{n+1}u(x+\theta h,y+\theta k),\quad 0<\theta<1.$$

We have here assumed that the line segment between $(x,y)$ and $(x+h,y+k)$ is included in $\Omega$.

**Derivation of the previous formula.** We look at a function of one variable $\mu(t)=u(x+th,y+tk)$ for fixed $x,y,h,k$. Using Taylor's expansion with reminder for the case of one variable for $\mu(t)$ around $t=0$, we obtain the two variables formula by setting $t=1$.

### 2.2.2 Big $\mathcal{O}$-notation

Let $\phi$ be a function of $h$ and $p$ a positive integer. Then we have

$$\phi(h)=\mathcal{O}(h^p)\quad\text{when}\quad h\to 0$$

if there exist two constants $C,H>0$ such that

$$|\phi(h)|\le C|h|^p\quad\text{when}\quad 0<|h|<H.$$

If this holds, we say that $\phi(h)$ is of *order* $p$ in the variable $h$.

The typical use of the big $\mathcal{O}$-notation is in connection with the local truncation error in numerical methods. For example in the Taylor expansion in one variable

$$|r_n|=\left|\frac{h^{n+1}}{(n+1)!}\,u^{(n+1)}(x+\theta h)\right|\le\frac{M}{(n+1)!}|h|^{n+1},\quad M=\max_{y\in I}|u^{(n+1)}(y)|$$

where we know that the maximum exists if $I$ is a closed, and bounded interval. So in this case we have $r_n=\mathcal{O}(h^{n+1})$.

Note that with the definition above for positive integers $p$ we have

$$\phi(h)=\mathcal{O}(h^{p+1})\quad\Rightarrow\quad\phi(h)=\mathcal{O}(h^p).$$

Therefore sometimes when we write $\phi(h)=\mathcal{O}(h^p)$ we mean that $p$ is the biggest possible integer such that this is true. Often it is convenient to write $\mathcal{O}(h^p)$ in formulae with sums, like for example in the Taylor expansion of $u$ above we replace $r_n$ with $\mathcal{O}(h^{n+1})$ such that

$$u(x+h)=\sum_{k=0}^{n}\frac{h^k}{k!}u^{(m)}(x)+\mathcal{O}(h^{n+1}).$$

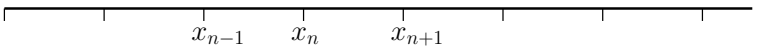
In general we have that $\phi(h)=\mathcal{O}(h^{p_\phi})$ and $\psi(h)=\mathcal{O}(h^{p_\psi})$, then

$$\psi(h)+\phi(h)=\mathcal{O}(h^q),\quad\text{where}\quad q=\min(p_\phi,p_\psi).$$

But sometimes one can get higher powers. An obvious example is when $\phi(h)=h^2$, $\psi(h)=h^3-h^2$, each of them is $\mathcal{O}(h^2)$ but their sum is $\mathcal{O}(h^3)$. If you multiply a function $\phi(h)$ by a constant ($\neq 0$), the order does not change.

### 2.2.3 Difference approximations to the derivatives

We introduce a *grid* on $\mathbf{R}$ i.e. a monotone sequence of real numbers $\{x_n\}$ where $x_n \in \mathbf{R}$.




Assume that $u(x)$ is a given function, $u \in C^q(I)$, for a $q$ which we will specify later. Let

$$u_n := u(x_n), \qquad u_n^{(m)} := u^{(m)}(x_n).$$

Assume the grid points $x_n$ are equidistant, i.e. $x_{n+1} = x_n + h$ for all $n$, where $h \in \mathbf{R}$ is called *step-size*. We want to approximate $u_n^{(m)}$ with expressions of the type

$$\sum_{\ell=p}^{q} a_\ell u_{n+\ell}$$

$p \le q$ are integers, and typically $p \le 0$ and $q \ge 0$.

**Truncation error.** We define

$$\tau_n(h) = \sum_{\ell=p}^{q} a_\ell u_{n+\ell} - u_n^{(m)}.$$

The strategy is to choose $p$ and $q$, and then compute the $q - p + 1$ parameters $a_p, \ldots, a_q$ such that $\tau_n$ is "small".

By Taylor expansion we obtain

$$u_{n+\ell} = u(x_n + \ell h) = \sum_{k=0}^{\nu} \frac{(\ell h)^k}{k!} u_n^{(k)} + r_\nu,$$

where $r_\nu = \mathcal{O}(h^{\nu+1})$ and $\nu \ge m$, such that

$$\tau_n = \sum_{\ell=p}^{q} a_\ell \sum_{k=0}^{\nu} \frac{1}{k!} (\ell h)^k u_n^{(k)} - u_n^{(m)} + \mathcal{O}(h^{\nu+1}),$$

which can be rearranged in the form

$$\tau_n = \sum_{k=0}^{\nu} \frac{h^k}{k!} \left( \sum_{\ell=p}^{q} a_\ell \ell^k \right) u_n^{(k)} - u_n^{(m)} + \mathcal{O}(h^{\nu+1}).$$

We want that $\tau_n = \mathcal{O}(h^r)$ with $r$ as big as possible. To approximate $u_n^{(m)}$ we need to impose conditions on $p$ and $q$. Set $j := q - p$. In order to get consistent approximation formulae (i.e. such that $\tau_n(h) \to 0$ when $h \to 0$), we must require $j \ge m$. We choose then $a_p, \ldots, a_q$ such that

$$\frac{h^k}{k!} \sum_{\ell=p}^{q} \ell^k \, a_\ell = \begin{cases} 0 & 0 \le k \le m-1, \\ 1 & k = m, \\ 0 & m+1 \le k \le j. \end{cases} \tag{2.5}$$

Note that we have $q-p+1 = j+1$ free parameters $a_p, \ldots, a_q$ we can use, and the conditions in (2.5) must be satisfied for $0 \le k \le j$, this means a total of $j+1$ conditions. The system of equations has a unique solution for $h \neq 0$. If we choose $a_\ell$ from (2.5), and assume $\nu \ge j$, we obtain the following truncation error

$$\tau_n = \sum_{k=j+1}^{\nu} \frac{h^k}{k!} u_n^{(k)} \sum_{\ell=p}^{q} a_\ell \, \ell^k + \mathcal{O}(h^{\nu+1}).$$

This methods is called *the method of undetermined coefficients.*

**Example.** $m = 1$ ($u'_n$). Choose $p = -1,\ q = 1,\ j = 2$. We want to find $a_{-1}$, $a_0$, $a_1$. We write $j + 1 = 3$ equations i.e. $k = 0, 1, 2$ i (2.5).

$$\left.\begin{array}{lrcrcrcl} k=0 & a_{-1} & + & a_0 & + & a_1 & = & 0 \\ k=1 & -h\,a_{-1} & + & 0\cdot a_0 & + & h\,a_1 & = & 1 \\ k=2 & h^2\,a_{-1} & + & 0\cdot a_0 & + & h^2\,a_1 & = & 0 \end{array}\right] \Rightarrow \begin{array}{lcr} a_{-1} & = & -\frac{1}{2h} \\ a_0 & = & 0 \\ a_1 & = & \frac{1}{2h} \end{array}$$

Looking at the first terms in the local truncation error we obtain

$$\tau_n = \sum_{k=3} \frac{h^k}{k!}\, u_n^{(k)} \left( -\frac{1}{2h}(-1)^k + \frac{1}{2h}\, 1^k \right) = \sum_{s=1} \frac{u_n^{(2s+1)}}{(2s+1)!}\, h^{2s}.$$

In the last equality, we have used the fact that the terms with even $k$ disappear, such that we can put $k = 2s + 1$ and let $s = 1, 2, \ldots$. We have omitted the upper limit value for the index $s$ on purpose because the number of terms we include in the remainder depend on the circumstances. Since the first term in the expression for $\tau_n$ is of type $h^2$, we say that the formula is of *order* 2.

**Some more formulae.** Other popular difference approximations are

$$\begin{array}{lcl} m=1 & \dfrac{u_{n+1}-u_n}{h} & = u'_n + \frac{1}{2!}h\,u''_n + \cdots \\ m=1 & \dfrac{u_n-u_{n-1}}{h} & = u'_n - \frac{1}{2!}h\,u''_n + \cdots \\ m=2 & \dfrac{u_{n+1}-2u_n+u_{n-1}}{h^2} & = u''_n + \frac{1}{12}h^2 u_n^{(4)} + \cdots \end{array}$$

Which order have these formulae?

**Exercises**

1. Assume $m = 1$, $p = -2$, $q = 0$ use the method of undetermined coefficients to obtain an approximation of $u_n^{(1)}$ of the second order in $h$.
2. Assume $m = 1$, $p = -2$, $q = 1$ use the method of undetermined coefficients to obtain an approximation of $u_n^{(1)}$ of the third order in $h$.
3. Consider $m = 2$, $p = 0$, $q = 1$, so that $j = q - p = 1 < m$. Try to construct an approximation formula for $u_n^{(2)}$ using the method of undetermined coefficients. What happens?

### 2.2.4 Difference operators and other operators

| | |
|---|---|
| Forward difference: | $\Delta u(x) = u(x + h) - u(x)$. |
| Backward difference: | $\nabla u(x) = u(x) - u(x - h)$. |
| Central difference: | $\delta u(x) = u(x + \frac{h}{2}) - u(x - \frac{h}{2})$. |
| Mean value: | $\mu u(x) = \frac{1}{2}\left(u(x + \frac{h}{2}) + u(x - \frac{h}{2})\right)$. |
| Shift: | $Eu(x) = u(x + h)$. |
| Unity operator | $1\,u(x) = u(x)$. |

**Linearity.** All the operators

$$\Delta,\ \nabla,\ \delta,\ \mu,\ E,\ 1,$$

are linear. This means that for $\alpha \in \mathbf{R}$, and with functions $u(x)$ and $v(x)$ we have

$$F(\alpha u(x) + v(x)) = \alpha F u(x) + F v(x),$$

where $F$ can be any of the operators above. Let us verify this for $F = \Delta$.

$$\begin{aligned} \Delta(\alpha u(x) + v(x)) &= (\alpha u(x+h) + v(x+h)) - (\alpha u(x) + v(x)) \\ &= \alpha(u(x+h) - u(x)) + (v(x+h) - v(x)) = \alpha \Delta u(x) + \Delta v(x). \end{aligned}$$

**Powers of the operators.** Let $F$ be one of the above defined operators. We can define powers of $F$ as follows

$$F^0 = 1, \qquad F^k u(x) = F(F^{k-1} u(x)).$$

**Example.**

$$\begin{aligned} \delta u(x) &= u(x + \tfrac{h}{2}) - u(x - \tfrac{h}{2}), \\ \delta^2 u(x) &= \delta(\delta u(x)) = \delta u(x + \tfrac{h}{2}) - \delta u(x - \tfrac{h}{2}) = u(x+h) - u(x) - (u(x) - u(x-h)), \\ &= u(x+h) - 2u(x) + u(x-h). \end{aligned}$$

Another interesting example is the shift operator. We observe that $E^k u(x) = u(x+kh)$. In this case it is easy to extend the definition to include all possible real powers, simply by defining $E^s u(x) = u(x+sh)$ for all $s \in \mathbf{R}$. For example we have $E^{-1} u(x) = u(x-h)$ and this is the inverse of $E$ since $Eu(x-h) = E^{-1} u(x+h) = u(x)$.

**Relations between the difference operators.**

$$\begin{aligned} \Delta u(x) &= u(x+h) - u(x) = Eu(x) - 1\,u(x) = (E-1)\,u(x), \\ \nabla u(x) &= u(x) - u(x-h) = 1\,u(x) - E^{-1} u(x) = (1 - E^{-1})\,u(x), \\ \delta u(x) &= u(x + \frac{h}{2}) - u(x - \frac{h}{2}) = (E^{1/2} - E^{-1/2})\,u(x), \\ \mu u(x) &= \frac{1}{2}\left( u(x + \frac{h}{2}) + u(x - \frac{h}{2}) \right) = \frac{1}{2}(E^{1/2} + E^{-1/2})\,u(x). \end{aligned}$$

In a more compact notation we have

$$\begin{aligned} \Delta &= (E-1), \\ \nabla &= (1 - E^{-1}), \\ \delta &= (E^{1/2} - E^{-1/2}), \\ \mu &= \frac{1}{2}(E^{1/2} + E^{-1/2}). \end{aligned}$$

And now we have for example

$$\Delta^k = (E-1)^k = \sum_{\ell=0}^{k} \binom{k}{\ell} (-1)^{k-\ell} E^\ell,$$

such that

$$\Delta^k u(x) = \sum_{\ell=0}^{k} \binom{k}{l} (-1)^{k-\ell} E^\ell u(x) = \sum_{\ell=0}^{k} \binom{k}{l} (-1)^{k-\ell} u(x + \ell h).$$

### 2.2.5 Differential operator.

Define

$$D = \frac{\mathrm{d}}{\mathrm{d}x} \quad \text{so that} \quad Du(x) = u'(x).$$

Let

$$D^m u(x) = u^{(m)}(x).$$

If $u(x)$ is analytic[1] in an interval containing $x, x+h$ we have

$$u(x+h) = \sum_{m=0}^{\infty} \frac{h^m}{m!} D^m u(x) = \left( \sum_{m=0}^{\infty} \frac{1}{m!} (hD)^m \right) u(x) = \mathrm{e}^{hD}\, u(x).$$

We think of this only as a *notation*. We have

$$Eu(x) = \mathrm{e}^{hD} u(x),$$

and then $E = \mathrm{e}^{hD}$.

**Relation between $D$ and the other operators.**

$$\Delta = E - 1 = \mathrm{e}^{hD} - 1 = \sum_{m=1}^{\infty} \frac{1}{m!} (hD)^m,$$

$$\Delta = hD + \frac{1}{2!}(hD)^2 + \cdots$$

We will see that under the extra assumption that $u$ is analytic we can make manipulations with analytic functions in the way we are used to. The meaning is always that the final result is expanded with a Taylor expansion and is interpreted as a sum of powers of operators which are applied to a smooth function. The analyticity requirement can always be relaxed by considering a Taylor expansion with reminder, and requiring the function to be differentiable only a finite number of times.

We consider powers of $\Delta$, and we obtain

$$\Delta^k = \left( \sum_{m=1}^{\infty} \frac{(hD)^m}{m!} \right)^k = h^k D^k + \frac{k}{2!} h^{k+1} D^{k+1} + \cdots$$

or

$$\Delta^k u(x) = h^k D^k u(x) + \frac{k}{2!} h^{k+1} D^{k+1} u(x) + \cdots$$

showing that $\Delta^k / h^k$ is a first order approximation (truncation error $\mathcal{O}(h)$) of the operator $D^k$.

Note that for $s \in \mathbf{R}$ we have

$$E^s u(x) = u(x + sh) = \sum_{k=0}^{\infty} \frac{(sh)^k}{k!} D^k u(x) = \mathrm{e}^{shD} u(x)$$

which reflects known computational rules. For central differences we can therefore write

$$\delta = E^{1/2} - E^{-1/2} = \mathrm{e}^{\frac{1}{2}hD} - \mathrm{e}^{-\frac{1}{2}hD} = 2 \sinh \frac{hD}{2}.$$

[1] By analytic function on an interval we simply mean that its Taylor expansion converges in a neighborhood of any point of the interval.

We can also compute

$$\delta^k = \left(2\sinh\frac{hD}{2}\right)^k = \left(hD + \frac{2}{3!}\left(\frac{hD}{2}\right)^3 + \cdots\right)^k = (hD)^k + \frac{k}{24}(hD)^{k+2} + \cdots$$

that is

$$\delta^k u(x) = h^k D^k u(x) + \frac{k}{24}h^{k+2}D^{k+2}u(x) + \cdots$$

this shows that $\delta^k/h^k$ is a second order approximation of $D^k$.

In particular we find as before that

$$\delta^2 u(x) = u(x+h) - 2u(x) + u(x-h) = h^2 u''(x) + \frac{1}{12}h^4 u^{(4)}(x) + \cdots \tag{2.6}$$

It is tempting to manipulate further with analytic functions. We have seen that

$$\frac{\delta}{2} = \sinh\frac{hD}{2}.$$

We write therefore formally

$$D = \frac{2}{h}\sinh^{-1}\frac{\delta}{2}.$$

It is possible to expand $\sinh^{-1} z$ in a Taylor expansion

$$\sinh^{-1} z = z - \frac{1}{6}z^3 + \frac{3}{40}z^5 - \frac{5}{112}z^7 + \cdots$$

so $z = \delta/2$ and by multiplying by $2/h$ we obtain

$$D = \frac{1}{h}\left(\delta - \frac{1}{24}\delta^3 + \frac{3}{640}\delta^5 - \frac{5}{7168}\delta^7 + \cdots\right).$$

Since we know that $\delta^k = \mathcal{O}(h^k)$ we see that we can find approximations to the differential operator $D$ of arbitrary high order by including enough terms in the expansion. The manipulation we have carried out is not rigorously justified here, but it turns out to be correct. For a detailed discussion on algebraic manipulations with differential operators, see the textbook by Arieh Iserles, *A first course in the numerical analysis of differential equations*, published by Cambridge University Press.

# Chapter 3

# Boundary value problems

## 3.1 A simple case example

We consider the boundary value problem

$$u_{xx} = f(x), \qquad 0 < x < 1, \quad u(0) = \alpha, \quad u(1) = \beta, \tag{3.1}$$

the exact solution can be obtained by integrating twice on both sides between 0 and 1, and then imposing the boundary conditions. We want to use this simple test problem to illustrate some of the basic features of finite difference discretization methods.

To obtain a finite difference discretization for this problem we consider the grid

$$x_m = m\,h,\ m = 0, \ldots, M+1, \quad h = \frac{1}{M+1},$$

and the notation $u_m := u(x_m)$ such that $u_0 = \alpha$ and $u_{M+1} = \beta$. We denote with capital letters $U_m \approx u_m$ the numerical approximation to $u(x)$ at the grid point $x = x_m$.

By replacing derivatives with central difference approximations to the left hand side of (3.1) we obtain the so called discrete problem whose solution is the numerical approximation that we are seeking, this is

$$\frac{1}{h^2}\left(U_{m-1} - 2U_m + U_{m+1}\right) = f_m, \quad m = 1, \ldots, M. \tag{3.2}$$

This is a linear system of equations

$$A_h \mathbf{U} = \mathbf{F}, \tag{3.3}$$

where $A_h$ is a $M \times M$ matrix $\mathbf{U} \in \mathbf{R}^M$ and $\mathbf{F} \in \mathbf{R}^M$ and

$$A_h := \frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}, \quad \mathbf{U} := \begin{bmatrix} U_1 \\ \vdots \\ U_M \end{bmatrix}, \quad \mathbf{F} := \begin{bmatrix} f_1 - \frac{\alpha}{h^2} \\ f_2 \\ \vdots \\ f_{M-1} \\ f_M - \frac{\beta}{h^2} \end{bmatrix}.$$

We know that $u''(x_m) = \frac{1}{h^2}\,\delta^2 u(x_m) + \mathcal{O}\left(h^2\right)$ and we want to deduce similar information about the error

$$e_m := U_m - u_m, \quad e_0 = 0,\ e_{M+1} = 0.$$

Let the error vector $\mathbf{e}_h$ be

$$\mathbf{e}_h := \mathbf{U} - \mathbf{u}, \qquad \mathbf{u} := \begin{bmatrix} u_1 \\ \vdots \\ u_M \end{bmatrix}.$$

We will in the sequel associate such vector to a piecewise constant function $e_h(x)$ defined on the interval $[0, 1]$ as follows

$$e_h(x) = e_m, \quad x \in [x_m, x_{m+1}), \quad m = 1, \ldots, M.$$

Because we are approximating a function, $u(x)$ solution of (3.1), it is appropriate to think of the numerical solution as a piecewise constant function approximating $u(x)$ and similarly for the error. We are therefore interested in measuring the norm of this piecewise constant error function rather than the norm of the corresponding error vector, however the two are closely related. In fact we can see that the following relationships hold:

- $\|e_h\|_\infty = \max_{0\le x\le 1} |e_h(x)| = \max_{1\le m\le M} |e_m| = \|\mathbf{e}_h\|_\infty$;
- $\|e_h\|_1 = \int_0^1 |e_h(x)|\, dx = \sum_{m=1}^M \int_{x_m}^{x_{m+1}} |e_h(x)|\, dx = h \sum_{m=1}^M |e_m| = h\|\mathbf{e}_h\|_1$;
- $\|e_h\|_2 = \left(\int_0^1 |e_h(x)|^2\, dx\right)^{\frac{1}{2}} = \left(h \sum_{m=1}^M |e_m|^2\right)^{\frac{1}{2}} = h^{\frac{1}{2}} \|\mathbf{e}_h\|_2$;

and we see that in these three popular cases the vector norm is related to the function norm of the corresponding piecewise constant function (with respect to the assumed grid) simply by a scaling factor. A similar result is true for the case of $\|\cdot\|_q$.

**Truncation error**. The truncation error is the vector that by definition has components

$$\tau_m := \frac{1}{h^2}\left(u_{m-1} - 2u_m + u_{m+1}\right) - f_m, \quad m = 1, \ldots, M, \quad \tau_h := \begin{bmatrix} \tau_1 \\ \vdots \\ \tau_M \end{bmatrix}.$$

By using that $u''(x_m) = \frac{1}{h^2}\,\delta^2 u(x_m) - \frac{1}{12} h^2 u_m^{(4)} + \mathcal{O}\left(h^4\right)$ and $u''_m = f_m$, we obtain

$$\tau_m = u''_m + \frac{1}{12} h^2 u_m^{(4)} + \mathcal{O}\left(h^4\right) - f_m = \frac{1}{12} h^2 u_m^{(4)} + \mathcal{O}\left(h^4\right).$$

**Equation for the error**. The relationship between the error $\mathbf{e}_h$ and the truncation error is easily obtained: recall that by definition

$$\tau_h = A_h \mathbf{u} - \mathbf{F},$$

rearranging and subtracting this from $A_h \mathbf{U} = \mathbf{F}$ we obtain the important relation

$$A\mathbf{e}_h = -\tau_h \tag{3.4}$$

which can be also written componentwise as

$$\frac{1}{h^2}\left(e_{m-1} - 2e_m + e_{m+1}\right) = -\tau_m, \quad m = 1, \ldots, M.$$

**Definition**. A method for the boundary value problem (3.1) is said to be *consistent* with the boundary value problem with respect to the norm $\|\cdot\|$ if and only if

$$\|\tau_h\| \to 0, \quad \text{when } h \to 0.$$

Consistence in the vector norm $\|\cdot\|$ implies that the corresponding piecewise constant function tends to zero as $h$ tends to zero in the corresponding function norm, this is because of the relationship between vector and function norms as we have seen for $\|\cdot\|_1$, $\|\cdot\|_\infty$ and $\|\cdot\|_2$.

**Definition**. A method for the boundary value problem (3.1) is said to be *convergent* in the (function) norm $\|\cdot\|$ if and only if

$$\|e_h\| \to 0, \quad \text{when } h \to 0.$$

**Definition**. *Stability.* Assume a difference method for the boundary value problem (3.1) is given by the discrete equation

$$A_h \mathbf{U} = \mathbf{F},$$

where $h$ is the step-size of discretization. The method is *stable* in the norm $\|\cdot\|$ if there exist constants $C > 0$ and $H > 0$ such that

1. $A_h^{-1}$ exists for all $h < H$;
2. $\|A_h^{-1}\| \leq C$ for all $h < H$.

The matrix norm in which we should prove stability is the one subordinate to the chosen function/vector norm in which we want to prove convergence.

**Proposition.** For the boundary value problem (3.1), stability and consistence with respect to the norm $\|\cdot\|$ imply convergence in the same norm.

*Proof.* We use (3.4) to obtain a bound for the norm of the error. Since we have stability $A_h$ is invertible for all $h < H$ and therfore

$$\mathbf{e}_h = -A_h^{-1}\tau_h.$$

Then

$$\|e_h\|_q \leq h^{\frac{1}{q}}\|A_h^{-1}\|_q\|\tau_h\|_q = \|A_h^{-1}\|_q\|\tau_h\|_q \leq C\|\tau_h\|_q,$$

and we conclude that $\|e_h\|_q \to 0$ as $h \to 0$ because so does $\|\tau_h\|_q$.

For the case $\|\cdot\|_\infty$ the vector norm of $\mathbf{v} \in \mathbf{R}^M$ and the function norm of the corresponding piecewise constant function defined on the grid coincide, so to include this case we can conveniently adopt the notation $h^{\frac{1}{\infty}} := \lim_{q\to\infty} h^{\frac{1}{q}} = 1$.

### 3.1.1 2-norm stability for the case example

Observe that the eigenvalues of the matrix (3.3) are

$$\lambda_m = \frac{2}{h^2}\left(\cos(m\,\pi\,h) - 1\right), \quad m = 1, \ldots, M,$$

and the corresponding eigenvectors $\mathbf{v}^m$ have components

$$v_j^m = \sin(m\,\pi\,j\,h), \quad j = 1, \ldots, M.$$

Since by definition $\|A_h\|_2 = \sqrt{\rho(A_h^T A_h)}$ then because $A_h$ is symmetric $\|A_h\|_2 = \rho(A_h)$. Denote with $\sigma(B)$ the spectrum of the $M \times M$ matrix $B$ (the collection of all the eigenvalues of $B$), then $\|A_h\|_2 = \max_{\lambda\in\sigma(A_h)} |\lambda|$. Analogously

$$\|A_h^{-1}\|_2 = \rho(A_h^{-1}) = \max_{\lambda\in\sigma(A_h)} |\lambda^{-1}| = \frac{1}{\min_{\lambda\in\sigma(A_h)} |\lambda|}.$$

Using the series expansion $\cos(x) = 1 - \frac{x^2}{2} + \frac{x^4}{4!} + \mathcal{O}(x^6)$, with $x = m\,\pi\,h$ in the expression for the $m$-th eigenvalue we get

$$\lambda_m = \frac{2}{h^2}\left(-\frac{(m\,\pi\,h)^2}{2} + \frac{(m\,\pi\,h)^4}{4!} + \mathcal{O}(h^6)\right) = -m^2\pi^2 + \mathcal{O}(h^2),$$

and

$$\min_{\lambda\in\sigma(A_h)} |\lambda| = |\lambda_1| = -\pi^2 + \mathcal{O}(h^2),$$

such that

$$\|A_h^{-1}\|_2 = \frac{1}{|\lambda_1|} \to \frac{1}{\pi^2}, \quad \text{when } h \to 0,$$

and so there exist $C > 0$ and $H > 0$ such that $\|A_h^{-1}\|_2 = \frac{1}{|\lambda_1|} < C$ for all $h < H$, which proves stability of the proposed difference scheme in the 2-norm.

**Exercise.** Using the estimates for $\|A_h^{-1}\|_2$ and $\|\tau_h\|_2$ obtained in this section, prove that

$$\|e_h\| \leq \frac{1}{\pi^2}\,\frac{h^{2.5}}{12}\,\|f''\|_2 + \mathcal{O}(h^{3.5}),$$

assume $f$ is twice differentiable.

### 3.1.2 Neumann boundary conditions

We consider now the boundary value problem

$$u_{xx} = f(x), \qquad 0 < x < 1, \quad u'(0) = \sigma, \quad u(1) = \beta, \tag{3.5}$$

and we propose three different discretizations of the left boundary condition which combined with the earlier consider discretization of the second derivative will lead to three different linear systems. In this case the matrices we obtain are no longer symmetric.

**Case 1**

We approximate $u'(0) = \sigma$ by

$$\frac{U_1 - U_0}{h} = \sigma,$$

and combine this discrete equation with (3.2), this leads to the $(M+1) \times (M+1)$ linear system

$$A_h \mathbf{U} = \mathbf{F},$$

where

$$A_h := \frac{1}{h^2} \begin{bmatrix} -h & h & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}, \quad \mathbf{U} := \begin{bmatrix} U_0 \\ \vdots \\ U_M \end{bmatrix}, \quad \mathbf{F} := \begin{bmatrix} \sigma \\ f_1 \\ \vdots \\ f_{M-1} \\ f_M - \frac{\beta}{h^2} \end{bmatrix}.$$

This leads to an overall method of order of consistency 1.

**Case 2**

To obtain order two we introduce a fictitious node external to the domain to the left of $x_0$ which we call $x_{-1}$. We approximate $u'(0) = \sigma$ by

$$\frac{U_1 - U_{-1}}{2h} = \sigma.$$

We use this equation together with (3.2) for $m = 0$, i.e.

$$\frac{1}{h^2} (U_{-1} - 2U_0 + U_1) = f_0,$$

and eliminate $U_{-1}$.

We are left with the equation

$$\frac{U_1 - U_0}{h} = \frac{f_0 h}{2} + \sigma$$

which we now combine with (3.2), this leads to a $(M+1) \times (M+1)$ linear system

$$A_h \mathbf{U} = \mathbf{F},$$

where $A_h$ is as in the previous case and

$$\mathbf{F} := \begin{bmatrix} \frac{f_0 h}{2} + \sigma \\ f_1 \\ \vdots \\ f_{M-1} \\ f_M - \frac{\beta}{h^2} \end{bmatrix}.$$

**Case 3**

In this discretization we want to obtain order 2 without the use of fictitious nodes. We then consider the approximation of $u'(0) = \sigma$ by

$$-\frac{\frac{3}{2}U_0 - 2U_1 + \frac{1}{2}U_2}{h} = \sigma,$$

where the left hand side gives a discretization of order two of the first derivative of $u$ at zero. Proceeding as in case 1, we obtain the $(M+1) \times (M+1)$ linear system

$$A_h \mathbf{U} = \mathbf{F},$$

where

$$A_h := \frac{1}{h^2} \begin{bmatrix} -\frac{3}{2}h & 2h & -\frac{h}{2} & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}, \quad \mathbf{U} := \begin{bmatrix} U_0 \\ U_1 \\ \vdots \\ U_M \end{bmatrix}, \quad \mathbf{F} := \begin{bmatrix} \sigma \\ f_1 \\ \vdots \\ f_{M-1} \\ f_M - \frac{\beta}{h^2} \end{bmatrix}.$$

**Exercise.** Prove with the method of undetermined coefficients that

$$-\frac{\frac{3}{2}u(0) - 2u(h) + \frac{1}{2}u(2h)}{h} = u'(0) + \mathcal{O}(h^2).$$

**Neumann boundary conditions on both sides**

We consider now the boundary value problem

$$u_{xx} = f(x), \qquad 0 < x < 1, \quad u'(0) = \sigma_0, \quad u'(1) = \sigma_1, \tag{3.6}$$

and we propose a second order discretization of the boundary conditions using two fictitious nodes external to $[0,1]$, $x_{-1}$ and $x_{M+2}$. The approximated boundary conditions become

$$\frac{U_1 - U_{-1}}{2h} = \sigma_0, \qquad \frac{U_{M+2} - U_M}{2h} = \sigma_1,$$

which we combine respectively with (3.2) for $m = 0$ and $m = M+1$ and obtain the linear system of equations $(M+2) \times (M+2)$ linear system

$$A_h \mathbf{U} = \mathbf{F},$$

where

$$A_h := \frac{1}{h^2} \begin{bmatrix} -h & h & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & 1 & -2 & 1 \\ & & & h & -h \end{bmatrix}, \quad \mathbf{U} := \begin{bmatrix} U_0 \\ \vdots \\ U_{M+1} \end{bmatrix}, \quad \mathbf{F} := \begin{bmatrix} \sigma_0 + \frac{h}{2} f_0 \\ f_1 \\ \vdots \\ f_M \\ f_{M+1}\frac{h}{2} - \sigma_1 \end{bmatrix}.$$

But now the matrix $A_h$ is singular, in fact it is easy to verify that $\mathbf{v} := [1, \ldots, 1]^T \in \mathbf{R}^{M+2}$ is an eigenvector of $A_h$ with eigenvalue 0. We want to determine the condition of existence of solutions for this linear system.

Before we do that we consider the solution of (3.6). Integrating on both sides between 0 and 1 we obtain

$$u'(1) - u'(0) = \int_0^1 f(x)\,dx,$$

that is

$$\sigma_1 - \sigma_0 = \int_0^1 f(x)\,dx, \tag{3.7}$$

this is a necessary condition for the existence of solutions of the boundary value problem (3.6).

