StuDocu.com

CSE2 - 2020 - Lecture Notes numeicalProgramming

Numerisches Programmieren (IN0019) (Technische Universität München)

StuDocu wird von keiner Universität gesponsert oder unterstützt.
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Numerical Programming 2
(MA 3306)
Summer Term 2020

Rainer Callies
Department of Mathematics M2
Technical University of Munich

Error messages/proposed corrections please email to
callies@ma.tum.de

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

These notes follow Prof. Callies's "Numerical Programming 2" course in the Summer term 2020. His course is constructed using many different resources like books, other professor's lecture material or codes, none of which are cited here as these are my personal notes.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Contents

This document is available free of charge on StuDocu.com

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

1 Repetition

1.1 Norms

Definition (special vector norms)

For $\vec{x} \in \mathbb{C}^n$ we define

$$\begin{aligned} \|\vec{x}\|_1 &:= |x_1| + |x_2| + \ldots + |x_n| \\ \|\vec{x}\|_2 &:= \sqrt{\sum_{i=1}^{n} |x_i|^2} \\ \|\vec{x}\|_\infty &:= \max\{|x_1|, |x_2|, \ldots, |x_n|\} \end{aligned}$$

Definition (vector norm)

A vector norm p is a mapping $\|\cdot\|_p : \mathbb{C}^n \to \mathbb{R}_0^+$ such that

$$\begin{aligned} \|\vec{x}\|_p &> 0 \quad \forall \vec{x} \neq \vec{0} \quad \wedge \quad \|\vec{x}\|_p = 0 \Leftrightarrow \vec{x} = \vec{0} \\ \|a\vec{x}\|_p &= |a| \cdot \|\vec{x}\|_p \quad \forall a \in \mathbb{C}, \forall \vec{x} \\ \|\vec{x} + \vec{y}\|_p &\leq \|\vec{x}\|_p + \|\vec{y}\|_p \quad \forall \vec{x}, \vec{y} \end{aligned}$$

□

Definition (matrix norm)

A matrix norm is a mapping $\|\cdot\| : \mathbb{R}^{n \times m} \to \mathbb{R}_0^+$ such that

$$\begin{aligned} \|A\| &> 0 \quad \forall A \neq 0 \quad \wedge \quad (\|A\| = 0 \Leftrightarrow A = 0) \\ \|\alpha A\| &= |\alpha| \|A\| \qquad \text{(homogenity)} \\ \|A + B\| &\leq \|A\| + \|B\| \qquad \text{(triangular inequality)} \end{aligned}$$

for all $A, B \in \mathbb{R}^{n \times m}, \alpha \in \mathbb{R}$. Generalization to $A, B \in \mathbb{C}^{n \times m}, \alpha \in \mathbb{C}$ possible!

A matrix norm $\|\cdot\|$ is called *sub-multiplicative*, if

$$\|C \cdot D\| \leq \|C\| \cdot \|D\| \quad \forall C \in \mathbb{R}^{n \times m}, D \in \mathbb{R}^{m \times q}$$

A matrix norm $\|\cdot\|$ is called *compatible or consistent* with the vector norm $\|\cdot\|_p$, if

$$\|A\vec{x}\|_p \leq \|A\| \|\vec{x}\|_p \quad \forall A \in \mathbb{R}^{n \times m}, \vec{x} \in \mathbb{R}^m$$

Definition

Let be $A \in \mathbb{R}^{n \times m}$, then the vector norm $\|\cdot\|_p$ can be used to define the following matrix norm

$$\|A\|_p := \sup_{\vec{x} \neq 0} \frac{\|A\vec{x}\|_p}{\|\vec{x}\|_p} = \sup_{\|\vec{x}\|_p = 1} \|A\vec{x}\|_p$$

This matrix norm is called *induced matrix norm*. □

Induced matrix norms are compatible. Among all compatible matrix norms, the induced matrix norm is the smallest one

$$\|A\vec{x}\|_p \leq \|A\| \|\vec{x}\|_p \quad \forall \vec{x} \quad \wedge \quad \|A\| \geq \|A\|_p$$

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

■ Examples: induced matrix norms

$$
\begin{aligned}
\|\vec{x}\|_1 &= |x_1| + \ldots + |x_n| \quad \Rightarrow \quad \|A\|_1 = \max_{j=1\ldots n} \left(\sum_{i=1}^{m} |a_{ij}| \right) \\
\|\vec{x}\|_2 &= \sqrt{\sum_{i=1}^{n} |x_i|^2} \quad \Rightarrow \quad \|A\|_2 = \sqrt{\lambda_{max}(A^H A)} \\
&\qquad\qquad\qquad\qquad = \max_{\vec{x} \neq 0} \sqrt{\vec{x}^H A^H A \vec{x} / (\vec{x}^H \vec{x})} \\
\|\vec{x}\|_\infty &= \max\{|x_1|, \ldots, |x_n|\} \quad \Rightarrow \quad \|A\|_\infty = \max_{i=1\ldots m} \left(\sum_{j=1}^{n} |a_{ij}| \right)
\end{aligned}
$$

$\|A\|_\infty$ is the maximum absolute row sum of the matrix and is called *row sum norm*, $\|A\|_1$ is the maximum absolute column sum of the matrix and is called *column sum norm*, $\|A\|_2$ is the *spectral norm*. □

■ Example

All induced norms are sub-multiplicative, the matrix norm $\|A\| := \max_{i,j} |a_{ij}|$ is not. The *Frobenius-norm* is a sub-multiplicative norm compatible with – but not induced by – the vector norm $\|\cdot\|_2$

$$\|A\|_F := \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} |a_{ij}|^2}$$

□

1.2 Fixed-Point Interations in Banach Spaces

When we talked about the solution of nonlinear equations in the last semester, we mainly focused on the scalar case: For $f \in \mathcal{C}^p([a,b], \mathbb{R})$ with $p \in \mathbb{N}$ we wanted to calculate x^* such that

$$f(x^*) = 0\,.$$

Mostly it is not possible to calculate an analytical solution. Therefore, we discussed efficient iterative methods, in each of which a sequence $\{x_k\}$ of approximations of x^* is generated such that

$$\lim_{k \to \infty} x_k = x^*\,.$$

The convergence properties of some of those methods have been analyzed in the framework of fixed point methods. For that we restricted our analysis to one-step methods $x_{k+1} = \phi(x_k)$ with $f, \phi \in \mathcal{C}^p(\mathbb{R}, \mathbb{R})$, p sufficiently large. Then

$$\lim_{k \to \infty} x_k = x^* \quad \Rightarrow \quad \phi(x^*) = x^*\,.$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

■ Example

What is the meaning of a fixed-point in Newton's method?

$$x_{k+1} = \phi(x_k) = x_k - \frac{f(x_k)}{f'(x_k)} \quad \xrightarrow{x_k = x_{k+1} = x^*} \quad 0 = -\frac{f(x^*)}{f'(x^*)} \quad \Rightarrow \quad f(x^*) = 0$$

□

Most of the statements derived for fixed-point iterations in the scalar case (e.g. Newton's method applied to the one-dimensonal case) can be generalized to the n-dimensional case (and more general also to Banach spaces) with only minor changes: Instead of the absolute value the vector norm is used.

By this, we can reuse the knowledge obtained here for the iterative solution of large linear systems.

Definition (Banach space = complete and normed vector space)

A Banach space B is a vector space with a metric (norm) that allows the computation of vector length and distance between vectors and it is complete in the sense that a Cauchy sequence of vectors always converges to a well defined limit vector that is within the space. □

■ Examples of Banach spaces

$(\mathbb{R}^n, \|\cdot\|_1)$, $(\mathbb{R}^n, \|\cdot\|_2)$, $(\mathbb{R}^n, \|\cdot\|_\infty)$

Remark

In $\mathbb{R}^n$ all norms are equivalent, i.e. let denote $\|\cdot\|_a, \|\cdot\|_b$ two different norms (e.g. $\|\cdot\|_2, \|\cdot\|_\infty, \ldots$), then there exist two constants $\alpha > 0, \beta > 0$

$$\ni \quad \alpha\|\vec{x}\|_a \leq \|\vec{x}\|_b \leq \beta\|\vec{x}\|_a \quad \forall \vec{x} \in \mathbb{R}^n .$$

Statements using one special norm are valid for all norms, only the values of constants (e.g. Lipschitz constant L) change, if the norm changes. □

Definition (Operators)

Let be X, Y normed spaces over $\mathbb{R}$. A function $T : D \to Y, D \subseteq X$, is called *operator*. Often we abbreviate Tx for $T(x)$ even if T is not linear.

A standard iteration scheme $x_{k+1} = \phi(x_k)$ for $X = \mathbb{R}$ can thus be written as $x_{k+1} = Tx_k$ with $T := \phi$.

The operator $T : D \to Y$ is *linear*, if D is a linear subspace of X and

$$T(x + \alpha y) = T(x) + \alpha T(y), \quad \forall x, y \in D, \alpha \in \mathbb{R}$$

The operator $T : D \to Y$ is *continuous* in $x_0 \in D$, if for $\{x_n\} \in D$ with $x_n \to x_0$ always follows $Tx_n \to Tx_0$.

$\varepsilon - \delta$-formulation as usual: $\|x - x_0\| < \delta \Rightarrow \|Tx - Tx_0\| < \varepsilon$. □

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Definition (Lipschitz condition)

Operator $T : D \to Y$ is in D *Lipschitz continuous* with *Lipschitz constant* L, if

$$\|Tx - Ty\| \leq L\|x - y\| \quad \forall x, y \in D$$

T is *contractive*, if $L < 1$ is possible.

If T is linear, then we can restrict our investigation to the case $y = 0$ and obtain

$$\|Tx\| \leq L\|x\| \quad \forall x \in D$$

In this case the smallest Lipschitz constant possible is called *operator norm* $\|T\|$ of T

$$\|T\| := \sup_{x \in D} \frac{\|Tx\|}{\|x\|}$$

□

Iteration methods in Banach spaces

Many problems in analysis – e.g. the iterative solution of nonlinear equations, the iterative solution of large and sparse linear systems – can be written as operator equations in a properly selected Banach space B

$$x = Tx \quad \text{with} \quad T : D \subseteq B \to B$$

A solution $x^* \in B$ of this equation is called *fixed-point* of T.

To calculate x^* often iteration methods are used

$$x_1 = Tx_0, x_2 = Tx_1, x_3 = Tx_2, \ldots, x_{n+1} = Tx_n, \ldots$$

■ **Important question:** Do we get convergence $x_n \to x^*$?

Necessary (!) condition:

The iteration method converges, if the corresponding operator T is contractive.

Banach's fixed-point theorem

- Let be B a Banach space, $D \subseteq B$ closed;
- let the operator $T : D \to B$ be a contractive mapping with Lipschitz constant $L < 1$ and
- let T map D into D: $T(D) \subseteq D$.

Then the equation $x = Tx$ admits a unique fixed-point $x^* \in D$.

If the sequence $\{x_n\}$ is iteratively generated by $x_{n+1} = Tx_n$, then it converges to x^*. The following estimate holds

$$\|x_n - x^*\| \leq \frac{1}{1-L}\|x_{n+1} - x_n\| \leq \frac{L^n}{1-L}\|x_1 - x_0\|$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

1.3 Error Propagation – Basics

The errors investigated in this context are unavoidable and problem-induced, even if a perfect numerical algorithm is chosen. To reduce the errors one has to reformulate the underlying mathematical problem.

Only input errors are considered, other error sources like rounding errors are neglected here.

Input errors may be measurement errors, but also all errors (e.g. rounding and discretization errors) made in previous iteration steps or in the solution of preceeding subproblems.