We now turn to the analysis of the solutions of the discrete boundary value problem $A_h\mathbf{U} = \mathbf{F}$. Recall that

$$\text{Range}(A_h) := \text{span}\{A_h\mathbf{e}_1, \dots A_h\mathbf{e}_{M+2}\},$$

where we denote with $\mathbf{e}_j$ $j = 1, \dots, M+2$ the canonical basis of $\mathbf{R}^{M+2}$ and then $A_h\mathbf{e}_1, \dots A_h\mathbf{e}_{M+2}$ are the columns of $A_h$.

Recall that

$$\text{Ker}(A_h^T) = \{\mathbf{z} \in \mathbf{R}^{M+2} \mid A_h^T\mathbf{z} = 0\},$$

and $\text{Ker}(A_h^T)$ is a vector space.

**Proposition.** $A_h\mathbf{U} = \mathbf{F}$ admits a solution if and only if $\mathbf{F} \in \text{Range}(A_h)$

**Proposition.** $\mathbf{F} \in \text{Range}(A_h)$ if and only if $\mathbf{F} \perp \text{Ker}(A_h^T)$.

*Proof* The proof of these two propositions is by exercise.

Since $\mathbf{F} \perp \text{Ker}(A_h^T)$ if and only if

$$\forall \mathbf{z} \in \mathbf{R}^{M+2}, \text{ s.t. } A_h^T\mathbf{z} = 0,\ \mathbf{F}^T\mathbf{z} = 0,$$

we need to characterize $\mathbf{z} = [z_0, \dots, z_{M+1}]^T$ generic element of $\text{Ker}(A_h^T)$, this amounts at finding the solutions of the homogeneous linear system $A_h^T\mathbf{z} = 0$,

$$\frac{1}{h^2}\begin{bmatrix} -h & 1 & & & & & \\ h & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & -2 & h & \\ & & & & 1 & -h \end{bmatrix}\begin{bmatrix} z_0 \\ \vdots \\ \vdots \\ \vdots \\ z_{M+1} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{bmatrix},$$

and we find that the solution is $\mathbf{z} = [z_0, hz_0, \dots, hz_0, z_0]^T \in \mathbf{R}^{M+2}$ for any value of $z_0 \in \mathbf{R}$. Then

$$\mathbf{F}^T\mathbf{z} = 0 \Leftrightarrow z_0(f_0\frac{h}{2} + \sigma_0) + hz_0\sum_{m=1}^{M} f_m + z_0(f_{M+1}\frac{h}{2} + \sigma_1) = 0,$$

for all $z_0$ and for $z_0 \neq 0$ we can divide out by $z_0$ to obtain the condition

$$\sigma_1 - \sigma_0 = f_0 \frac{h}{2} + h \sum_{m=1}^{M} f_m + f_{M+1} \frac{h}{2}.$$

We easily recognize that this condition is the discrete analogous of (3.7) and in fact the right hand side is the composite trapezoidal rule applied to the integral of $f(x)$ on the discretization grid.

## 3.2 Linear boundary value problems

Consider the general linear boundary value problem

$$a(x)u''(x) + b(x)u'(x) + c(x)u(x) = f(x), \quad a \leq x \leq b, \quad u(a) = \alpha, \quad u(b) = \beta,$$

a discretization of this problem on the grid $x_m = a + m \cdot h$, $h = \frac{b-a}{M+1}$ is

$$a_m \frac{U_{m-1} - 2U_m + U_{m+1}}{h^2} + b_m \frac{U_{m+1} - U_{m-1}}{2h} + c_m U_m = f_m, \quad m = 1, \ldots, M$$

and $a_m := a(x_m)$, $b_m := b(x_m)$, $c_m := c(x_m)$ and $U_m \approx u_m := u(x_m)$. The corresponding linear system of equations has the form

$$\frac{1}{h^2} \begin{bmatrix} h^2c_1 - 2a_1 & a_1 + h\frac{b_1}{2} & & & & \\ a_2 - h\frac{b_2}{2} & h^2c_2 - 2a_2 & a_2 + h\frac{b_2}{2} & & & \\ & \ddots & \ddots & \ddots & & \\ & & & & a_{M-1} + h\frac{b_{M-1}}{2} & \\ & & & a_m - h\frac{b_M}{2} & h^2c_M - 2a_M & \end{bmatrix} \begin{bmatrix} U_1 \\ \\ \vdots \\ \\ U_M \end{bmatrix} = \begin{bmatrix} f_1 - \alpha(\frac{a_1}{h^2} - \frac{b_1}{2h}) \\ f_2 \\ \vdots \\ f_{M-1} \\ f_M - \beta(\frac{a_M}{h^2} + \frac{b_M}{2h}) \end{bmatrix}$$

The linear system we obtain will in general be not symmetric.

### 3.2.1 A self-adjoint case

Suppose now we want to discretize

$$(\kappa(x)u'(x))' = f(x), \quad a \leq x \leq b, \quad u(a) = \alpha, \quad u(b) = \beta,$$

we could differentiate out the left hand side to obtain a problem of the class described above, i.e

$$\kappa(x)u''(x) + \kappa'(x)u'(x) = f(x)$$

this is however not the best strategy because, as we have seen, this leads in general to non symmetric linear systems. Consider now the differential operator $\mathcal{L}$ defined as follows

$$\mathcal{L}u(x) := (\kappa(x)u'(x))'$$

this operator is self-adjoint with respect to the $L_2([a, b])$ inner product,

$$\langle f, g \rangle := \int_a^b f(x)g(x)\, dx, \quad f, g \in L_2([a, b]).$$

This motivates the search for a symmetric discretization $A_h$ of $\mathcal{L}$.

To achieve this we consider the mid-points of the grid $x_{m-\frac{1}{2}} = x_m - \frac{h}{2}$ and $x_{m+\frac{1}{2}} = x_m + \frac{h}{2}$, we then consider the following approximations

$$\kappa_{m-\frac{1}{2}} u'_{m-\frac{1}{2}} = \kappa_{m-\frac{1}{2}} \frac{u_m - u_{m-1}}{h} + \mathcal{O}(h^2)$$

and

$$\kappa_{m+\frac{1}{2}} u'_{m+\frac{1}{2}} = \kappa_{m+\frac{1}{2}} \frac{u_{m+1} - u_m}{h} + \mathcal{O}(h^2).$$

We use these two to produce

$$(\kappa\, u')'\big|_{x_m} = \frac{1}{h^2}\left[\kappa_{m-\frac{1}{2}} u_{m-1} - (\kappa_{m-\frac{1}{2}} + \kappa_{m+\frac{1}{2}}) u_m + \kappa_{m+\frac{1}{2}} u_{m+1}\right] + \mathcal{O}(h^2).$$

The methods is of second order and the linear system we have to solve is

$$\frac{1}{h^2}\begin{bmatrix} -(\kappa_{\frac{1}{2}} + \kappa_{\frac{3}{2}}) & \kappa_{\frac{3}{2}} & & & & \\ \kappa_{\frac{3}{2}} & -(\kappa_{\frac{3}{2}} + \kappa_{\frac{5}{2}}) & \kappa_{\frac{5}{2}} & & & \\ & \ddots & \ddots & \ddots & & \\ & & & & \kappa_{M-\frac{1}{2}} & \\ & & & \kappa_{M-\frac{1}{2}} & -(\kappa_{M-\frac{1}{2}} + \kappa_{M+\frac{1}{2}}) & \end{bmatrix} \begin{bmatrix} U_1 \\ \vdots \\ \vdots \\ U_M \end{bmatrix} = \begin{bmatrix} f_1 - \alpha \frac{\kappa_{\frac{1}{2}}}{h^2} \\ f_2 \\ \vdots \\ f_{M-1} \\ f_M - \beta \frac{\kappa_{M+\frac{1}{2}}}{h^2} \end{bmatrix}.$$

## 3.3 A nonlinear example

We consider the pendulum equations

$$\theta''(t) + \sin(\theta(t)) = 0, \quad 0 < t < T, \quad \theta(0) = \alpha,\ \theta(T) = \beta, \tag{3.8}$$

with numerical discretization on the grid $t_m = m\,h$, $m = 0, \ldots, M+1$ and $h = \frac{T}{M+1}$ leading to the discretized equations

$$\frac{1}{h^2}(\Theta_{m-1} - 2\Theta_m + \Theta_{m+1}) + \sin(\Theta_m) = 0, \quad m = 1, 2, \ldots, M.$$

The system of equations is now nonlinear and we denote it in a vector form as

$$G_h(\Theta) = 0, \quad \Theta = \begin{bmatrix} \Theta_1 \\ \vdots \\ \Theta_M \end{bmatrix},$$

where $G_h(\Theta) = A_h\Theta + \sin(\Theta)$, $\sin(\Theta) := [\sin(\Theta_1), \ldots, \sin(\Theta_M)]^T$ and $A_h$ is the same as in (3.3).

The numerical solution of this system must be addressed iteratively, for example with a Newton method:

$$\Theta^{[k+1]} = \Theta^{[k]} - J_h(\Theta^{[k]})^{-1} G_h(\Theta^{[k]}),$$

where $J_h(\Theta)$ is the Jacobian of $G_h(\Theta)$, and it is easily seen that $J_h(\Theta)$ is a tridiagonal symmetric matrix such that

$$(J_h)_{i,j}(\Theta) = \begin{cases} \frac{1}{h^2} & \text{for} \quad j = i-1 \text{ or } j = i+1, \\ \frac{-2}{h^2} + \cos(\Theta_i) & \text{for} \quad j = i, \\ 0 & \text{otherwise}, \end{cases}$$

so

$$J_h(\Theta) = A_h + C(\Theta), \qquad C(\Theta) := \mathrm{diag}(\cos(\Theta_1), \ldots, \cos(\Theta_M)).$$

The truncation error is

$$\tau_h := G_h(\overrightarrow{\theta}), \qquad \overrightarrow{\theta} := \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_M \end{bmatrix}, \quad \theta_j := \theta(t_j).$$

It is easily shown by Taylor expansion that the components of $\tau_h$ satisfy

$$\tau_j = \frac{1}{12} h^2 \theta^{(4)}(t_j) + \mathcal{O}(h^4), \quad j = 1, \ldots, M,$$

this ensures that the method is consistent of order 2. As usual we want to use the connection between error $E_h := \Theta - \overrightarrow{\theta}$ and truncation error $\tau_h$ in order to prove convergence. In general this connection is a bit less manageable in the case of nonlinear problems, but in this particular example it is not too complicated.

We combine the discrete equation $G_h(\Theta) = 0$ and the equation for the truncation error $G_h(\overrightarrow{\theta}) = \tau_h$ to obtain

$$G_h(\Theta) - G_h(\overrightarrow{\theta}) = -\tau_h. \tag{3.9}$$

From (3.9), using $G_h(\Theta) = A_h\Theta + \sin(\Theta)$ we get

$$A_h E_h + \sin(\Theta) - \sin(\overrightarrow{\theta}) = -\tau_h,$$

and by Taylor expansion

$$\sin(\Theta) = \sin(\overrightarrow{\theta}) + C(\hat{\theta}) E_h,$$

where $C(\hat{\theta})$ is the diagonal matrix earlier defined, and the components $\hat{\theta}_i$ of the vector $\hat{\theta}$ belong to the open intervals $(\Theta_i, \theta(t_i))$.

Due to the stability already proven for the linear case, $A_h$ is invertible for all $h < H$ and so

$$E_h + A_h^{-1} C(\hat{\theta}) E_h = -A_h^{-1} \tau_h.$$

We here use the same notation for the vectors $E_h$ and $\tau_h$ and the corresponding piecewise constant functions defined on the discretization grid, the norms we consider in the sequel are function norms. Assuming we operate in the 2-norm, we get

$$\|E_h\|_2 \leq \|A_h^{-1}\|_2 \left[ \|C(\hat{\theta}) E_h\|_2 + \|\tau_h\|_2 \right] \leq \|A_h^{-1}\|_2 \left[ \max_{1 \leq m \leq M} |\cos(\hat{\theta}_m)| \, \|E_h\|_2 + \|\tau_h\|_2 \right]$$

so

$$(1 - \|A_h^{-1}\|_2) \, \|E_h\|_2 \leq \|A_h^{-1}\|_2 \|\tau_h\|_2.$$

We know from earlier analysis that $\|A_h^{-1}\|_2 = \frac{1}{|\lambda_1|}$ where $\lambda_1$ is the eigenvalue of $A_h$ with minimum absolute value. We also obtained the estimate $|\lambda_1| = \pi^2 + \mathcal{O}(h^2)$ and so we can also deduce that $\|A_h^{-1}\|_2 = \frac{1}{\pi^2} + \mathcal{O}(h^2)$ and for $h$ small enough $\|A_h^{-1}\|_2 < 1$ and we get

$$\|E_h\|_2 \leq \frac{\|A_h^{-1}\|_2}{1 - \|A_h^{-1}\|_2} \|\tau_h\|_2.$$

Using again the estimate for $\|A_h^{-1}\|_2$ we can obtain

$$\frac{\|A_h^{-1}\|_2}{1 - \|A_h^{-1}\|_2} = \frac{1}{\pi^2 - 1} + \mathcal{O}(h^2).$$

$$\|E_h\|_2 \leq \frac{1}{\pi^2 - 1}\|\tau_h\|_2 + \mathcal{O}(h^2)$$

and recalling that $\|\tau_h\|_2 = \frac{1}{12}h^2\|\theta^{(4)}\|_2 + \mathcal{O}(h^4)$ we get finally

$$\|E_h\|_2 \leq \frac{1}{\pi^2 - 1}\left(\frac{1}{12}h^2\|\theta^{(4)}\|_2\right) + \mathcal{O}(h^4)$$

which guarantees convergence when $h$ goes to zero, as it is easy to see that $\theta^{(4)}$ is bounded by differentiating the equation twice.

# Chapter 4

# Discretization of the heat equation

## 4.1 On the derivation of the heat equation

We will use a standard example to describe the numerical schemes for parabolic differential equations throughout the course. We are talking about the linear heat equation in one space dimension, which for example can be used to model the flow of heat on a straight homogeneous rod over time. The rod is insulated everywhere except at the two ends.


The rod in the picture has length $L = 1$, and let the coordinate $x$ describes a point along the rod. At time $t \geq 0$ the rod has temperature $u(x,t)$ at the point $x$. We can derive the differential equation by using Fourier's law. The flux of heat $\phi$ through a cross section of the rod at $x$ is proportional to the temperature gradient, such that

$$\phi = -\lambda u_x, \quad \lambda > 0,$$

and the following conservation law holds true

$$\rho c u_t + \phi_x = 0, \quad \rho \text{ is the rod's density.}$$

These two equations imply together that

$$u_t = a\, u_{xx}, \quad a = \frac{\lambda}{\rho c}.$$

By introducing scales for time, space and temperature

$$w = \frac{u}{u_0}, \quad y = \frac{x}{L}, \quad \tau = \frac{at}{L^2},$$

where $w$, $y$ and $\tau$ are dimensionless variables, $u_0$ is the characteristic temperature,and $L$ is the rod's length, we get

$$w_\tau = w_{yy}, \quad 0 < y < 1.$$

We have seen that after scaling it is always possible to assume that the space interval is $[0,1]$ and the coefficient $a$ can be set equal to 1. From now on we will usually look at the problem

$$u_t = u_{xx}, \quad 0 < x < 1,\ t > 0.$$

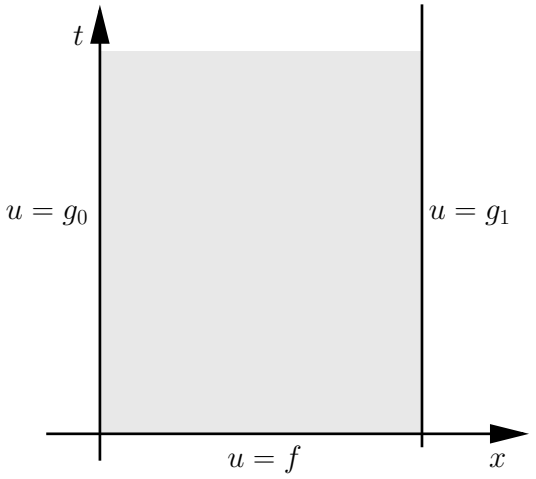
Together with the differential equation we need to provide appropriate boundary conditions and initial conditions. The kind of boundary and initial conditions necessary and sufficient to get a unique solution vary from differential equation to differential equation. We will consider few options for the heat equations.

**Pure initial value problem.** In this case we assume the rod is infinitely long.

$$\begin{aligned} u_t &= u_{xx}, && x \in \mathbf{R}, \quad t > 0, \\ u(x,0) &= f(x), && x \in \mathbf{R}. \end{aligned}$$

**Initial/Boundary value problem (I/BVP).** This case includes the situation of heat transport in a homogeneous rod of length 1. We must consider an initial function and boundary conditions at the two ends of the rod.

$$\begin{aligned} u_t &= u_{xx}, && 0 < x < 1, \ t > 0, \\ u(x,0) &= f(x), && 0 \le x \le 1, \\ u(0,t) &= g_0(t), && t > 0, \\ u(1,t) &= g_1(t), && t > 0. \end{aligned} \tag{4.1}$$




## 4.2 Numerical solution of the initial/boundary value problem

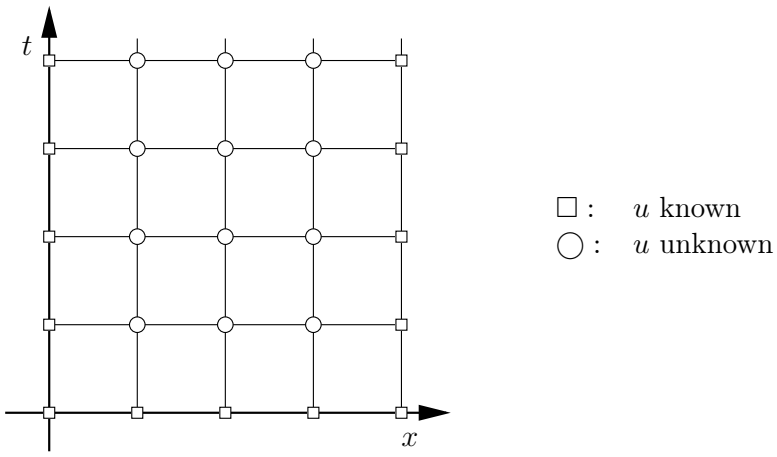
### 4.2.1 Numerical approximation on a grid.

We adopt a *step-size* $h$ in the $x$-direction, and one in the $t$-direction which we denote $k$. We assume at first that $h = 1/(M+1)$ for a given integer $M$.

We then define *gridpoints* or *nodes* $(x_m, t_n)$ by

$$x_m = mh, \ 0 \le m \le M+1, \qquad t_n = nk, \quad n = 0, 1, 2, \ldots.$$

Observe that this means that $x_0 = 0$ and $x_{M+1} = 1$ are the boundary points. The solution in the point $(x_m, t_n)$ is denoted $u_m^n := u(x_m, t_n)$. And from now on we denote with $U_m^n$ the approximation to the solution in $(x_m, t_n)$ produced by the numerical method.




### 4.2.2 Euler, Backward Euler and Crank–Nicolson

We present three different difference schemes for the heat equation.

**The Euler' s method.** We adopt a simpler notation for the derivatives and set

$$\partial_x u = u_x = \frac{\partial u}{\partial x}, \quad \partial_x^k u = \frac{\partial^k u}{\partial x^k}, \quad \partial_t u = u_t = \frac{\partial u}{\partial t}.$$

We expand $u_m^{n+1} = u(x_m, t_n + k)$ for constant $x = x_m$, around $t = t_n$, and get

$$u_m^{n+1} = u_m^n + k\partial_t u_m^n + \varphi_m^n, \qquad \varphi_m^n = \frac{1}{2}k^2\partial_t^2 u_m^n + \cdots .$$

But we can now use the heat equation ensuring $\partial_t\, u_m^n = \partial_x^2\, u_m^n$, we then approximate this second derivative with central differences as in (2.6)

$$u_m^{n+1} = u_m^n + \frac{k}{h^2}\delta_x^2 u_m^n - \psi_m^n + \varphi_m^n$$

where the index on $\delta$ means that we apply this operator in the $x$-direction i.e.

$$\delta_x^2\, u_m^n = u_{m+1}^n - 2u_m^n + u_{m-1}^n.$$

From the expression in (2.6) we find that

$$\psi_m^n = \frac{1}{12}\, k\, h^2\, \partial_x^4\, u_m^n + \cdots .$$

Summarizing we have

$$u_m^{n+1} = u_m^n + \frac{k}{h^2}\, \delta_x^2\, u_m^n + k\tau_m^n = u_m^n + \frac{k}{h^2}(u_{m+1}^n - 2u_m^n + u_{m-1}^n) + k\tau_m^n$$

where

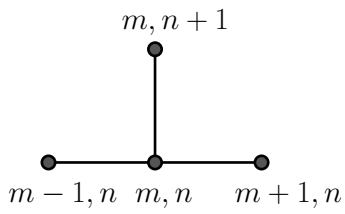
$$k\tau_m^n = \varphi_m^n - \psi_m^n = \left(\frac{1}{2}k^2\, \partial_t^2 - \frac{1}{12}k\, h^2\, \partial_x^4\right)\, u_m^n + \cdots$$

the *Euler's formula* is obtained by replacing all the exact $u$-values (small letters) with approximate values $U$ (capital letters) in the above formula and discard the term $k\tau_m^n$.

**Euler's method**

$$U_m^{n+1} = U_m^n + r\,\delta_x^2 U_m^n \qquad (4.2)$$

$$r = \frac{k}{h^2}$$




The picture above to the right is called computational molecule, it is a sort of local chart over the grid, indicating which grid-points are used in the formula. The idea is now to start at $n = 0$, corresponding to $t_0 = 0$ where $u(x, t_0) = u(x, 0) = f(x)$ which is known. It is then possible to order the values $U_m^0 = f(x_m),\ m = 0, \ldots, M+1$. Then we set $n = 1$ and use first the boundary values to get $U_0^1 = g_0(k)$ and $U_{M+1}^1 = g_1(k)$. For the remaining values we use the formula (4.2) above. It is possible to see that the grid-point at level $t_{n+1} = t_1$ in the computational molecule can be computed using known values.

**Algorithm (Euler's method for the heat equation)**

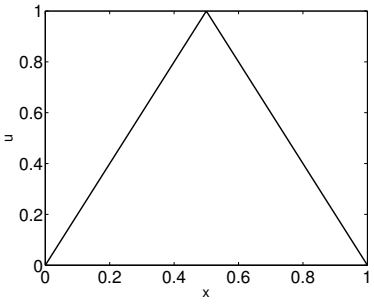
$U_m^0 := f(x_m), \quad m = 0, \ldots, M+1$
**for** $n = 0, 1, 2, \ldots$
  $U_0^{n+1} := g_0(t_{n+1})$
  $U_{M+1}^{n+1} := g_1(t_{n+1})$
  $U_m^{n+1} := U_m^n + r\,\delta_x^2 U_m^n,\ m = 1, \ldots, M$
**end**

**Example.**

$$u_t = u_{xx} \qquad 0 < x < 1,\ t > 0,$$

$$f(x) = \begin{cases} 2\,x, & 0 \le x \le \frac{1}{2}, \\ 2\,(1-x), & \frac{1}{2} < x \le 1, \end{cases}$$

$$g_0(t) = g_1(t) = 0, \qquad t > 0.$$




In the picture above we display the initial function. By running a simulation in Matlab based on this example, where we let $h = 0.1$ ($M = 9$), and $k = 0.0045$. The reason why we take $k$ so small compared to $h$ will be explained later on. Figure 4.1 shows the numerical solution in the grid-points as small rings, at time $t = 0, 5, 10$ and 20.

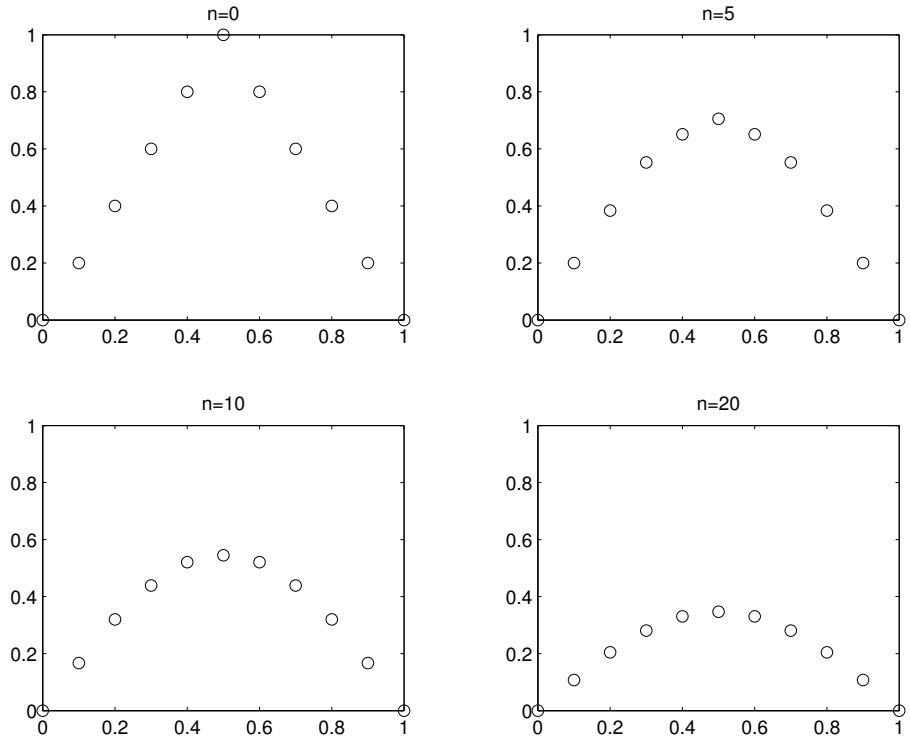



Figure 4.1: Matlab simulation of the heat equation using the Euler method

**Backward Euler.** We expand now instead $u_m^n$ around $x = x_m$, $t = t_{n+1}$, and obtain

$$\begin{aligned} u_m^n &= u(x_m, t_{n+1} - k) \\ &= u_m^{n+1} - k\,\partial_t\, u_m^{n+1} + \tfrac{1}{2}\,k^2\,\partial_t^2\, u_m^{n+1} + \cdots \\ &= u_m^{n+1} - k\,\partial_x^2\, u_m^{n+1} + \tfrac{1}{2}\,k^2\,\partial_t^2\, u_m^{n+1} + \cdots \\ &= u_m^{n+1} - k\,(\tfrac{1}{h^2}\delta_x^2\, u_m^{n+1} - \tfrac{1}{12}h^2\partial_x^4\, u_m^{n+1} + \cdots) + \tfrac{1}{2}\,k^2\,\partial_t^2\, u_m^{n+1} + \cdots \\ &= u_m^{n+1} - r\delta_x^2\, u_m^{n+1} + k\tau_m^n, \end{aligned}$$

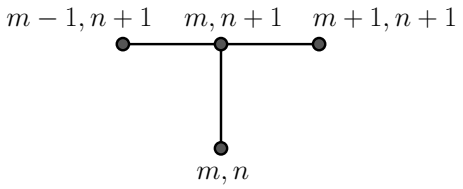
where

$$k\tau_m^n = \left(\frac{1}{12}\,k\,h^2\,\partial_x^4 + \frac{1}{2}\,k^2\,\partial_t^2\right)\, u_m^{n+1} + \cdots .$$

By replacing the $u$'s med $U$'s and discard the term $k\tau_m^n$ we obtain

**Backward Euler's method**

$$U_m^{n+1} - r\,\delta_x^2\, U_m^{n+1} = U_m^n, \qquad r = \tfrac{k}{h^2}. \tag{4.3}$$




The Backward Euler method is an implicit method. This means that at each time step we must solve a system of linear equations to compute $U_m^{n+1}$, $m = 1, \ldots, M$. We

will discuss later on the solution of linear systems. We now will present another implicit method.

**Crank–Nicolsons method.** This method is based on the trapezoidal rule, consider the following expansion of the error for the trapezoidal rule for quadrature

$$\int_0^k f(t)\,\mathrm{d}t = \frac{1}{2}\,k\,\left(f(0)+f(k)\right) - \frac{1}{12}\,k^3\,f''\left(\frac{k}{2}\right) + \cdots .$$

To derive the method we use the obvious formula

$$u(x_m,t_{n+1}) - u(x_m,t_n) = \int_{t_n}^{t_{n+1}} u_t(x_m,t)\,\mathrm{d}t,$$

and approximate the integral with the trapezoidal rule where we use the notation $u_m^{n+1/2} = u(x_m, t_n + \frac{1}{2}k)$.

$$\begin{aligned} u_m^{n+1} &= u_m^n + \frac{1}{2}\,k\,(\partial_t u_m^n + \partial_t u_m^{n+1}) - \frac{1}{12}\,k^3\,\partial_t^3\,u_m^{n+1/2} + \cdots \\ &= u_m^n + \frac{1}{2}\,k\,(\partial_x^2 u_m^n + \partial_x^2 u_m^{n+1}) - \frac{1}{12}\,k^3\,\partial_t^3\,u_m^{n+1/2} + \cdots \\ &= u_m^n + \frac{1}{2}\,k\,\left(\frac{1}{h^2}\,\delta_x^2 u_m^n + \frac{1}{h^2}\,\delta_x^2 u_m^{n+1}\right) - \frac{1}{2}\,k\,\left(\frac{1}{12}\,h^2\,\partial_x^4 u_m^n + \frac{1}{12}h^2\partial_x^4 u_m^{n+1} + \cdots\right) \\ &\quad -\frac{1}{12}\,k^3\,\partial_t^3\,u_m^{n+1/2} + \cdots . \end{aligned}$$

We simplify and summarize

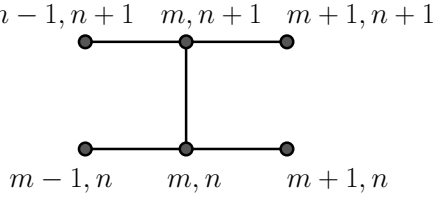
$$\begin{aligned} u_m^{n+1} &= u_m^n + \frac{r}{2}\,(\delta_x^2 u_m^n + \delta_x^2 u_m^{n+1}) + k\tau_m^n, \\ k\tau_m^n &= -\frac{1}{12}\,k^3\partial_t^3\,u_m^{n+1/2} - \frac{1}{12}\,k\,h^2\,\partial_x^4 u_m^{n+1/2} + \cdots , \end{aligned}$$

where we have used that

$$\frac{1}{2}(\partial_x^4 u_m^n + \partial_x^4 u_m^{n+1}) = \partial_x^4 u_m^{n+1/2} + \mathcal{O}(k^2).$$

**Crank–Nicolsons method**

$$(1-\tfrac{r}{2}\,\delta_x^2)\,U_m^{n+1} = (1+\tfrac{r}{2}\,\delta_x^2)\,U_m^n, \qquad r = \tfrac{k}{h^2}. \tag{4.4}$$




We summarize by writing all the formulae in a compact form

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **(E)** | $U_m^{n+1}$ | $=$ | $(1+r\,\delta_x^2)\,U_m^n$ | or | $\frac{1}{k}\Delta_t U_m^n$ | $=$ | $\frac{1}{h^2}\delta_x^2 U_m^n$, |
| **(BE)** | $(1-r\,\delta_x^2)\,U_m^{n+1}$ | $=$ | $U_m^n$ | or | $\frac{1}{k}\nabla_t U_m^{n+1}$ | $=$ | $\frac{1}{h^2}\delta_x^2 U_m^{n+1}$, |
| **(CN)** | $(1-\frac{r}{2}\,\delta_x^2)\,U_m^{n+1}$ | $=$ | $(1+\frac{r}{2}\,\delta_x^2)\,U_m^n$ | or | $\frac{1}{k}\delta_t U_m^{n+1/2}$ | $=$ | $\frac{1}{h^2}\,\delta_x^2\mu_t U_m^{n+1/2}$. |

Note that **(E)** is explicit while both **(BE)** and **(CN)** are implicit.

**The local the truncation error.** In the methods presented we have used the symbol $\tau_m^n$, this is the local truncation error in the point $(x_m, t_n)$. Given a finite difference formula, to find the corresponding local truncation error one inserts the exact solution evaluated in the grid points in the finite difference formula. Since the exact solution does not satisfy the finite difference formula, then an error term appears, this is the local truncation error. So for example for the forward Euler method applied to the heat equation we have the finite difference formula

$$\frac{U_m^{n+1} - U_m^n}{k} = \frac{U_{m-1}^n - 2U_m^n + U_{m+1}^n}{h^2},$$

replacing $U_m^n$ with the exact solution in $(x_m, t_n)$, $u_m^n$ we get

$$\frac{u_m^{n+1} - u_m^n}{k} = \frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2} + \tau_m^n,$$

so

$$\tau_m^n := \frac{u_m^{n+1} - u_m^n}{k} - \frac{u_{m-1}^n - 2u_m^n + u_m^n}{h^2}.$$

The local truncation error can expressed in powers of the step-sizes $h$ and $k$ and the derivatives of the exact solution by using the Taylor expansion.