Definition of the problem

Let be $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ normed vector spaces and consider a mapping $f : X \to Y$ (often called "problem" or "subproblem").

For $x, \delta x \in X$ we define

$$\delta f := f(x + \delta x) - f(x)$$

Important question

How large is the perturbation δf of the solution compared to the perturbation δx of the input data?

Definition: absolute and relative condition for the norm error

The *absolute condition number* of f in $x \in X$ is

$$\kappa_{abs}(f, x) := \lim_{\delta \to 0} \left(\sup_{\|\delta x\| < \delta} \frac{\|\delta f\|_Y}{\|\delta x\|_X} \right)$$

Using the norm allows an approach from different directions!

The *relative condition number* of f in $x \in X$ is

$$\kappa_{rel}(f, x) := \lim_{\delta \to 0} \left(\sup_{\|\delta x\| < \delta} \frac{\|\delta f\|_Y / \|f\|_Y}{\|\delta x\|_X / \|x\|_X} \right)$$

A problem is *ill-conditioned*, if $\kappa_{abs}(f, x) \gg 1$ or $\kappa_{rel}(f, x) \gg 1$. □

Remark

If the condition number is large, a small perturbation in the input data causes large perturbations in the final result. The condition numbers compress the information about error amplification into one scalar. □

Remark

Introducing condition numbers can be seen as a linearization of the original problem. This leads to an equivalent definition of the condition numbers. □

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

■ Example

Consider $\vec{f} \in \mathcal{C}^1(D, \mathbb{R}^m)$, $D \subseteq \mathbb{R}^n$. Multidimensional Taylor expansion with truncation after the linear term yields

$$\delta\vec{f} \doteq Df(\vec{x}) \cdot \delta\vec{x}, \quad Df(\vec{x}) \text{ Jacobian}$$

For the calculation of the condition numbers for this problem we use matrix norms and obtain

$$\kappa_{abs}(\vec{f}, \vec{x}) = \|Df(\vec{x})\|, \quad \kappa_{rel}(\vec{f}, \vec{x}) := \frac{\|Df(\vec{x})\|}{\|\vec{f}(\vec{x})\|/\|\vec{x}\|}$$

□

■ Example (perturbed linear system)

Definition of possible perturbations:

Let us consider the perturbation

$$A\vec{x} = \vec{b} \quad \rightarrow \quad (A + \delta A)(\vec{x} + \delta\vec{x}) = \vec{b} + \delta\vec{b}, \quad A \in \mathbb{R}^{n\times n}, \vec{x}, \vec{b} \in \mathbb{R}^n$$

Does a unique solution of the perturbed system exist?

We assume that the matrix A is non-singular and that the perturbation δA small enough such that also $\det(A + \delta A) \neq 0$. How small is small enough?

To answer that question, we analyze the kernel of $A + \delta A$. We know that the singular case $\det(A + \delta A) = 0$ is equivalent to the existence of at least one $\vec{x} \neq \vec{0}$ such that $(A + \delta A)\vec{x} = \vec{0}$. In that case

$$\begin{aligned}(A + \delta A)\vec{x} = \vec{0} \quad &\Rightarrow \quad \vec{x} = -A^{-1}\delta A\,\vec{x} \quad \Rightarrow \quad \|\vec{x}\| \leq \|A^{-1}\| \cdot \|\delta A\| \cdot \|\vec{x}\| \\ &\Rightarrow \quad (1 - \|A^{-1}\| \cdot \|\delta A\|) \cdot \|\vec{x}\| \leq 0 \quad \Rightarrow \quad \|\delta A\| \geq \frac{1}{\|A^{-1}\|}\end{aligned}$$

If $\|\delta A\| < 1/\|A^{-1}\|$, the norm estimate in the first line can be valid only for $\vec{x} = \vec{0}$. Thus $\vec{x} = \vec{0}$ is the only solution of $(A + \delta A)\vec{x} = \vec{0}$ and therefore $A + \delta A$ is non-singular.

Calculation of the condition numbers:

Consider the vector function $\vec{f} : \mathbb{R}^n \to \mathbb{R}^n$ with $\vec{f}(\vec{b}) := A^{-1}b$; using the definition of the total derivative we get

$$\begin{aligned}\kappa_{abs}(\vec{f}, \vec{b}) &= \|\vec{f}'(b)\| = \|A^{-1}\| \\ \kappa_{rel}(\vec{f}, \vec{b}) &= \frac{\|A^{-1}\|}{\|A^{-1}\vec{b}\|/\|\vec{b}\|} = \|A^{-1}\|\frac{\|A\vec{x}\|}{\|\vec{x}\|} \leq \|A^{-1}\|\|A\|\end{aligned}$$

Analogously we get for $\vec{g} : \{A \in \mathbb{R}^{n\times n} \mid \det A \neq 0\} \to \mathbb{R}^n$ with $\vec{g}(A) := A^{-1}\vec{b}$

$$\kappa_{rel}(\vec{g}, A) \leq \|A^{-1}\|\|A\|$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Therefore this expression is defined as the ”condition of the linear system” and denoted by $\kappa_{rel}(A) := \|A^{-1}\|\|A\|$ (rough estimate!).

For an *induced matrix norm*, we may further rewrite this as

$$\|A\|_p\|A^{-1}\|_p = \frac{\max_{\|\vec{x}\|_p=1} \|A\vec{x}\|_p}{\min_{\|\vec{z}\|_p=1} \|A\vec{z}\|_p}$$

Properties of $\kappa_{rel}(A)$:

- $\kappa_{rel}(A) \geq 1$.
- $\kappa_{rel}(A) = \kappa_{rel}(\alpha A) \quad \forall \alpha \in \mathbb{R}, \alpha \neq 0$.
- $\kappa_{rel}(A) = \infty \quad \Leftrightarrow \quad \det(A) = 0$

In contrast to the determinant, the condition of the linear system is invariant under scaling. □

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

2 Iterative Solution of Linear Systems

Problem

We want to solve the (large) linear system

$$Ax = b, \quad A \in \mathbb{C}^{n\times n}, b \in \mathbb{C}^n$$

with $n^2 >$ available storage and $a_{ik} = 0$ for almost all i, k ("sparse matrix").

A special structure of A (e.g. banded matrix with $a_{ik} = 0$ for $|i-k| >$ const) is not necessary.

The direct methods for the solution of linear systems investigated up to now are not suited. Even in case of sufficient storage, *fill-in occurs* when performing e.g. Gaussian elimination on a sparse matrix: Many entries of the matrix change from zero to a non-zero value in the execution of the algorithm.

■ Example

- Finite Elements, Finite Differences for the solution of PDEs
- Control theory, stationary state: $\dot{x} = Ax + Bu \;\Rightarrow\; Ax + B(Fx) = 0$

■ **Example** (discretization of Laplace's equation)

Consider Laplace's equation (an elliptic PDE)

$$u_{xx} + u_{yy} = 0 \;\; \forall (x,y) \in [0,1] \times [0,1]$$

with the boundary conditions $u(0,y) = 0, u(1,y) = 1, u(x,0) = 0, u(x,1) = 0$.

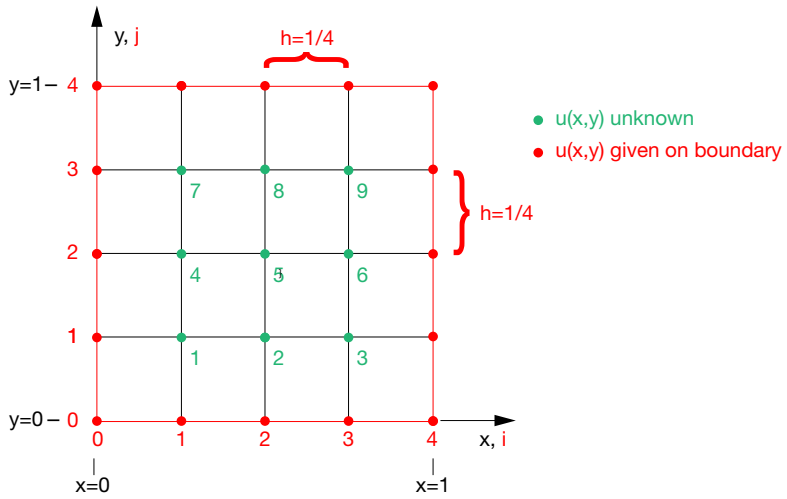


Figure 1: Grid for finite difference approach to Laplace's equation

In the simplest case the differential quotient is approximated by the difference quotient ($\rightarrow$ five-point stencil of a point in the grid made up of the point itself together with its four "neighbors")

$$u_{xx} = \frac{u(x+h,y) - 2u(x,y) + u(x-h,y)}{h^2}, \quad u_{yy} = \ldots$$

© Rainer Callies TUM 2020

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

We make use of the uniform grid $x_i = i \cdot h, y_j = j \cdot h$ and define $u_{ij} := u(x_i, y_j)$ to obtain a linear system

$$u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1} - 4u_{i,j} = 0$$

The result is a sparse linear system for the unknowns $u_{ij}, i = 1,\ldots n, j = 1,\ldots,m$. These u_{ij} are sorted in the z-vector of the system $Az = b$; the sorting strategy affects the structure of A.

For the example system sketched in the last figure we obtain using row-wise numbering (3 rows with 3 variables each)

$$A = \begin{pmatrix} T & B & 0 \\ B & T & B \\ 0 & B & T \end{pmatrix}, \quad T = \begin{pmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{pmatrix}_{3\times 3}, \quad B = \mathrm{diag}(-1,-1,-1)$$

and

$$Az = A \begin{pmatrix} u_{11} \\ u_{21} \\ u_{31} \\ u_{12} \\ \vdots \\ u_{33} \end{pmatrix} = \begin{pmatrix} u_{10} + u_{01} \\ u_{20} \\ u_{30} + u_{41} \\ u_{02} \\ \vdots \\ u_{34} + u_{4,3} \end{pmatrix} = b$$

The five-point stencil at the grid point with the (green) number k leads to the k-th row of the linear system.

If system size increases, the block structure remains unchanged, but the size and the number of the blocks increase accordingly and the ratio of the zero elements increases too. □

Solution strategy: iterative methods

The objective of the algorithms is to limit fill-in and storage requirements.

The memory requirement per non-zero element is larger than in case of direct methods, because not only the value of a_{ik} has to be stored, but also its address i, k. Fortunately, the numer of non-zero elements is (relatively) low. □

2.1 Linear Iterative Methods – Stationary Methods

2.1.1 Introduction

Definition

Let us consider the problem $Ax = b, \;\; A \in \mathbb{C}^{n\times n}, b \in \mathbb{C}^n$.

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

An iterative method is called convergent, if for each initial guess $x^{(0)} \in \mathbb{C}^n$ (or $\mathbb{R}^n$) a sequence $\{x^{(m)}\}$ of approximations is generated that converges to the solution x^* of the linear system $Ax = b$

$$\lim_{m\to\infty} x^{(m)} = x^*$$

The iterative method is called consistent, if x^* is a fixed-point of the iteration.

The iterative method is called linear, if $x^{(m+1)}$ depends linearly on $x^{(m)}$ and b

$$x^{(m+1)} = Mx^{(m)} + Nb$$

with the square matrices $M, N \in \mathbb{C}^{n\times n}$ properly chosen. □

■ Example

The *modified Richardson iteration* is

$$x^{(m+1)} = x^{(m)} + \omega\left(b - Ax^{(m)}\right)$$

where ω is a scalar parameter that has to be chosen such that the sequence $\{x^{(m)}\}$ converges. It is easy to see that the method has the correct fixed point x^* and thus is consistent.

In our notation, $N = \omega I$ and $M = I - \omega A$.

If there are both positive and negative eigenvalues of A, the method will diverge for any ω if the initial error $x^{(0)} - x^*$ has nonzero components in the corresponding eigenvectors.

We observe that in each iteration step the main effort is one matrix-vector multiplication only corresponding to $\mathcal{O}(N^2)$ operations for general matrices and $\mathcal{O}(N)$ operations for sparse matrices. □

2.1.2 Classical linear iterative methods

Let denote D the diagonal part of $A \in \mathbb{R}^{n\times n}$, E its strictly lower triangular part and F its strictly upper triangular part. We get the additive decomposition

$$A = E + D + F$$

■ Example

$$A := \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 11 & 12 & 13 & 14 \\ 15 & 16 & 17 & 18 \end{pmatrix} \Rightarrow$$

$$E = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 \\ 11 & 12 & 0 & 0 \\ 15 & 16 & 17 & 0 \end{pmatrix}, D = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 \\ 0 & 0 & 13 & 0 \\ 0 & 0 & 0 & 18 \end{pmatrix}, F = \begin{pmatrix} 0 & 2 & 3 & 4 \\ 0 & 0 & 7 & 8 \\ 0 & 0 & 0 & 14 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

The following iterations again are all *fixed-point iterations* and thus consistent.

Iterations schemes for different classical methods

Choose an initial guess $x^{(0)}$ to the solution x^* of the linear system $Ax = b$.

for $m = 0$ **to** p **do**

$$A_1 x^{(m+1)} + A_2 x^{(m)} = b \quad \Rightarrow \quad A_1 x^{(m+1)} = b - A_2 x^{(m)}$$

$x^{(m)}$ denotes the result after the m-th iteration cycle. Comparing the above decomposition with the general definition, we get $M = -A_1^{-1} A_2$ and $N = A_1^{-1}$.

Now we choose for the

Jacobi method (J)	$: A_1 = D$	, $A_2 = E + F$
Gauß-Seidel method (GS)	$: A_1 = D + E$	, $A_2 = F$
Successive over-relaxation (SOR)	$: A_1 = D/\omega + E$	, $A_2 = (1 - 1/\omega)D + F$

We always decompose such that $A_1 + A_2 = A$ (additive decomposition).

Motivation for the choice of those A_1:
The resulting linear system $A_1 x^{(m+1)} = \ldots$ can be efficiently solved.

The initial guess $x^{(0)}$ of course should be as good as possible, but in principle there is no restriction for the choice.

Iteration stops as soon as $\|x^{(p+1)} - x^*\| < tol$. □

What basic idea do these methods have in common?

We start with the residual vector $r^{(m)} := b - Ax^{(m)}$ into the $(m+1)$th cycle.

In the i-th substep ($i = 1, \ldots, n$) we try to make one (!) component $r_i^{(m)}$ precisely zero: For that we choose an index pair (i, k) with $a_{ik} \neq 0$ and modify

$$x_k^{(m)} \to x_k^{(m)} + \delta x_k^{(m)} =: x_k^{(m+1)} \quad \text{such that}$$

$$r_i^{(m)} - a_{ik}\delta x_k^{(m)} = b_i - \sum_{j=1}^{n} a_{ij} x_j^{(m)} - a_{ik}\delta x_k^{(m)} = 0 \;\Rightarrow\; \delta x_k^{(m)} = \ldots \quad (*)$$

From the linear equation $(*)$ the correction $\delta x_k^{(m)}$ can be calculated. After that iteration step the i-th equation is exactly fulfilled for one moment, but already in the next step that property is destroyed again and another equation is exactly fulfilled (i.e. another component of the residual is precisely zero).

We have to assure that each row – and thus each component of the residual – is reached.

The different methods mainly differ in their strategy to choose the sequence of index pairs (i, k). *We hope*, that for a proper choice of the index pairs (i, k) the iteration converges: $x^{(m)} \to x^*$. That remains to be studied in detail.

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

In the *Gauss-Seidel method* (GS) we cyclically choose $i = k = 1, 2, \ldots, n$ and thus carry out m times complete cycles of n steps to obtain $x^{(m)}$. Permutations may be necessary to achieve $a_{ii} \neq 0$.

Successive Over-Relaxation (SOR) refines that process by choosing a relaxation factor $\omega \neq 0$ (mostly $\omega \in]1, 2[$). Therefore, the older method (GS) is a special case of (SOR).

Algorithm 1: Method of successive over-relaxation (SOR)

■ Start: $x^{(0)}$ arbitrary initial guess

for $m = 1$ **to** p **do**

  Choose relaxation factor $\omega = \omega(m) \neq 0$

  for $i = 1$ **to** n **do**

$$x_i := x_i + \omega \left(b_i - \sum_{j=1}^{n} a_{ij} x_j \right) / a_{ii}$$

The *Jacobi method* (J) modifies the inner loop:

...

for $i = 1$ **to** n **do**

$$x_i' := \left(b_i - \sum_{j=1, j \neq i}^{n} a_{ij} x_j \right) / a_{ii} = x_i + \left(b_i - \sum_{j=1}^{n} a_{ij} x_j \right) / a_{ii}$$

for $i = 1$ **to** n **do**

$$x_i := x_i'$$

...

In step $i = 1, \ldots n$ of the $(m+1)$-th cycle the i-th component $x_i^{(m)}$ is modified such that $r_i^{(m)} = 0$; the modification is not immediately applied, but only stored. After the complete cycle has been finished all stored modifications are applied simultaneously ($x_i^{(m)} \to x_i^{(m+1)}$ for $i = 1, \ldots, n$). This method nowadays is not used very often. □

Remark

For sparse A, the iteration steps in (GS), (SOR) and (J) are cheap!

We prove: The (J) algorithm is compatible with the matrix formulation.

By construction of the (J) algorithm we get

$$x^{(m+1)} = x^{(m)} + D^{-1}(b - Ax^{(m)}) = D^{-1}(b - Ex^{(m)} - Fx^{(m)})$$

Then we multiply the complete equation by the matrix D. □

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

We prove: The (SOR) algorithm is compatible with the matrix formulation.

Let us analyze step i of the inner loop in the $(m+1)$-th cycle.
In that case the components $x_1^{(m+1)}, \ldots, x_{i-1}^{(m+1)}$ in the inner loop already have been updated. For $i \in \{1, \ldots, n\}$ step i can be written in detail

$$\begin{aligned} x_i^{(m+1)} &= x_i^{(m)} + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(m+1)} - \sum_{j=i}^{n} a_{ij} x_j^{(m)} \right) \\ &= x_i^{(m)} + \omega \left(D^{-1}(b - Ex^{(m+1)} - Dx^{(m)} - Fx^{(m)}) \right)_{i-\text{th component}} \end{aligned}$$

Multiplication by D/ω from the left yields the desired result. □

Convergence Theorem 1

If $\{x^{(m)}\}$ converges at all, then it converges to x^*.

Proof:

If $x^* = \lim_{m\to\infty} x^{(m)}$ exists, then we get by insertion into the matrix formulation

$$A_1 x^* + A_2 x^* = b \;\Rightarrow\; x^* = (A_1 + A_2)^{-1} b = A^{-1} b$$

and x^* is uniquely determined for $\det(A) \neq 0$. □

Definition

Let denote $\varepsilon^{(m)} := x^{(m)} - x^*$ the error after the m-th iteration cycle
and $\varrho := \max\limits_{1\le i\le n} |\lambda_i(A_1^{-1}A_2)|$ the *spectral radius* (= largest absolute value of the eigenvalues) of the matrix $A_1^{-1}A_2$. $(-A_1^{-1}A_2)$ is called *iteration matrix*. □

Convergence Theorem 2

It is $\varepsilon^{(m+1)} = -A_1^{-1}A_2 \cdot \varepsilon^{(m)}$ and consequently $\varepsilon^{(m)} = (-A_1^{-1}A_2)^m \varepsilon^{(0)}$.

Proof:

$$A_1 x^{(m+1)} + A_2 x^{(m)} = b = A_1 x^* + A_2 x^* \;\Rightarrow\; A_1 \varepsilon^{(m+1)} + A_2 \varepsilon^{(m)} = 0$$

The second formula can be proven by induction. □

Convergence Theorem 3

$$\lim_{m\to\infty} x^{(m)} = x^* = A^{-1} b \;\; \forall x^{(0)} \quad \Leftrightarrow \quad \varrho < 1$$

Moreover, convergence is "linear": $\varepsilon^{(m)} = \mathcal{O}(\varrho^m)$

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Proof: (only "⇒", by contradiction)

If $\varrho \geq 1$, then there exists at least one EV v_{max} for the maximum EW λ_{max} with $|\lambda_{max}| \geq 1$. If $x^{(0)} = x^* - v_{max}$, then by definition $\varepsilon^{(0)} = v_{max}$.

From convergence theorem 2 we get (using the EW/EV definition)

$$\varepsilon^{(m)} = (-A_1^{-1}A_2)^m v_{max} \overset{EV}{=} (-1)^m \lambda_{max}^m v_{max}$$

Taking the norm gives $\|\varepsilon^{(m)}\| = |\lambda_{max}|^m \|v_{max}\| \geq \|v_{max}\|$ and thus no convergence $\lightning$. □

Remark

Per decimal digit of precision therefore $-1/\log_{10}(\varrho)$ cycles have to be performed. □

Convergence Theorem 4

- $A \in \mathbb{C}^{n\times n}$:
  (SOR) converges – if at all – only for $\omega \in]0,2[$.
- $A \in \mathbb{C}^{n\times n}$ positive definite:
  (GS) converges, (SOR) converges for $\omega \in]0,2[$ fixed,
  convergence of (J) not granted
- $A \in \mathbb{C}^{n\times n}$ strictly diagonal-dominant, i.e. $|a_{ii}| > \sum_{j=1,\neq i}^{n} |a_{ij}|, \;\; i = 1,\ldots,n$:
  (J) and (GS) converge.

Convergence rates of the methods (GS), (SOR) and (J) are investigated later.

■ Example

Jacobi method and 5-point stencil for Laplace's equation

$$A = \left(\begin{array}{cccc|cccc|c}
4 & -1 & & 0 & -1 & & & 0 & \\
-1 & \ddots & \ddots & & & \ddots & & & \\
 & \ddots & \ddots & -1 & & & \ddots & & \\
0 & & -1 & 4 & 0 & & & -1 & \\
\hline
-1 & & & 0 & 4 & -1 & & 0 & \\
 & \ddots & & & -1 & \ddots & \ddots & & \\
 & & \ddots & & & \ddots & \ddots & -1 & \\
0 & & & -1 & 0 & & -1 & 4 & \\
\hline
 & & & & & & & & \\
\end{array}\right)_{N^2\times N^2}$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

$$-A_1^{-1}A_2 \;=\; -D^{-1}(E+F) = -\frac{1}{4}I(A-4I) = I-\frac{1}{4}A$$

$$= \frac{1}{4}\left(\begin{array}{cccc|cccc|c}
0 & 1 & & 0 & 1 & & & 0 & \\
1 & \ddots & \ddots & & & \ddots & & & \\
 & \ddots & \ddots & 1 & & & \ddots & & \\
0 & & 1 & 0 & 0 & & & 1 & \\
\hline
1 & & & 0 & 0 & 1 & & 0 & \\
 & \ddots & & & 1 & \ddots & \ddots & & \\
 & & \ddots & & & \ddots & \ddots & 1 & \\
0 & & & 1 & 0 & & 1 & 0 & \\
\hline
 & & & & & & & &
\end{array}\right)_{N^2\times N^2}$$

Claim:

$-A_1^{-1}A_2$ has the N^2 eigenvectors $z^{(k,l)}$, $k,l=1,\ldots,N$, with the components

$$z^{(k,l)}_{(i-1)\cdot N+j} := \sin\left(\frac{k\pi i}{N+1}\right)\sin\left(\frac{l\pi j}{N+1}\right)$$

and the correspondig eigenvalues

$$\lambda^{(k,l)} := \frac{1}{2}\left(\cos\frac{k\pi}{N+1}+\cos\frac{l\pi}{N+1}\right)$$

The index i refers to the i-th block row and the index j to the number of the single row within this block row.