### 4.2.3 Solution of the linear systems in Backward Euler's method and Crank–Nicolson

Our starting point is that we know $U_m^n$, $0 \le m \le M+1$, together with $U_0^{n+1}$ and $U_{M+1}^{n+1}$ given by the boundary conditions. We need to compute $U_m^{n+1}$, $1 \le m \le M$. Let us consider Crank–Nicolson first. The right hand side (r.h.s.) of the equations is known, we set

$$d_m^{n+1} = \left(1 + \frac{r}{2}\delta_x^2\right) U_m^n = \frac{r}{2} U_{m-1}^n + (1-r)\, U_m^n + \frac{r}{2} U_{m+1}^n, \quad 1 \le m \le M.$$

For the left hand side (l.h.s.) we get component-wise

$$\left(1 - \frac{r}{2}\,\delta_x^2\right) U_m^{n+1} = -\frac{r}{2} U_{m-1}^{n+1} + (1+r)\, U_m^{n+1} - \frac{r}{2} U_{m+1}^{n+1}, \quad 1 \le m \le M,$$

where we substitute $U_0^{n+1} = g_0^{n+1} = g_0(t_{n+1})$ and $U_{M+1}^{n+1} = g_1^{n+1} = g_1(t_{n+1})$. We can now express the equation in matrix-vector form

$$\begin{bmatrix} 1+r & -\frac{r}{2} & & & \\ -\frac{r}{2} & 1+r & -\frac{r}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{r}{2} & 1+r & -\frac{r}{2} \\ & & & -\frac{r}{2} & 1+r \end{bmatrix} \begin{bmatrix} U_1^{n+1} \\ U_2^{n+1} \\ \vdots \\ U_{M-1}^{n+1} \\ U_M^{n+1} \end{bmatrix} = \begin{bmatrix} d_1^{n+1} + \frac{r}{2}\, g_0^{n+1} \\ d_2^{n+1} \\ \vdots \\ d_{M-1}^{n+1} \\ d_M^{n+1} + \frac{r}{2}\, g_1^{n+1} \end{bmatrix}.$$

In a similar way we get the following linear system for Bakward-Euler

$$\begin{bmatrix} 1+2r & -r & & & \\ -r & 1+2r & -r & & \\ & \ddots & \ddots & \ddots & \\ & & -r & 1+2r & -r \\ & & & -r & 1+2r \end{bmatrix} \begin{bmatrix} U_1^{n+1} \\ U_2^{n+1} \\ \vdots \\ U_{M-1}^{n+1} \\ U_M^{n+1} \end{bmatrix} = \begin{bmatrix} U_1^n + r\, g_0^{n+1} \\ U_2^n \\ \vdots \\ U_{M-1}^n \\ U_M^n + r\, g_1^{n+1} \end{bmatrix}.$$

The two matrices in Crank-Nicolson and Backward Euler are examples of *tridiagonal* matrices. These special tridiagonal matrices are also called *Toeplitz matrices*, i.e. the elements along each of the three diagonals are equal. In Toeplitz matrices we need only to specify the first row and the first column to determine the whole matrix. If in addition we know that the matrix is symmetric, it is enough to specify the first row (or column).

In general when we solve differential equations with difference methods we obtain sparse matrices. PDEs in one space dimension and at most second order derivatives give rise typically to (at least) tridiagonal matrices. Higher order derivatives imply a larger bandwidth in the matrix. There is a clear correspondence between the bandwidth of the matrix and the number of neighboring values $U_{m-p}, \ldots, U_m, U_{m+1}, \ldots, U_{m+q}$ used to approximate the highest order derivative of $u$ with respect to $x$, at the grid point $x_m$ (see also the method of undetermined coefficients form chapter 1). In the case of several space dimensions we obtain a block structure in the matrices, for example the heat equation in two space dimensions will typically give a block-tridiagonal matrix. Toeplitz structure is lost when considering a heat equation where the rod's material is inhomogeneous (see the chapter 4.1), then the differential equation becomes

$$u_t = a(x)u_{xx}$$

where $a(x)$ is a given function.

There are special algorithms which can be used to solve linear systems with tridiagonal matrices. Among the direct methods a variant of Gaussian elimination called Thomas algorithm. We are not going to discuss this algorithm in detail here.

Instead we consider some simple examples on how it is possible to use Matlab to assemble and solve the above equations.

### 4.2.4 Solution of linear systems in Matlab

When working with sparse matrices, that is with matrices the majority of whose elements are zero, it is important to store the matrix in a cheap way in the computer. For a $M \times M$-matrix as the one considered in the previous section it is unpractical to store all the elements, it is possible instead to store a list of all the indexes corresponding to nonzero elements and the corresponding value. If you choose $M = 1000$ in the last example, the matrix will have $10^6 =$ one million elements in total, while there are about 3000 elements which are different from zero. Another issue is that if we multiply a large and sparse matrix with a vector, we will make a lot of multiplications by and additions with zero which could be avoided.

Matlab has built in facilities for this. Start Matlab and try typing `> help sparse` or `> help spdiags`. The function *sparse* converts a full matrix into a sparse matrix, menaning that only the non-zero elements are stored. The conversion of a matrix from sparse format to full format can be achieved using the function `full`. The function `spdiags` is used to generate matrices in sparse format using their diagonals. To create the matrix for the Crank–Nicolson method in sparse format, we proceed as follows. We assume $M = 10$ so that $h = 1/11$ and choose $k$ such that $r = k/h^2 = 1$. Try the following sequence of commands

```
> M=10;
> r=1;
> e=ones(M,1);
> A=spdiags([-r/2*e, (1+r)*e, -r/2*e], -1:1,M,M);
```

Try to remove ';' in the last command to see how Matlab shows a matrix in sparse format. Matlab assigns indexes to the diagonals in a matrix by giving the index 0 to the main

diagonal, the sub-diagonal gets index $-1$, the super-diagonal gets index 1 and so on. The second input argument of `spdiags` is a vector `d` of integer numbers such that the coulmn $j$ from the matrix given as the first input argument becomes the diagonal `d(j)` in the result. The two last input arguments in the call to the function specify that the result must be a $M \times M$ matrix.

Let us see how a time-step with Crank–Nicolson can be implemented in Matlab. Assume that the variable `U0` has $M+1$ elements and stores the numerical approximation at time $t = t_n$. If for example $n = 0$, `U0` is generated from the given starting values. It can be a good idea to let `U0` have dimension $M+2$ and store the boundary points respectively in `U0(1)` and `U0(M+2)`

Let us now set $n = 0$, and use the hat function as a starting value, i.e. $f(x) = 1-|2x-1|$. We define `U0` by

```
> h=1/(M+1);      % definer romskrittlengde h
> X=(0:h:1)';     % Gitterpunktene i x-retning er en kolonnevektor
> U0 = 1-abs(2*X-1); % Definer hattfunksjonen som startverdi
```

Assume the above commands are executed, such that `r,A,U0,X` are all defined.

We assume also that the boundary values are $g_0(t) = g_1(t) = 0$. Then we can make a step with Crank–Nicolson as follows

```
> d=r/2*U0(1:end-2)+(1-r)*U0(2:end-1)+r/2*U0(3:end);
> U1=[0;A\d;0];  % Numerisk losning ved tidsskritt n+1
> plot(X,U1,'o') % Plott resultatet
```

Observe that the definition of $A$ is done once and for all, and it is used in all the following time-steps. It would have been even better to LU-factorize $A$ at the beginning, so that in the subsequent steps one uses just the backward substitution algorithm.

Try to plot or make a movie of the results in time, look at `> help movie`, for example.You can also try with a bigger value of $M$ to improve accuracy and decrease the numerical error.

### 4.2.5 The $\theta$-method

It is possible to present all the three methods defined in the previous sections in a unified format by writing

$$(1 - \theta\, r\, \delta_x^2)\, U_m^{n+1} = (1 + (1-\theta)\, r\, \delta_x^2)\, U_m^n.$$

One gets then

| | |
|---|---|
| **(E)** | $\theta = 0$ |
| **(BE)** | $\theta = 1$ |
| **(CN)** | $\theta = \frac{1}{2}$ |

**Local truncation error for the $\theta$-method**

$$k\,\tau_m^n = (1 - \theta\, r\, \delta_x^2)\, u_m^{n+1} - (1 + (1-\theta)\, r\, \delta_x^2)\, u_m^n = (1 - \theta\, r\, \delta_x^2)(u_m^{n+1} - u_m^n) - r\, \delta_x^2\, u_m^n.$$

We expand all the expressions around $(x_m, t_n)$ and obtain

$$\begin{aligned} k\,\tau_m^n &= \left(1-\theta\,k\,(\partial_x^2+\frac{1}{12}\,h^2\,\partial_x^4+\cdots)\right)\left(k\,\partial_t+\frac{1}{2}\,k^2\,\partial_t^2+\frac{1}{6}\,k^3\,\partial_t^3\right)\,u_m^n-k\,\left(\partial_x^2+\frac{1}{12}\,h^2\,\partial_x^4+\cdots\right)u_m^n \\ &= \left(k\,\partial_x^2+\frac{1}{2}\,k^2\,\partial_t^2+\frac{1}{6}\,k^3\,\partial_t^3-\theta\,k^2\,\partial_t^2-\frac{1}{2}\,\theta\,k^3\,\partial^3-k\,\partial_x^2-\frac{1}{12}\,k\,h^2\,\partial_x^4+\cdots\right)\,u_m^n+\cdots \\ &= (\frac{1}{2}-\theta)\,k^2\,\partial_t^2 u_m^n-\frac{1}{12}\,k\,h^2\partial_x^4 u_m^n+(\frac{1}{6}-\frac{1}{2}\,\theta)\,k^3\,\partial_t^3 u_m^n+\cdots. \end{aligned}$$

We conclude that

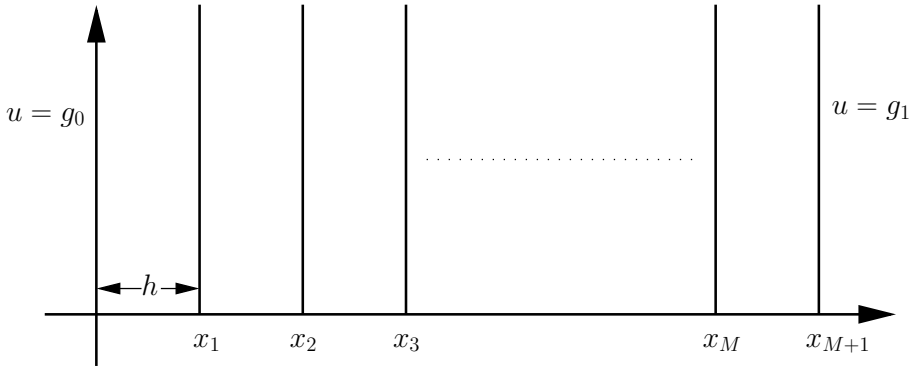
$$\begin{aligned} k\,\tau_m^n &= \mathcal{O}(k^2+k\,h^2) \quad \text{when } \theta\neq\frac{1}{2}, \\ k\,\tau_m^n &= \mathcal{O}(k^3+k\,h^2) \quad \text{when } \theta=\frac{1}{2}. \end{aligned}$$

And we then expect that **(CN)** is more accurate than **(E)** and **(BE)**.

## 4.3 Semi-discretization

### 4.3.1 Semi-discretization of the heat equation

We look again at the (I/BV) problem for the heat equation (4.1). Let us now draw vertical grid-lines as shown in the picture.




The lines are parallel to the $t$-axis and cross the $x$-axes in $x = x_m,\ m = 0, \ldots, M+1$. We consider the differential equation along such lines, this means

$$\begin{aligned} \partial_t u(x_m,t) &= \partial_x^2 u(x_m,t) \\ &= \frac{1}{h^2}\delta_x^2 u(x_m,t)+\varphi(x_m,t), \\ \varphi(x_m,t) &= -\frac{1}{12}\,h^2\partial_x^4 u(x_m,t)+\cdots. \end{aligned}$$

We introduce the functions of one variable $v_m(t),\ m = 0, \ldots, M+1$, as approximations to $u(x_m, t)$. We require that

$$\begin{aligned} v_0(t) &= g_0(t), \\ v_{M+1}(t) &= g_1(t), \\ \dot{v}_m(t) &= \frac{1}{h^2}\,\delta_x^2\, v_m(t), \quad v_m(0) = f(x_m), \quad m = 1, \ldots, M, \end{aligned}$$

where $\dot{v}_m(t) = \dfrac{\mathrm{d}v_m(t)}{\mathrm{d}t}$. For a more compact notation, let $\boldsymbol{v}(t) := [v_1(t), \ldots, v_M(t)]^T$. We get

$$\dot{\boldsymbol{v}} = \begin{bmatrix} \dot{v}_1 \\ \dot{v}_2 \\ \vdots \\ \dot{v}_{M-1} \\ \dot{v}_M \end{bmatrix} = \underbrace{\frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}}_{A} \cdot \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{M-1} \\ v_M \end{bmatrix} + \underbrace{\frac{1}{h^2}\begin{bmatrix} g_0(t) \\ 0 \\ \vdots \\ 0 \\ g_1(t) \end{bmatrix}}_{\boldsymbol{b}(t)}.$$

So we obtain a (linear) system of ordinary differential equations (ODEs) of the type

$$\dot{\boldsymbol{v}} = A\boldsymbol{v} + \boldsymbol{b}(t), \qquad \boldsymbol{v}(0) = \boldsymbol{v}_0 = [f(x_1), \ldots, f(x_M)]^T. \tag{4.5}$$

This system is a special case of the general format for ODEs which is used in standard numerical codes for ODEs. The general format is

$$\dot{\boldsymbol{v}} = \boldsymbol{F}(t, \boldsymbol{v}), \quad \boldsymbol{v}(0) = \boldsymbol{v}_0, \tag{4.6}$$

where $\boldsymbol{v}$ and $\boldsymbol{F}(t, \boldsymbol{v})$ are vectors in $\mathbf{R}^M$. Three of the simplest methods for the numerical solution of (4.6) with time-step $k$ are

| | | |
|---|---|---|
| **(E)** | Euler : | $\boldsymbol{V}^{n+1} = \boldsymbol{V}^n + k\,\boldsymbol{F}(t_n, \boldsymbol{V}^n)$ |
| **(BE)** | Backward Euler : | $\boldsymbol{V}^{n+1} = \boldsymbol{V}^n + k\,\boldsymbol{F}(t_{n+1}, \boldsymbol{V}^{n+1})$ |
| **(T)** | Trapezoidal rule : | $\boldsymbol{V}^{n+1} = \boldsymbol{V}^n + \frac{k}{2}\,\left(\boldsymbol{F}(t_n, \boldsymbol{V}^n) + \boldsymbol{F}(t_{n+1}, \boldsymbol{V}^{n+1})\right).$ |

By setting $\boldsymbol{F}(t, \boldsymbol{v}) = A\boldsymbol{v} + \boldsymbol{b}(t)$ from (4.5), it is possible to reproduce the three methods presented earlier. In particular **(T)** becomes **(CN)**.

### 4.3.2 Semidiscretization principle in general

This strategy can be applied also to differential equations other than the heat equation. For such equations we replace all derivatives with difference approximations, while the time, $t$, remains continuous. The result is

$$\text{PDE} \quad \longrightarrow \quad \text{System of ODEs}$$

An advantage with this approach is that now it is possible to exploit off-the-shelf software for ODEs, where advanced routines for error and step-size control are already incorporated.

In particular the method can be of interest in case of nonlinear PDEs, because in that case also the semi-discrertized ODE problem will be nonlinear, and standard ODE software is designed to solve numerically also such problems in an efficient way.

A problem often encountered is that the resulting ODE system is *stiff*. For the linear semi-discrete system (4.5) this means typically that the eigenvalues $\lambda_1, \ldots, \lambda_M$ of $A$ have negative real part and the quotient

$$\alpha = \frac{\max_i |\mathrm{Re}\,\lambda_i|}{\min_i |\mathrm{Re}\,\lambda_i|},$$

is relatively large. If $A$ is the tridiagonal matrix reported above then

$$\lambda_s = -\frac{4}{h^2}\,\sin^2\frac{s\pi}{2(M+1)}, \quad m = 1, \ldots, M,$$

that is all the eigenvalues are real. For small values of $x$, $\sin x \approx x$, such that, for big $M$, the eigenvalue with the smallest absolute value is

$$|\lambda_1| \approx \frac{4}{h^2}\frac{\pi^2 h^2}{2^2} = \pi^2,$$

and the eigenvalue with the biggest absolute value is

$$|\lambda_M| = -\frac{4}{h^2}\sin^2\frac{M\pi}{2(M+1)} \approx \frac{4}{h^2}\sin^2\frac{\pi}{2} = \frac{4}{h^2}.$$

Then we get $\alpha \approx \frac{4}{\pi^2 h^2} \gg 1$ when $h$ is small.

Later on we will see that this fact implies that **(BE)** and **(CN)** perform better than **(E)**.

### 4.3.3 General approach

Abstractly we can write a partial differential equation (evolution equation) in the form

$$\partial_t u = Lu,$$

where $L$ is a differential operator with space derivatives. For example the heat equation is obtained by considering $L = \partial_x^2$. Generally, in one space dimension, we have

$$L = L(x, t, \partial_x, \partial_x^2, \ldots).$$

The semi-discretization leads to

$$L \qquad \longrightarrow \qquad L_h,$$

that is, $L_h$, is a discrete operator now acting on the components of a vector of functions in one variable instead of functions of two variables. For each component we write

$$\partial_t u(x_m, t) = L_h\, u(x_m, t) + \varphi(x_m, t),$$

where $\varphi(x_m, t)$ is the truncation error in the space direction. We let now $v_m(t) \approx u(x_m, t)$ and define

$$\dot{v}_m(t) = L_h v_m(t), \quad \text{(including the boundary conditions)}.$$

We will next look at the truncation error due to the time-discretization, that is *after* having chosen an integration method for ODEs. Assume we use the trapezoidal rule for example.

Let $y(t)$ be the solution of

$$\dot{y} = F(t, y), \qquad y \in \mathbf{R}^M.$$

It is possible to show that with step-size $k$ this gives

$$y_m^{n+1} := y_m(t_{n+1}) = y_m^n + \frac{k}{2}\left(F_m(t_n, y^n) + F_m(t_{n+1}, y^{n+1})\right) + \psi_m^n,$$

where

$$\psi_m^n = -\frac{1}{12}\, k^3\, y_m^{(3)}(t_n) + \cdots .$$

But $u(x_m, t)$ is such that $\partial_t u(x_m, t) = L_h u(x_m, t) + \varphi(x_m, t)$. Let us insert $y_m(t) = u(x_m, t)$ in the general ODE-formulation above, such that

$$F_m(t, y) = L_h u(x_m, t) + \varphi(x_m, t),$$

i.e., a system of ODEs whose solution is the solution of the PDE problem along the vertical lines $(x = x_m, t)$. We get then

$$u_m^{n+1} = u_m^n + \frac{k}{2}\left(L_h u_m^n + \varphi_m^n + L_h u_m^{n+1} + \varphi_m^{n+1}\right) + \psi_m^{n+1} = u_m^n + \frac{k}{2}\left(L_h u_m^n + L_h u_m^{n+1}\right) + k\,\tau_m^n,$$

where

$$k\,\tau_m^n = \frac{k}{2}(\varphi_m^n + \varphi_m^{n+1}) + \psi_m^n.$$

Note that this is true in general for the semi-discretization principle. In particular the result is consistent with what we know for the three methods **(E)**, **(BE)** and **(CN)** applied to the heat equation.

### 4.3.4 $u_t = Lu$ with different choices of $L$

**Case A.**

$$u_t = \underbrace{a(x)\, u_{xx} + b(x)\, u_x + c(x)\, u}_{Lu}.$$

We can also write

$$L = a\partial_x^2 + b\partial_x + c.$$

Requirements: $a, b$ og $c$ are continuous in $[0, 1]$,
$a(x) > 0$ in $[0, 1]$.

Space-discretization

| | | Discretization | Truncation error |
|---|---|---|---|
| $u_{xx}$ | $\longrightarrow$ | $\frac{1}{h^2}\delta_x^2 u$ | $\mathcal{O}(h^2)$ |
| $u_x$ | $\longrightarrow$ | $\frac{1}{h}\Delta_x u$ | $\mathcal{O}(h)$ |
| | | $\frac{1}{h}\nabla_x u$ | $\mathcal{O}(h)$ |
| | | $\frac{1}{2h}(u(x+h) - u(x-h)) = \frac{1}{h}\,\mu\,\delta_x u$ | $\mathcal{O}(h^2)$ |

We set therefore

$$L_h u = a\,\frac{1}{h^2}\delta_x^2 u + b\left\{\begin{array}{l} \frac{1}{h}\Delta_x u \\ \frac{1}{h}\nabla_x u \\ \frac{1}{h}\,\delta_x\,\mu\,u \end{array}\right\} + c\,u.$$

We get $Lu = L_h u + \varphi$ where

$$\varphi = -\frac{1}{12}\,a\,h^2\partial_x^4 u - b\left\{\begin{array}{c} -\frac{1}{2}\,h\,\partial_x^2 u \\ \frac{1}{2}\,h\,\partial_x^2 u \\ -\frac{1}{6}\,h^2\,\partial_x^3 u \end{array}\right\}.$$

The choice of $\Delta_x$ versus $\nabla_x$ is called upwind/downwind- differencing. One of these two is chosen in the case of the so called convection dominated problems. The sign of $b$ determines if one should use $\Delta_x$ or $\nabla_x$. $b > 0 \to \Delta$, $b < 0 \to \nabla$.

**Case B.** Consider now the equation

$$u_t = \underbrace{(a(x)\,u_x)_x}_{Lu}, \qquad L = \partial_x(a\partial_x).$$

$L$ is *self-adjoint*. In particular this means that if you use the inner product between differentiable functions which are 0 in the endpoints 0 and 1, defined by

$$\langle u, v\rangle = \int_0^1 u(x)v(x)\,\mathrm{d}x,$$

so we get $\langle Lu, v\rangle = \langle u, Lv\rangle$ for all $u, v$. If we look at an analogous situation with the inner product on $\mathbf{R}^n$

$$\langle x, y\rangle = y^T x$$

and replace $L$ with a matrix, then the analogous condition becomes

$$y^T Ax = \langle Ax, y\rangle = \langle x, Ay\rangle = y^T A^T x,$$

for all $x, y \in \mathbf{R}^n$, implying that $A = A^T$, i.e. that $A$ is symmetric.

A possible idea is to expand $L$ by using the product rule of differentiation,

$$u_t = au_{xx} + a'\,u_x.$$

So that we get an equation of the same type as in the case A with $b = a'$ and $c = 0$. Another (and usually better) possibility is to discretize directly the original form, we let

$$\begin{aligned}\partial_x(a\,\partial_x)u(x_m, t) \quad\longrightarrow\quad & \frac{1}{h}\,\delta_x(a\,\frac{1}{h}\,\delta_x U)_m = \frac{1}{h^2}\,\delta_x(a_m(U_{m+1/2} - U_{m-1/2})) \\ & = \frac{1}{h^2}\left(a_{m+1/2}(U_{m+1} - U_m) - a_{m-1/2}(U_m - U_{m-1})\right).\end{aligned}$$

The truncation error is $\mathcal{O}(h^2)$.

*A method by Tikhonov and Samarski.* Write the equation in conservative form

$$1)\ u_t + w_x = 0, \qquad 2)\ w = -au_x.$$

We discretize 1) by

$$\partial_t u_m = -\partial_x w_m \approx -\frac{1}{h}\delta_x w_m = -\frac{1}{h}(w_{m+1/2} - w_{m-1/2}).$$

For the other equation we get $u_x = -w/a$ and

$$\int_{x_m}^{x_{m+1}} u_x \,\mathrm{d}x = -\int_{x_m}^{x_{m+1}} \frac{w}{a}\,\mathrm{d}x \approx -w_{m+1/2}\int_{x_m}^{x_{m+1}} \frac{\mathrm{d}x}{a}.$$

Set now

$$A_m = \frac{1}{h\int_{x_m}^{x_{m+1}} \frac{\mathrm{d}x}{a}}.$$

And therefore

$$\frac{1}{h}w_{m+1/2} \approx -A_m(u_{m+1} - u_m), \quad \frac{1}{h}w_{m-1/2} \approx -A_{m-1}(u_m - u_{m-1}),$$

such that in the discretization of 1) we get

$$\partial_t u_m \approx A_m(u_{m+1} - u_m) - A_{m-1}(u_m - u_{m-1}),$$

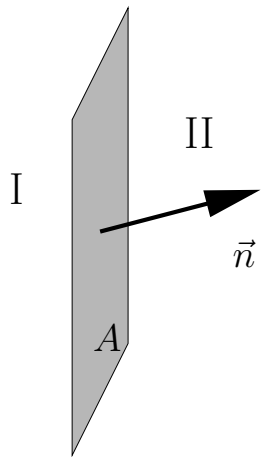
and the semi-discrete system is

$$\dot{v}_m = A_m(v_{m+1} - v_m) - A_{m-1}(v_m - v_{m-1}).$$

## 4.4 Boundary conditions involving derivatives

### 4.4.1 Different types of boundary conditions

We look at boundary conditions used for the heat equation in 3 space dimensions.




The picture illustrates the flux of heat $\phi$ through the surface $A$, with normal vector $\vec{n}$ from the side I to the side II. This flux is proportional to the directional derivative of the temperature in the direction of a normal vector pointing *outwards* the domain. We write

$$\phi = -\lambda\frac{\partial u}{\partial n} = -\lambda\,\vec{n}\cdot\nabla u.$$

You can think that the surface $A$ is a part of the surface (boundary) of the domain of definition of the problem in $\mathbf{R}^3$ where we solve the equations. We name the domain in space $\Omega$, and the boundary $\partial\Omega$.

Physical situations with boundary conditions including derivatives.

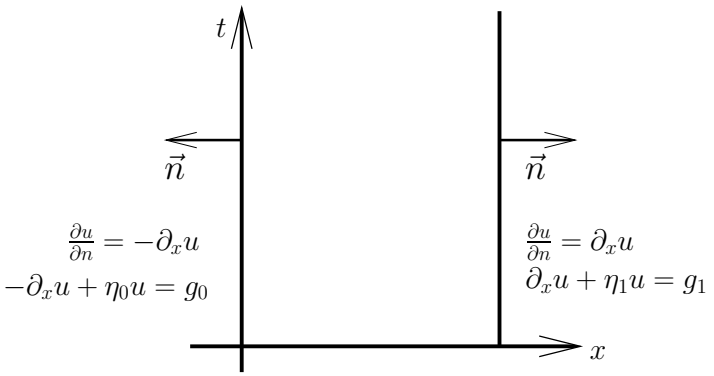
1. The heat flux given (specified) on $\partial\Omega$

$$-\lambda\frac{\partial u}{\partial n} = \phi \quad \text{given.}$$

2. Convection
$$-\lambda \frac{\partial u}{\partial n} = \alpha(u - u_0).$$
where we do not consider the boundary layer.

3. Radiation (Plancks radiation law from statistical mechanics)
$$-\lambda \frac{\partial u}{\partial n} = \sigma(u^4 - u_0^4).$$

In what follows we will consider the model
$$\frac{\partial u}{\partial n} + \eta\, u = g,$$
where $\eta \in \mathbf{R}$ and the function $g$ is defined on the boundary $\partial\Omega$. We require $\eta > 0$, and recall that the normal vector $\vec{n}$ is pointing *outwards* $\Omega$. We consider the case of one space dimension.




In one space dimension $\frac{\partial u}{\partial n} = \pm u_x$, the sign depends on where the normal vector is pointing (to the right or to the left). We call the relative Initial/Boundary value problem (I/BVP).

$$\begin{aligned}
&u_t = u_{xx},\\
&u(x,0) = f(x),\\
&-u_x(0,t) + \eta_0\, u(0,t) = g_0(t),\\
&u_x(1,t) + \eta_1\, u(1,t) = g_1(t),\\
&\eta_0, \eta_1 > 0.
\end{aligned}$$

**NB!** $u(0,t)$ and $u(1,t)$ are unknown.

Semi-discretization: we choose now $h = 1/M$ and $x_m = mh,\ 0 \le m \le M$. We get all together $M+1$ unknowns $v_0, \ldots, v_M$ where $v_m(t) \approx u(x_m, t)$.

### 4.4.2 Discretization of the boundary conditions

We look at how the derivatives in the boundary conditions can be discretized. A useful technique is to introduce "fictitious grid lines"; one to the left for the left boundary, i.e. the line $x = -h,\ t > 0$, and one to the right for the right boundary, that is the line $x = 1 + h,\ t > 0$.

**Left boundary.** We want to use a difference approximation with truncation error $\mathcal{O}(h^2)$ and try the use of central differences, which will involve the solution at the fictitious grid line outside the domain of definition of the solution of the equation.

$$
\begin{gathered}
-\partial_x u(0,t) + \eta_0\, u(0,t) = g_0(t), \\
\downarrow \\
-\frac{u_1 - u_{-1}}{2h} + \eta_0\, u_0 = g_0 + \theta_0,
\end{gathered}
$$

where

$$
\theta_0 = -\frac{1}{6}\, h^2\, \partial_x^3 u_0 + \cdots = \text{truncation error.}
$$

Here we have $u_{-1} = u(x_{-1}, t) = u(-h, t)$ outside the domain where we look for $u(x,t)$. This might seem a bit dubious, but later on we will see that $u_{-1}$ is eliminated from the discrete equations.

**Right boundary.** Similarly we get

$$
\begin{gathered}
\partial_x u(0,t) + \eta_1\, u(1,t) = g_1(t), \\
\downarrow \\
\frac{u_{M+1} - u_{M-1}}{2h} + \eta_1\, u_M = g_1 + \theta_1,
\end{gathered}
$$

where

$$
\theta_1 = \frac{1}{6}\, h^2\, \partial_x^3 u_M + \cdots .
$$

The semi-discretization is then

$$
\begin{cases}
\dot{v}_m = \dfrac{1}{h^2}\, \delta_x^2 v_m, \quad 0 \le m \le M, \\
-\dfrac{v_1 - v_{-1}}{2h} + \eta_0\, v_0 = g_0, \\
\dfrac{v_{M+1} - v_{M-1}}{2h} + \eta_1\, v_M = g_1,
\end{cases}
\tag{4.7}
$$

that is $M+3$ equations for the $M+3$ unknowns $v_{-1}, v_0, \ldots, v_M, v_{M+1}$. We eliminate $v_{-1}$ and $v_{M+1}$. From the two last equations in (4.7) we obtain the following equations for the values relative to the fictitious grid lines

$$
\begin{aligned}
v_{-1} &= v_1 - 2\, h\, \eta_0\, v_0 + 2hg_0, \\
v_{M+1} &= v_{M-1} - 2h\eta_1 v_M + 2hg_1.
\end{aligned}
$$

We substitute the above expressions in the first equation of (4.7) for $m = 0,\ m = M$

$$
\begin{aligned}
\dot{v}_0 &= \frac{1}{h^2}\, \delta_x^2 v_0 = \frac{1}{h^2}\, (v_{-1} - 2\, v_0 + v_1) \\
&= \frac{1}{h^2}\, \big( -2\, (h\, \eta_0 + 1)\, v_0 + 2\, v_1 \big) + \frac{2}{h}\, g_0,
\end{aligned}
$$

$$
\begin{aligned}
\dot{v}_M &= \frac{1}{h^2}\, \delta_x^2 v_M = \frac{1}{h^2}\, (v_{M-1} - 2\, v_M + v_{M+1}) \\
&= \frac{1}{h^2}\, \big( 2\, v_{M-1} - 2\, (h\, \eta_1 + 1)\, v_M \big) + \frac{2}{h}\, g_1.
\end{aligned}
$$

We can write this in matrix-vector notation. With $\boldsymbol{v}(t) = [v_0(t), \ldots, v_M(t)]^T$ we get

$$\dot{\boldsymbol{v}} = \frac{1}{h^2}\, Q\, \boldsymbol{v} + \frac{2}{h}\, \boldsymbol{d}, \tag{4.8}$$

where

$$Q = \begin{bmatrix} -2(h\eta_0+1) & 2 & & & & \\ 1 & -2 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 & -2 & 1 & \\ & & & 2 & -2(h\eta_1+1) & \end{bmatrix}, \qquad \boldsymbol{d} = \begin{bmatrix} g_0 \\ 0 \\ \vdots \\ 0 \\ g_1 \end{bmatrix}. \tag{4.9}$$

We note that this matrix is not symmetric, but it can be obtained from a symmetric matrix via a similarity transformation. With the diagonal matrix $D = \mathrm{diag}(\sqrt{2}, 1, \ldots, 1, \sqrt{2})$ we find in fact that $\tilde{Q} = D^{-1}QD$ is symmetric and it has therefore real eigenvalues. $\tilde{Q}$ is also *negative definite*[1].