Proof:

For e.g. the 1^{st} component we get using trigonometric angle sum identities

$$\begin{aligned}
& 4(-A_1^{-1}A_2 z^{(k,l)})_1 \\
= \;& z_2^{(k,l)} + z_{N+1}^{(k,l)} = z^{(k,l)}_{i=1,j=2} + z^{(k,l)}_{i=2,j=1} \\
= \;& \sin\frac{k\pi}{N+1}\sin\frac{2l\pi}{N+1} + \sin\frac{2k\pi}{N+1}\sin\frac{l\pi}{N+1} \\
= \;& \sin\frac{k\pi}{N+1}\cdot 2\sin\frac{l\pi}{N+1}\cos\frac{l\pi}{N+1} + 2\sin\frac{k\pi}{N+1}\cos\frac{k\pi}{N+1}\cdot\sin\frac{l\pi}{N+1} \\
= \;& 4\lambda^{(k,l)} z_1^{(k,l)}
\end{aligned}$$

Conclusion: For the convergence rate we get

$$\varrho(-A_1^{-1}A_2) = \max_{k,l}|\lambda^{(k,l)}| = \cos\frac{\pi}{N+1}$$

This is a typical behavior often observed when iteration methods are applied to linear systems resulting from the discretization of PDEs by mult-point stencils: If system size increases, not only the effort per iteration cycle increases, but the number of necessary cycles increases too! □

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Definition (consistently ordered matrix)

Given $A = D + E + F = D(I + L + U) \in \mathbb{C}^{n\times n}$ with $L := D^{-1}E$ and $U := D^{-1}F$. We define $J(\alpha) := -(\alpha L + \alpha^{-1}U)\,,\ \alpha \in \mathbb{C} \setminus \{0\}$.

A is *consistently ordered*, if the EWs of the matrix $J(\alpha)$ are independent of α.

□

Remark

By *reordering* the variables $x_1, \ldots x_n$ of a linear system $Ax = b$, the resulting and new matrix can be consistently *ordered* (that is the reason for the notion).

□

■ **Example** (tridiagonal matrices)

$$\begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \quad \Rightarrow \quad J(\alpha) = -\frac{1}{2}\begin{pmatrix} 0 & 1/\alpha & 0 \\ \alpha & 0 & 1/\alpha \\ 0 & \alpha & 0 \end{pmatrix}$$

$$\Rightarrow \quad J(\alpha) - \lambda I = -\frac{1}{2}\begin{pmatrix} 2\lambda & 1/\alpha & 0 \\ \alpha & 2\lambda & 1/\alpha \\ 0 & \alpha & 2\lambda \end{pmatrix}$$

Expanding the determinant $\det(J(\alpha) - \lambda I) = 0$ along a column yields

$$2\lambda((2\lambda)^2 - \alpha \cdot (1/\alpha)) - \alpha\frac{2\lambda}{\alpha} = 0$$

and thus the characteristic polynomial – and with it the EWs – are independent of α. □

Convergence Theorem 5 (for consistently ordered matrices)

Assumption:	A consistently ordered
Claim:	$\varrho_{Gauß-Seidel} = (\varrho_{Jacobi})^2$
Remark:	(J) needs approximately twice as many iterations as (GS).

Convergence Theorem 6 (for consistently ordered matrices)

Assumption:	A consistently ordered EWs of $J = J(\alpha)$ are real-valued $\varrho_J := \varrho_{Jacobi} < 1$ $H(\omega) := -A_1^{-1}A_2 = -(D/\omega + E)^{-1}((1 - 1/\omega)D + F)$ defines the iteration matrix of the SOR method

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Claim: For the optimal relaxation parameter ω_b we get

$$\omega_b := \arg\min_{\omega\in]0,2[} \varrho(H(\omega)) = \frac{2}{1+\sqrt{1-\varrho_J^2}}, \quad \varrho(H(\omega_b)) = \omega_b - 1$$

and in general

$$\varrho(H(\omega)) = \begin{cases} \omega - 1 & , \ \omega \in [\omega_b, 2] \\ 1-\omega+\omega^2\varrho_J^2/2+\omega\varrho_j\sqrt{1-\omega+\omega^2\varrho_J^2/4}, & \omega\in]0,\omega_b] \end{cases}$$

Remark: If only a coarse estimate of ω_b is known, it is better to choose ω_b a little bit larger than the resulting ω_b.

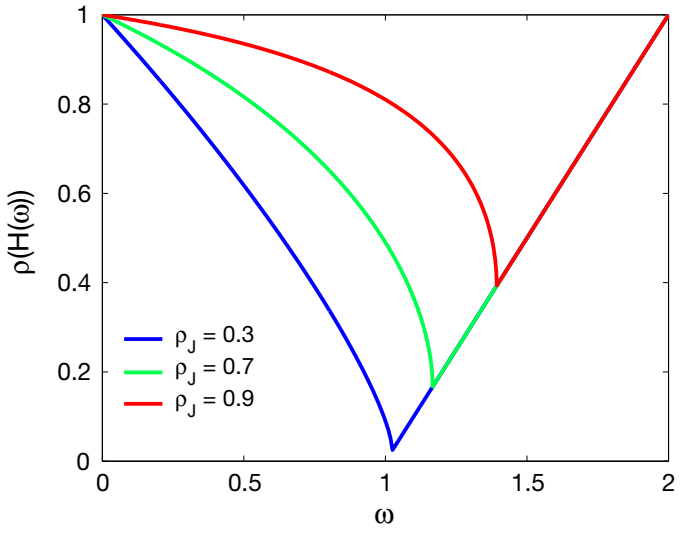


Figure 2: Spectral radius ϱ of SOR as a function of the relaxation parameter ω

■ **Example** (Laplace's eq., 5-point stencil, example continued)

For this example we have proven $\varrho_J = \cos\left(\frac{\pi}{N+1}\right)$. With that information we get for (SOR) and for (GS) – as a special case of (SOR) –

$$\begin{aligned} \varrho_{GS} &= \varrho(H(1)) = \varrho_J^2 = \cos^2\left(\frac{\pi}{N+1}\right) \\ \omega_b &= \frac{2}{1+\sin\left(\frac{\pi}{N+1}\right)} \\ \varrho(H(\omega_b)) &= \omega_b - 1 = \left(\frac{\varrho_J}{1+\sqrt{1-\varrho_J^2}}\right)^2 = \frac{\cos^2\left(\frac{\pi}{N+1}\right)}{\left(1+\sin\left(\frac{\pi}{N+1}\right)\right)^2} \end{aligned}$$

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Question:
How man cycles of (J) do we need instead of one *optimal* (SOR) cycle?

Answer:

$$\varrho_J^k = \varrho(H(\omega_b)) \;\Rightarrow\; k = \frac{\ln \varrho(H(\omega_b))}{\ln \varrho_J}$$

We want to estimate that expression using several Taylor expansions up to $\mathcal{O}(N^{-3})$ to get an idea of the order of magnitude of that effect:

$$\begin{aligned}
\ln \varrho(H(\omega_b)) &= 2\left(\ln \varrho_J - \ln\left(1+\sin\frac{\pi}{N+1}\right)\right) \\
\cos\left(\frac{\pi}{N+1}\right) &= 1-\frac{1}{2}\left(\frac{\pi}{N+1}\right)^2+\mathcal{O}(N^{-4}) \qquad \text{for } N \gg 1 \\
\ln(1+z) &= z+\mathcal{O}(z^2) = z - z^2/2 + \mathcal{O}(z^3) \;\Rightarrow \\
\\
\ln \varrho_J &= \ln\left(\underbrace{1-\frac{1}{2}\left(\frac{\pi}{N+1}\right)^2+\mathcal{O}(N^{-4})}_{=:z}\right) \\
&= -\frac{1}{2}\left(\frac{\pi}{N+1}\right)^2+\mathcal{O}(N^{-4}) \\
1+\sin\left(\frac{\pi}{N+1}\right) &= 1+\frac{\pi}{N+1}+\mathcal{O}(N^{-3}) \\
\ln\left(1+\sin\left(\frac{\pi}{N+1}\right)\right) &= \frac{\pi}{N+1}+\mathcal{O}(N^{-3})-\frac{1}{2}\left(\frac{\pi}{N+1}\right)^2+\mathcal{O}(N^{-6}) \\
\ln \varrho(H(\omega_b)) &= 2\left(\ln \varrho_J - \ln\left(1+\sin\frac{\pi}{N+1}\right)\right) = -\frac{2\pi}{N+1}+\mathcal{O}(N^{-3}) \\
\Rightarrow\; k(N) &\approx \frac{4}{\pi}(N+1)
\end{aligned}$$

In our example, the optimal SOR method is more than N times faster than (J) !!

□

Block Iteration Methods

Block iteration schemes are generalizations of the ”point” iteration schemes described above. They update a whole set of components at each time, typically a subvector of the solution vector, instead of only one component. The matrix A and the right-hand side and solution vectors of $Ax = b$ are partitioned as follows:

$$A \;\rightarrow\; \begin{pmatrix} A_{11} & \cdots & A_{1M} \\ \vdots & & \vdots \\ A_{M1} & \cdots & A_{MM} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_M \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_M \end{pmatrix}$$

in which the partitionings of b and x into subvectors b_i and x_i are identical and compatible with the partitioning of A.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

We assume that the A_{ij} are square matrices with $\det(A_{ii}) \neq 0$.

Now we define, similarly to the scalar case, the splitting $A = D + E + F$ with

$$D = \begin{pmatrix} A_{11} & & & \\ & A_{22} & & \\ & & \ddots & \\ & & & A_{MM} \end{pmatrix}, \quad E = \begin{pmatrix} 0 & & & \\ A_{12} & 0 & & \\ \vdots & \vdots & \ddots & \\ A_{M1} & A_{M2} & \cdots & 0 \end{pmatrix}, \quad F = \dots$$

With these definitions, it is easy to generalize the iterative methods defined earlier – e.g. Jacobi, Gauss-Seidel, and SOR –, which made one scalar component $r_i^{(m)}$ of the residual equal to zero in each substep.

For example, the block Jacobi iteration is now defined as a technique in which the new subvectors $x_i^{(m+1)}$ are all calculated according to

$$\begin{aligned} Dx^{(m+1)} &+ (A-D)x^{(m)} = b \quad \Rightarrow \\ A_{jj}x_j^{(m+1)} &= b_j - \sum_{k=1,k\neq j}^{M} A_{jk}x_k^{(m)}, \quad j = 1 \dots M \end{aligned}$$

The iterative method simply is applied blockwise instead of componentwise.

The (much smaller) linear subsystems $A_{jj}x_j^{(m+1)} = \dots$ can be solved by a direct method (LU, QR, ...) each.