To get an example of a fully discrete method, we can use the trapezoidal rule, and obtain Crank–Nicolson. Let $U^n = (U_0^n, \ldots, U_M^n)^T$, then

$$U^{n+1} = U^n + \frac{k}{2}\left(\frac{1}{h^2}\, Q\, U^n + \frac{2}{h}\, d^n + \frac{1}{h^2}\, Q\, U^{n+1} + \frac{2}{h}\, d^{n+1}\right),$$

or in matrix-vector notation

$$(I - \frac{r}{2}\, Q)\, U^{n+1} = (I + \frac{r}{2}\, Q)\, U^n + \frac{k}{h}(d^n + d^{n+1}).$$

**Alternative discretization of the boundary conditions.** It is possible to avoid fictitious grid lines and use difference approximations of lower order, obtaining a less accurate approximation

$$-\partial_x u_0 + \eta_0\, u_0 = g_0,$$

$$\downarrow$$

$$-\frac{u_1 - u_0}{h} + \eta_0\, u_0 = g_0 + \tilde{\theta}_0,$$

where $\tilde{\theta}_0 = \frac{1}{2} h \partial_x^2 u_0 + \cdots$, and respectively backward differences for the right boundary,

$$\partial_x u_M + \eta_1\, u_M = g_1 \quad \rightarrow \quad \frac{u_M - u_{M-1}}{h} + \eta_1\, u_M = g_1 + \tilde{\theta}_1.$$

## 4.5 Nonlinear parabolic differential equations

In general one could consider equations of the form

$$u_t = f(x, t, u, u_x, u_{xx}), \quad \frac{\partial f}{\partial u_{xx}} > 0, \quad + \text{ initial conditions \& boundary conditions.}$$

We semi-discretize the equation by introducing $x_m = mh$ and $v_m(t) \approx u(x_m, t)$, $h = 1/(M+1)$.

$$\dot{v}_m = f\left(x_m, t, v_m, \frac{1}{2h}(v_{m+1} - v_{m-1}), \frac{1}{h^2}\delta_x^2 v_m\right).$$

[1] a matrix $A$ is *negative definite* if and only if $(-A)$ is positive definite

If we include the boundary conditions, we get a system of ordinary differential equations (ODEs). Taking $v = [v_1, v_2, \ldots, v_M]^T$ and $F = [F_1, F_2, \ldots, F_M]^T$ where

$$F_m = f\left(x_m, t, v_m, \frac{1}{2h}(v_{m+1} - v_{m-1}), \frac{1}{h^2}\delta_x^2 v_m\right).$$

We have found a nonlinear system of ODEs

$$\dot{v} = F(t, v),$$

which can be solved with suitable ODE integration codes (for example in Matlab).

**Burgers' equation.**

$$u_t = \varepsilon\, u_{xx} - u u_x$$

It is possible to semi-discretize by taking

$$F_m = \frac{\varepsilon}{h^2}\,(v_{m+1} - 2v_m + v_{m-1}) - v_m\,\frac{1}{2h}\,(v_{m+1} - v_{m-1}).$$

Here it is possible for example to use a Runge-Kutta method applied to $\dot{v} = F(v)$, see for example `> help ode45` in Matlab. Note that Burgers' equations can be written in the form

$$\partial_t u = \varepsilon\,\partial_x^2 u - \frac{1}{2}\partial_x u^2,$$

which can be discretized directly with central differences, to obtain

$$F_m = \frac{\varepsilon}{h^2}\,(v_{m+1} - 2v_m + v_{m-1}) - \frac{1}{4h}(v_{m+1}^2 - v_{m-1}^2).$$

**A particular equation type.** Sometimes nonlinear partial differential equations are given in the following form

$$b(u)\,u_t = (a(u)\,u_x)_x, \quad b(u) > 0, \quad a(u) > 0.$$

Here it is possible to use the same strategy as above on the following problem

$$u_t = \frac{a(u)}{b(u)}\,u_{xx} + \frac{a'(u)}{b(u)}\,u_x^2.$$

But a better approach is to let

$$((a(u)\,u_x)_x)_m \longrightarrow \frac{1}{h^2}(\delta_x(a\,\delta_x v))_m = \frac{1}{h^2}(a_{m+1/2}(v_{m+1} - v_m) - a_{m+1/2}(v_m - v_{m-1}))$$

where

$$a_{m\pm1/2} = a(v_{m\pm1/2}) = a(v(x_m \pm h/2)),$$

these are quantities which are not defined on the grid. But we can use the following approximation

$$u_{m\pm1/2} = \frac{1}{2}(u_m + u_{m\pm1}) + \mathcal{O}(h^2),$$

so such approximation has a truncation error of the same order as the truncation error due to the approximation of the derivatives. We next define

$$\alpha_{m\pm1/2} = a\left(\frac{v_m + v_{m\pm1}}{2}\right).$$

The semi-discretization becomes now

$$b(v_m)\dot{v}_m = \frac{1}{h^2}\left(\alpha_{m+1/2}(v_{m+1} - v_m) - \alpha_{m-1/2}(v_m - v_{m-1})\right), \quad \text{truncation error: } \mathcal{O}(h^2).$$

**Crank–Nicolson for a nonlinear parabolic problem.**

$$U_m^{n+1} = U_m^n + \frac{k}{2}\left(F_m(U^n) + F_m(U^{n+1})\right),$$

where

$$F_m(U^n) = \frac{1}{b(U^n)}(\alpha_{m+1/2}\Delta_x U_m^n - \alpha_{m-1/2}\nabla_x U_m^n).$$

This means that we must solve a nonlinear equation to compute $U_m^{n+1}$. If we require the same accuracy as for Crank–Nicolson, but we would like to avoid solving nonlinear systems of equations at each time step, we could try and apply a 3-level formula. We use central differences in time and get

$$\frac{U_m^{n+1} - U_m^{n-1}}{2k} = F_m(U^n),$$

or

$$U_m^{n+1} = U_m^{n-1} + 2k\, F_m(U^n).$$

**NB! This formula is always unstable for this equation.**

**Modification.** In the unstable formula we have

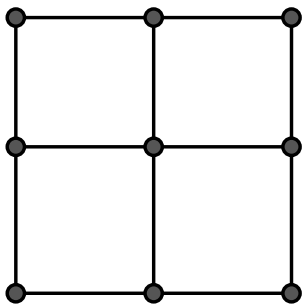
$$F_m(U^n) = \frac{1}{b(U_m^n)}\,\frac{1}{h^2}\left(\alpha_{m+1/2}\Delta_x U_m^n - \alpha_{m-1/2}\nabla_x U_m^n\right).$$

Replace

$$\begin{aligned}
\Delta_x U_m^n &\longrightarrow \frac{1}{3}(\Delta_x U_m^{n-1} + \Delta_x U_m^n + \Delta_x U_m^{n+1}),\\
\nabla_x U_m^n &\longrightarrow \frac{1}{3}(\nabla_x U_m^{n-1} + \nabla_x U_m^n + \nabla_x U_m^{n+1}).
\end{aligned}$$

We get the method

$$\begin{aligned}
U_m^{n+1} \;=\; U_m^{n-1} + \frac{2}{3}r\,\frac{1}{b(U_m^n)}\big(&\alpha_{m+1/2}(\Delta_x U_m^{n-1} + \Delta_x U_m^n + \Delta_x U_m^{n+1})\\
&-\alpha_{m-1/2}(\nabla_x U_m^{n-1} + \nabla_x U_m^n + \nabla_x U_m^{n+1})\big),
\end{aligned}$$



which originally was proposed by Lees. The picute to the left shows the computational molecule for this scheme. The formula is *linearly implicit*, i.e. only one linear system is solved per time-step. Here, as for multi-step methods for ODEs, we need a starting procedure: we need to compute $U^1$ with another method, which should have the same local truncation error as the method used for the rest of the integration. Starting procedures are needed in general in $p$-level formulae when $p > 2$.

# Chapter 5

# Stability, consistency and convergence

We are now going to analyze the numerical solution of a partial differential equation. Part of this theory is valid for general PDEs, but the examples are based on the heat equation and the methods introduced so far for such equation.

## 5.1 Properties of the continuous problem

When approximating the solution of a partial differential equation it is important that the PDE itself has a solution. A *well posed PDE problem* satisfyes the following three criteria

1. A solution of the problem exists.
2. The solution is unique.
3. The solution depends continuously on initial and boundary data.

**Example.** We consider again the I/BVP for the heat equation as an example

$$\begin{aligned} u_t &= u_{xx}, \quad 0 < x < 1,\ t > 0, \\ u(x,0) &= f(x), \quad 0 \le x \le 1, \\ u(0,t) &= g_0(t), \quad t > 0, \\ u(1,t) &= g_1(t), \quad t > 0. \end{aligned}$$

Initial and boundary data are the functions $f$, $g_0$ and $g_1$.

Assume $f$, $g_0$ and $g_1$ are continuous and that $f(0) = g_0(0), f(1) = g_1(0)$. Then the I/BVP for the heat equation has a unique solution $u(x,t)$ which is continuous for $0 \le x \le 1,\ t \ge 0$ and fulfilling the maximum principle

$$\max_{0\le x\le 1} |u(x,t)| \le \max\left\{ \max_{0\le x\le 1} |f(x)|, \quad \max_{s\le t} |g_0(s)|, \quad \max_{s\le t} |g_1(s)| \right\}. \tag{5.1}$$

More general and advanced results of this type can be found in the literature. We are not going to report the proof of this result, but rather observe how the maximum

principle (5.1) implies the properties (2) and (3) listed above. Let $u_1$ and $u_2$ be solutions with data

$$f^{(i)},\ g_0^{(i)},\ g_1^{(i)}, \quad i = 1, 2.$$

Consider

$$w = u^{(1)} - u^{(2)}, \quad \phi = f^{(1)} - f^{(2)}, \quad \gamma_0 = g_0^{(1)} - g_0^{(2)}, \quad \gamma_1 = g_1^{(1)} - g_1^{(2)}.$$

We obtain $w_t = u_t^{(1)} - u_t^{(2)} = u_{xx}^{(1)} - u_{xx}^{(2)} = w_{xx}$ so $w$ is a solution of the heat equation with data

$$w(x,0) = \phi(x), \quad w(0,t) = \gamma_0(t), \quad w(1,t) = \gamma_1(t),$$

and from (5.1) we obtain that

$$\max_{0\le x\le 1} |w(x,t)| \le \max\left\{ \max_{0\le x\le 1} |\phi(x)|, \quad \max_{s\le t} |\gamma_0(s)|, \quad \max_{s\le t} |\gamma_1(s)| \right\}. \tag{5.2}$$

So uniqueness (2) follows now, because two solutions with the same initial and boundary data will have $\phi, \gamma_0, \gamma_1$ identically equal to zero, and therefore also $w(x,t)$ will be zero. But also the property (3) follows from (5.2). The property (3) is often referred to as *stability of the differential equation.* Can we transfer this concept to the numerical solution?

## 5.2 Convergence of a numerical method

Let $u$ be the solution of a PDE problem (for example I/BVP as above) on a rectangle

$$\Omega_T = [0,1] \times [0,T] = \{(x,t) : 0 \le x \le 1,\ 0 \le t \le T\}.$$

We introduce a grid

$$G = \{(x_m, t_n),\ 0 \le m \le M,\ 0 \le n \le N\},$$

where

$$x_m = mh,\ h = \frac{1}{M}, \quad t_n = nk,\ k = \frac{T}{N}.$$

Let $U$ be defined on the grid such that $U_m^n \approx u(x_m, t_n)$. The discretization error is

$$e_m^n = u_m^n - U_m^n, \qquad u_m^n = u(x_m, t_n).$$

We say that $U \to u$ in $\Omega_T$ when $h \to 0$, $k \to 0$ if

$$\max_{0\le n\le T/k} \quad \max_{0\le m\le 1/h} |e_m^n| \to 0, \quad h \to 0,\ k \to 0.$$

In more general terms. Consider

$$U^n = [U_0^n, \ldots, U_M^n]^T,\ u^n = [u_0^n, \ldots, u_M^n]^T,\ e^n = [e_0^n, \ldots, e_M^n]^T,\ \text{vectors in } \mathbf{R}^{M+1}.$$

We choose now a vector-norm $\|\cdot\|$ defined on $\mathbf{R}^{M+1}$ for all $M \geq 0$, and say that $U \to u$ i $\Omega_T$ if

$$\max_{0 \leq n \leq T/k} \|e^n\| \to 0 \quad \text{when } k \to 0,\ h \to 0.$$

Examples of norms are

$$\|e^n\|_\infty = \max_m |e_m^n|,$$

and a scaled variant of the usual $\|\cdot\|_2$-norm

$$\|e^n\|_{2,h} = \left( h \sum_{m=0}^{M} |e_m^n|^2 \right)^{1/2} = \frac{1}{\sqrt{M}} \|e^n\|_2.$$

**Remarks.**

1. The concept of norm is a bit tricky in this context. We know that all norms on a finite-dimensional space are *equivalent* , a fact that implies in turn that convergence in one norm is equivalent to convergence in another norm. But now we must take into account that the dimension of this space goes towards infinity when $h \to 0$. We have for example the relationship

   $$\|x\|_{2,h} \leq \|x\|_\infty \leq \sqrt{M} \|x\|_{2,h}.$$

   And the vector with 1 in the first component and 0 on the others will converge to 0 in the norm $\|\cdot\|_{2,h}$ but not in the norm $\|\cdot\|_\infty$ as $h \to 0$ (and $M \to \infty$).

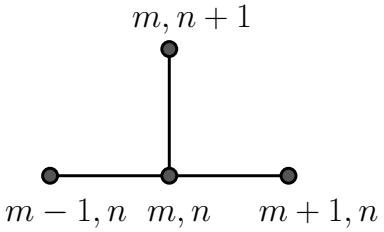
2. The scaled norm $\|\cdot\|_{2,h}$ can be interpreted as an approximation of the $L_2$-norm of an underlying continuous function,

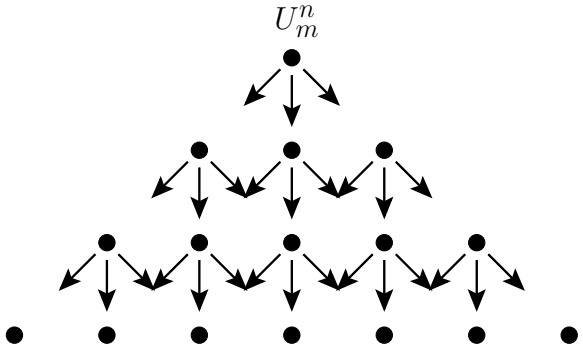
   $$\|\boldsymbol{f}\|_{2,h} = \left( h \sum_{m=0}^{M} |e_m^n|^2 \right)^{1/2} \approx \left( \int_0^1 |f(x)|^2 \, \mathrm{d}x \right)^{1/2} = \|f\|_{L_2},$$

   where $\boldsymbol{f}_m = f(mh)$; or as the exact $L_2$-norm of an underlying piecewise-constant function.

## 5.3 Domain of dependence of a numerical method

**Domain of dependence of the Euler method** We look at the computational molecule for the Euler method earlier presented, we note that the approximation $U_m^n$ in the point $(x_m, t_n)$ depends on $U_{m-1}^{n-1}$, $U_m^{n-1}$ and $U_{m+1}^{n-1}$. Each of these depends on three grid-points at the previous time level and so on. Continuing downwards in the same way, we find that $U_m^n$ depends indirectly on the points $U_{m-n}^0, \ldots, U_{m+n}^0$ if one solves a pure initial value problem (or if $0 \leq m - n,\ m + n \leq M$). The domain of dependence is in this case the triangle with vertices $U_{m-n}^0, U_{m+n}^0, U_m^n$. Generally the domain of dependence of $U_m^n$ includes all $U_\mu^\nu$ values which have been included in the computation of $U_m^n$.




In the (I/BV) problem the domain of dependence will look like in the picture below.

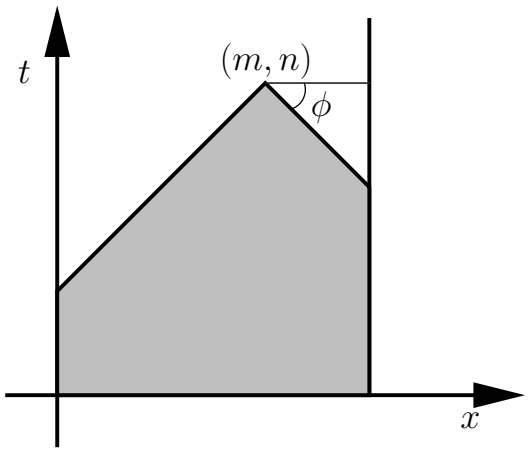
The angle

$$\phi = \arctan \frac{k}{h},$$

characterizes the domain of dependence for the Euler method. The solution of the PDE $u(x_m, t_n)$ has a domain of dependence including the whole rectangle with corners $(0,0), (0,t_n), (1,t_n), (1,0)$. If we let $r = \frac{k}{h^2}$ be constant when $h \to 0$ then

$$\phi = \arctan \frac{k}{h} = \arctan rh \to 0,$$

such that in the limit we will get the whole domain of dependence for the exact solution.




## 5.4 Proof of convergence for the Euler's method on the (I/BVP) with $r \leq \frac{1}{2}$

We consider the problem (4.1) and recall that Euler's method is

$$\begin{aligned} U_m^{n+1} &= U_m^n + r\,\delta_x^2 U_m^n, \quad 1 \leq m \leq M, n \geq 0, \\ U_0^n &= g_0^n, \quad U_{M+1}^n = g_1^n, \quad n > 0, \\ U_0^m &= f_m, \qquad 0 \leq m \leq M+1. \end{aligned} \tag{5.3}$$

The exact solution satisfies

$$u_m^{n+1} = u_m^n + r\,\delta_x^2 u_m^n + k\tau_m^n, \tag{5.4}$$

and $\tau_m^n$ is the local truncation error. We define now $e_m^n = u_m^n - U_m^n$, and subtract (5.3) from (5.4). We get

$$e_m^{n+1} = e_m^n + r\,\delta_x^2 e_m^n + k\tau_m^n, \; n > 0, 1 \leq m \leq M. \tag{5.5}$$

Moreover we have $e_m^0 = 0$ and $e_0^n = e_{M+1}^n = 0$. We know that for the Euler's method it holds that

$$k\tau_m^n = \frac{1}{2}\,k^2\,\partial_t^2 u_m^n - \frac{1}{12}\,k\,h^2\,\partial_x^4 u_m^n + \cdots,$$

and it seems reasonable assuming enough regularity of the exact solution, to assume that there exists a constant $A$ such that

$$|\tau_m^n| \le A(k+h^2), \quad \text{for all } m,n.$$

We write out explicitely (5.5) and get

$$e_m^{n+1} = r\, e_{m-1}^n + (1-2r)e_m^n + r\, e_{m+1}^n + k\tau_m^n.$$

When we later on take the absolute value, we will make use of the hypothesis that $0 \le r \le \frac{1}{2}$ because this implies that both $r$ and $1-2r$ are non-negative. We obtain therefore

$$\begin{aligned} |e_m^{n+1}| &\le r\,|e_{m-1}^n| + (1-2r)\,|e_m^n| + r\,|e_{m+1}^n| + A\,(k^2+kh^2) \\ &\le \max_\ell |e_\ell^n|(r+(1-2r)+r) + A\,(k^2+kh^2) \\ &= \max_\ell |e_\ell^n| + A\,(k^2+kh^2). \end{aligned}$$

Denoting now with $E^n = \max_\ell |e_\ell^n|$ we find that

$$E^{n+1} \le E^n + A\,(k^2+kh^2),$$

and since $E^0 = 0$ is

$$E^n \le n\,k\,A\,(k+h^2) = t_n\,A\,(k+h^2) \le T\,A\,(k+h^2),$$

that is

$$\max_\ell |e_\ell^n| \le T\,A\,(k+h^2) \quad \text{for all } n \le T/k.$$

We conclued that Euler's method converges, that is $U \to u$ when $h \to 0$ and $k \to 0$ for constant $r \le \frac{1}{2}$.

## 5.5 Stability on unbounded time interval ($F$-stability)

Let us write the $\theta$-method (including **(E)**, **(BE)** **(CN)**) in a matrix form. We get

$$(1-\theta\, r\, \delta_x^2)U_m^{n+1} = (1+(1-\theta)\, r\, \delta_x^2)U_m^n.$$

Let $S$ be the matrix

$$S = \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}. \tag{5.6}$$

We define the vector $U^n = [U_1^n, \ldots, U_m^n]^T$ for $n = 0, 1, \ldots$ we can write the $\theta$-method in a vector-form as

$$(I - \theta\, r\, S)U^{n+1} = (I + (1-\theta)\, r\, S)U^n + d^n,$$

where

$$d^n = [\theta\, r\, g_0^{n+1} + (1-\theta)\, r\, g_0^n,\ 0,\ \ldots,\ 0,\ \theta\, r\, g_1^{n+1} + (1-\theta)\, r\, g_1^n]^T.$$

Typically the difference scheme can be written as

$$A\,U^{n+1} = B\,U^n + c^n. \tag{5.7}$$

$A$ and $B$ will depend on $h$ and $k$ since the elements are functions of $h$ and $k$, but also because the matrix dimension typically is $M \times M$ where $M = 1/h$ (or $M = 1/h - 1$).

$A$ and $B$ *can* depend on $n$, but we assume here this does not happen.

A method is called a *one-step* method (ODE-terminology) or alternatively *two-level-method* (PDE-terminology) if it can be written in the form

$$U^{n+1} = C\,U^n + q^n. \tag{5.8}$$

If $A$ in (5.7) is invertible, we can take $C = A^{-1}B$ and $q^n = A^{-1}d^n$.

**Definition of $F$-stability.** (stability on $(0, \infty)$)

For an arbitrary vector $w^0$, compute the sequence $w^{n+1} = C\,w^n,\ n = 0, 1, \ldots$. Choose a vector norm $\|\cdot\|$. We say that (5.8) is $F$-stable if there exists a constant $L$ independent on $n$ such that

$$\|w^n\| \le L\|w^0\| \qquad \text{for all } \ w^0.$$

Note that the stability concept here considered has nothing to do with the solution of a differential equation, but is merely a property of the difference scheme.

**Criterion for $F$-stability.**

$$\rho(C) < 1 \qquad \Rightarrow \qquad F\text{-stability} \qquad \Rightarrow \qquad \rho(C) \le 1.$$

*Without Proof.*
You can find more on $F$-stability in the Norwegian version of the note.

## 5.6 Stability on $[0, T]$ when $h \to 0$, $k \to 0$

We now change viewpoint, and consider the stability of a method approximating a PDE on a rectangle $[0, 1] \times [0, T]$. We still look at the process $n \to \infty$, but now simultaneously $k \to 0$, such that we always have an upper bound $T = nk$. We assume moreover that $h \to 0$ at the same time, so that the dimensions of the matrices increase in the process.

We call this concept simply *stability* in this case. We analyze also now a computational scheme of the type(5.8).

**Definition of stability.** We say that (5.8) is stable if and only if it exists a constant $L$ independent on $h$ and $k$ such that $w^{n+1} = Cw^n$ satisfies

$$\|w^n\| \le L\,\|w^0\| \quad \text{for all } n \le \frac{T}{k} \quad \text{and starting vectors } w^0, \tag{5.9}$$

where $\|\cdot\|$ is a vector-norm.

**Equivalent definition.** The scheme defined by (5.8) is stable if and only if there exists a constant $L$ independent of $h$ and $k$ such that

$$\|C^n\| \le L \quad \text{for all } n \le \frac{T}{k}, \tag{5.10}$$

where the matrix norm is subordinate to the vector-norm in the previous definition.

**Example.** If we use the norm

$$\|w\|_\infty = \max_m |w_m| \qquad \text{i (5.9)},$$

the corresponding subordinate norm is

$$\|C\|_\infty = \max_\ell \sum_m |C_{\ell m}| \qquad \text{i (5.10)}.$$

*Proof* of the equivalence of (5.9) and (5.10). Assume that (5.9) holds true. Let $h$, $k$ and $n$ be arbitrary. Since the matrix norm in (5.10) is subordinate, we can find $w^0$ such that $\|C^n w^0\| = \|C^n\|\,\|w^0\|$. And so we get

$$\|C^n\|\,\|w^0\| = \|C^n w^0\| = \|w^n\| \le L\,\|w^0\| \quad \Rightarrow \quad \|C^n\| \le L$$

Assume on the other hand that (5.10) holds true, and let $w^0$, $h$, $k$ and $n$ be arbitrary, then

$$\|w^n\| = \|C^n w^0\| \le \|C^n\|\,\|w^0\| \le L\,\|w^0\|. \quad \square$$

If not specified otherwise, we will assume from now on that the matrix-norm which is used is subordinate to the vector norm in the definition (5.9).

**Sufficient criterion for stability.** If there exists a $\mu \ge 0$ independent of $h$ and $k$ such that

$$\|C\| \le 1 + \mu\,k,$$

then (5.8) is stable.

*Proof*

$$\|C^n\| \le \|C\|^n \le (1+\mu\,k)^n \le (1+\mu\,k)^{T/k} = \Big((1+\mu\,k)^{1/(\mu\,k)}\Big)^{\mu\,T} \le \mathrm{e}^{\mu\,T}.$$

We observe that the seqence $x_n = (1 + 1/n)^n$ is monotone increasing converges to $e = 2.717\cdots$ when $n \to \infty$. Such that we can write $L = \mathrm{e}^{\mu T}$ in (5.9).

**Necessary condition for stability.** If (5.8) is stable, there exists $\nu \geq 0$ independent of $h$ and $k$ such that

$$\rho(C) \leq 1 + \nu\, k, \tag{5.11}$$

where $\rho(C)$ is the spectral radius of $C$.

*Proof.* Since we assume that (5.8) is stable, it exists a constant $L$ such that $\|C^n\| \leq L$ for $n \leq T/k$. Moreover we have from (2.3) that

$$\rho(C)^n = \rho(C^n) \leq \|C^n\|.$$

So we get $\rho(C) \leq L^{1/n},\ n \leq T/k$, and in particular for $n = T/k$ we have

$$\rho(C) \leq L^{k/T} = \mathrm{e}^{k/T\, \ln L}.$$

We apply Taylor's formula with reminder to the right hand side of this inequality and we get

$$\rho(C) \leq 1 + \frac{k}{T} \ln L\, \mathrm{e}^{k/T\, \theta \ln L}, \quad \text{where } 0 < \theta < 1.$$

We use now that $\mathrm{e}^x$ is a monotone increasing function and that $\theta < 1$ and $k \leq T$, so we get

$$\rho(C) \leq 1 + \frac{k}{T} \ln L \mathrm{e}^{\ln L} = 1 + \frac{k}{T}\, L \ln L,$$

such that (5.11) is fulfilled with $\nu = \dfrac{L \ln L}{T}$. □

**Remarks.** The condition $\rho(C) \leq 1 + \nu\, k$ is in general not sufficient for stability. A somewhat artificial counterexample is obtained by considering $C = C(h) = I + F \in \mathbf{R}^{M \times M}$, $M = 1/h$, and assume that $C$ has elements equal to 1 in the elements $(i, i)$ and $(i, i-1)$ and 0 otherwise (as in a Jordan block). It is quite easy to see, for example, that $\|C^n\|_\infty = 2^n$ when $0 \leq n \leq M - 1$. This fact is sufficient to conclude that (5.8) with such choice of $C$ can not be stable. But on the other hand since $C$ is triangular and its eigenvalues coincide with its diagonal elements, then it must be $\rho(C) = 1$, and (5.11) is fulfilled.

Note also that stability is depending on the norm, it is possible that a difference method is stable in one norm, but not stable in another norm.

**A common mistake.** The following argument is wrong. Assume that the considered subordinate matrix norm is such that for diagonal matrices we have $\rho(D) = \|D\|$ (this is true for the usual norms). Assume also that $C$ is diagonalizable, $C = P \Lambda P^{-1}$.

$$\begin{aligned} \|C^n\| &= \|P \Lambda^n P^{-1}\| \leq \|P\|\, \|\Lambda^n\|\, \|P^{-1}\| = \|P\|\, \|P^{-1}\|\, \rho(C)^n \\ &\leq \|P\|\, \|P^{-1}\|\, (1 + \nu k)^n \leq \|P\|\, \|P^{-1}\|\, \mathrm{e}^{\nu T}, \end{aligned}$$

so we have at the same time found a bound for $\|C^n\|$. The problem is however that $P$ can depend on $h, k$ in such a way that $\|P\|\ (\|P^{-1}\|) \to \infty$ when $h, k \to 0$.

**Important special case.** If $C$ is symmetric the condition $\rho(C) \leq 1 + \nu\, k$ is both necessary and sufficient for stability when we use $\|\cdot\|_{2,h}$. For symmetric matrices we have namely $\|C\|_{2,h} = \rho(C)$.

**Stability of the $\theta$-method for (I/BVP).** We recall that the $\theta$-method has the form

$$(I - \theta\, r\, S)U^{n+1} = (1 + (1-\theta)\, r\, S)U^n + d^n,$$

such that

$$C = (I - \theta\, r\, S)^{-1}\,(I + (1-\theta)\, r\, S).$$

Here we have $r = k/h^2$, and $S$ is the symmetric matrix defined in (5.6). And we have therefore the diagonalization $S = P\Lambda P^T$ where $P^T P = I$. We get also

$$\begin{aligned} I - \theta\, r\, S &= P(I - \theta\, r\, \Lambda)P^T, \\ (I + (1-\theta)\, r\, S) &= P(I + (1-\theta)\, r\, \Lambda)P^T. \end{aligned}$$

By substituting in the expression for $C$ we get

$$C = P\,\underbrace{(I - \theta\, r\, \Lambda)^{-1}(I + (1-\theta)\, r\, \Lambda)}_{\Delta}\,P^T = P\Delta P^T.$$

Now $\Delta$ is a diagonal matrix with real elements

$$\Delta_m = \frac{1 + (1-\theta)\, r\lambda_m}{1 - \theta\, r\lambda_m}$$

From the diagonalization it is clear that $C$ is also symmetric so it is enough to require that $\rho(C) \leq 1 + \nu\, k$ for a $\nu \geq 0$. From before we know that

$$\lambda_m = -4\sin^2\phi_m, \quad \phi_m = \frac{m\pi}{2(M+1)}, \quad m = 1, \ldots, M,$$

and therefore

$$\Delta_m = \frac{1 - 4(1-\theta)\, r \sin^2\phi_m}{1 + 4\,\theta\, r \sin^2\phi_m}.$$

We assume that $0 \leq \theta \leq 1$, such that the expression in the numerator is $\leq 1$, while the denominator is such that $\geq 1$. So we have $\Delta_m \leq 1$ for all $m$. We try and require that $\Delta_m \geq -1$, substitute the expression for $\Delta_m$ in the above inequality, and multiply each side with the denominator (which is positive)

$$\begin{aligned} 1 - 4(1-\theta)\, r \sin^2\phi_m &\geq -1 - 4\,\theta\, r \sin^2\phi_m \\ &\Downarrow \\ 2\,(1 - 2\theta)\, r\, \sin^2\phi_m &\leq 1. \end{aligned}$$

If $\frac{1}{2} \leq \theta \leq 1$ the left hand side is $\leq 0$, so the inequality is satisfied for all values of $r \geq 0$. But if $0 \leq \theta < \frac{1}{2}$ we must require

$$r \leq \frac{1}{2\,(1 - 2\theta)\sin^2\phi_m}, \quad m = 1, \ldots, M.$$

The right hand side is minimal when $m = M$, i.e. $\phi_m = \frac{M\pi}{2(M+1)} = \frac{\pi}{2} - \frac{h\pi}{2}$. We get

$$\sin^2\left(\frac{\pi}{2} - \frac{h\pi}{2}\right) = \cos^2\frac{h\pi}{2},$$

so the condition must become

$$r \le \frac{1}{2(1-2\theta)\cos^2(h\pi/2)}.$$

For small values of $h$ we have $\cos^2(h\pi/2) \approx 1$ and we get a sufficient condition by substituting it with 1. In summary we get

**Stability criterium for $\theta$-method applied to (I/BVP).**

$$0 \le \theta < \tfrac{1}{2} \quad \Rightarrow \quad \text{Stable if} \quad 0 \le r \le \frac{1}{2(1-2\theta)},$$
$$\tfrac{1}{2} \le \theta \le 1 \quad \Rightarrow \quad \text{Stable for all} \quad r \ge 0.$$

**Stability of the $\theta$-method for (I/BVPD).** (Case of boundary conditions involving derivatives). If we apply the $\theta$-method on the semi-discretized system (4.8) we get a difference method which can be written in the form (5.8) with

$$C = (I - \theta\, r\, Q)^{-1}(I + (1-\theta)\, r\, Q),$$

where $Q$ is given by (4.9). Since $Q$ is not symmetric also $C$ will not be symmetric, and it is not possible to use the condition $\rho(C) \le 1 + \nu\, k$ for non symmetric matrices (because this is only necessary for stability). However in this particular case this condition is also sufficient. Here is the explanation. We have seen earlier that by using the matrix $D = \mathrm{diag}(\sqrt{2}, 1, \ldots, 1, \sqrt{2})$ we get that $\tilde{Q} = D^{-1}\, Q\, D$ is symmetric. This implies that also $\tilde{C} = D^{-1}CD$ is symmetric. We obtain that

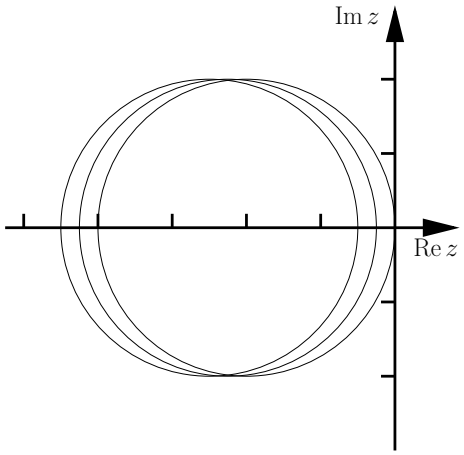
$$\|C^n\| \le \|D\|\, \|D^{-1}\|\, \|\tilde{C}^n\|.$$

For the most common norms we have that $\|D\| = \sqrt{2}$ and $\|D^{-1}\| = 1$. Since $\tilde{C}$ is symmetric we have stability if $\rho(\tilde{C}) \le 1$. But $C$ and $\tilde{C}$ are similar matrices, so they have the same eigenvalues. The stability is then guaranteed for **(I/BVPD)** if $\rho(C) \le 1$.

The matrix $C$ has eigenvalues

$$\Delta_m = \frac{1 + (1-\theta)\, r\, \lambda_m}{1 - \theta\, r\, \lambda_m}, \tag{5.12}$$

where $\lambda_m$ are the eigenvalues of $Q$. We can not find an explicit expression for the eigenvalues of $Q$, but we know they are real because $\tilde{Q}$ is symmetric. We can use the Gershgorin's theorem to get a sufficient criterion for stability.




,

The Gershgorin's discs for all rows except the first and the last are identical, see the disc furthest to the right in the picture. The first and the last discs are drawn on the left. Their center is $-2(1+\eta_i\, h),\ i = 0, 1$ and their left intersection with the real axis is $-4 - 2\eta_i\, h,\ i = 0, 1$. The eigenvalues of $Q$ are therefore on the real axis and $-4 - 2\eta\, h \le \lambda \le 0$ where $\eta = \max\{\eta_0, \eta_1\}$. From (5.12) it follows now that $\Delta_m \le 1$ for $0 \le \theta \le 1$. The condition $\Delta_m \ge -1$ gives us moreover

$$r\, \lambda_m\, (2\theta - 1) \le 2,$$

which is satisfied for $\theta \ge \frac{1}{2}$, for any $r > 0$. Let $\theta < \frac{1}{2}$.

If we had had $\lambda_m = 0$ the inequality would have hold unconditionally. For $\lambda_m < 0$ we divide on both sides by a the positive value $-\lambda_m(1 - 2\theta)$. The critical value occurs for the eigenvalue placed furthest to the left so we substitute the limit value instead $\lambda_m \ge -4 - 2\eta\, h$. In the end we get

**Stability criterion for the $\theta$-method applied to (I/BVPD).**

$$0 \le \theta < \tfrac{1}{2} \quad \Rightarrow \quad \text{Stable if} \quad 0 \le r \le \frac{1}{2(1-2\theta)(1+\frac{\eta h}{2})},$$

$$\tfrac{1}{2} \le \theta \le 1 \quad \Rightarrow \quad \text{Stable for all} \quad r \ge 0.$$

where $\eta = \max\{\eta_0, \eta_1\}$.

## 5.7 Stability and roundoff error

In numerical computations on the computer one should always take into account roundoff error, because the real numbers in a computer are represented only with a fixed and finite number of digits. When we try and compute $U^{n+1}$ from (5.8), it is in fact another quantity we really find, and it is given by

$$\tilde{U}^{n+1} = C\, \tilde{U}^n + q^n + s^n.$$

The vector $s^n$ contains the round-off error produced at step $n$. If we define the error due to round-off by $R^n = \tilde{U}^n - U^n$ we get

$$R^{n+1} = C\, R^n + s^n.$$

We can use the formula recursively and we get

$$R^n = C^n\, R^0 + C^{n-1}\, s^1 + \cdots + C\, s^{n-2} + s^{n-1}.$$

Assume $R^0 = 0$. Then we get

$$\|R^n\| \le \|C^{n-1}\|\, \|s^0\| + \cdots + \|C\|\, \|s^{n-2}\| + \|s^{n-1}\|.$$

Further we assume to have a bound $\sigma$ such that $\|s^\ell\| \le \sigma$ for all $\ell$. If (5.8) is stable we get then

$$\|R^n\| \le \sigma + \sum_{j=1}^{n-2} L\,\sigma = (1 + (n-2)L)\sigma,$$

so stability guarantees that the round-off error increases at most linearly with $n$.

## 5.8 Consistency and Lax' equivalence theorem

We recall that (5.7)

$$AU^{n+1} = BU^n + c^n, \tag{5.13}$$

which holds for difference methods applied to a linear PDE. If we substitute the exact solution of the PDE in the formula, we obtain the local truncation error as residual, we set $\tau^n = [\tau_1^n, \ldots, \tau_m^n]^T$, and have by definition

$$Au^{n+1} = Bu^n + c^n + k\,\tau^n. \tag{5.14}$$

**Consistency.** The difference method (5.13) is *consistent* (with the differential equation) if

$$\tau_m^n \to 0, \quad \text{for all } m, n \text{ når } h \to 0,\ k \to 0.$$

**Remark.** The literature is not consistent about the definition of local truncation error, and this fact influences also the definition of consistency. Alternatively it is common to define truncation error as $\widehat{\tau}_m^n = k\,\tau_m^n$ such that the condition for consistency is $\frac{1}{k}\widehat{\tau}_m^n \to 0$.

**Lax' equivalence theorem.** A consistent difference scheme is convergent if and only if it is stable.

The proof of Lax' equivalence theorem is outside the scope of this course. We will instead juts show a simpler result, namely that consistency and stability imply convergence.

**Proposition** Assume the two level difference scheme (5.13) is consistent (i.e. for $\tau_m^n$ defined by (5.14), $\tau_m^n \to 0$ when $k \to 0$ and $h \to 0$) and there exsist constants $K > 0$, $\tilde{K} > 0$ and $H > 0$ such that the inverse $A^{-1}$ exitst for all $h < H$ and $k < K$ and $\|A^{-1}\| \le \tilde{K}$ for all $h < H$ and $k < K$ in a given norm subordinate to the vector norm $\|\cdot\|$.
Assume the difference scheme can be written in the form (5.8) with $C = A^{-1}B$ and it is stable.
Then the difference scheme converges.

*Proof.* Subtracting (5.14) from (5.13) we obtain the following equation for the error

$$AE^{n+1} = BE^n - k\tau^n.$$

Exchanging $n+1$ with $n$ and multiplying by the inverse of $A$ we get

$$E^n = A^{-1}BE^{n-1} - kA^{-1}\tau^{n-1},$$

we set $q^{n-1} := -kA^{-1}\tau^{n-1}$ and simplify the above equation to obtain

$$E^n = CE^{n-1} + q^{n-1}$$

and by using the formula recursively we obtain

$$E^n = C^nE^0 + C^{n-1}q^0 + C^{n-2}q^1 + \cdots + Cq^{n-2} + q^{n-1}.$$

Taking the norm on both sides and using the assumed stability of (5.8) we get

$$\|E^n\| \leq L\sum_{s=0}^{n-1}\|q^s\|,$$

now we use that

$$\|q^s\| \leq k\,\|A^{-1}\|\,\|\tau^s\| \leq k\,\tilde{K}\,\|\tau^s\|\,,$$

and we obtain

$$\|E^n\| \leq L\,\tilde{K}\,n\,k\max_{0\leq s\leq n-1}\|\tau^s\| \leq L\,\tilde{K}\,T\max_{0\leq s\leq n-1}\|\tau^s\|,$$

which we substitute in the bound for the norm of the error and we get

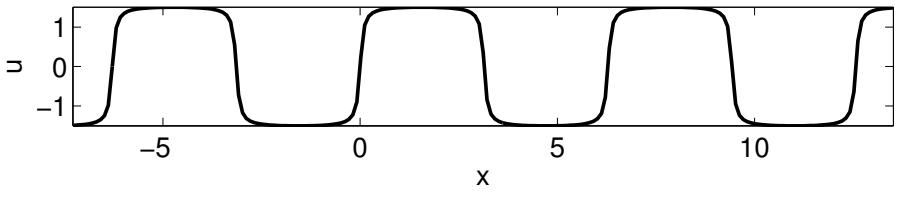
$$\|E^n\| \leq L\,T\,\tilde{K}\max_{0\leq s\leq n-1}\|\tau^s\| \to 0,$$

when $k\to 0$, $h\to 0$, and we have proved convergence.

## 5.9 von Neumann's stability criterion

We consider now again the heat equation, but we use *periodic* boundary conditions

$$\begin{aligned} u_t &= u_{xx}, \quad -\infty < x < \infty,\ t > 0,\\ u(x,0) &= f(x), \quad -\infty < x < \infty,\\ f(x+2\pi) &= f(x), \quad x\in\mathbf{R},\\ u(x+2\pi,t) &= u(x,t), \quad x\in\mathbf{R}. \end{aligned}$$




,

We can expand $f(x)$ in a Fourier series

$$f(x) = \sum_{\beta=-\infty}^{\infty} A_\beta \, \mathrm{e}^{\mathrm{i}\beta x}, \quad A_\beta = \frac{1}{2\pi} \int_0^{2\pi} f(x) \, \mathrm{e}^{-\mathrm{i}\beta x} \, \mathrm{d}x.$$

Using separation of variables, we get solutions of the form

$$u_\beta(x,t) = \mathrm{e}^{-\beta^2 t} \, \mathrm{e}^{\mathrm{i}\beta x}, \quad \beta \in \mathbf{Z}.$$

and by using the initial function $f(x)$ this gives

$$u(x,t) = \sum_{\beta=-\infty}^{\infty} A_\beta \, \mathrm{e}^{-\beta^2 t} \, \mathrm{e}^{\mathrm{i}\beta x}.$$

Let us consider an analogous analysis for the numerical solution, we use first the Euler method.

$$\begin{aligned} U_m^{n+1} &= (1 + r\,\delta_x^2) U_m^n, \quad h = \frac{2\pi}{M}, \\ U_{m+M}^n &= U_m^n \quad \text{for all } m \in \mathbf{Z}, \\ U_m^0 &= f(x_m) \quad \text{for all } m \in \mathbf{Z}. \end{aligned}$$

We can now try and write

$$U_m^n = \sum_{\beta=-\infty}^{\infty} A_\beta \, \xi^n \, \mathrm{e}^{\mathrm{i}\beta x_m}, \tag{5.15}$$

this formula fits for $U_m^0$. We check if it is possible to choose $\xi$ such that this is satisfied also for $n > 0$. It is enough to check one general term in the series, so we set

$$U_m^n = \xi^n \, \mathrm{e}^{\mathrm{i}\beta x_m}.$$

Which substituted in the Euler's method gives

$$\xi^{n+1} \, \mathrm{e}^{\mathrm{i}\beta x_m} = \xi^n \, \mathrm{e}^{\mathrm{i}\beta x_m} + r \, (\xi^n \, \mathrm{e}^{\mathrm{i}\beta x_{m-1}} - 2\,\xi^n \, \mathrm{e}^{\mathrm{i}\beta x_m} + \xi^n \, \mathrm{e}^{\mathrm{i}\beta x_{m+1}}).$$

We can assume $\xi \neq 0$ and use that $x_m = mh$, we can divide each side by $\xi^n \, \mathrm{e}^{\mathrm{i}\beta x_m}$ and we get

$$\xi = 1 + r \, (\mathrm{e}^{-\mathrm{i}\beta h} - 2 + \mathrm{e}^{\mathrm{i}\beta h}) = 1 + 2r \, (\cos \beta h - 1) = 1 - 4r \, \sin^2 \frac{\beta h}{2}.$$

We can take $\xi = \xi_\beta$ from this expression in the sum (5.15) and obtain

$$U_m^n = \sum_{\beta=-\infty}^{\infty} A_\beta \, \xi_\beta^n \, \mathrm{e}^{\mathrm{i}\beta x_m}, \quad \xi_\beta = 1 + 2r \, (\cos \beta h - 1),$$

which is the exact solution of the differential equation. $\xi_\beta$ corresponds the factor $\mathrm{e}^{-\beta^2 \, h}$ in the expression for the exact solution We should therefore require that $|\xi_\beta| \leq 1$ for all $\beta$ to ensure that the numerical solution is stable. In this particular case $\xi_\beta$ is real, and we note immediately that $\xi_\beta \leq 1$ for all $\beta$. When we require $\xi_\beta \geq -1$ we get as a condition

$$r \leq \frac{1}{2 \sin^2 \frac{\beta h}{2}}, \quad \text{for all } \beta,$$

so we must again require $r \leq \frac{1}{2}$.

**General case.** We consider 2-level difference formulae written in the form

$$\sum_{p=-r}^{r} a_p U_{m+p}^{n+1} = \sum_{p=-s}^{s} b_p U_{m+p}^{n}.$$

We look for solutions of the form

$$U_m^n \;=\; \xi^n \, \mathrm{e}^{\mathrm{i}\beta x_m},$$

which substituted into the difference formula give

$$\xi^{n+1} \, \mathrm{e}^{\mathrm{i}\beta x_m} \sum_{p=-r}^{r} a_p \, \mathrm{e}^{\mathrm{i}\beta p h} = \xi^n \, \mathrm{e}^{\mathrm{i}\beta x_m} \sum_{p=-s}^{s} b_p \, \mathrm{e}^{\mathrm{i}\beta p h},$$

and so

$$\xi \;=\; \frac{\displaystyle\sum_{p=-s}^{s} b_p \, \mathrm{e}^{\mathrm{i}\beta p h}}{\displaystyle\sum_{p=-r}^{r} a_p \, \mathrm{e}^{\mathrm{i}\beta p h}}.$$

**Von Neumann's stability criterion.** There is a constant $\mu \geq 0$ such that

$$|\xi| \leq 1 + \mu \, k.$$

**Example.** Let us now use the differential equation

$$u_t = u_{xx} - \lambda u_x.$$

We use Euler's method and central differences on $u_x$

$$U_m^{n+1} = U_m^n + \frac{k}{h^2} \, \delta_x^2 U_m^n - \lambda \, \frac{k}{2h} \, (U_{m+1}^n - U_{m-1}^n).$$

Assume $U_m^n = \xi^n \, \mathrm{e}^{\mathrm{i}\beta x_m}$ and note that $r = \dfrac{k}{h^2}$ implies that $\dfrac{k}{2h} = \dfrac{1}{2} r h$.

$$\begin{aligned} \xi \;&=\; 1 + r \, (\mathrm{e}^{-\mathrm{i}\beta h} - 2 + \mathrm{e}^{\mathrm{i}\beta h}) - \frac{\lambda r h}{2} (\mathrm{e}^{\mathrm{i}\beta h} - \mathrm{e}^{-\mathrm{i}\beta h}) \\ &=\; 1 - 4 \, r \, \sin^2 \frac{\beta h}{2} - i \, \lambda \, r \, h \, \sin \beta h \end{aligned}$$

We compute $|\xi|^2 = (\operatorname{Re} \xi)^2 + (\operatorname{Im} \xi)^2$,

$$|\xi|^2 = \left( 1 - 4 \, r \, \sin^2 \frac{\beta h}{2} \right)^2 + \lambda^2 \, r \, k \sin^2 \beta h, \qquad \text{siden } r^2 \, h^2 = r \, k.$$

In order to get $|\xi| \leq 1 + \mu\, k$ we need to have

$$\left|1 - 4\, r\, \sin^2 \frac{\beta h}{2}\right| \leq 1 + \tilde{\mu}\, k,$$

and from the Euler's method applied to $u_t = u_{xx}$ we know already that this implies $r \leq \frac{1}{2}$. But if $r \leq \frac{1}{2}$ we get

$$|\xi|^2 \leq 1 + \frac{1}{2}\lambda^2\, k,$$

and obtain by using the mean-value theorem that von Neumann's stability criterion is satisfied with $\mu = \frac{1}{4}\lambda^2$ when $r \leq \frac{1}{2}$.

**Relationship between von Neumann and the earlier definition of stability.** Assume

1. the differential equation has constant coefficients and one dependent variable; (only one $u$-component).
2. we are given a pure initial value problem;
3. we apply a 2-level difference formula;

then von Neumann's stability is necessary and sufficient for stability (as previously defined).

One might note that von Neumann's stability analysis is used as an indication of stability or instability in much more general cases.

**Example.**

$$u_t = a(x,u)\, u_{xx}, \qquad (+\text{randkrav \& startkrav}).$$

Using Euler's method we get

$$U_m^{n+1} = U_m^n + r\, a(x, U_m^n)\, \delta_x^2 U_m^n.$$

With $a$ constant the von Neumann's stability criterion has the form

$$|\xi| = \left|1 - 4\, a\, r\, \sin^2 \frac{\beta h}{2}\right| = 1 + \mathcal{O}(k).$$

Let us for example assume it is possible to bound $a$ such that $1 \leq a(x,u) \leq 2$ for all $x, u$ we are interested in. Then we could require

$$r \leq \frac{1}{2\, \max a} = \frac{1}{4},$$

without having performed any rigorous analysis.

# Chapter 6

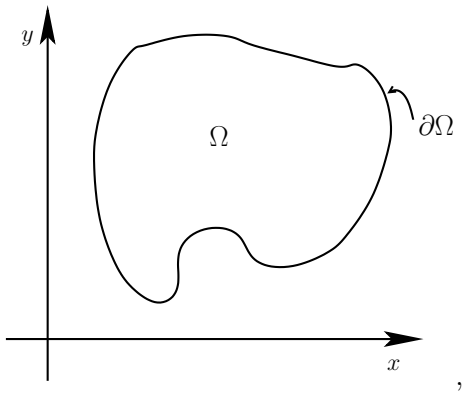
# Elliptic differential equations

## 6.1 Elliptic equation on the plane

We consider partial differential equations of the type

$$a\,u_{xx} + 2b\,u_{xy} + c\,u_{yy} = d(x, y, u, u_x, u_y), \qquad (x, y) \in \Omega$$

where $a$, $b$, $c$ and $d$ can be functions of $x$ and $y$.




,

**Ellipticity** If the functions $a$, $b$ and $c$ for all $(x, y) \in \Omega$ satisfy

$$ac - b^2 > 0$$

the differential equation is elliptic on $\Omega$.

**Example.** If $a = c = 1, b = d = 0$ we get the Laplace equation

$$u_{xx} + u_{yy} = 0.$$

**Boundary conditions.** There are three types boundary conditions

1. Dirichlet boundary conditions: $u = f$ on $\partial\Omega$.
2. Neumann boundary conditions: $\frac{\partial u}{\partial n} = \vec{n} \cdot \nabla u = g$ on $\partial\Omega$.

3. Robin boundary conditions: $\alpha u + \beta \frac{\partial u}{\partial n} = \gamma$ on $\partial\Omega$.

The boundary conditions can often be a mixture of 1–3, such that the boundary $\partial\Omega$ can be divided in subparts, and with different boundary conditions on the different parts.

**Maximum principle.** Let $\Omega$ be a open connected subset of $\mathbf{R}^2$. Define the differential operator $L$ by

$$Lu = a\,u_{xx} + 2b\,u_{xy} + c\,u_{yy} + d\,u_x + e\,u_y$$

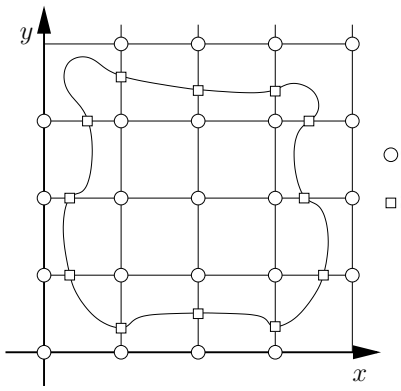
where $a$, $b$, $c$, $d$ and $e$ are functions of $x$ and $y$, and $L$ is elliptic ($ac > b^2$) in $\Omega$.

If $Lu = 0$ on $\Omega$, then $u$ can not assume a strict local maximum or minimum on $\Omega$ unless $u =$ constant on $\Omega$.

*Alternative formulation.* If $u$ is continuous on $\bar{\Omega} = \Omega \cup \partial\Omega$ and $Lu = 0$ on $\Omega$ $u$ will assume its maximum/minimum on the boundary $\partial\Omega$.

## 6.2 Difference methods derived using Taylor series

We start considering a regular grid.




Grid-lines: $x = x_\ell,\ y = y_m$

Grid-points: $P = (x_\ell, y_m)$

We look for an approximation of the solution of the elliptic PDE on a *net* made of grid-pints and points of intersection between the boundary and the grid-lines.

We define some subsets of the net

$G = \{(x_\ell, y_m)\}$ : the whole grid

$\mathcal{N}^o$: $G \cap \Omega$, set of the internal grid-points

$D = \{(x, y) \in \partial\Omega : x = x_\ell \text{ or } y = y_m\}$

$\mathcal{N} = \mathcal{N}^o \cup D$

The intersection points (between the boundary and the grid-lines) can cause problems, because often they reduce the accuracy of the numerical approximation and can also destroy the matrix structure (e.g. symmetry) and produce linear systems of equations which are difficult to solve numerically via iterative techniques.

**Grid-like net.** A net is grid-like if all the points in the set $D$ are grid-points. In the sequel we consider grid-like nets with constant step-size, i.e. $x_\ell = x_0 + \ell h,\ \ell = 0, 1, \ldots$ and $y_m = y_0 + mh,\ m = 0, 1, \ldots$.

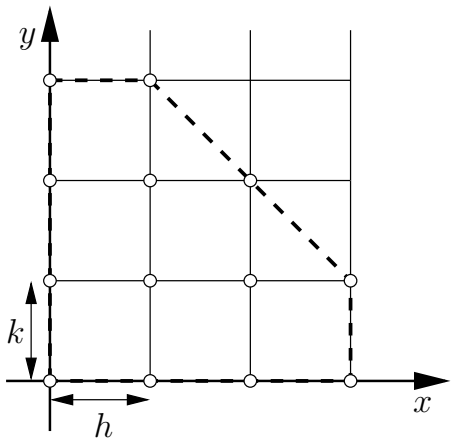
**Poisson's equation.**

$\Delta u = u_{xx} + u_{yy} = f$ i $\Omega$, $\qquad u = g(x, y)$ på $\partial\Omega$.

$$\partial_x^2 u_p = \frac{1}{h^2}\,\delta_x^2 u_p + \mathcal{O}(h^2)$$

$$\partial_y^2 u_p = \frac{1}{k^2}\,\delta_y^2 u_p + \mathcal{O}(k^2)$$

where $u_p$ is the solution of the differential equation evaluated in the point $p = (x, y)$.




If we discretize the differential equation in the point $p = (x, y)$ we get

$$\Delta u_p = f_p \longrightarrow \frac{1}{h^2}\delta_x^2 u_p + \frac{1}{k^2}\delta_y^2 u_p = f_p + \tau_p$$

where $f_p$ is the function $f$ evaluated in the point $p$, and the local truncation error $\tau_p$ is

$$\tau_p = \frac{1}{12}\,h^2\,\partial_x^4 u_p + \frac{1}{12}\,k^2\,\partial_y^4 u_p + \cdots \tag{6.1}$$

We let $U_p$ be the approximation of $u_p$ and we get the linear system

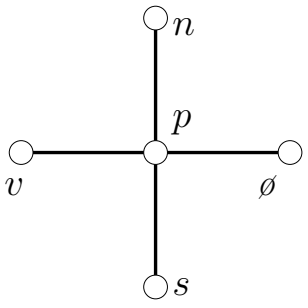
$$\begin{aligned} &\tfrac{1}{h^2}\delta_x^2 U_p + \tfrac{1}{k^2}\delta_y^2 U_p = f_p, && p \in \mathcal{N}^o, \\ &U_p = g_p && p \in D. \end{aligned}$$

**Classic 5-points formula for the Poisson equation.** In the case $k = h$ we get

$$\delta_x^2 U_p + \delta_y^2 U_p = h^2\, f_p.$$

We can rewrite this formula by using the four directions: north (N), south (S), east (E) and west (W)

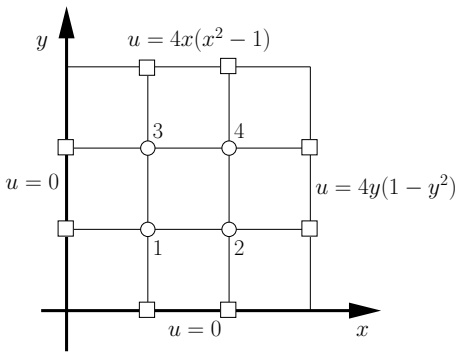
$$U_w + U_s + U_e + U_n - 4U_p = h^2\, f_p$$




**Example.** We compute in detail a concrete case.

$$\Delta u = 0, \quad (x, y) \in \Omega = (0, 1) \times (0, 1), \qquad u(x, y) = g(x, y), \quad (x, y) \in \partial\Omega$$

where $g(x, y)$ is as as described in the picture.




$$
\begin{aligned}
g(0,y) &= 0, && 0 \le y \le 1\\
g(x,0) &= 0, && 0 \le x \le 1\\
g(1,y) &= 4y(1-y^2), && 0 \le y \le 1\\
g(x,1) &= 4x(x^2-1), && 0 \le x \le 1
\end{aligned}
$$

$$
\begin{array}{lrrrrcr}
p=1: & -4\,U_1 & +U_2 & +U_3 & & = & 0\\
p=2: & U_1 & -4\,U_2 & & +U_4 & = & -\frac{32}{27}\\
p=3: & U_1 & & -4\,U_3 & +U_4 & = & \frac{32}{27}\\
p=4: & & U_2 & +U_3 & -4\,U_4 & = & 0.
\end{array}
$$

By solving these equations we get

$$U_1 = U_4 = 0, \quad U_2 = \frac{8}{27}, \quad U_3 = -\frac{8}{27}.$$

You can verify that the solution of the partial equation is

$$u(x,y) = 4xy\,(x^2 - y^2).$$

From (6.1) we see that the local truncation error is identically equal to zero because $\partial_x^4 u \equiv 0$ and $\partial_y^4 u \equiv 0$ and therefore all the higher order derivatives. So in this case the formula gives the exact solution of the problem. This is a very special case, with different boundary conditions we would get $\tau_p \neq 0$.

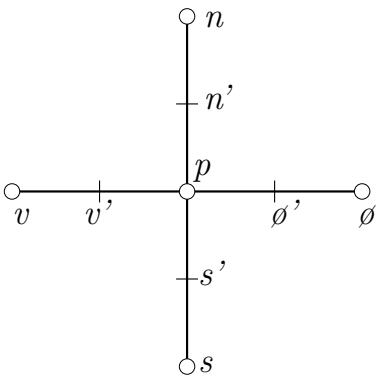
### 6.2.1 Discretization of a self-adjoint equation

We consider the problem

$$Lu = f, \qquad \text{where} \quad Lu = \partial_x(a\partial_x u) + \partial_y(c\partial_y u), \quad a = a(x,y),\ c = c(x,y).$$

$$\partial_x(a\partial_x u) = \frac{1}{h^2}\,(a_{ø'}(u_ø - u_p) - a_{v'}(u_p - u_v)) + \mathcal{O}(h^2)$$

$$\partial_y(c\partial_y u) = \frac{1}{k^2}\,(c_{n'}(u_n - u_p) - c_{s'}(u_p - u_s)) + \mathcal{O}(k^2)$$




So $Lu = f$ can be discretized to

$$\alpha_ø U_ø + \alpha_n U_n + \alpha_v U_v + \alpha_s U_s - \alpha_p U_p = f_p,$$

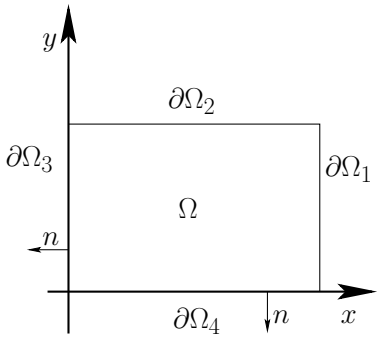
$$\text{where} \quad \alpha_p = \alpha_ø + \alpha_n + \alpha_v + \alpha_s$$

and

$$\alpha_ø = \frac{1}{h^2}a_{ø'},\ \alpha_n = \frac{1}{k^2}c_{n'},\ \alpha_v = \frac{1}{h^2}a_{v'},\ \alpha_s = \frac{1}{k^2}c_{s'}.$$

## 6.3 Boundary conditions of Neumann and Robin type

**Example.**




$$\Delta u = 0 \qquad \text{i } \Omega$$

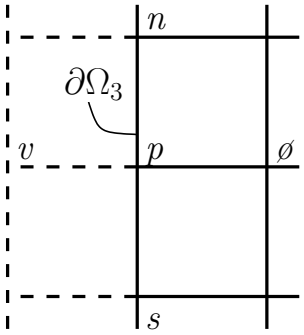
$$u = g(x,y) \qquad \text{on } \partial\Omega_1 \cup \partial\Omega_2$$

$$\partial_x u - q^{(0)}\, u = f^{(0)} \qquad \text{on } \partial\Omega_3$$

$$\partial_y u - q^{(1)}\, u = f^{(1)} \qquad \text{on } \partial\Omega_4$$

We require that $q^{(0)} \geq 0$ and $q^{(1)} \geq 0$. Now we need an equation for $U_p$ for all $p \in \mathcal{N}^o$, and for all $p$ on $\partial\Omega_3$ and $\partial\Omega_4$. Let us for the sake of simplicity use the same step in both space directions, i. e. $k = h$. Let us use the left boundary $\partial\Omega_3$ as an example,

$$\partial_x u - qu = f.$$




**Alternative 1.**

$$\partial_x u_p = \frac{u_\text{ø} - u_p}{h} + \mathcal{O}(h) \quad \longrightarrow \quad \frac{U_\text{ø} - U_p}{h} - q_p^{(0)}\, U_p = f_p^{(0)}$$

where $q_p^{(0)}$ and $f_p^{(0)}$ are the given functions $q^{(0)}$ and $f^{(0)}$ evaluated in the point $p$.

**Alternative 2.** Use the fictitious point $v$ lying outside the domain see picture.

$$\partial_x u_p = \frac{u_\text{ø} - u_v}{2h} + \mathcal{O}(h^2) \quad \longrightarrow \quad \frac{U_\text{ø} - U_v}{2h} - q_p^{(0)}\, U_p = f_p^{(0)} \tag{6.2}$$

The extra unknown $U_v$ requires an extra equation, and we use the scheme for the approximation of the differential equation in the point $p$

$$U_\text{ø} + U_n + U_v + U_s - 4U_p = 0, \tag{6.3}$$

we eliminate $U_v$ using the discretized boundary condition. From (6.2) we find

$$U_v = U_\text{ø} - 2h\,(q_p^{(0)}\, U_p + f_p^{(0)}),$$

which substituted in (6.3) gives

$$2\,U_\text{ø} + U_n + U_s - (4 + 2\,h\,q_p^{(0)})U_p = 2\,h\,f_p^{(0)}.$$

This discretization is more accurate than alternative 1.