■ Example

With finite difference approximations of PDEs, it is standard to block the variables and the matrix by partitioning along whole lines of the mesh. More general, a block can also correspond to the unknowns associated with a few consecutive lines in the plane. One such blocking is illustrated for a 6×6 grid:

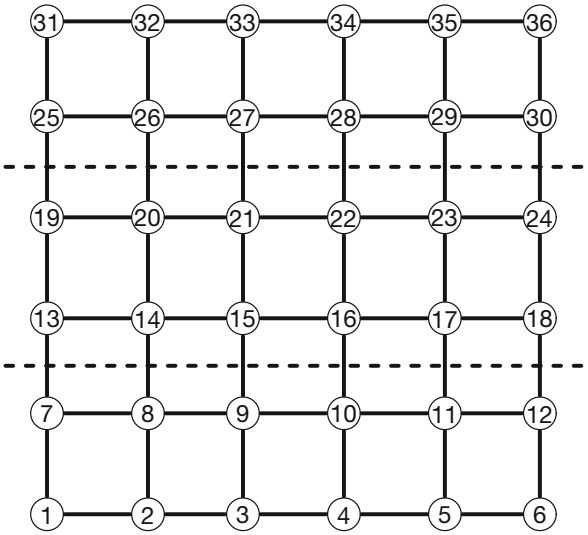


Figure 3: Blocking example for a 6×6 mesh: partitioning into three subdomains

The corresponding matrix has the following block structure:

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

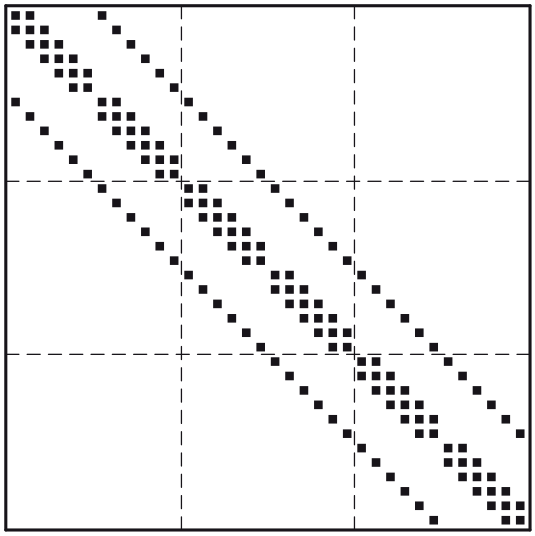


Figure 4: Block structure of the matrix A associated with that mesh

□

The advantage of block iterations is the smaller number of iteration cycles (in our benchmark example of the 5-point stencil the number of iterations depends on N and in case of block iterations on the much smaller number $M \ll N$). So the number of cycles required to achieve convergence often decreases rapidly as the block-size increases.

The disadvantage is, that the effort per cycle significantly increases, because the subproblems – linear systems with A_{ii} – have to be solved directly. Moreover fill-in may occur in the subproblems.

Finally, block techniques can be defined in more general terms. First, by using blocks that allow us to update arbitrary groups of components, and second, by allowing the blocks to overlap. This is a form of the domain-decomposition method.

Theorem

Let be A a matrix in block-tridiagonal form

$$A = \begin{pmatrix} D_1 & A_{12} & & & \\ A_{21} & D_2 & A_{23} & & \\ & A_{32} & \ddots & \ddots & \\ & & \ddots & \ddots & A_{M-1,M} \\ & & & A_{M,M-1} & D_M \end{pmatrix}$$

and let the $A_{ii} = D_i$ be square diagonal matrices.

Then A is called *T*-matrix and A is consistently ordered.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

2.2 Methods Based on Minimization – Krylov Subspace Methods

2.2.1 Fundamental Idea

Let us solve $Ax = b,\ A \in \mathbb{C}^{n\times n} \ \wedge \ A$ positive definite.

This problem is substituted by the minimization problem

$$\min_{x\neq 0} f(x) \quad \text{with} \quad f(x) := \frac{1}{2}x^T Ax - b^T x$$

2.2.2 Simplest realization: Gradient method

The gradient is the direction of steepest descent; naively one might think that this qualifies the method to find the minimum most quickly.

Algorithm 2: Gradient method

- Start: $x^{(0)}$ arbitrary initial guess

for $k = 0, 1, 2 \ldots$ **do**

(A) Determine the search direction (gradient)

$$d_k := -\nabla f(x^{(k)}) = b - Ax^{(k)}$$

(B) One-dimensional minimization along the search direction

$$\alpha_k := \arg\min_{t\geq 0}\{f(x^{(k)} + t\cdot d_k)\}$$

We can explicitly determine the argument α_k from

$$\begin{aligned} 0 &\stackrel{!}{=} \frac{\mathrm{d}}{\mathrm{d}t} f(x^{(k)} + t\cdot d_k) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{1}{2}(x^{(k)} + t\cdot d_k)^T A(x^{(k)} + t\cdot d_k) - b^T(x^{(k)} + t\cdot d_k)\right] \\ &= t d_k^T A d_k + \left({x^{(k)}}^T A d_k - b^T d_k\right) = t d_k^T A d_k - d_k^T d_k \\ \Rightarrow \quad t &= \frac{d_k^T d_k}{d_k^T A d_k} =: \alpha_k \end{aligned}$$

(C) Update: $x^{(k+1)} := x^{(k)} + \alpha_k d_k$

Remarks

- The method converges to the solution x^*, if A is positive definite.
- The method is converging locally for α sufficiently small and $\nabla f(x^{(k)}) \neq 0$:

$$f(x^{(k)} + \alpha d_k) = f(x^{(k)}) + \nabla f(x^{(k)})^T\left(-\alpha \nabla f(x^{(k)})\right) + \mathcal{O}(\alpha^2 d_k^2) < f(x^{(k)})$$

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

■ Caution necessary if $\nabla f(x^{(k)})$ is determined numerically e.g. from finite difference approximation (error in search direction!). □

Rate of convergence

Let be $A \in \mathbb{R}^{n\times n}$ positive definite with EWs $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$; let us define $\kappa := \operatorname{cond}_2 A = \lambda_n/\lambda_1$ and $f(x) := \frac{1}{2}x^T Ax - b^T x$.

Minimization of the quadratic function $f(x)$ by the Gradient method produces a sequence $\left\{x^{(k)}\right\}_{k\in\mathbb{N}_0}$ with

$$\left\|x^{(k)} - x^*\right\|_A \leq \left(\frac{\kappa-1}{\kappa+1}\right)^k \left\|x^{(0)} - x^*\right\|_A$$

Here $\|x\|_A := \sqrt{x^T Ax}$ denotes the so-called "energy norm" and x^* the exact solution of $Ax = b$.

Remark

If the linear system is ill-conditioned, then we get for the rate of convergence because of $\kappa \gg 1$:

$$\frac{\kappa-1}{\kappa+1} \approx 1$$

After a few iteration steps the iteration almost terminates.

Linear systems which result from the discretization of elliptic PDEs are often ill-conditioned. □

■ Example

Let us apply the Gradient method to the function $f(x,y) := \frac{1}{2}(x^2 + ay^2)$, $a \gg 1$, with initial values $x^{(0)} = (x_0, y_0) = (a, 1)$.

$$\Rightarrow \quad A = \begin{pmatrix} 1 & 0 \\ 0 & a \end{pmatrix}, \quad \kappa = a$$

$$d_0 = -a\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \alpha_0 = \frac{2}{1+a} \quad \Rightarrow \quad \begin{pmatrix} x^{(1)} \\ y^{(1)} \end{pmatrix} = \varrho\begin{pmatrix} a \\ -1 \end{pmatrix}, \quad \varrho := \frac{a-1}{a+1} \approx 1$$

$$d_1 = -\varrho a\begin{pmatrix} 1 \\ -1 \end{pmatrix}, \alpha_0 = \frac{2}{1+a} \quad \Rightarrow \quad \begin{pmatrix} x^{(2)} \\ y^{(2)} \end{pmatrix} = \varrho^2\begin{pmatrix} a \\ 1 \end{pmatrix} = \varrho^2\begin{pmatrix} x^{(0)} \\ y^{(0)} \end{pmatrix}$$

$$\ldots$$

by induction we prove: $\quad \begin{pmatrix} x^{(k)} \\ y^{(k)} \end{pmatrix} = \varrho^k \begin{pmatrix} a \\ (-1)^k \end{pmatrix}$

gradient: $\quad d_k = -\begin{pmatrix} x^{(k)} \\ ay^{(k)} \end{pmatrix}, d_{k+1} = -\varrho\begin{pmatrix} x^{(k)} \\ -ay^{(k)} \end{pmatrix}$

$\Rightarrow \quad d_k \perp d_{k+1}$ for this special case.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

We observe that always after two iterations d_k is parallel to d_{k+2}: We are searching for a better approximation of the solution in a direction, in which we already have searched two steps before. That is not very efficient! □

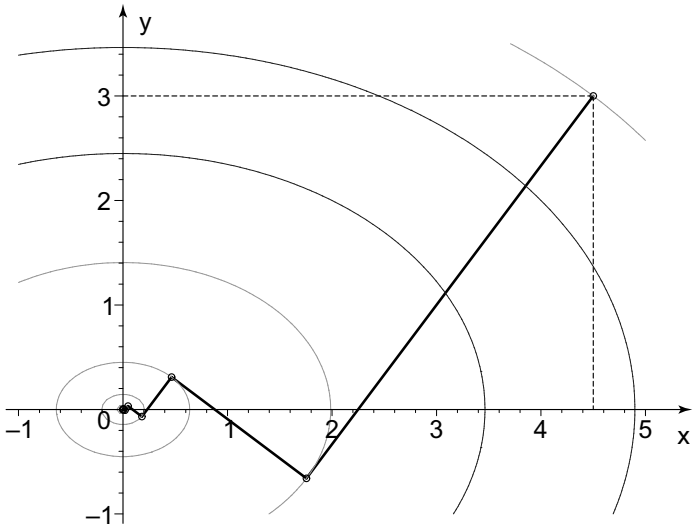


Figure 5: Gradient method applied to $f(x,y) := \frac{x^2}{2} + y^2$, $x^{(0)} = (4.5, 3)$. Iterates shown in a contour picture with contour lines at $19.125, 12, 6, 1.976, 0.204, 0.021, 0.0022$.

2.2.3 Scalar Product

An inner product space is a vector space with an additional structure called an inner product (= scalar product). This additional structure associates each pair of vectors in the space with a scalar quantity. Inner products allow the generalized definition of orthogonality.

Definition (scalar product, Hilbert space)

Consider the vector space $X = \mathbb{K}^n$ ($\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$). A mapping

$$\langle \cdot, \cdot \rangle_X : X \times X \to \mathbb{K}$$

is called *scalar product*, if $\forall\, x, y, z \in X, \forall\, \alpha \in \mathbb{K}$ we get

(1) $\langle x, x \rangle_X \geq 0 \ \wedge \ (x = 0 \Leftrightarrow \langle x, x \rangle_X = 0)$

(2) $\langle x + \alpha y, z \rangle_X = \langle x, z \rangle_X + \alpha \langle y, z \rangle_X$

(3) $\langle x, y \rangle_X = \overline{\langle y, x \rangle_X}$

X with the scalar product $\langle \cdot, \cdot \rangle_X$ is called *pre-Hilbert space*.

An inner product naturally induces an associated norm by the definition $\|x\|_X := \sqrt{\langle x, x \rangle_X}$, thus an inner product space is also a normed vector space.

If X is complete with reference to *this special* norm (Banach space), then X is a *Hilbert space*.

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

The lower index in $\langle \cdot, \cdot \rangle_X$ reminds us, that the scalar product has to be specified exactly . □

■ Example

$X := \mathbb{R}^n$ with

$$\langle x, y \rangle_2 := \sum_{i=1}^{n} x_i y_i = x^T y$$

is a Hilbert space, if $\|x\|_2 := \sqrt{\langle x, x \rangle_2}$. □

Definition (orthogonality)

Let X be Hilbert space, then

$$x, y \in X \text{ orthogonal (or } x \perp y) \quad :\Leftrightarrow \quad \langle x, y \rangle_X = 0$$

□

2.2.4 CG Method (Conjugate Gradient Method)

Fundamental idea

In contrast to the gradient method we want to avoid searching in the same direction several times!