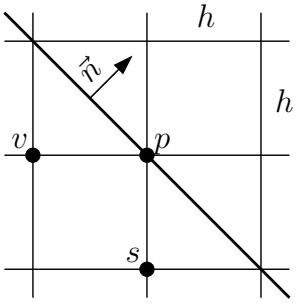
The same approach is used on the boundary $\partial\Omega_4$. For this boundary we see only at alternative 2 with fictitious boundary point $s$, and $p$on $\partial\Omega_4$:

$$\partial_y u_p = \frac{u_n - u_s}{2h} + \mathcal{O}(h^2) \quad \longrightarrow \quad \frac{U_n - U_s}{2h} - q_p^{(1)} U_p = f_p^{(1)}.$$

Here we can also use (6.3) and eliminate the fictitious value $U_s$, the result is

$$2\,U_n + U_v + U_\phi - (4 + 2\,h\,q_p^{(1)})U_p = 2\,h\,f_p^{(1)}.$$

**Boundary along the grid-diagonal**




$$\partial_n u + qu = \vec{n} \cdot \nabla u + qu = f$$

$$\vec{n} = [n_x, n_y]^T$$

$$n_x = n_y = \tfrac{1}{\sqrt{2}} \quad \text{because } k = h$$

$$\Downarrow$$

$$\partial_n u = n_x \partial_x u + n_y \partial_y u = \tfrac{1}{\sqrt{2}}(\partial_x u + \partial_y u)$$

$$\partial_n u_p = \frac{1}{\sqrt{2}} \left( \frac{u_p - u_v}{h} + \frac{u_p - u_s}{h} \right) + \mathcal{O}(h)$$

such that a first order approximation is

$$(2 + \sqrt{2}\,h\,q_p)U_p - U_v - U_s = \sqrt{2}\,h\,f_p.$$

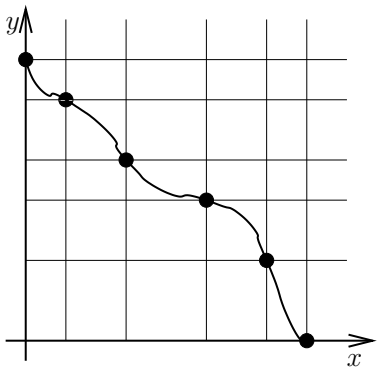
It is usual to consider such low order approximations for this kind of boundary. In principle it is possible to use fictitious grid-points as in alternative 2, but the formulae become quite complicated.

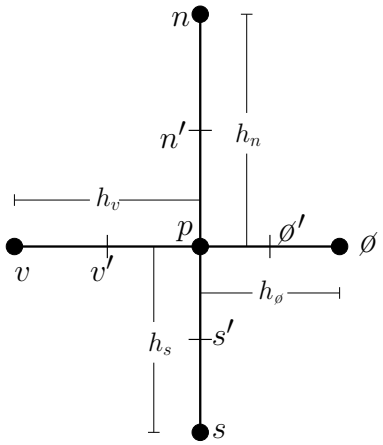
**A problem with pure Neumann boundary conditions for the Laplace's equation.** The problem

$$\begin{aligned} \Delta u &= 0, && \text{in } \Omega \\ \partial_n u &= g, && \text{on } \partial\Omega \end{aligned}$$

has a solution only if $\int_{\partial\Omega} g = 0$. This can be shown by using the divergence theorem on $\int_\Omega \Delta u$. If $u$ is a solution, we see immediately that also $u + c$ is a solution for an arbitrary constant $c$. Therefore we have not a *well posed* PDE-problem. This fact has a counterpart in the discrete problem where we solve a linear system of equations of the type $AU = b$ where $A$ is a square singular matrix. If $b$ belongs to the range of $A$ (the span of the columns of $A$) there is at least a solution (not necessarily unique as we can add any arbitrary non zero vector belonging to the null space of $A$), otherwise there is no solution of the linear system.

## 6.4 Grid-like net and variable step-size




We consider the approximation of

$$Lu = \partial_x(a\partial_x u) + \partial_y(c\partial_y u).$$

We let

$$\partial_x(a\partial_x u) \quad \longrightarrow \quad L_h^{(x)}U_p = \frac{2}{h_v + h_ø}\left(a_{ø'}\,\frac{U_ø - U_p}{h_ø} - a_{v'}\,\frac{U_p - U_v}{h_v}\right)$$

and

$$\partial_y(c\partial_y u) \quad \longrightarrow \quad L_h^{(y)}U_p = \frac{2}{h_s + h_n}\left(c_{n'}\,\frac{U_n - U_p}{h_n} - c_{s'}\,\frac{U_p - U_s}{h_s}\right)$$

We refer to the picture above to the right for the definition of the points $p, v, v', s, s', ø, ø', n, n'$ and the relative step-sizes. We approximate $L$ by

$$L_h = L_h^{(x)} + L_h^{(y)}$$

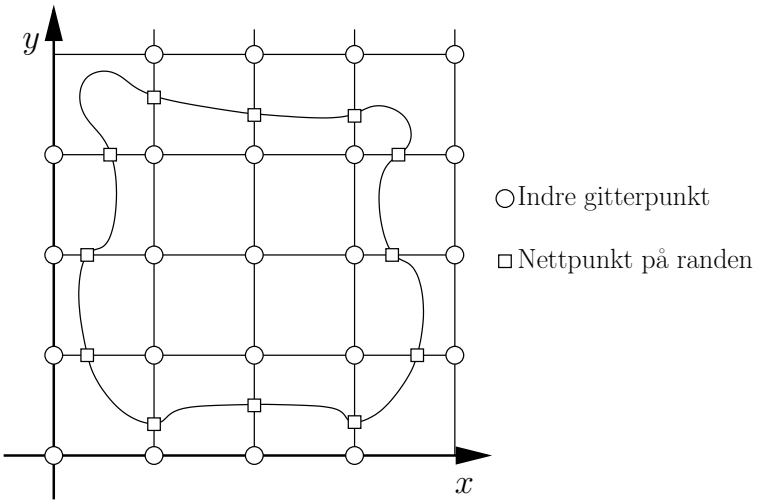
and find

$$\begin{aligned} L_h^{(x)}u_p &= \frac{2}{h_v + h_ø}\Big((1 + \frac{h_ø}{2}\partial_x + \frac{h_ø^2}{8}\partial_x^2 + \frac{h_ø^3}{48}\partial_x^3 + \cdots)(a_p\,(\partial_x + \frac{1}{24}h_ø^2\partial_x^3 + \cdots)u_p) \\ &\qquad -(1 - \frac{h_v}{2}\partial_x + \frac{h_v^2}{8}\partial_x^2 - \frac{h_v^3}{48}\partial_x^3 + \cdots)(a_p\,(\partial_x + \frac{1}{24}h_v^2\partial_x^3 + \cdots)u_p)\Big) \\ &= \partial_x(a\partial_x)u_p + \frac{1}{3}(h_ø - h_v)\partial_x^2(a\partial_x)u_p + \frac{1}{24}\frac{h_ø^3 + h_v^3}{h_ø + h_v}(\partial_x a\partial_x^3 + \partial_x^3 a\partial_x)u_p + \ldots \end{aligned}$$

By a similar calculation for $L_h^{(y)}u_p$ eventually we get

$$Lu - L_h u = \begin{cases} \mathcal{O}((h_ø - h_v) + (h_n - h_s) + h_ø^2 + h_n^2) & \text{generelt} \\ \mathcal{O}(h_ø^2 + h_n^2) & h_ø = h_v,\ h_n = h_s. \end{cases}$$

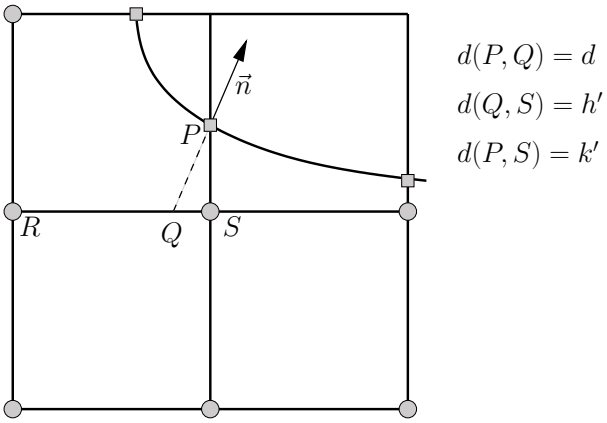
## 6.5 General rectangular net




We consider the equation $\Delta u = f$ as an example.

**Dirichlet problem.** We can use the general 5-points-formula with $h_\emptyset$, $h_v$, $h_n$, $h_s$ for all the grid-points internal to the domain.

**Robin problem.** The difficulty here is $\partial_n u = \vec{n} \cdot \nabla u$. We let the grid be rectangular with step-sizes $h$ and $k$ respectively in the $x$- and $y$-direction.




We have

$$\partial_n u_P = \frac{u_P - u_Q}{d} + \mathcal{O}(d), \qquad d = \sqrt{h'^2 + k'^2}.$$

The problem is that $Q$ is not a grid-point, but we can approximate the solution in the point $Q$ using linear interpolation. We find

$$u_Q = u_R \frac{h'}{h} + u_S \frac{h - h'}{h} + \mathcal{O}(h^2).$$

Therefore we get

$$\partial_n u_P = \frac{1}{d}\left(u_P - (u_R \frac{h'}{h} + u_S \frac{h - h'}{h})\right) + \mathcal{O}(h^2/d) + \mathcal{O}(d).$$

But we note that this type of discrete problem is generally difficult to derive and handle.

## 6.6 Discretization using Taylor expansion on a completely general net

We want to set up a discrete version of the operator

$$Lu = au_{xx} + 2bu_{xy} + cu_{yy} + du_x + eu_y + fu.$$

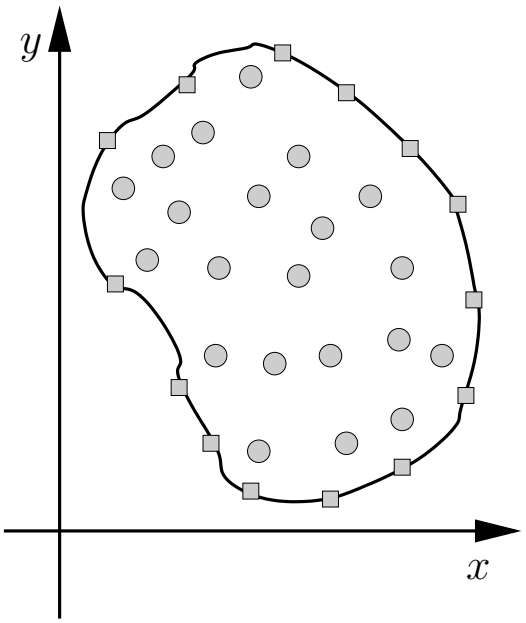
Let $P$ be an arbitrary node internal to the domain. We chose $s$ nodes among the rest of the net-points, typically the $s$ closest adjacent points to $P$. Let $h$ be a characteristic grid spacing. We describe le position of $Q_i$ relative to $P$ by coordinates

$$\overline{PQ}_i = (\xi_i h, \eta_i h).$$

We approximate the operator $L$ by $L_h$ where we assume

$$L_h U_P = \sum_{i=0}^{s} \alpha_i U_{Q_i} - \alpha_0 U_P.$$

for a choice of constants $\alpha_0, \ldots, \alpha_s$.




As usual we insert the exact solution $u$ in this formula and using Taylor expansion in 2 dimensions as in (2.4), we get

$$u_{Q_i} = u_P + \xi_i h \partial_x u_P + \eta_i h \partial_y u_P + \frac{1}{2}\xi_i^2 h^2 \partial_x^2 u_P + \xi_i \eta_i h^2 \partial_x \partial_y u_P + \frac{1}{2}\eta_i^2 h^2 \partial_y^2 u_P + \cdots,$$

which substituted in the espression for $L_h u_P$ gives

$$\begin{aligned} L_h u_p &= \left(\sum_{i=1}^{s} \alpha_i - \alpha_0\right) u_P + \left(h \sum_{i=1}^{s} \xi_i \alpha_i\right) \partial_x u_P + \left(h \sum_{i=1}^{s} \eta_i \alpha_i\right) \partial_y u_P + \left(\frac{1}{2} h^2 \sum_{i=1}^{s} \xi_i^2 \alpha_i\right) \partial_x^2 u_P \\ &\quad + \left(h^2 \sum_{i=1}^{s} \xi_i \eta_i \alpha_i\right) \partial_x \partial_y u_P + \left(\frac{1}{2} h^2 \sum_{i=1}^{s} \eta_i^2 \alpha_i\right) \partial_y^2 u_P \end{aligned}$$

Since this should be "as similar as possible to $Lu_p$", we should require

$$\begin{aligned}
\sum_{i=1}^{s} \alpha_i &= \alpha_0 + f, \\
\sum_{i=1}^{s} \xi_i \alpha_i &= \frac{d}{h}, \\
\sum_{i=1}^{s} \eta_i \alpha_i &= \frac{e}{h}, \\
\sum_{i=1}^{s} \xi_i^2 \alpha_i &= \frac{2a}{h^2}, \\
\sum_{i=1}^{s} \xi_i \eta_i \alpha_i &= \frac{2b}{h^2}, \\
\sum_{i=1}^{s} \eta_i^2 \alpha_i &= \frac{2c}{h^2}.
\end{aligned}$$

Moreover the following equation should be satisfied for as many values of $\ell$ as possible

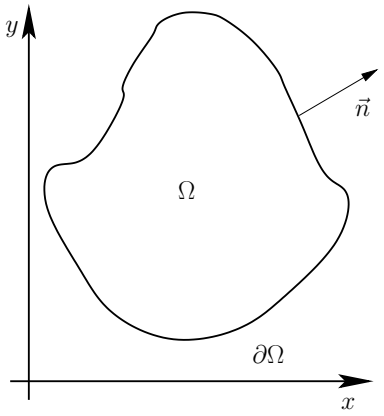
$$\sum_{i=1}^{s} \xi_i^{\ell} \, \eta_i^{\ell-m} \, \alpha_i = 0, \;\; m = 0, \ldots, \ell, \quad \ell = 3, 4, \ldots .$$

This can be justified by considering the rest of the Taylor expansion of $L_h u_P$, the terms $h^3$, of order 3 in $h$, and higher

$$\sum_{\ell=3}^{\cdots} \frac{h^{\ell}}{\ell!} \sum_{m=0}^{\ell} \binom{\ell}{m} \sum_{i=1}^{s} (\alpha_i \, \xi_i^m \, \eta_i^{\ell-m}) \, \partial_x^m \partial_y^{\ell-m} u_P.$$

## 6.7 Difference formulae derived by integration

This technique is called often box-integration, and is closely related to what in the literature goes under the name of finite-volume methods.



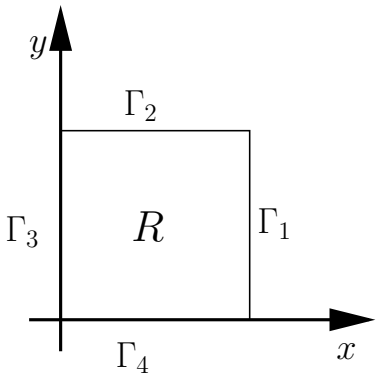

**Gauss' divergence theorem.** Given a domain $\Omega$ as in the picture with boundary $\partial\Omega$ and such that in each point the normal vector $\vec{n}$ is pointing outwards. Let $\vec{p} = \vec{p}(x, y)$ be a vector field on $\Omega$. Then it is true that

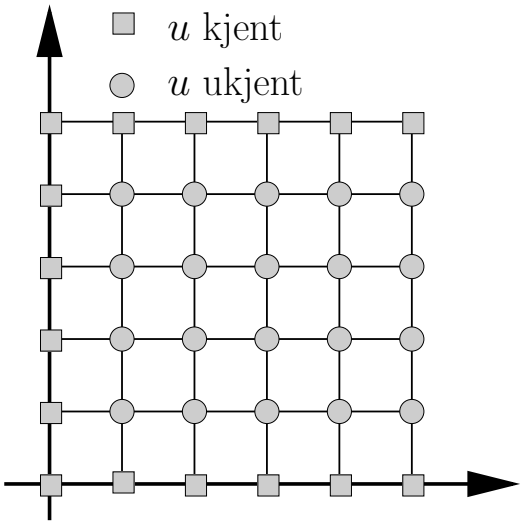
$$\int_{\Omega} \operatorname{div} \vec{p} \, \mathrm{d}A = \oint_{\partial\Omega} \vec{p} \cdot \vec{n} \, \mathrm{d}s.$$

We denote by $\operatorname{div} \vec{p} = \nabla \cdot \vec{p}$ the divergence of the vector field $\vec{p}$. Specially if $\vec{p} = \nabla u$ (gradient vector field) we get that $\operatorname{div} \vec{p} = \Delta u = \partial_x^2 u + \partial_y^2 u$. So we obtain

$$\int_{\Omega} \Delta u \, \mathrm{d}A = \oint_{\partial\Omega} \nabla u \cdot \vec{n} \, \mathrm{d}s = \oint_{\partial\Omega} \partial_n u \, \mathrm{d}s.$$

We illustrate the box-integration via a special example. Let $\Omega = R$ be a rectangle with boundaries $\Gamma_1, \ldots, \Gamma_4$ as in the picture.



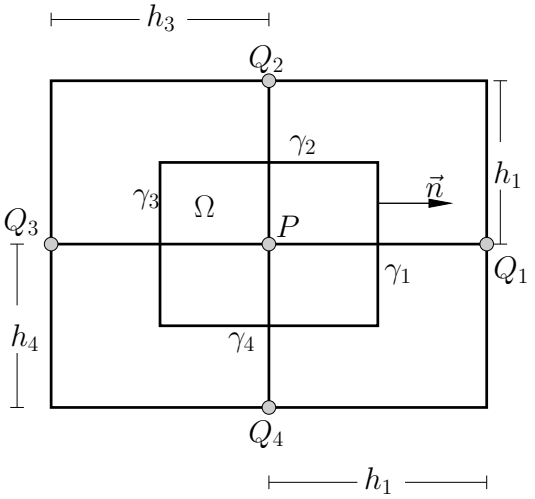

We consider the problem

$$\begin{aligned} \Delta u &= f, && \text{i } R. \\ \partial_n u + au &= d, && \text{on } \Gamma_1, \\ u &= g, && \text{on } \Gamma_2 \cup \Gamma_3 \cup \Gamma_4. \end{aligned}$$

Here we have introduced a rectangular grid on $R$, note that it is possible that the step-sizes are variable in both directions.

Let now $P$ be an internal grid-point in $R$. We first look at the rectangle bounded by the gridlines neighbouring the gridlines on which $P$ lies (see the picture). Next we look at a new rectangle placed exactly in the center of the previous one as in the picture, we call this $\Omega$. The edge boundaries of $\Omega$ are centered between the gridlines and are called $\gamma_i$, $i = 1, 2, 3, 4$. The lenght of $\gamma_i$ is called $\ell_i$. We let the step-size out of $P$ be $h_1$ (to the right), $h_2$ (upwards), $h_3$ (to the left), $h_4$ (downwards). So as in the picture we get

$$\ell_1 = \ell_3 = \frac{h_2 + h_4}{2}, \ \ell_2 = \ell_4 = \frac{h_1 + h_3}{2}.$$




The area $\Omega$ becomes $A = \frac{1}{4}(h_1 + h_3)(h_2 + h_4)$. Now we use Gauss' divergence theorem on the small rectangle $\Omega$ and get

$$\Delta u = f \quad \Rightarrow \quad \int_\Omega \Delta u \, \mathrm{d}A = \underbrace{\oint_{\partial\Omega} \partial_n u \, \mathrm{d}s}_{I} = \underbrace{\int_\Omega f \, \mathrm{d}A}_{II}.$$

We try and approximate I and II.

$$\text{I:} \quad \oint_{\partial\Omega} \partial_n u \, \mathrm{d}s = \sum_{i=1}^{4} \int_{\gamma_i} \partial_n u \, \mathrm{d}s \approx \sum_{i=1}^{4} \ell_i \, \frac{u_{Q_i} - u_P}{h_i}$$

$$\text{II:} \quad \int_{\Omega} f \, \mathrm{d}A \approx f_P \, A = \frac{1}{4} f_P (h_1 + h_3)(h_2 + h_4).$$

So our difference formula becomes

$$\sum_{i=1}^{4} \frac{\ell_i}{A h_i} (U_{Q_i} - U_P) = f_P.$$

We can alternatively write the formula in the "old format"

$$\sum_{i=1}^{4} \alpha_i U_{Q_i} - \alpha_0 U_P = f_P$$

where

$$\alpha_1 = \frac{2}{h_1(h_1+h_3)}, \; \alpha_2 = \frac{2}{h_2(h_2+h_4)}, \; \alpha_3 = \frac{2}{h_3(h_1+h_3)}, \; \alpha_4 = \frac{2}{h_4(h_2+h_4)}, \; \alpha_0 = \frac{2}{h_1 h_3} + \frac{2}{h_2 h_4}.$$

This formula can be used for all internal points in $R$. But we must have equations also for the unknowns on the boundary $\Gamma_1$.
Now $\gamma_1$ coincides with the exterior boundary $\Gamma_1$
where we have

$$\partial_n u + a u = d$$

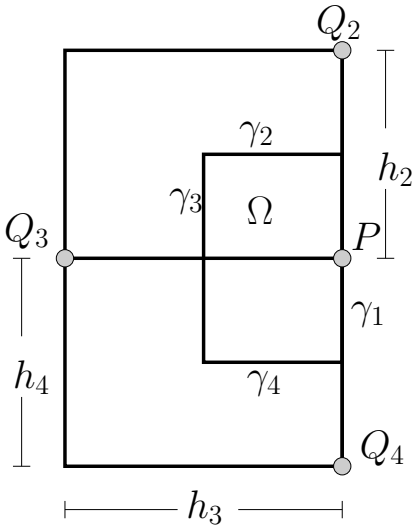
Gauss' divergence theorem on $\Omega$ gives

$$\sum_{i=1}^{4} \int_{\gamma_i} \partial_n u \, \mathrm{d}s = \int_{\Omega} f \, \mathrm{d}A.$$

Looking in particular at the boundary $\gamma_1$ where $\partial_n u$ is given as $d - au$ we obtain

$$\int_{\gamma_1} \partial_n u \, \mathrm{d}s = \int_{\gamma_1} (d - au) \, \mathrm{d}s \approx \ell_1 (d_P - a_P u_P)$$

where $a_P$ and $d_P$ are functions of $a$ and $d$ evaluated in the point $P$. On the other boundaries we set




$$\int_{\gamma_1} \partial_n u \, \mathrm{d}s \approx \ell_i \frac{u_{Q_i} - u_P}{h_i}, \; i = 2, 3, 4.$$

We end up with the following difference formula for the point $P$ belonging to the boundary $\Gamma_1$,

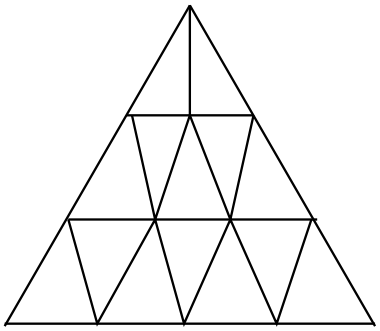
$$\ell_1 (d_P - a_P U_P) + \sum_{i=2}^{4} \frac{\ell_i}{h_i} (U_{Q_i} - U_P) = f_P A$$

where

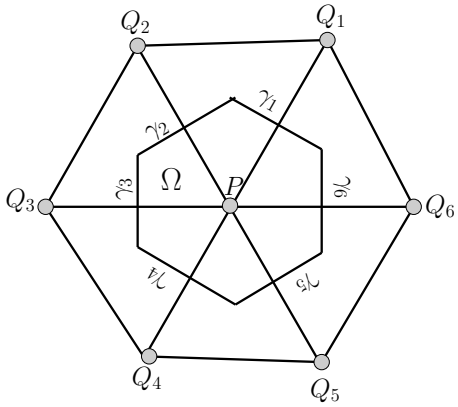
$$A = \frac{1}{4} h_3 (h_2 + h_4), \quad \ell_2 = \ell_4 = \frac{1}{2} h_3, \quad \ell_3 = \frac{h_2 + h_4}{2}.$$

## 6.8 Net based on triangles

The nice thing about box-integration is that it is not based on the assumption that the domain is subdivided in rectangles, in fact triangles can be used with the same ease. It is often simpler to subdivide a domain which has not a rectangular shape using triangles rather then rectangles. The triangles need not be equal or similar (same angles), but we require that all the angles are less than 90 degrees. We also assume we do not have the so called "hanging nodes", we require that no node can be placed along the edge of a triangle (except on the corner points).



In the next picture we have depicted a part of the triangular net where $s$ triangles ($s = 6$ in the picture) have the internal grid-point $P$ as a common corner. The edge $\gamma_i$ in the internal polygon intersects with a 90 degrees angle in the midpoint of the segment $\overline{PQ_i}$. We let $\gamma_i$ have length $\ell_i$ and the segments $\overline{PQ_i}$ length $h_i$. We assume also that $\Omega$ has area $A$. We use Gauss' divergence theorem on $\Omega$, and get




$$\sum_{i=1}^{s} \int_{\gamma_i} \partial_n u \, \mathrm{d}s = \int_{\Omega} f \, \mathrm{d}A.$$

Now we approximate as before

$$\int_{\gamma_i} \partial_n u \, \mathrm{d}s \approx \ell_i \frac{U_{Q_i} - U_P}{h_i}$$

such that the final formula becomes

$$\sum_{i=1}^{s} \frac{\ell_i}{h_i A} (U_{Q_i} - U_P) = f_P.$$

We have not given specific formulae for the computation of $A$ or any relation between $\ell_i$ and $h_i$ (this is not possible in the general case we have considered). There is an important alternative to box-integration, namely the so called finite element method. Which rely on completely different mathematical fundaments compared to the one presented in this and the previous chapter. The course *TMA4220 Numerical solution of partial differential equations with the finite element method* treats all the details abut these methods.

## 6.9 Difference equations

Let us first write an elliptic PDE with boundary conditions in an abstract form as

$$\begin{aligned} Lu &= f, \quad \text{in } \Omega, \\ Bu &= g, \quad \text{on } \partial\Omega. \end{aligned}$$

We introduce the following discretization of this problem

$$\alpha_0 U_P - \sum_{i=1}^{s} \alpha_i U_{Q_i} = \beta_P. \tag{6.4}$$

We let $P$ range over all the points where we want to approximate $u$. Here $Q_i$ is the neighbouring point of $P$ for $i = 1, \ldots, s$. The coefficients $\alpha_1, \ldots, \alpha_s$ can depend on $P$, and similarly does $s$. Using these formulae one determines the linear system of algebraic equations approximating the elliptic PDE. The entries of the matrix defining such linear system (coefficient matrix) are the coefficients $\alpha_i$, $i = 0, \ldots s$ for all $P$ where the numerical solution is sought.

**Useful properties for** (6.4)**.**

i.
$$\left.\begin{array}{l}\alpha_0 > 0\\ \alpha_i \geq 0\end{array}\right\} \quad \text{for all } P.$$

ii.
$$\alpha_0 \geq \sum_{i=1}^{s} \alpha_i \qquad \text{for all } P,$$

if this property hols the coefficient matrix is called *diagonally dominant*. We also require strict inequality for *at least* one $P$.

iii. The coefficient matrix is symmetric.


If we assume that the coefficient used in node $P$ in front of $Q$ is $\alpha_{P,Q}$, the symmetry of the matrix means that $\alpha_{P,Q} = \alpha_{Q,P}$.

We recall the maximum principle discussed in chapter 6.1. An analogous principle holds true for difference equations of the type described above.

**Discrete maximum principle.** Assume that the difference equation satisfies (i) and (ii) given above. Assume the quantity $V_P$ is defined for all $P \in \Omega \cup \partial\Omega$ and that

$$\alpha_0 V_P - \sum_i \alpha_i V_{Q_i} \leq 0$$

for every $P \in \Omega$ (internal grid-point).
Then we have

$$V_P \leq \max_{S \in \partial\Omega} V_S \quad \text{for all } P \in \Omega.$$

*Proof.* Assume the converse is true, namely that there is a $P^* \in \Omega$ such that

$$V_{P^*} = \max_{P \in \Omega} V_P > \max_{S \in \partial\Omega} V_S. \tag{6.5}$$

By using the hypothesis of the theorem we get

$$\alpha_0 V_{P^*} \leq \sum_i \alpha_i V_{Q_i}$$

$$\Updownarrow \text{ (i)}$$

$$V_{P^*} \leq \sum_i \frac{\alpha_i}{\alpha_0} V_{Q_i} \overset{\text{(ii)}}{\leq} \frac{1}{\sum_j \alpha_j} \sum_i \alpha_i V_{Q_i} = \sum_i \gamma_i V_{Q_i}$$

where

$$\gamma_i = \frac{\alpha_i}{\sum_j \alpha_j} \qquad \text{such that } \sum_i \gamma_i = 1.$$

We get therefore that

$$\sum_i \gamma_i V_{Q_i} \leq \sum_i \gamma_i \max_i V_{Q_i} = \max_i V_{Q_i}$$

and then $V_{P^*} \leq \max_i V_{Q_i}$. By (6.5), we get $V_{P^*} = \max_i V_{Q_i} = V_{Q_{i_1}}$, so $(\alpha_0 - \alpha_{i_1})V_{P^*} \leq \sum_{i \neq i_1} \alpha_i V_{Q_i}$. So either there is only one neighbouring point $Q_{i_1}$, or $\alpha_0 - \alpha_{i_1} > 0$ so by a similar argument we get $V_{P^*} = \max_{i \neq i_1} V_{Q_i}$ and we can proceed til we cover all the neighbouring points so that $V_{Q_i} = V_{P^*}$ for all $i$. We can next move in turn to each one of the neighbouiring points $Q_i$ and repeat the same argument till one of the neighbouring points is a boundary point $S$. Therefore we get that $V_S = V_{P^*}$ for $S \in \partial\Omega$ which is a contradiction (see (6.5)) and we conclude that the theorem holds.

## 6.10 Convergence of the methods for elliptic equations

### 6.10.1 Convergence for the 5-point formula on a Dirichlet problem

Consider the problem

$$\begin{aligned} -\Delta u &= f \quad \text{i } R \\ u &= g \quad \text{på } \partial R \end{aligned}$$

where $R$ is the square $(0,1) \times (0,1)$.

Set $h = \frac{1}{M}$ and apply the 5-point formula $L_h U_p = f_p$ where

$$\begin{aligned} L_h U_p &= \frac{1}{h^2}(4U_p - U_\o - U_v - U_n - U_s), \quad p \in R \\ U_p &= g_p, \quad p \in \partial R. \end{aligned}$$

The truncation error on the node $p$ is given as $\tau_p = L_h u_p + f_p$. Using Taylor expansion we can show that

$$|\tau_p| = \frac{1}{6} h^2 K = \bar{\tau}$$

where

$$K = \max_{p \in R} \{|\partial_x^4 u_p|, |\partial_y^4 u_p|\},$$

this definition requires that these fourth derivatives are bounded on $R$. The discretization error (global error) is defined as

$$e_p = u_p - U_p,$$

and we find that

$$\begin{aligned} L_h e_p &= L_h u_p - L_h U_p = f_p + \tau_p - f_p = \tau_p, \quad p \in R \\ L_h e_p &= 0, \qquad\qquad\qquad\qquad\qquad\qquad\quad p \in \partial R. \end{aligned}$$

So we have

$$|L_h e_p| \leq \bar{\tau} \qquad \text{for alle } p \in R.$$

We find now that the function $\varphi(x,y) = \frac{1}{2}x^2$ and we apply the operator $L_h$ on this function

$$L_h \varphi_p = \frac{1}{h^2}\left(4\,\frac{1}{2}x_p^2 - \frac{1}{2}x_\emptyset^2 - \frac{1}{2}x_v^2 - \frac{1}{2}x_s^2 - \frac{1}{2}x_n^2.\right)$$

Where we have

$$\left.\begin{aligned} x_o &= x_p + h \\ x_n &= x_s = x_p \\ x_v &= x_p - h \end{aligned}\right\} \Rightarrow L_h\varphi_p = \frac{1}{2h^2}\left(4x_p^2 - (x_p+h)^2 - (x_p - h^2) - x_p^2 - x_p^2\right) = -1.$$

We set now $V_p = e_p + \bar{\tau}\varphi_p$ and get

$$L_h V_p = L_h e_p + \bar{\tau} L_h \varphi_p = L_h e_p - \bar{\tau} \leq 0.$$

So $L_h$ satisfies the hypothesis in the discrete maximum principle with $V_p = e_p + \bar{\tau}\varphi_p$. Therefore we get

$$e_p + \bar{\tau}\varphi_p \leq \max_{S \in \partial R}(e_S + \bar{\tau}\varphi_S) \leq \frac{1}{2}\bar{\tau} \quad \text{for alle } p \in R,$$

since $e_S = 0$ and $R$ is the square $(0,1)\times(0,1)$ so $\varphi(x,y) = \frac{1}{2}x^2 \leq \frac{1}{2}$ for $(x,y) \in R$.