What are the consequences for a mathematical algorithm?

If $x^{(k)}$ is optimal with respect to the search direction $p \neq 0$, then the search direction q in the next iteration step $x^{(k)} \to x^{(k+1)}$ is chosen such that also $x^{(k+1)}$ is optimal in the direction p.

How can we realize this idea?

"$x^{(k)}$ optimal in the direction $p \neq 0$" means, that (at least locally) in the direction p no further improvement is possible ($\to$ definition of "contour line")

$$\nabla f(x^{(k)})^T \cdot p = 0$$

The same property should be valid also for $x^{(k+1)}$

$$\begin{aligned} 0 &\stackrel{!}{=} \nabla f(x^{(k+1)})^T \cdot p = \nabla f(x^{(k)} + \alpha q)^T \cdot p, \qquad \alpha \neq 0 \\ \Rightarrow \quad 0 &= (Ax^{(k+1)} - b)^T p = (A(x^{(k)} + \alpha q) - b)^T p \\ &= (Ax^{(k)} - b)^T p + \alpha (Aq)^T p = \underbrace{\nabla f(x^{(k)})^T \cdot p}_{=0} + \alpha (Aq)^T p \\ \Rightarrow \quad 0 &= (Aq)^T p, \quad \text{because } \alpha \neq 0 \end{aligned}$$

Definition (conjugate vectors)

Let be A positive definite.

Vectors with the property $q^T A p = 0$ are called *conjugate with respect to* A. In the special scalar product defined by $\langle p, q \rangle_A := p^T A q$ the conjugate vectors are perpendicular (= orthogonal). □

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Remark (linear independence)

Let be $\{p_1, \ldots, p_k\} \subset \mathbb{R}^n$ mit $k \leq n$ pairwise conjugate vectors with respect to A (A positive definite) with $p_j \neq 0$.

Then the p_j are linearly independent, orthogonal in the special scalar product $\langle \cdot, \cdot \rangle_A$ and span a k-dimensional (sub-)space, because

$$\sum_{j=1}^{k} \alpha_j p_j = 0 \;\Rightarrow\; 0 = \left(\sum_{j=1}^{k} \alpha_j p_j \right)^T A p_i = \alpha_i (p_i^T A p_i) \;\Rightarrow\; \alpha_i = 0,\ i = 1, \ldots, k$$

The last step holds because $p_i^T A p_i \neq 0$ for A positive definite and $p_i \neq 0$.

Algorithm 3: CG method (core algorithm)

- Given: $A \in \mathbb{R}^{n \times n}$ positive definite, $b \in \mathbb{R}^n$
  $\{p_0, \ldots, p_{n-1}\} \subset \mathbb{R}^n$ pairwise conjugate w.r.t. A, $p_j \neq 0$
- Start: $x^{(0)} \neq 0$ arbitrary initial guess

for $k = 0, 1, 2, \ldots, n-1$ **do**

(A) Calculate the original search direction: $d_k := b - A x^{(k)}$

(B) Determine α_k from

$$\alpha_k = \frac{d_k^T p_k}{p_k^T A p_k} = \frac{\langle d_k, p_k \rangle_2}{\langle p_k, p_k \rangle_A}$$

(C) Update with modified search direction: $x^{(k+1)} := x^{(k)} + \alpha_k p_k$

Detailed analysis of step (B)

Let be $x^* := A^{-1} b$. Because the $\{p_j\}$ are a basis of the $\mathbb{R}^n$ w.r.t $\langle \cdot, \cdot \rangle_A$, we get

$$x^* - x^{(0)} = \sum_{j=0}^{n-1} \alpha_j p_j \iff x^* = x^{(0)} + \sum_{j=0}^{n-1} \alpha_j p_j \,.$$

From the step $x^{(k+1)} := x^{(k)} + \alpha_k p_k$ in part (C) we see that in the $(k+1)$th iteration step the correct component $\alpha_k p_k$ in the direction of one basis vector p_k is added (recursive scheme) to get the improved approximation $x^{(k+1)}$ of the true solution x^*. The direction of the basis vector p_k is only used once.

How do we obtain the α_k in this algorithm? We investigate $\langle p_k, x^* - x^{(0)} \rangle_A$:

$$\begin{aligned} p_k^T A (x^* - x^{(0)}) &= p_k^T (b - A x^{(0)}) = p_k^T \left(b - A x^{(0)} - \sum_{i=0}^{k-1} \alpha_i A p_i \right) \\ &= p_k^T (b - A x^{(k)}) = p_k^T d_k \end{aligned}$$

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

We have subtracted the sum in the bracket, because in the steps before the components of $x^* - x^{(0)}$ in the direction of $p_0, \ldots, p_{k-1}$ already have been added and the scalar product is not changed by that operation (the $\{p_j\}$ are conjugate). On the other hand we can use the basis representation and get

$$p_k^T A(x^* - x^{(0)}) = p_k^T \left(\sum_{j=0}^{n-1} \alpha_j A p_j \right) = \alpha_k \cdot (p_k^T A p_k) = \alpha_k \cdot \langle p_k, p_k \rangle_A$$

From these two equations we can calculate the unknown α_k. □

Important remarks and properties

- The CG method strongly resembles the Gradient method.
  Here we have directly derived the core algorithm via the basis representation. It can be shown that this algorithm minimizes the residuum in a properly chosen subspace too (see below).
- *In an exact calculation without numerical errors (!!)* after n steps all α_j have been determined and the true solution $x^{(n)} = A^{-1}b = x^*$ has been calculated.
- The algorithm is an iterative algorithm because of the numerical errors that accumulate for $n \gg 1$. The iteration is continued until $\|d_k\| \leq tol$.
  Often we do not calculate all p_k, but already obtain a good approximation to x^* after applying a few iteration steps: $p_k, k = 1, \ldots, N < n$. This allows us to approximately solve systems where n is so large that the direct method would take too much time.
- The CG method is numerically stable even in presence of rounding errors.
- Using very tricky programming techniques only one of the expansive matrix-vector multiplications is needed per iteration step ($\rightarrow$ Hestenes/Stiefel)
- In the core algorithm the basis was assumed to be given. That is no realistic situation.
  In a real algorithm, the basis $\{p_j\}$ is constructed simultaneously with the main iteration. For that, we add at the end of algorithm 3 an additional step (D) which – here without proof – can be written as

$$p_0 = d_0 \quad \text{and} \quad p_{k+1} = d_{k+1} + \frac{d_{k+1}^T d_{k+1}}{d_k^T d_k} p_k \, .$$

  Of course, then d_{k+1} already has to be calculated before. □

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Theorem (minimum property of the iterate)

Let be $A \in \mathbb{R}^{n\times n}$ positive definite, $V_k := \operatorname{span}\{p_0, \ldots, p_{k-1}\}$ and apply the CG method from algorithm 3.

Then the approximation $x^{(k)}$ of x^* minimizes the function $f(x) := \frac{1}{2}x^T Ax + bx$ not only along the line $\{x^{(k-1)} + \alpha p_{k-1}, \alpha \in [0,1]\}$ (analogously to the Gradient method), but in the total subspace $x^{(0)} + V_k$. □

Theorem (rate of convergence of the CG method)

Assumption analogously to the Gradient method. Let denote $\kappa = \operatorname{cond}_2 A$.

For the CG method we obtain

$$\|x^{(k)} - x^*\|_A \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \|x^{(0)} - x^*\|_A$$

□

Remarks

- Comparison with Gradient method $\Rightarrow$ instead of κ now $\sqrt{\kappa}$
- Generalization to arbitrary matrices possible (GMRES $\rightarrow$ generalized minimum residuum, MINRES)
- Increasing efficiency by **preconditioning** $\rightarrow$ κ is changed

  Idea of preconditioning:

  Given $Ax = b$ with $A \in \mathbb{R}^{n\times n}$ positive definite. Choose $B \in \mathbb{R}^{n\times n}$ positive definite too and solve instead of the original problem

$$\begin{aligned} & \tilde{A}\tilde{x} = \tilde{b} \quad \text{with} \quad \tilde{A} := BAB, \tilde{x} := B^{-1}x, \tilde{b} := Bb \\ \text{or} \quad & \tilde{A}x = \tilde{b} \quad \text{with} \quad \tilde{A} := BA, \tilde{b} := Bb \\ \text{or} \quad & \tilde{A}\tilde{x} = b \quad \text{with} \quad \tilde{A} := AB, \tilde{x} := B^{-1}x \end{aligned}$$

  Choose B such that $\kappa(\tilde{A}) \ll \kappa(A)$ and that B can be cheaply applied. A good preconditioner concentrates the EWs.

□

■ Example

A simple matrix B for preconditioning is a diagonal matrix, the diagonal elements of which are the inverse of the roots of the diagonal elements of the original matrix.

This idea is motivated by the following theorem:

For a positive definite matrix the minimum EW is smaller or equal the minimum diagonal element, the maximum EW is greater or equal the maximum diagonal element. All EWs are positive. □

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

3 Numerical Solution of Ordinary Differential Equations

3.1 Basic Definitions and Transformations

Let be $U \subseteq \mathbb{R} \times \mathbb{R}^n$ a domain (i.e. an open and connected subset), $f : U \to \mathbb{R}^n$ sufficiently often differentiable (theory says: at least continuous) and the initial value $(t_0, x_0) \in U$.

We want to determine a function $x \in \mathcal{C}^1(I, \mathbb{R}^n)$ on an open and connected interval $I =]t_0, t_f[\subset \mathbb{R}$ such that

$$x'(t) = f(t, x(t)) \quad \wedge \quad t_0 \in I \ \wedge \ x(t_0) = x_0 \ \wedge \ (t, x(t)) \in U \ \ \forall t \in I$$

Analytically such a solution often does not exist. Thus we want to calculate numerically the solution of the above described initial value problem (IVP) of an ordinary differential equation (ODE).

Remarks

- The t-argument can be formally removed by introducing an additional variable $x_{n+1}(t) := t$ together with $f_{n+1}(t, x) := 1$: The ODE then is called autonomous.
- The IVP is equivalent to the following integral equation

  $$x(t) = x_0 + \int_{t_0}^{t} f(\xi, x(\xi)) \, \mathrm{d}\xi$$

  This formula we will use for the construction of numerical methods.
- The interval I can always be transformed to $]0, 1[$.
- We only solve explicit ODEs.
- Any ODE of higher order can be transformed to a system of ODEs of first order, e.g.

$$\begin{array}{rl} & x''' = f(t, x, x', x''), \quad f \in \mathcal{C}^2(I \times \mathbb{R}^n \times \mathbb{R}^n, \mathbb{R}^n) \\ \Rightarrow & \begin{array}{rcl} z_1(t) & := & x(t) \\ z_2(t) & := & x'(t) \\ z_3(t) & := & x''(t) \end{array} \quad \Rightarrow \quad z' = \begin{pmatrix} z_1' \\ z_2' \\ z_3' \end{pmatrix} = \begin{pmatrix} z_2 \\ z_3 \\ f(t, z_1, z_2, z_3) \end{pmatrix} \end{array}$$

■ Example

$$\begin{pmatrix} y_1'(x) \\ y_2''(x) \end{pmatrix} = \begin{pmatrix} x \cdot y_2^2(x) \\ y_2'(x) + x \cdot y_1(x) \end{pmatrix} = f(x, y(x), y(x)')$$

$$y(4) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \ \ y_2'(4) = 7, \ \ I =]4, 13[$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Transformation of the ODE into an autonomous system of first order with the new independent variable ξ instead of x leads to

$$\begin{array}{rcl} z_1 &=& y_1 \\ z_2 &=& y_2 \\ z_3 &=& y_2' \\ z_4 &=& x \end{array} \quad \Rightarrow \quad z'(\xi) = \begin{pmatrix} z_1' \\ z_2' \\ z_3' \\ z_4' \end{pmatrix} = \begin{pmatrix} z_4 \cdot z_2^2 \\ z_3 \\ z_3 + z_4 \cdot z_1 \\ 1 \end{pmatrix} = \tilde{f}(z(\xi)),\ \xi \in]4, 13[$$

After that we transform to the standard "time" interval $]0, 1[$ by

$$\xi \to t := \frac{\xi - 4}{13 - 4} \quad \Rightarrow \quad t \in]0, 1[,\ \frac{\mathrm{d}}{\mathrm{d}t} = (13-4)\frac{\mathrm{d}}{\mathrm{d}\xi} \ \leftrightarrow \ \frac{\mathrm{d}}{\mathrm{d}\xi} = \frac{1}{9} \cdot \frac{\mathrm{d}}{\mathrm{d}t}$$

So we obtain the equivalent and final system in standard form

$$z'(t) = \frac{\mathrm{d}}{\mathrm{d}t} z(t) = \begin{pmatrix} 9(z_4(t) \cdot z_2(t)^2) \\ 9 z_3(t) \\ 9(z_3(t) + z_4(t) \cdot z_1(t)) \\ 9 \end{pmatrix}, \quad z(0) = \begin{pmatrix} 1 \\ 2 \\ 7 \cdot 9 \\ 4 \end{pmatrix}, \quad t \in]0, 1[$$

Prior to the numerical solution always transform a problem into this standard form! □

3.2 Summary of Important Theorems

Existence theorem of Peano

Let be $U \subseteq \mathbb{R} \times \mathbb{R}^n$ domain, $f \in \mathcal{C}^0(U, \mathbb{R}^n)$ and $(t_0, x_0) \in U$.

Then the IVP (not the ODE alone!)

$$x'(t) = f(t, x(t)), \quad x(t_0) = x_0$$

has *at least one* (no uniqueness!) solution, which can be extended to the boundary of U in both directions (i.e. $t < t_0$ and $t > t_0$).

Remark

"To extend to the boundary" means to come as close as we want to the boundary of U: either $x(t)$ contains the respective boundary point or $\|x(t)\|$ is unbounded at the boundary.

To extend to the boundary does not mean that a solution exists on the total interval $[a, b]$. The solution might leave U before reaching a or b. □

Definition (Lipschitz condition)

Let be $U \subseteq \mathbb{R}^{n+1}$ and $f \in \mathcal{C}^0(U, \mathbb{R}^n)$.

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

$f(t,x)$ satisfies a *(global) Lipschitz condition* on U with respect to x with *Lipschitz constant* L , if

$$\exists L > 0 \quad \ni \quad \|f(t,x_1) - f(t,x_2)\| \leq L\|x_1 - x_2\| \quad \forall (t,x_1),(t,x_2) \in U$$

$f(t,x)$ satisfies a *local Lipschitz condition* on U w.r.t. x, if for every $(t_1,x_1) \in U$ there exists $\delta_1 = \delta_1(t_1,x_1) > 0$ such that $f(t,x)$ satisfies a global Lipschitz condition on

$$U_{\delta_1}((t_1,x_1)) \cap U$$

with Lipschitz constant $L = L(\delta_1,t_1,x_1)$ (the constant may differ at different points). U_{δ_1} denotes the open ball with radius δ_1 around the point (t_1,x_1). □

Theorem (global Lipschitz condition, sufficient condition)

Let be $U \subseteq \mathbb{R}^{n+1}$ convex domain, $f \in \mathcal{C}^1(U,\mathbb{R}^n)$. If

$$\left|\frac{\partial f_i}{\partial x_j}\right| \leq K\,, \quad i,j = 1,\ldots,n\,, \forall (t,x) \in U$$

– i.e. all partial derivatives are *(continuous and) bounded* –, then f satisfies a global Lipschitz condition in U (with $L \leq K \cdot n$, if for the norm we have chosen $\|\cdot\| = \|\cdot\|_\infty$).

Theorem (local Lipschitz condition, sufficient condition)

Let be $U \subseteq \mathbb{R}^{n+1}$ domain, $f \in \mathcal{C}^0(U,\mathbb{R}^n)$ and Jacob matrix $\left(\frac{\partial f}{\partial x}\right) \in \mathcal{C}^0(U,\mathbb{R}^{n\times n})$.

Then f satisfies a *local* Lipschitz condition in U.

Remark: no boundedness, no convexity, less smoothness (w.r.t. t) necessary

---

Theorem (global existence and uniqueness)

Let be $U = [a,b] \times \mathbb{R}^n$ domain and $f \in \mathcal{C}^0(U,\mathbb{R}^n)$; let f satisfy a global Lipschitz condition in U w.r.t. x.

For every $(t_0,x_0) \in U$ there exists *exactly one solution* $x(t)$ of the IVP

$$x'(t) = f(t,x(t))\,, \quad x(t_0) = x_0$$

defined on the full interval $a \leq t \leq b$.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Remark

In this case it cannot happen that $|x(t)| \to \infty$ for $t \to t_1 \in]t_0, b[$; the solution exists and is uniquely defined on the full interval $[a, b]$.

Sufficient condition: $f \in \mathcal{C}^1$ and $U = Q$ (Q cuboid) sufficiently large. □

Theorem (local existence and uniqueness)

Let be $U = [a, b] \times \mathbb{R}^n$ domain and $f \in \mathcal{C}^0(U, \mathbb{R}^n)$; let f satisfy a local Lipschitz condition in U w.r.t. x.

For every $(t_0, x_0) \in U$ there exists *exactly one solution* $x(t)$ of the IVP

$$x'(t) = f(t, x(t)), \quad x(t_0) = x_0.$$

This solution can be extended to the boundaries of U in both directions.

Remark

In may happen that $|x(t)| \to \infty$ for $t \to t_1 \in]t_0, b[$. The solution exists and is unique on an interval $]c, d[\subseteq]a, b[$ with $t_0 \in]c, d[$.

Sufficient condition: Jacobian w.r.t. $\vec{x}$ is continuous. □

Remark

The exact calculation of the Lipschitz constant is mostly impossible, we can only estimate it. □

Theorem (continuous dependency of a solution)

Assumptions:

- $I \subset \mathbb{R}$ interval, $U \subseteq I \times \mathbb{R}^n$
- $f \in \mathcal{C}^0(U, \mathbb{R}^n)$ satisfies global Lipschitz condition in U w.r.t. x with Lipschitz constant L.
- $x \in \mathcal{C}^1(I, \mathbb{R}^n)$ is solution of IVP $\quad x(t)' = f(t, x(t)),\ x(t_0) = x_0,\ t_0 \in I \quad (*)$
- Let $z \in \mathcal{C}^1(I, \mathbb{R}^n)$ denote an approximation to the solution of the IVP $(*)$ with

$$\|z(t_0) - x(t_0)\| \leq \gamma \quad \wedge \quad \|z'(t) - f(t, z(t))\| \leq \delta, \quad \gamma, \delta > 0 \ \text{ const.}$$

- $\operatorname{graph}(x(t)) \subset U,\ \operatorname{graph}(z(t)) \subset U$

Claim:

$$\|x(t) - z(t)\| \ \leq \ \gamma e^{L|t-t_0|} + \frac{\delta}{L}\left(e^{L|t-t_0|} - 1\right)$$

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

3.3 Numerical Methods: Basic Idea and Notation

Definition

Numerical methods always use a discretization, i.e. we subdivide the integration interval $I = [t_0, t_f]$

$$t_0 < t_1 < t_2 < \ldots < t_N = t_f$$

The t_i are the *grid points*, $h_m := t_{m+1} - t_m$ is the *stepsize* and $I_h := \{t_0, t_1, \ldots, t_N\}$ is the *mesh of nodes*. Grid points are also called discretization nodes. If $h_m = h \ \forall m$, then I_h is called *equidistant*.

Let denote $x_i := x(t_i)$ the *exact solution* of the IVP $x' = f(t,x),\ x(t_0) = x_0$ at the point t_i.

Let denote $\eta_i := \eta(t_i)$ the *approximation of the solution obtained by a numerical method* at the same grid point; therefore, η is only defined at $t = t_i, i = 0, 1, \ldots, N$. □

Basic idea to obtain a numerical method: formal integration

$$x'(t) = f(t, x(t)) \ \Rightarrow$$

$$\frac{x(t+h) - x(t)}{h} = \frac{1}{h}\int_t^{t+h} f(\xi, x(\xi))\,\mathrm{d}\xi =: \Delta(t,x,h) \overset{!}{\approx} \phi(t,x,h)$$

Here $\phi(t,x,h)$ denotes the numerical approximation to $\Delta(t,x,h)$.

$\phi(t,x,h)$ is the *increment function*; the notation is a formal one only.

The exact solution

$$\Delta(t,x,h) = \begin{cases} \dfrac{x(t+h) - x(t)}{h} & , \ h \neq 0 \\ f(t,x) & , \ h = 0 \end{cases}$$

is the *exact relative increment*. □

Definition (discretization method, grid function)

A *discretization method* for the approximation of the solution $x(t)$ of the IVP $x' = f(t,x),\ x(t_0) = x_0$, is a numerical rule that tells us how to assign a *grid function* $\eta : I_h \to \mathbb{R}^n$ to a mesh of nodes I_h. □

■ **Example**

Explicit Euler:

$$\eta_{m+1} = \eta_m + h_m f(t_m, \eta_m) \ \Rightarrow \ \phi(t,x,h) = f(t,x)$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Implicit Euler:

$$\eta_{m+1} = \eta_m + h_m f(t_{m+1}, \eta_{m+1}) \quad \Rightarrow \quad \phi(t,x,h) = f(t+h, x(t+h))$$

More complicated formulae are possible, e.g. in case of an equidistant grid

$$\begin{aligned} \eta_{m+2} &= \eta_m + h/2 \cdot \big(f(t_m,\eta_m) + 2f(t_{m+1},\eta_{m+1}) + f(t_{m+2},\eta_{m+2})\big) \\ \Rightarrow \quad \phi(t,x,h) &= \frac{1}{4}\Big(f(t,x(t)) + 2f(t+h/2, x(t+h/2)) + f(t+h, x(t+h))\Big) \end{aligned}$$

or the explicit mid-point rule

$$\eta_{m+2} = \eta_m + 2h f(t_{m+1}, \eta_{m+1}) \quad \Rightarrow \quad \phi(t,x,h) = f(t+h/2, x(t+h/2))$$

□

Definition (one-step method)

A one-step method is a discretization method which for the calculation of η_{m+1} only uses η_m, but not e.g. $\eta_{m-1}, \eta_{m-2}, \ldots$

Therefore, a one-step method can be written as

■ Initial value: $\eta_0 = y_0$

for $i = 1$ **to** $N-1$ **do**

$$\eta_{i+1} = \eta_i + h_i \phi(t_i, \eta_i, h_i)$$
$$t_{i+1} = t_i + h_i$$

□

Remark

With this definition we also can write the implicit Euler as a one-step method

$$\phi(t_m, \eta_m, h_m) := f(t_m + h_m, \eta_m + h_m \phi(t_m, \eta_m, h_m))$$

□

Remark

If not otherwise stated, we will restrict ourselves to one-step methods! □

3.4 Consistency and Convergence of One-Step Methods

Consider the IVP $x' = f(t,x)$, $x(t_0) = x_0$ from chap. 3.1 on the closed interval $I = [t_0, t_f]$. Let ϕ be a one-step method which we want to analyze.