If we repeat the same argument with $V_p = -e_p + \bar{\tau}\varphi_p$ we find respectively that

$$-e_p + \bar{\tau}\varphi_p \leq \frac{1}{2}\bar{\tau} \quad \text{for all } p \in R.$$

Since $\varphi_p \bar{\tau} \geq 0$ can we conclude that

$$|e_p| \leq \frac{1}{2}\bar{\tau} \leq \frac{1}{12}Kh^2 \quad \text{for alle } p \in R.$$

### 6.10.2 Some general comments on convergence

We look at the general difference scheme of the type

$$\alpha_{pp} u_p - \sum_q \alpha_{pq} u_q = \beta_p + \tau_p.$$

The discretization error is $e_p = u_p - U_p$, and by substituting into the formula we obtain

$$\alpha_{pp} e_p - \sum_q \alpha_{pq} e_q = \tau_p, \quad p \in \Omega.$$

If we arrange all the values of the error and the local truncation error in the grid-points $e_p$, $\tau_p$, $p \in R$ in vectors $\boldsymbol{e}$ and $\boldsymbol{\tau}$, we can write the system in the form

$$A\boldsymbol{e} = \boldsymbol{\tau}$$

If $A$ is invertible we get

$$\boldsymbol{e} = A^{-1}\boldsymbol{\tau}$$

implying

$$\|\boldsymbol{e}\|_\infty \leq \|A^{-1}\|_\infty \cdot \|\boldsymbol{\tau}\|_\infty.$$

Stability: the difference methods is called *stable* if there is a constant $C$ such that

$$\|A^{-1}\| \leq C, \qquad \text{for all step-sizes } h.$$

It is typically possible to prove (by Taylor expansion) that he truncation error $\boldsymbol{\tau}$ satisfies

$$\|\boldsymbol{\tau}\|_\infty = \mathcal{O}(h^\sigma), \quad \sigma \text{ integer number.}$$

This together with stability will imply that

$$\|\boldsymbol{e}\|_\infty = \mathcal{O}(h^\sigma).$$

It can happen that for example $\tau_p = \mathcal{O}(h^2)$ for some $p$, while for others (typically those close to the boundary or on the boundary) we have $\tau_p = \mathcal{O}(h)$. So in general the global error $\|\boldsymbol{e}\|_\infty$ can not be expected to be of higher order than 1 in $h$, $\mathcal{O}(h)$. It can happen however that in such situations one can get $\|\boldsymbol{e}\|_\infty = \mathcal{O}(h^2)$ anyway.

## 6.11 Some remarks on the solution of linear systems of algebraic equations

The solution of linear systems of the type (6.4) is an important sub-area of numerical analysis. How complex PDE problems we can handle and how accurate numerical solutions we can produce at a computer, depends greatly on how large linear systems of equations we can solve. It is outside the scope of this course to discuss this subject in detail, we will give here some general information.

When we choose a method to solve numerically a linear system of equations we have two main classes of methods we can consider, i.e. *direct methods* and *iterative* methods. The first class includes Gaussian elimination, or more specifically the Cholesky factorisation if the system is symmetric and positive definite. The iterative methods include Jacobi, Gauss–Seidel, and SOR (succsessive over-relaxation). But the type of linear iterative equation-solvers which has got most success in the last decades, are the so called *Krylov subspace methods* and for symmetric matrices the *conjugate gradient method*. It is not easy to say precisely when it is best to use direct methods rather than direct methods. Typical advantages of iterative methods are:

1. the linear systems are very large and sparse; a system is sparse if the non zero entries of the matrix are relatively small portion of the total number of entries;

2. the matrices are not banded, i.e. there are indexes, corresponding to non zero elements, relatively far away form the diagonal position ($ij$-element with large value of $|i-j|$); the bandwidth of a matrix can be for example defined as follows

$$b(A) = \max\{|i-j| : a_{ij} \neq 0\};$$

3. the system can be preconditioned effectively; this means that it is possible to find matrices $T, S$, such that the system

$$\begin{array}{cccc} T^{-1}AS & S^{-1}x & = & T^{-1}b \\ \widehat{A} & \hat{x} & = & \hat{b} \end{array}$$

is "easier" to solve compared to the original system $Ax = b$.

In practice direct methods and iterative methods are both effective when working with partial differential equations in 2 space dimensions, while in 3 space dimensions we might expect an advantage in using iterative methods. Typically iterative solvers use matrix-vector multiplications (by $A$ or respectively $\widehat{A}$) as building blocks. If property 1. above is satisfied the multiplications $Ax$ can be executed efficiently provided the matrix is easy to access in the computer memory. The property 2. disfavours the use of Gaussian elimination, which can be executed efficiently for banded matrices. The factors $L$ and $U$ in the $LU$-factorization have in this case the same bandwidth as $A$ (if no pivoting is performed). On the other hand if the bandwidth is large the number of non zero elements in $L$ and $U$ can be much larger than in $A$. This effect is called "fill-in".

A real difference in performance between iterative methods and direct methods can be seen when we use appropriate preconditioning techniques, property 3. above. Note that the transformation $\widehat{A} = T^{-1}AS$ is not carried out in practice but only in theory. For the sake of simplicity, let us assume $S = I$. As mentioned above the iterative methods are built on operations of the type $y = Ax$ for arbitrary vectors $x$. For the preconditioned system we get $y = \widehat{A}x = T^{-1}Ax$. The preconditioning is therefore an internal procedure in the method: each time we compute $\widehat{A}x$ we do it by first computing $\tilde{y} = Ax$ and then$y = T^{-1}\tilde{y}$. The last operation is not performed as an explicit matrix-vector multiplication, but as a process, a function applied to $\tilde{y}$. An example of such a process is the computation of parts of the Gaussian elimination algorithm on the system $Ay = \tilde{y}$. Such approach is called incomplete LU-factorization and is done by limiting or neglecting the fill-in. This approach can also be efficiently parallelized, meaning that the computational work can be divided over many processors on a supercomputer and can be executed very efficiently. Another preconditioning technique can be obtained by splitting the domain of definition of the partial differential equation in many small sub-domains. In turn, the differential equation is split in many different differential equations, while discarding or approximating the coupling between the various sub-domains. Each of the linear systems arising form one of the differential equations is then solved on a different processor by, for example, Gaussian elimination.

The course *TMA4205 Numerical Linear Algebra* presents these issues and methods in detail.

# Chapter 7

# Hyperbolic equations

## 7.1 Examples

1. The most famous example of a hyperbolic differential equation is the linear second order wave equation
$$u_{tt} = c^2 u_{xx}$$
or in several space dimensions
$$u_{tt} = c^2 \Delta u.$$

2. Consider the following general second order linear PDEs in two space dimensions
$$a\, u_{xx} + 2b\, u_{xy} + c u_{yy} + d u_x + e u_y + f\, u = 0,$$
where $a, b, c, d, e, f$ are functions of $x, y$, such equation is said to be hyperbolic when
$$b^2 - ac > 0.$$
Note that $y$ plays here the role of $t$ in the equations above.

3. Another equation appearing in many applications is the so called conservation law
$$u_t + c(x, t, u)\, u_x = 0.$$
This is a scalar quasi-linear PDE, of first order which is also said to be hyperbolic. Such an equation can be used to describe the road traffic flow .

4. Systems of first order linear equations with constant coefficients can be written in the following form
$$u_t + A\, u_x = 0 \tag{7.1}$$
where $u \in \mathbf{R}^n$ and $A$ is a real $n \times n$-matrix. Such a system is hyperbolic if $A$ is diagonalisable with real eigenvalues.

5. Let us consider a nonlinear hyperbolic system of equations describing an important application namely the shallow water equations. We consider this set of equations in one space dimension, and let $v(x, t)$ be the velocity of the fluid in the point $x$ at time $t$, while $z(x, t)$ measures the (vertical) wave height relative to an equilibrium position.

| | |
|---|---|
| Mass conservation | $z_t + (vz)_x = 0.$ |
| Momentum conservation | $v_t + (\frac{1}{2}v^2 + z)_x = 0.$ |

6. Next we write all these equations in general conservation form

$$u_t + (f(u))_x = 0. \tag{7.2}$$

Here $u \in \mathbf{R}^n$ while $f : \mathbf{R}^n \to \mathbf{R}^n$ is a nonlinear mapping, and (7.1) is the special case where $f(u) = Au$ for a $n \times n$ matrix $A$. To ensure (7.2) is hyperbolic we require that the Jacobian matrix $Df = f'(u)$ is diagonalisable with real eigenvalues.

## 7.2 Characteristics

We consider a model problem which we will use a lot in the sequel

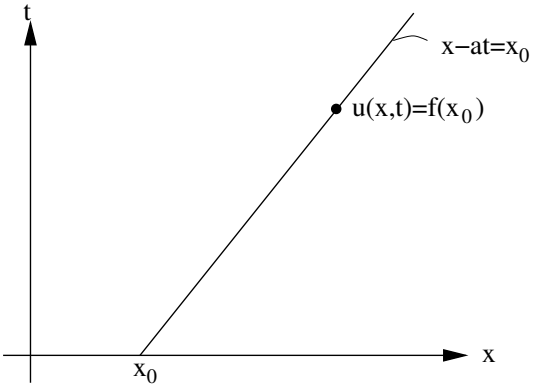
$$u_t + au_x = 0, \quad -\infty < x < \infty, \quad t \geq 0, \qquad a > 0 \tag{7.3}$$

where $a$ is a constant. The initial function must be given on all of $\mathbf{R}$

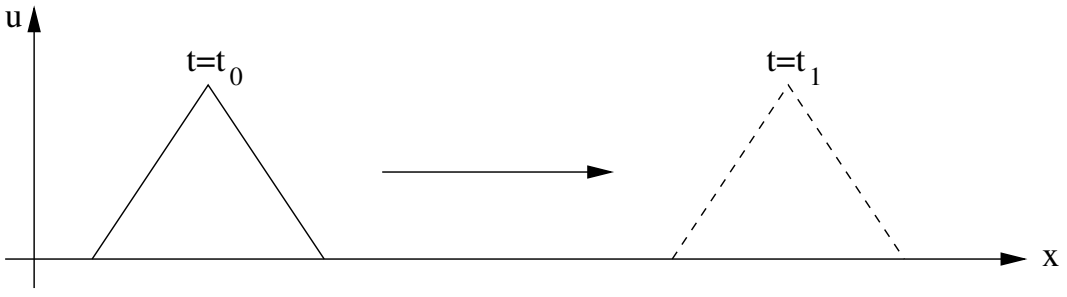
$$u(x, 0) = f(x), \qquad -\infty < x < \infty.$$

For this simple model problem it is possible to find the exact solution, that is

$$u(x, t) = f(x - at).$$




You can verify that by differentiating appropriately and substituting into the equation $u(x, t) = f(x-at)$ satisfies (7.3). This means that in the $(x, t)$-plane there is a line where $u(x, t)$ is constant, i.e. the line $x = x_0 + at$ where $u(x, t) = u(x_0 + at, t) = f(x_0 + at - at) = f(x_0)$. This line is called *characteristic*.
Constant values of $u$ are transported from the initial value $u(x_0, 0) = f(x_0)$ and forward in time. The velocity of propagation is the reciprocal of the slope of the curve. When the line is vertical the slope is infinite, and the velocity is 0. If the characteristic is oriented towards right and upwards the value $u(x_0, 0)$ is transported in the same direction. In a $xu$-plot we can draw the solution at different values of time




Let us now consider the more general case

$$u_t + a(x, t)\, u_x = b(x, t) \tag{7.4}$$

**Characteristic equation for** (7.4)

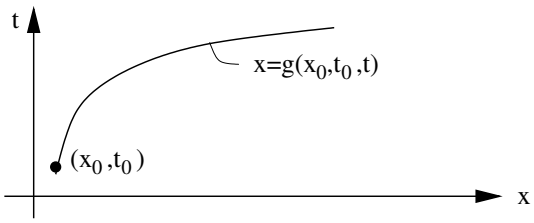
$$\frac{dx}{dt} = a(x, t). \tag{7.5}$$

Assume now that $x_0$ and $t_0$ is given and let

$$x = g(x_0, t_0, t)$$

be a solution of (7.5) satisfying $x(t_0) = x_0$. Let $u(x,t)$ be a solution of (7.4) and consider

$$v(t) := u(x,t)|_{x=g} = u(g(x_0,t_0,t),t)$$



Characteristic $\gamma$ through $(x_0, t_0)$

Here $v$ gives the solution $u(x,t)$ along the curve $\gamma$. We can differentiate $v$ with respect to time using the chain rule of derivation

$$v'(t) = \left(u_t + u_x \frac{\mathrm{d}x}{\mathrm{d}t}\right)\Big|_{x=g} = (u_t + a(x,t)u_x)|_{x=g} = b(x,t)|_{x=g}$$

We see that along the characteristic, $v'(t)$ is a known function of $t$, so we can find $v$ by integration

$$v(t) = v(t_0) + \int_{t_0}^{t} b(x,s)|_{x=g(x_0,t_0,s)}\,\mathrm{d}s.$$

If we know $u(x_0,t_0)$ we can find $u(x,t)$ for $(x,t) \in \gamma$ from the formula

$$u(x,t) = u(x_0,t_0) + \int_{t_0}^{t} b(g(x_0,t_0,s),s)\,\mathrm{d}s. \tag{7.6}$$

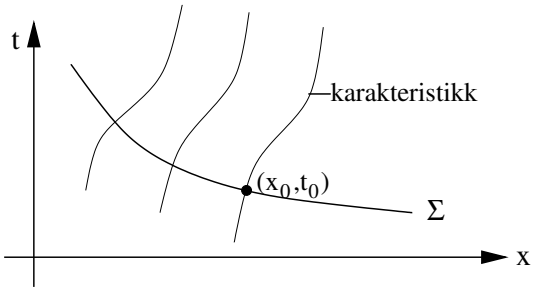
Special case: $b(x,t) \equiv 0$ implies $u(x,t) = u(x_0,t_0)$ for all $(x,t) \in \gamma$.

**Initial value problem for** (7.4)**.** Assume that $u(x,t)$ satisfies the PDE (7.4) and that

$$u(x,t) = f(x,t) \qquad \text{along the curve } \Sigma:\ x = \sigma(t)$$

NB! $\Sigma$ should *not* be a characteristic.

Solution strategy: Compute the characteristic $x = g(x_0,t_0,t)$ through the point $(x_0,t_0)$ belonging to $\Sigma$. Use than (7.6) with $u(x_0,t_0) = f(x_0,t_0)$.




**A simple example.** Let us consider again the simplest example

$$u_t + au_x = 0 \qquad a \in \mathbf{R}.$$

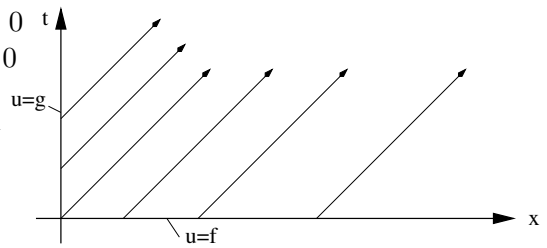
The characteristics are

$$\frac{dx}{dt} = a \qquad \Leftrightarrow \qquad x = g(x_0,t_0,t) = x_0 + a(t - t_0).$$

Let us for example assume that the curve $\Sigma$ is the $x$-axis where $u(x,0) = f(x)$ i.e. we have $t_0 = 0$. The characteristic through $(x,t)$ intersects the $x$-axis in $x_0 = x - at$ where the solution is $u(x_0,0) = f(x_0) = f(x-at)$ and we have reproduced the exact solution which we have presented earlier.

**Initial/boundary valule problem for $u_t + au_x = 0$.** We formulate the problem for $0 \le x \le \infty$ and $t \ge 0$. Assume now for simplicity that $a > 0$.

| | | |
|---|---|---|
| Initial value | $u(x,0) = f(x),$ | $x \ge 0$ |
| Boundary value | $u(0,t) = h(t),$ | $t \ge 0$ |




The exact solution of this equation is given by

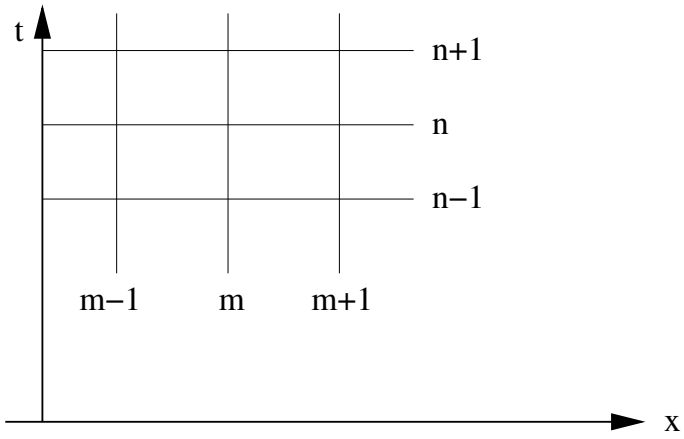
$$u(x,t) = \begin{cases} f(x-at), & x - at \ge 0 \\ h(t - x/a), & x - at < 0. \end{cases}$$

*Remark.* We could have used, for example, boundaries $x = 0$ and $x = 1$, but than it would be wrong to specify the solution along the line $x = 1$.

## 7.3 Explicit difference formulae for $u_t + au_x = 0$

$x_m = mh, \quad t_n = nk$

$u_m^n = u(x_m, t_n)$




Different discretizations

$$\partial_t u_m^n = \frac{1}{k}(u_m^{n+1} - u_m^n) + \mathcal{O}(k) \tag{7.7}$$

$$\partial_x u_m^n = \frac{1}{h}(u_{m+1}^n - u_m^n) + \mathcal{O}(h) \tag{7.8}$$

$$\partial_x u_m^n = \frac{1}{h}(u_m^n - u_{m-1}^n) + \mathcal{O}(h) \tag{7.9}$$

$$\partial_x u_m^n = \frac{1}{2h}(u_{m+1}^n - u_{m-1}^n) + \mathcal{O}(h^2) \tag{7.10}$$

If we choose (7.7) and (7.9) we get

$$\frac{1}{k}(u_m^{n+1} - u_m^n) + \frac{a}{h}(u_m^n - u_{m-1}^n) + \mathcal{O}(k) + \mathcal{O}(h) = 0$$

and therefore for the numerical method

$$U_m^{n+1} = (1 - ap)U_m^n + apU_{m-1}^n, \qquad p = \frac{k}{h} \tag{7.11}$$

The truncation error bacomes $\tau_m^n = \mathcal{O}(k^2) + \mathcal{O}(kh)$. If we choose (7.10) for $u_x$ we obtain instead

$$U_m^{n+1} = U_m^n - \frac{ap}{2}(U_{m+1}^n - U_{m-1}^n) \tag{7.12}$$

Here we obtain $\tau_m^n = \mathcal{O}(k^2) + \mathcal{O}(kh^2)$, looking as an improvement compared to (7.11) but we will later show that (7.12) *is always unstable for this problem*!

**Lax-Wendroff's formula.** From the differential equation we obtain

$$u_{tt} = (-au_x)_t = -a(u_t)_x = -a(-au_x)_x = a^2 u_{xx}.$$

We construct a difference discretization by Taylor expanding $u_m^{n+1}$ to the second order, use the differential equation, and then discretize the spatial derivatives with central differences. We obtain

$$\begin{aligned} u_m^{n+1} &= u_m^n + k\,\partial_t u_m^n + \frac{1}{2}\,k^2\,\partial_t^2 u_m^n + \mathcal{O}(k^3) \\ &= u_m^n - ak\,\partial_x u_m^n + \frac{1}{2}\,(ak)^2\,\partial_x^2 u_m^n + \mathcal{O}(k^3) \\ &= u_m^n - ak\,\frac{1}{2h}\,(u_{m+1}^n - u_{m-1}^n) + +\frac{1}{2}\,(ak)^2\,\frac{1}{h^2}\,\delta_x^2 u_m^n + \mathcal{O}(kh^2) + \mathcal{O}(k^3) \end{aligned}$$

From here we obtain

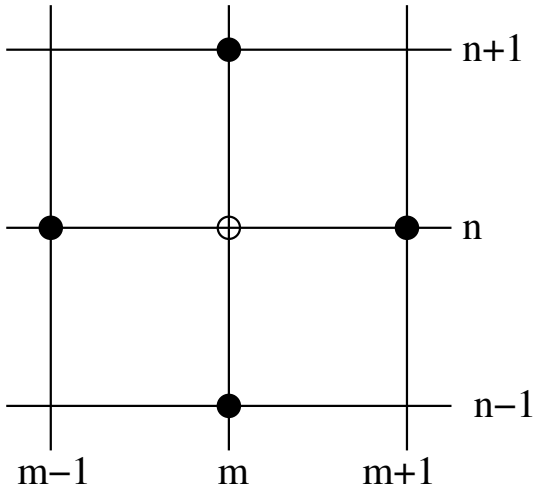
**Lax-Wendroff's formula for** $u_t + au_x = 0$

$$U_m^{n+1} = U_m^n - \frac{ap}{2}(U_{m+1}^n - U_{m-1}^n) + \frac{1}{2}\,(ap)^2\,\delta_x^2 U_m^n \tag{7.13}$$

Truncation error

$$k\tau_m^n = \mathcal{O}(k^3) + \mathcal{O}(kh^2)$$

**Leap frog formula (Hoppe-bukk formel).**




$$\partial_t u_m^n = \frac{1}{2k}(u_m^{n+1} - u_m^{n-1}) + \mathcal{O}(k^2)$$

$$\partial_x u_m^n = \frac{1}{2h}(u_{m+1}^n - u_{m-1}^n) + \mathcal{O}(h^2)$$

We get therefore the **leap-frog formula (hoppe-bukk formel) for** $u_t + au_x = 0$

$$U_m^{n+1} = U_m^{n-1} - ap(U_{m+1}^n - U_{m-1}^n)$$

Truncation error: $k\tau_m^n = \mathcal{O}(k^3) + \mathcal{O}(kh^2)$

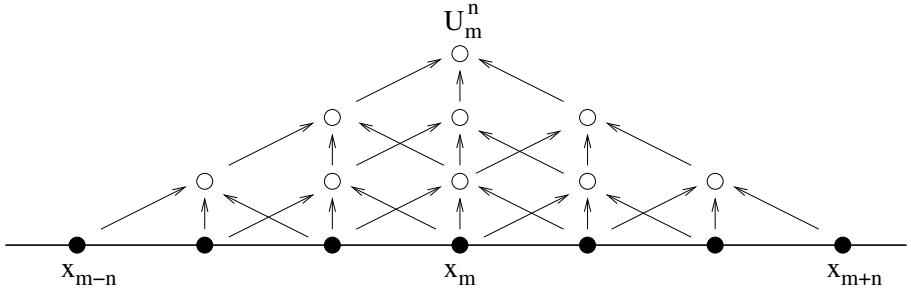
Note that this is a two-step (three levels) method in time. For this reason it is neceaasry to provide a starting method, analogously to what happens for multistep methods for ordinary differential equations.

## 7.4 Stability

**Courant-Friedrichs-Levy condition.** We consider again the equation $u_t + au_x = 0$. Assume we have a difference formula of the type

$$U_m^{n+1} = \alpha_{-1} U_{m-1}^n + \alpha_0 U_m^n + \alpha_1 U_{m+1}^n d \tag{7.14}$$

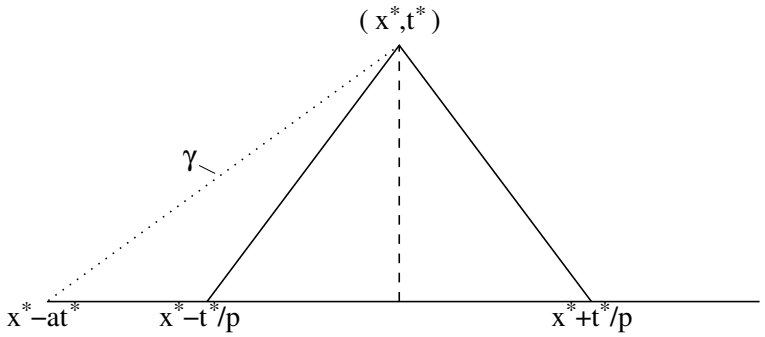
We study the domain of dependence for $U_m^n$. $U_m^n$ depends on $U_{m+\ell}^0$ der $-n \leq \ell \leq n$. The dependence intervall for $U_m^n$ on the $x$-axis is $I_m^n = [x_{m-n}, x_{m+n}]$.




Let us fix $x_m = x^* = mh$ and $t_n = t^* = nk$ while sending $h$ and $k$ to zero and $m, n$ to infinity simultaneously. Assume also that this is done in such a way that $p = k/h =$ constant. The end points for $I_m^n$ are then

$$x_{m\pm n} = x^* \pm nh = x^* \pm t^* \frac{h}{k} = x^* \pm \frac{t^*}{p}$$

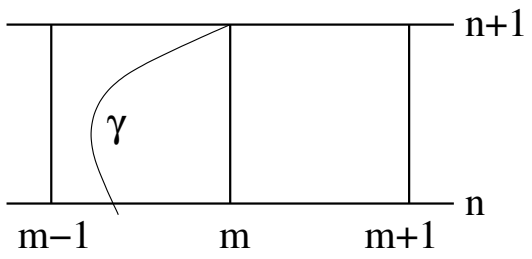
So $I_m^n = [x^* - \frac{t^*}{p}, x^* + \frac{t^*}{p}]$. Note that this interval is fixed when $h$ and $k$ go to zero as described above. Let $\gamma$ be a characteristic for $u_t + au_x = 0$ passing though the point $(x^*, t^*)$. The inclination with respect to the $x$-axis is $a$ and the line intersects the $x$-axis in $x^* - at^*$ so $u(x^*, t^*) = f(x^* - at^*)$. If $x^* - at^* \notin I_m^n$ this means that the computed approximation $U_m^n$ is built upon initial data that do not include $f(x^* - at^*)$ regardless of how small $h$ and $k$ are. In the situation depicted in the figure we can not have convergence towards the exact solution, when $h, k \to 0$, for all initial values.




Necessary condition for convergence: CFL-condition. The characteristic though $(x^*, t^*)$ must intersect the $x$-axis in a point of the domain of dependencefor $x_m = x^*,\ t_n = t^*$. CFL is a short for *Courant-Friedrichs-Levy.*

The same principle holds also for curved characteristic curves, as those one gets when $a = a(x, t)$.

The characteristic though $(x_m, t_{n+1})$ intersects the line $t = t_n$ between the two external points corresponding to $m - 1$ and $m+1$ (whose corresponding numerical approximations are featuring in the difference formula), as described in the figure. The characteristics must never leave the domain of dependence.



Let us see which is the necessary criterion for convergence for the case of constant $a$ and a difference formula of the type considered above. We have the condition

$$x^* - t^*/p \leq x^* - at^* \leq x^* + t^*/p$$

giving $|a|p \leq 1$. This is the necessary condition for convergence for all the numerical formulae of the type (7.14).

**Von Neumann-condition.** We remind that this condition, discussed earlier in this course, is based on Fourier analysis. The method consists in substituting

$$U_m^n = \xi^n \, \mathrm{e}^{\mathrm{i}\beta x_m}, \quad \mathrm{i} = \sqrt{-1} \tag{7.15}$$

in the difference equation, and then solve it writh respect to th *amplification factor* $\xi$. The stability condition is then

$$|\xi| \leq 1 \quad \text{for any } \beta \in \mathbf{R}.$$

**Stability of the Lax-Wendroff's scheme.** We remind the formula valid for problems of the type $u_t + au_x = 0$

$$U_m^{n+1} = U_m^n - \frac{1}{2}ap(U_{m+1}^n - U_{m-1}^n) + \frac{1}{2}(ap)^2(U_{m+1}^n - 2U_m^n + U_{m-1}^n).$$

If we substitute (7.15), simplified with $\xi^n \mathrm{e}^{\mathrm{i}\beta x_m}$ on both sides and we use $\mathrm{e}^{\mathrm{i}\beta h} = \cos\beta h + \mathrm{i}\sin\beta h$ we get

$$\xi = 1 - \mathrm{i}\,ap\,\sin\beta h + (ap)^2(\cos\beta h - 1) = 1 - 2(ap)^2\sin^2\frac{\beta h}{2} - \mathrm{i}\,ap\,\sin\beta h.$$

We now use the trigonometric identity $\cos\beta h = 1 - 2\sin^2\frac{\beta h}{2}$ here and below. Let us define $r = ap$ og $q = \sin\frac{\beta h}{2}$.

$$\begin{aligned} |\xi|^2 &= (1 - 2r^2q^2)^2 + r^2(1 - \cos^2\beta h) = (1 - 2r^2q^2)^2 + r^2(1 - (1 - 2q^2)^2) \\ &= 1 - 4r^2q^2 + 4r^4q^4 + r^2 + r^2 - r^2(1 - 4q^2 + 4q^4) \\ &= 1 - 4r^2\,(1 - r^2)\,q^4 \end{aligned}$$

We require then

$$\begin{gathered} 1 - 4r^2\,(1 - r^2)\,q^4 \leq 1, \qquad 0 \leq q \leq 1 \\ \Updownarrow \\ 4r^2\,(1 - r^2) \geq 0 \\ \Updownarrow \\ |r| \leq 1 \\ \Updownarrow \\ |a|\,p \leq 1 \end{gathered}$$

then we obtain exactly the same condition obtained from the CFL-condition.

A simpler fomula to check is the "naive" formula (7.12) which we DO NOT recommend

$$U_m^{n+1} = U_m^n - \frac{ap}{2}(U_{m+1}^n - U_{m-1}^n).$$

We obtain here

$$\xi = 1 - \mathrm{i}\,ap\,\sin\beta h \tag{7.16}$$

so that

$$|\xi|^2 = 1 + (ap)^2 \sin^2 \beta h > 1$$

for almost all the $\beta$, that is the formula is unstable for all step-sizes. An interesting modification of this bad method can be obtained by replacing

$$U_m^n \quad \text{with} \quad \frac{1}{2}\left(U_{m-1}^n + U_{m+1}^n\right).$$

In von Neumann analysis the expression for $\xi$ in (7.16) is replaced by

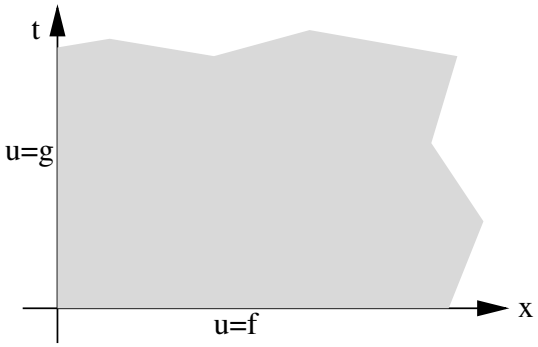
$$\xi = \cos \beta h - \mathrm{i}\, ap \sin \beta h,$$

and we find that $|\xi|^2 = 1 + (1 - (ap)^2) \sin^2 \beta h$ so that we get stability if $|ap| \leq 1$. This method is called *Lax-Friedrichs* method.

## 7.5 Implicit methods for $u_t + au_x = 0$

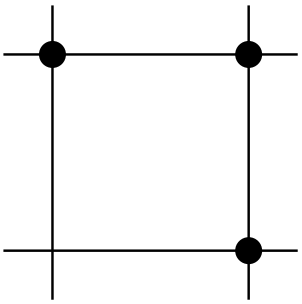
Implicit methods can only be used on initial/boundary value problems. We consider problems of the type

$$\begin{aligned} u_t + au_x &= 0, \qquad x \geq 0,\ t \geq 0, \\ u(x,0) &= f(x), \quad x \geq 0, \\ u(0,t) &= g(t), \quad t \geq 0. \end{aligned}$$




**The easiest implicit formula.** We try the easiest possible discretization of the derivative in $(x_m, t_{n+1})$

$$\begin{aligned} \partial_t u_m^{n+1} &= \frac{1}{k}(u_m^{n+1} - u_m^n) + \mathcal{O}(k) \\ \partial_x u_m^{n+1} &= \frac{1}{h}(u_m^{n+1} - u_{m-1}^{n+1} + \mathcal{O}(h) \end{aligned}$$



which gives the formula

$$U_m^{n+1} - U_m^n + ap(U_m^{n+1} - U_{m-1}^{n+1}) = 0.$$

We can solve explicitely with respect to $U_m^{n+1}$ and obtain

$$U_m^{n+1} = \frac{ap}{1+ap} U_{m-1}^{n+1} + \frac{1}{1+ap} U_m^n.$$

Even if the superscript $n+1$ appears at the right hand side of this equation, the method is in practice explicit if we compute the solution values at time level $n+1$ from left to right. We observe that $U_1^{n+1}$ depends only on $U_0^{n+1}$ at time level $n+1$ and the latter is given as $g(t_{n+1})$. The computed value $U_1^{n+1}$ is subsequently used to compute $U_2^{n+1}$ and so on.

We find that the local truncation error for this method is

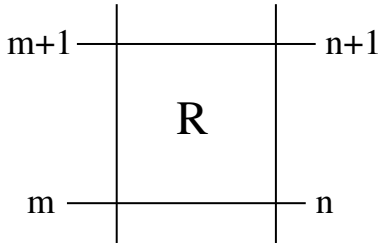
$$k\tau_m^n = -\frac{1}{2}\left(a^2 k^2 + a\, hk\right) \partial_x^2 u_m^n + \cdots = \mathcal{O}(k^2 + hk)$$

**Wendroff's method.** We consider a procedure similar to the box-integration

$$u_t + au_x = 0$$

$$\Downarrow$$

$$\int_{t_n}^{t_{n+1}} \int_{x_m}^{x_{m+1}} (u_t + au_x)\, \mathrm{d}x\, \mathrm{d}t = 0.$$




If we exchange the order of integration for the first term and we use the fundamental theorem of calculus, we find

$$\int_{x_m}^{x_{m+1}} (u^{n+1} - u^n)\, \mathrm{d}x + a \int_{t_n}^{t_{n+1}} (u_{m+1} - u_m)\, \mathrm{d}t = 0 \tag{7.17}$$

where

$$u^n = u(x, t_n), \quad \text{and} \quad u_m = u(x_m, t).$$

Now we recall the trapezoidal rule for quadrature. Given a function $f(x)$ we have

$$\int_r^{r+d} f(s)\, \mathrm{d}s = \frac{d}{2}(f(r) + f(r+d)) - \frac{1}{12} d^3 f''(r + \frac{d}{2}) + \cdots$$

So if we apply the trapezoidal rule to both the integrals in (7.17) we obtain

$$\frac{h}{2}\Big((u_m^{n+1} - u_m^n) + (u_{m+1}^{n+1} - u_{m+1}^n)\Big) + \frac{ak}{2}\Big((u_{m+1}^n - u_m^n) + (u_{m+1}^{n+1} - u_m^{n+1})\Big) + \mathcal{O}(k^3 + h^3) = 0$$

Wendroff's method.

$$(1 + ap)\, U_{m+1}^{n+1} + (1 - ap)\, U_m^{n+1} - (1 - ap)\, U_{m+1}^n - (1 + ap)\, U_m^n = 0$$

The truncation error can be expanded around the midpoint $(x_m + h/2, t_n + k/2)$ of the rectangle $R$, and we get then

$$k\tau_m^n = \frac{1}{6}\Big(a^3\, k^3 - a\, kh^2\Big)\partial_x^3 u_{m+1/2}^{n+1/2} + \cdots = \mathcal{O}(k^3 + kh^2).$$

To study the stability of the method we use the Von Neumann-condition again, with $\gamma = (1 - ap)/(1 + ap)$ we can write

$$U_{m+1}^{n+1} + \gamma\, U_m^{n+1} - \gamma\, U_{m+1}^n - U_m^n = 0.$$

By setting $U_m^n = \mathrm{e}^{\mathrm{i}\beta mh}$ we get

$$\xi(\mathrm{e}^{\mathrm{i}\beta h} + \gamma) = \gamma \mathrm{e}^{\mathrm{i}\beta h} + 1,$$

which we solve with respect to $\xi$ and obtain

$$\xi = \mathrm{e}^{\mathrm{i}\beta h}\, \frac{\gamma + \mathrm{e}^{-\mathrm{i}\beta h}}{\gamma + \mathrm{e}^{\mathrm{i}\beta h}}.$$

The first factor $\mathrm{e}^{\mathrm{i}\beta h}$ has absolute value 1, and the fraction is an expression of the type $z^*/z$ so this has also absolut value 1. Therefore we have $|\xi| = 1$ for all $\beta$ and we say that Wendroeff's method is unconditionally stable, that is stable for all $h$ and $k$.

## 7.6 Hyperbolic systems of first order equations

**Definition of hyperbolicity, characteristics.** We consider systems of partial differential equations of first order with two independent variables $x$ and $t$ and $\ell$ dependent variables

$$\begin{aligned} &u_t + Au_x = 0, \\ &u = (u_1, \ldots, u_\ell)^T,\ A \in \mathbf{R}^{\ell\times\ell}. \end{aligned} \tag{7.18}$$

To begin with we let $A$ be constant. If $A$ is diagonalizable and has real egenvalues, we call (7.18) hyperbolic. In this case $A$ can be factorized as

$$A = T\Lambda T^{-1}, \quad \Lambda = \mathrm{diag}(\lambda_i)_{i=1:\ell}, \quad \lambda_i \quad \text{real.}$$

We perform a variable transformation

$$\begin{aligned} u = Tv \Leftrightarrow v = T^{-1}u \\ u_t = Tv_t \\ u_x = Tv_x \end{aligned}$$

where $T$ is a constant matrix. By substituting in (7.18) we get

$$Tv_t + (T\Lambda T^{-1})Tv_x = T(v_t + \Lambda v_x) = 0.$$

We have therefore decoupled (7.18) into

$$\begin{aligned} v_t + \Lambda v_x = 0 \\ \Updownarrow \\ (v_i)_t + \lambda_i (v_i)_x = 0, \quad i = 1, \ldots, \ell. \end{aligned} \tag{7.19}$$

For each equation we have a characteristic equation

$$\frac{dx}{dt} = \lambda_i, \quad i = 1, \ldots, \ell. \tag{7.20}$$

We observe that to (7.18) pertain $\ell$ characteristic equations and $\ell$ families of characteristics,

$$x = \lambda_i t + \text{konstant}, \quad i = 1, \ldots, \ell. \tag{7.21}$$

**Example.** We rewrite the wave equation

$$\phi_{tt} = c^2 \phi_{xx}$$

into a system of first order equations, as follows

$$\begin{aligned} &u_1 = \phi_t, \quad u_2 = \phi_x, \\ &(u_1)_t = \phi_{tt} = c^2\phi_{xx} = c^2(u_2)_x \\ &(u_2)_t = \phi_{xt} = \phi_{tx} = (u_1)_x, \end{aligned}$$

and in the form (7.18)

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}_t + \begin{pmatrix} 0 & -c^2 \\ -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}_x = 0. \tag{7.22}$$

So the matrix $A$ in (7.18) has the form

$$A = -\begin{pmatrix} 0 & c^2 \\ 1 & 0 \end{pmatrix},$$

$A$ has eigenvalues and eigenvectors

$$\lambda_1 = c, \quad t_1 = \begin{pmatrix} -c \\ 1 \end{pmatrix}, \quad \lambda_2 = -c, \quad t_2 = \begin{pmatrix} c \\ 1 \end{pmatrix}$$

and so it is diagonalizable:

$$A = T\Lambda T^{-1},$$
$$\Lambda = \begin{pmatrix} c & 0 \\ 0 & -c \end{pmatrix}, \quad T = \begin{pmatrix} -c & c \\ 1 & 1 \end{pmatrix}, \quad T^{-1} = \begin{pmatrix} -1/(2c) & 1/2 \\ 1/(2c) & 1/2 \end{pmatrix}.$$

The characteristic equation is $dx/dt = \pm c$, and (7.22) has two families of characteristics $x = \pm ct +$ konst. The tranfromation to the form (7.19) is

$$v = T^{-1}u \Leftrightarrow v_1 = (-u_1/c + u_2)/2, \; v_2 = (u_1/c + u_2)/2,$$
$$(v_1)_t - c(v_1)_x = 0,$$
$$(v_2)_t + c(v_2)_x = 0.$$

**General linear system of first order equations.**

$$u_t + A(x,t)u_x + B(x,t)u = f(x,t) \tag{7.23}$$

The equation (7.23) is hyperbolic if $A$ is diagonalizable with real eigenvalues

$$A(x,t) = T(x,t) \cdot \Lambda(x,t) \cdot T^{-1}(x,t)$$
$$\Lambda(x,t) = \mathrm{diag}(\lambda_i(x,t))_{i=1:l}.$$

We perform the change of variables

$$u = T\,v$$
$$u_t = Tv_t + T_t v, \quad u_x = Tv_x + T_x v$$

(7.23) $\Rightarrow$

$$Tv_t + T_t v + T\Lambda T^{-1}(Tv_x + T_x v) + BTv = f$$

giving

$$v_t + \Lambda v_x + \Gamma v = g,$$
$$\Gamma = T^{-1}T_t + \Lambda T^{-1}T_x + T^{-1}BT, \tag{7.24}$$
$$g = T^{-1}f.$$

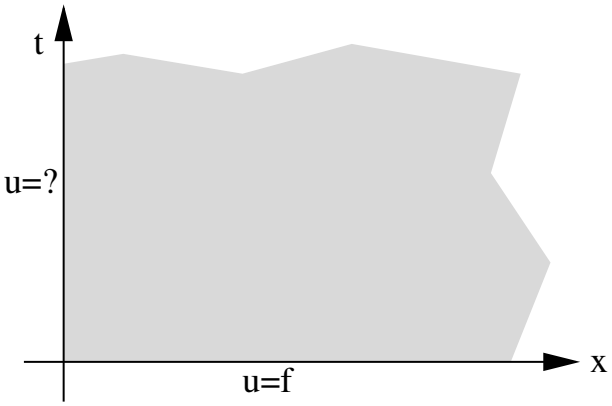
The equations (7.24) are not any longer a decoupled system as in the case (7.21), but lead to charachteristic equations for (7.23)

$$\frac{dx}{dt} = \lambda_i(x,t), \quad i = 1, \ldots, \ell.$$

**Initial/boundary value problem.** An initial/boundary value problem is given as

$$\begin{aligned} &u_t + Au_x = 0, \quad -\infty < x < \infty,\ t > 0 \\ &u(x,0) = f(x), \quad -\infty < x < \infty. \end{aligned}$$

**Initial/boundary value probelm of type I.**



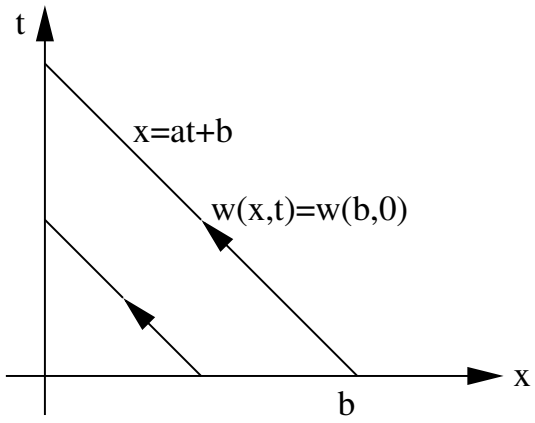

$$\begin{aligned} &\mathcal{R} = \{(x,t) : x \geq 0, t \geq 0 \\ &u_t + Au_x = 0 \text{ i } \mathcal{R} \\ &u(x,0) = f(x),\ x \geq 0 \end{aligned}$$

Moreover we need a set of boundary conditions for $u$ along the $t$-axis as we will se in the sequel.

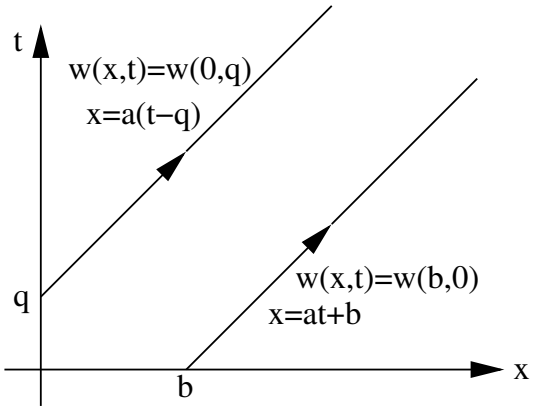
Consider first a scalar equation

$$w_t + aw_x = 0.$$

The characteristic equation is $dx/dt = a$. The characteristics are: $x = at + \text{konst}$.




$a \leq 0$. $w(x,t)$ is uniquely determined in $\mathcal{R}$ if $w(x,0)$ is known for $x \geq 0$.




$a > 0$. To be able to determine $w(x,t)$ in $\mathcal{R}$ we need to know $w(x,0),\ x \geq 0$ and $w(0,t),\ t \geq 0$.

Consider a system

$$\left.\begin{aligned} &u_t + Au_x = 0,\ \text{i } \mathcal{R}, \\ &u(x,0) = f(x),\ x \geq 0 \end{aligned}\right\} \quad \begin{aligned} &\Longleftrightarrow \\ &A = T\Lambda T^{-1}, \\ &u = Tv, \\ &h = T^{-1}f. \end{aligned} \quad \left\{\begin{aligned} &v_t + \Lambda v_x = 0 \text{ i } \mathcal{R}, \\ &v(x,0) = h(x), \quad x \geq 0. \end{aligned}\right.$$

Let $\Lambda = \text{diag}(\lambda_i)_{i=1:\ell}$ and assume the eigenvalues are ordered such that $\lambda_i > 0,\ i = 1, \ldots, k$ and $\lambda_i \leq 0,\ i = k+1, \ldots, \ell$ (for a certain $k$). We get scalar equations

$$(v_i)_t + \lambda_i (v_i)_x = 0 \text{ i } \mathcal{R}, \quad v_i(x,0) = h_i(x),\ x \geq 0.$$

In this case $v_i(x,t)$ will be determined by $v_i(x,0) = h_i(x)$ for all $i \geq k+1$ (characteristics pointing upwards towards left). To compute $v_i(x,t)$ in $\mathcal{R}$ with $i \leq k$ we need values of $v_i$ along the $t$-axis. We can get such values by imposing $k$ boundary conditions.

$$\sum_{j=1}^{l} \gamma_{ij} v_j(0,t) = g_i(t), \ i = 1:k.$$

The equations must be solvable for $\{v_i, i = 1, \ldots, k\}$ this means $C = [\gamma_{ij}]_{i,j=1:k}$ must be non singular. This leads to the following initial/boundary value problem:

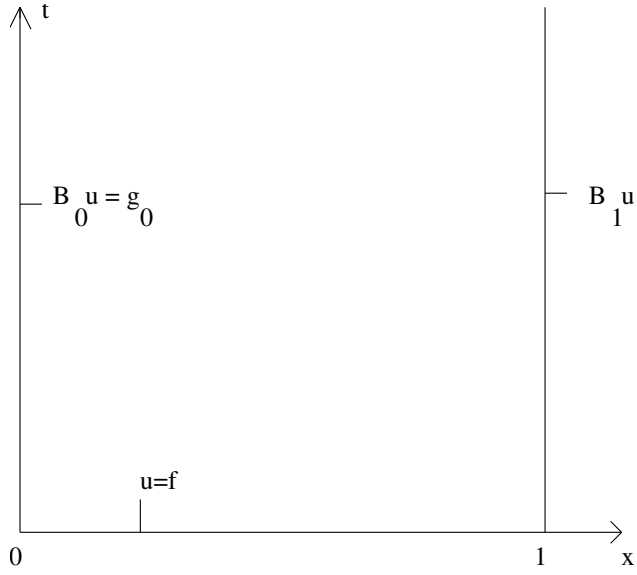
| | |
|---|---|
| $u_t + Au_x = 0$ i $\mathcal{R}$, | $At_i = \lambda_i t_i, \ i = 1, \ldots, \ell,$ |
| $u(x,0) = f(x), \ x \geq 0,$ | $\lambda_i > 0, \ i = 1, \ldots, k,$ |
| $Bu(0,t) = g(t).$ | $B$ er $k \times \ell$, |
| | $C = B \cdot [t_1, \cdots, t_k]$ is non singular. |

**Example.**

$$\left.\begin{array}{l}(u_1)_t - c^2(u_2)_x = 0 \\ (u_2)_t - (u_1)_x = 0\end{array}\right\} \text{ i } \mathcal{R}, \quad \left.\begin{array}{l}u_1(x,0) = f_1(x) \\ u_2(x,0) = f_2(x)\end{array}\right\}, \quad x \geq 0.$$

$$\beta_1 u_1(0,t) + \beta_2 u_2(0,t) = g(t), \quad t \geq 0, \qquad -\beta_1 c + \beta_2 \neq 0.$$

**Initial/boundary vaule problem of type II.**
$\mathcal{R} = \{(x,t) : 0 \leq x \leq 1, \ t \geq 0\}$. Assume $A$ has $k$ positive and $m$ negative eigenvalues $(k + m \leq l)$.




$u_t + Au_x = 0$ i $\mathcal{R}$,
$u(x,0) = f(x), \quad 0 \leq x \leq 1,$
$B_0 u(0,t) = g_0(t), \quad t \geq 0,$
$B_1 u(1,t) = g_1(t), \quad t \geq 0.$

Here $B_0$ $k \times \ell$ and $B_1$ is $m \times \ell$.

**Lax-Wendroff and Wendroff for systems.** We recall the Lax-Wendroff for the scalar equation $u_t + au_x = 0$

$$U_m^{n+1} = U_m^n - ap\mu_x \delta_x U_m^n + \frac{1}{2}(ap)^2 \delta_x^2 U_m^n$$

where $\mu_x u(x,t) = \frac{1}{2}(u(x + h/2, t) + u(x - h/2, t))$ is a averaging operator and $\delta_x u(x,t) = u(x + h/2, t) - u(x - h/2, t)$ as always. Particularly we note that

$$\mu_x \delta_x U_m^n = \mu_x (U_{m+1/2}^n - U_{m-1/2}^n) = \frac{1}{2}(U_{m+1}^n + U_m^n) - \frac{1}{2}(U_m^n + U_{m-1}^n) = \frac{1}{2}(U_{m+1}^n - U_{m-1}^n).$$

We formulate now the Lax-Wendroff method for a system of equations $u_t + Au_x = 0$ where the matrix $A \in \mathbf{R}^{\ell\times\ell}$ is constant and $U_m^n \in \mathbf{R}^\ell$ for all $m$ and $n$

$$U_m^{n+1} = U_m^n - pA\mu_x\delta_x U_m^n + \frac{p^2}{2}A^2\delta_x^2 U_m^n, \tag{7.25}$$

this is a straigtforward generalization. But we let now $A = A(x,t)$ depend on $x$ and $t$, and get

Lax-Wendroff for systems with variable $A$, $u_t + A(x,t)u_x = 0$

$$U_m^{n+1} = U_m^n - pA_m^{n+1/2}\mu_x\delta_x U_m^n + \frac{p^2}{2}A_m^{n+1/2}\delta_x\left(A_m^{n+1/2}\delta_x U_m^n\right)$$

where $A_m^{n+1/2} = A(x_m, t_n + k/2)$.

We consider now stability for constant $A$, and generalize the Von Neumann-condition to systems. We present only the procedure which is based on taking

$$U_m^n = \mathrm{e}^{\mathrm{i}\beta x_m}\, G^n U^0$$

and substitute it in the difference method. Now $G \in \mathbf{R}^{\ell\times\ell}$ is a matrix, often called amplification matrix, $\beta \in \mathbf{R}$ is an arbitrary frequency, and $U^0 \in \mathbf{R}^\ell$ an arbitrary vector. We substitute this expression into (7.25), and we get

$$G = I - \mathrm{i}pA\sin\theta + p^2(\cos\theta - 1)A^2, \quad \theta = \beta h. \tag{7.26}$$

A necessary condition for stability is then

$$\rho(G) \le 1 \qquad \text{for all } \theta \in \mathbf{R}.$$

Since $A$ is diagonalizable, we can write

$$A = T\Lambda T^{-1}, \quad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_\ell).$$

If we use this into (7.26) we get

$$G = T\left(I - \mathrm{i}p\sin\theta\Lambda + p^2(\cos\theta - 1)\Lambda^2\right)T^{-1}$$

then the matrix $T$ diagonalizes both $A$ and $G$. So the eigenvalues of $G$ are on the mid diagonal factor, they are

$$\mu_j = 1 - \mathrm{i}p\sin\theta\lambda_j + p^2(\cos\theta - 1)\lambda_j^2.$$

The expression for $\mu_j$ is then exactly the same as for $\xi$ in the scalar stability analysis for the Lax-Wendroff method, only $a$ is replaced by $\lambda_j$ on the right hand side. The same analysis goes through when we require $|\mu_j| \le 1$, and we obtain the condition $p|\lambda_j| \le 1$ for all $j$, i.e.

$$\rho(A)\, p \le 1,$$

which is the stability condition for Lax-Wendroff applied to systems with $A$ constant.

A necessary condition for stability for variable $A$ is that $\rho(A(x,t))\, p \le 1$ for all $(x,t)$ where the method is used.

We consider now a *generalization of Wendroff's method to the case of systems.* We recall the method for the scalar equation

$$(1+ap)\,U_{m+1}^{n+1}+(1-ap)\,U_m^{n+1}-(1-ap)\,U_{m+1}^n-(1+ap)\,U_m^n=0.$$

By using forward differences in space we can rewrite this expression as

$$\left(1+\frac{1}{2}(1+ap)\Delta_x\right)U_m^{n+1}=\left(1+\frac{1}{2}(1-ap)\Delta_x\right)U_m^n.$$

So we obtain

Wendroff's method for systems with variable $A$, $u_t+A(x,t)u_x=0$

$$\left(I+\frac{1}{2}(I+p\,A_{m+1/2}^{n+1/2})\Delta_x\right)U_m^{n+1}=\left(I+\frac{1}{2}(I-p\,A_{m+1/2}^{n+1/2})\Delta_x\right)U_m^n$$

where $A_{m+1/2}^{n+1/2}=A(x_m+h/2,t_n+k/2)$.

If we can compute $A$ only in the gridpoints, we can replace

$$A_{m+1/2}^{n+1/2}\quad\text{with}\quad\frac{1}{4}(A_m^n+A_{m+1}^n+A_m^{n+1}+A_{m+1}^{n+1})$$

without loss of accuracy. The stability analysis for constant $A$ follows the same technique as before, and the result is that Wendroff's method is stable for all $p$.

**Initial/boundary value problem for systems of two equations.** The wave equation $\phi_{tt}=\phi_{xx}$ can be rewritten in the form of a system as in (7.22), with $c=1$ the equation becomes

$$\begin{pmatrix}u_1\\u_2\end{pmatrix}_t+\begin{pmatrix}0&-1\\-1&0\end{pmatrix}\cdot\begin{pmatrix}u_1\\u_2\end{pmatrix}_x=0.$$

Here $A$ has eigenvalues $\lambda_1=1$ and $\lambda_2=-1$, so a family of characteristics points towards left and the other points towards right as in the picture at page 92. So it is necessary to specify only one boundary condition. The boundary/initial conditions can for example be

$$\begin{aligned}u(x,0)&=f(x),\ x\geq 0\\v(x,0)&=g(x),\ x\geq 0\\u(0,t)&=\phi(t),\ t\geq 0.\end{aligned}$$

Let us for example chose the Lax-Wendroff method as difference method. We let

$$W_m^n=\begin{pmatrix}U_m^n\\V_m^n\end{pmatrix},\qquad A=\begin{pmatrix}0&-1\\-1&0\end{pmatrix}.$$

The method becomes then

$$W_m^{n+1}=\left(I-pA\mu_x\delta_x+\frac{p^2}{2}A^2\,\delta_x^2\right)W_m^n. \tag{7.27}$$

Assume now that $W_m^n$ is known for all $m\geq 0$ and a $n\geq 0$. We can use (7.27) to find $W_m^{n+1}$ for $m\geq 1$. What about $W_0^{n+1}$?

$$U_0^{n+1}=\phi(t_{n+1})\quad\text{OK,}\quad\text{the problem is } V_0^{n+1}.$$

To obtain $V_0^{n+1}$ we propose two alternatives

1. From the first differential equations, $u_t - v_x = 0$, we find

$$v_x(0,t) = u_t(0,t) = \phi'(t).$$

We approximate $v_x(0,t) = \dfrac{1}{h}(v(h,t) + v(0,t) + \mathcal{O}(h)$. Then we get the approximation

$$V_0^{n+1} = V_1^{n+1} + h\phi'(t_{n+1}).$$

Note that $V_1^{n+1}$ is obtained from (7.27). Higher order approximations of the derivative can also be used,

2. Consider the second differential equation, $v_t - u_x = 0$ and use box-integration

$$\int_0^h \int_{t_n}^{t_{n+1}} v_t \,\mathrm{d}t\,\mathrm{d}x = \int_{t_n}^{t_{n+1}} \int_0^h u_x \,\mathrm{d}x\,\mathrm{d}t.$$

From the fundamental theorem of calculus we get then

$$\int_0^h (v(x,t_{n+1}) - v(x,t_n))\,\mathrm{d}x = \int_{t_n}^{t_{n+1}} (u(h,t) - u(0,t))\,\mathrm{d}t.$$

We approximate both integrals with trapezoidal rule

$$\frac{h}{2}(V_0^{n+1} - V_0^n + V_1^{n+1} - V_1^n) = \frac{k}{2}(U_1^n - U_0^n + U_1^{n+1} - U_0^{n+1}).$$

The equation is solved then with respect to $V_0^{n+1}$

$$V_0^{n+1} = -V_1^{n+1} + V_0^n + V_1^n + p(U_1^n + U_1^{n+1} - U_0^n - U_0^{n+1}).$$

## 7.7 Dissipation and dispersion

Let us now consider again the pure initial value problem

$$\begin{aligned} u_t + au_x = 0, \quad -\infty < x < \infty, \quad t \geq 0 \\ u(x,0) = f(x), \quad -\infty < x < \infty. \end{aligned}$$

The Fourier transform of the initial function $f(x)$ is

$$\hat{f}(\beta) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(x)\,\mathrm{e}^{-\mathrm{i}\beta x}\,\mathrm{d}x.$$

The inverse transform is

$$f(x) = \int_{-\infty}^{\infty} \hat{f}(\beta)\,\mathrm{e}^{\mathrm{i}\beta x}\,\mathrm{d}\beta.$$

The solution of the pure initial value problem can be written in the form

$$u(x,t) = \int_{-\infty}^{\infty} \hat{f}(\beta)\,\mathrm{e}^{\mathrm{i}\beta(x-at)}\,\mathrm{d}\beta.$$

If we let $f(x) = \mathrm{e}^{\mathrm{i}\beta x}$ we get the solution

$$u(x,t) = \mathrm{e}^{\mathrm{i}\beta(x-at)}.$$

The solution $u$ is a wave with amplitude 1 and wave length $\lambda = 2\pi/\beta$, moving with speed $a$ along the $x$-axis. Let us now apply a difference method to this equation with the same initial function. This gives a numerical approximation

$$U_m^n = \xi^n \, \mathrm{e}^{\mathrm{i}\beta x_m},$$

as we have already seen in the case of Von Neumann stability. Here $\xi$ is a complex number depending on $\beta$, as one can find by substituting the expression above in the differential equation. In the sequel it is useful to write $\xi$ in polar form

$$\xi = |\xi| \, \mathrm{e}^{-\mathrm{i}\varphi}.$$

Let us define the number $\alpha$ by requiring $\varphi = \beta \, \alpha \, k$ where $k$ is the time-step, so $\alpha = \varphi/(\beta k)$. We find

$$\xi^n = |\xi|^n \, \mathrm{e}^{-\mathrm{i}\, n\, \varphi} = |\xi|^n \, \mathrm{e}^{-\mathrm{i}\, n\, k\, \alpha\, \beta} = |\xi|^n \, \mathrm{e}^{-\mathrm{i}\, \beta\, \alpha\, t_n}$$

and so we obtain the following expression for the numerical and the exact solution

$$U_m^n = |\xi|^n \mathrm{e}^{\mathrm{i}\,\beta(x_m - \alpha t_n)},$$

$$u_m^n = |1|^n \mathrm{e}^{\mathrm{i}\,\beta(x_m - a t_n)}.$$

In general one can use an arbitrary initial function $f(x)$ and in this case

$$U_m^n = \int_{-\infty}^{\infty} \hat{f}(\beta) \, |\xi|^n \, \mathrm{e}^{\mathrm{i}\,\beta(x_m - \alpha t_n)} \, \mathrm{d}\beta.$$

Earlier we have used $|\xi| \leq 1$ as stability condition. The strict inequality is considered in the following definition.

**Dissipation.** If there is an upper bound for the time-step $k_0 > 0$ and a constant $\sigma > 0$ such that

$$|\xi| \leq 1 - \sigma(\beta h)^{2s} \qquad \text{for } |\beta h| \leq \pi$$

and all $k \leq k_0$, the difference formula is dissipative of order $2s$.

**Dispersion.** If $\alpha$ depends on $\beta$ the difference method is *dispersive*. Then different frequencies are transported at different speed. In other words we can say that dissipation measures the error in the amplitude, and dispersion measures the error in the phase.

**Example.** Let us use the definitions on the Lax-Wendroff method. From earlier analysis we have (with $\theta = \beta h$)

$$\xi = 1 - \mathrm{i}\, ap \, \sin\theta - (ap)^2(1 - \cos\theta) = 1 - 2r^2 \, \sin^2\left(\frac{\theta}{2}\right) - \mathrm{i}\, r \, \sin\theta, \quad r = ap.$$

From the stability analysis for the Lax-Wendroff method we found the expression

$$|\xi|^2 = 1 - q \, \sin^4\left(\frac{\theta}{2}\right), \qquad q = 4r^2(1 - r^2). \tag{7.28}$$

As earlier shown, we see that $|\xi| \leq 1$ if and only if $|r| \leq 1$, but the question is how $|\xi|$ behaves for $0 < |r| < 1$, case when $0 < q \leq 1$. Let us assume that if $x$ is a number such that $x \leq 1$ then

$$1 - x \leq 1 - x + \frac{1}{4}x^2 = (1 - \frac{x}{2})^2 \qquad \Rightarrow \qquad \sqrt{1 - x} \leq 1 - \frac{x}{2}$$

Taking the square root of (7.28) and letting $x = q\,\sin^4\left(\frac{\theta}{2}\right)$ we get

$$|\xi| \leq 1 - \frac{q}{2}\sin^4\left(\frac{\theta}{2}\right) = 1 - \frac{q}{2}\left(\frac{\sin(\theta/2)}{\theta}\right)^4 \theta^4.$$

As by the definition of dispersion we need only to look at the interval $-\pi \leq \theta \leq \pi$. The smallest value of $\sin(\theta/2))/\theta$ is $\sin(\pi/2))/\pi = 1/\pi$. Then we get

$$|\xi| \leq 1 - \frac{q}{2}\left(\frac{1}{\pi}\right)^4 \theta^4 = 1 - \frac{q}{2\pi^4}\,\theta^4, \quad |\theta| \leq \pi.$$

We conclude that Lax-Wendreoff is dissipative of order 4 with $\sigma = \dfrac{2(ap)^2(1-(ap)^2)}{\pi^4}$.
If we analyze now the dispersion, we find that

$$\varphi = \alpha\,\beta\,k = \arctan\left(\frac{r\sin\theta}{1 - 2r^2\sin^2(\theta/2)}\right).$$

Solving with respect to $\alpha$ and using that $\beta k = \frac{\theta r}{a}$ we get

$$\alpha = a\,\frac{1}{r\theta}\arctan\left(\frac{r\sin\theta}{1 - 2r^2\sin^2(\theta/2)}\right),$$

so in general $\alpha$ definitely depends on $\beta$ and Lax-Wendroff is therefore dispersive. It is interesting to see that at the stability limit $r = 1$ the scheme will not be dispersive, but one gets $\alpha = a$ for all frequencies $\beta$.