Definition (local discretization error)

Let be $\tilde{\eta}_{m+1}$ the result of a single (!) step of the one-step method with the exact initial value $\eta_m = x(t_m)$, i.e.

$$\tilde{\eta}_{m+1} = x(t_m) + h_m \phi(t_m, x(t_m), h_m)$$

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Then $T(t_m, x(t_m), h_m) := x(t_{m+1}) - \tilde{\eta}_{m+1}$ is called the *local discretization error* of the one-step method at t_{m+1}.

This really is the error of the method after one step only! □

Definition (consistent method)

A method is called *consistent*, if the local discretization error per unit step $T(t,x,h)/h$ converges uniformly to zero $\forall t, x$ for $h \to 0$

$$\frac{\|T(t,x,h)\|}{h} \leq \sigma(h) \quad \wedge \quad \lim_{h\to 0} \sigma(h) = 0 \quad \forall t \in I, \ \forall x$$

A method is *consistent of order* p, if

$$\|T(t,x,h)\| \leq C \cdot |h|^{p+1} =: \mathcal{O}(h^{p+1}) \quad \forall t \in I, \ \forall x, \ \forall h \in]0, h_{max}]$$

□

The order of consistency describes the quality of the approximation and allows to compare different discretization methods.

Theorem

$$\phi \text{ consistent} \quad \Longleftrightarrow \quad \lim_{h\to 0} \phi(t,x,h) = f(t,x)$$

uniformly, i.e. $\forall t \in I, \forall x$ and $\forall f \in \mathcal{C}^1(I \times \mathbb{R}^n, \mathbb{R}^n)$.

■ Example

Because of $\phi(t,x,h) = f(t,x)$ the explicit Euler is consistent.

Because $\eta_{m+1} = x_m + hf(t_m, x_m)$ with (in general) $f(t_m, x_m) \neq 0$ it is consistent of order $p = 1$:

$$\begin{aligned} T(t_m, x(t_m), h_m) &= x(t_{m+1}) - \tilde{\eta}_{m+1} \\ &= x(t_m + h) - x(t_m) - hf(t_m, x(t_m)) \\ &\overset{Taylor}{=} \frac{h^2}{2}\left(f_t(t_m, x_m) + f_x(t_m, x_m) \cdot f(t_m x_m)\right) + \mathcal{O}(h^3) \\ \Rightarrow \quad \frac{\|T\|}{h} &\leq h^1 \cdot const = \sigma(h) \ \to \ 0 \text{ for } h \to 0 \ \Rightarrow \ p = 1 \end{aligned}$$

In the autonomous case the discretization method is simplified to

$$\eta_{m+1} = \eta_m + hf(x_m)$$

□

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Definition (global discretization error)

W.l.o.g. we simply the situation and use a constant stepsize h, i.e. $t_m = t_0 + m \cdot h$.

The *global discretization error*

$$e(h,t) := \eta(t) - x(t) \quad \text{for} \quad t := t_m = t_0 + m \cdot h$$

directly describes the difference between the true solution and its numerical approximation. Because of the use of η it is only defined at discrete values (grid points).

Definition (convergent method)

A method is *convergent*, if the global discretization error $e(h,t)$ uniformly converges to zero $\forall t \in I$ for $h \to 0$.

A method is *convergent of order* p if

$$\|e(h,t)\| \leq s(t) \cdot |h|^p \quad \forall t \in I$$

and $s : I \to \mathbb{R}$ is a bounded function. □

Remark

In contrast to consistency it is very difficult to analyze convergence directly. □

3.5 Construction of One-Step Methods

3.5.1 Strategy

Consider the IVP from chap. 3.1 with $n = 1$. We want to construct a *one-step method with maximum order of consistency* for a given number of evaluations of the right-hand side $f(t, x(t))$.

We analyze a single integration step and carry out the Taylor expansion using $x' = f(t,x)$

$$\begin{aligned}
x(t+h) &= x(t) + hx'(t) + \frac{h^2}{2}x''(t) + \frac{h^3}{6}x'''(t) + \ldots = x(t) + h \cdot \Delta(t,x,h) \\
\Delta(t,x,h) &= f(t,x(t)) + \frac{h}{2}\frac{\mathrm{d}}{\mathrm{d}t}f(t,x(t)) + \frac{h^2}{6}\frac{\mathrm{d}^2}{\mathrm{d}t^2}f(t,x(t)) + \mathcal{O}(h^3) \\
&= f(t,x(t)) + \frac{h}{2}\Big(f_t(t,x(t)) + f_x(t,x(t))f(t,x(t))\Big) \\
&+ \frac{h^2}{6}\Big(f_{tt} + 2f_{tx} \cdot f + f_{xx} \cdot f^2 + f_x \cdot (f_t + f_x f)\Big) + \mathcal{O}(h^3)
\end{aligned}$$

If we would be able to calculate the necessary derivatives of $f(t,x(t))$ and $x(t)$, then we immediately would obtain a method which is consistent of order p.

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

As an example, by

$$\phi(t,x,h) := x'(t) + \frac{h}{2}x''(t) = f(t,x) + \frac{h}{2}\Big(f_t(t,x) + f_x(t,x)f(t,x)\Big)$$

we would construct a one-step method of order $p = 2$, for

$$\begin{aligned} & x(t+h) - \tilde{\eta}(t+h) \\ = & \left(x(t) + hx'(t) + \frac{h^2}{2}x''(t) + \frac{h^3}{6}x'''(t) + \ldots\right) - \left(x(t) + hx'(t) + \frac{h^2}{2}x''(t)\right) \\ = & \mathcal{O}(h^3) \end{aligned}$$

Unfortunately the explicit calculation of derivatives for real problems either is impossible or too expensive.

We make a general ansatz for the new method instead, e.g.

$$\phi(t,x,h) = \alpha_1 f(t,x) + \alpha_2 f(t + \beta_1 h, x + \beta_2 h f(t,x))$$

with the free parameters $\alpha_1, \alpha_2, \beta_1, \beta_2$.

The free parameters $\alpha_1, \alpha_2, \beta_1, \beta_2$ are chosen such that we obtain a method of maximum order of consistency. Because for the calculation of ϕ solely $(t, x(t))$ is needed, this ansatz again leads to a one-step method.

We now want to construct a method which is consistent of order $p = 2$. For that we again carry out a two-dimensional Taylor expansion of $\phi(t,x,h)$ and obtain

$$\begin{aligned} & f(t + \beta_1 h, x + \beta_2 h f(t,x)) \\ = & f(t,x) + \left(\beta_1 h \frac{\partial}{\partial t} + \beta_2 h f(t,x)\frac{\partial}{\partial x}\right) f(t,x) \\ & + \frac{1}{2!}\left(\beta_1 h \frac{\partial}{\partial t} + \beta_2 h f(t,x)\frac{\partial}{\partial x}\right)^2 f(t,x) + \ldots \\ = & f + (\beta_1 h f_t + \beta_2 h f_x f) + \frac{1}{2}\left(\beta_1 h \frac{\partial}{\partial t} + \beta_2 h f \frac{\partial}{\partial x}\right)(\beta_1 h f_t + \beta_2 h f_x f) + \ldots \end{aligned}$$

With $f := f(t,x)$ we get

$$\begin{aligned} \phi(t,x,h) &= \alpha_1 f + \alpha_2 f + \alpha_2 h(\beta_1 f_t + \beta_2 f_x \cdot f) \\ &+ \frac{\alpha_2}{2} h^2 \left(\beta_1^2 f_{tt} + 2\beta_1\beta_2 f f_{tx} + \beta_1\beta_2 f_x f_t + \beta_2^2 f f_{xx} f + \beta_2^2 f f_x^2\right) + \ldots \end{aligned}$$

Now we choose the free parameters such that as many h-terms as possible from the expansion of this ansatz $\phi(t,x,h)$ match those from $\Delta(t,x,h)$.

We get a method of order $p = 2$ if

$$\begin{aligned} 1 &= \alpha_1 + \alpha_2 \\ 1/2 &= \alpha_2\beta_1 \\ 1/2 &= \alpha_2\beta_2 \end{aligned}$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

The solution of the nonlinear system is not unique.

For the choice $\alpha_1 = \alpha_2 = \frac{1}{2}, \beta_1 = 1, \beta_2 = 1$ we get the "method of Heun", for

$$\alpha_1 = 0, \alpha_2 = 1, \beta_1 = \frac{1}{2}, \beta_2 = \frac{1}{2}$$

we get the "modified Euler". Both are consistent of order $p = 2$.

In a similar way a method of order $p = 3, \ldots$ can be constructed, if the ansatz contains a sufficient number of free parameters. □

3.5.2 Explicit Runge-Kutta Methods

We obtain an important class of methods by the Runge-Kutta ansatz

$$\phi(t,x,h) = \sum_{k=1}^{s} b_k \cdot f_k(t,x,h), \quad f_k(t,x,h) := f\left(t + c_k h, x + h \sum_{j=1}^{k-1} \alpha_{kj} f_j(t,x,h)\right)$$

with the free parameters b_k, α_{kj}, c_k. □

The definition specifies an explicit Runge-Kutta method with s stages. For e.g. $s = 4$ this method is called RK4 method. The method is called explicit, because f_k can be calculated using $f_1, \ldots, f_{k-1}$ only; these values have been calculated before and thus are already known.

Algorithm

In the RK method per integration step $t \to t + h$ (or $t_m \to t_{m+1} = t_m + h_m$) the following algorithm is executed

$$\begin{aligned} f_k(t,\eta(t),h) &:= f\left(t + c_k h, \eta(t) + h \sum_{j=1}^{k-1} \alpha_{kj} f_j(t,\eta(t),h)\right), \quad k = 1, \ldots s, \\ \eta(t+h) &= \eta(t) + h \sum_{k=1}^{s} b_k \cdot f_k(t,\eta(t),h) \end{aligned}$$

In a compact way the parameter set for an s-stage explicit Runge-Kutta method can be arranged in a *Butcher tableau*

$$\begin{array}{c|ccccc} 0 & 0 & & & & \\ c_2 & \alpha_{21} & 0 & & & \\ c_3 & \alpha_{31} & \alpha_{32} & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \ddots & \\ c_s & \alpha_{s1} & \alpha_{s2} & \cdots & \alpha_{s,s-1} & 0 \\ \hline & b_1 & b_2 & \cdots & b_{s-1} & b_s \end{array}$$

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

with the *nodes* $(0, c_2, \ldots, c_s)^T$, the *Runge-Kutta matrix* $A := (\alpha_{kj})$ and the *weights* $(b_1, b_2, \ldots, b_s)^T$.

These $\frac{s(s+1)}{2}$ parameters mostly are determined such that the resulting explicit RK method has maximum order of consistency.

Remark

A nice insight into the basic ideas of RK methods is obtained if we use the following equivalent reformulation of the classical RK method (i.e. a special RK4 method):

$$\eta_{m+1} = \eta_m + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4), \quad t_{m+1} = t_m + h$$

with

$$\begin{aligned} k_1 &= f(t_m, \eta_m), \\ k_2 &= f\left(t_m + \frac{h}{2}, \eta_m + \frac{h}{2}k_1\right), \\ k_3 &= f\left(t_m + \frac{h}{2}, \eta_m + \frac{h}{2}k_2\right), \\ k_4 &= f(t_m + h, \eta_m + hk_3). \end{aligned}$$

and the Butcher tableau

0				
$1/2$	$1/2$			
$1/2$	0	$1/2$		
1	0	0	1	
	$1/6$	$1/3$	$1/3$	$1/6$

Here a sequence of four Euler steps with stepsizes $c_i h$ is performed, all starting at η_m. After the i-th Euler step ($i = 1, 2, 3, 4$), the (t, x)-values of the resulting point are used to calculate an updated slope k_i specified by the right-hand side $f(t, x)$ of the differential equation. This slope is used for the next Euler step in case of $i = 1, 2, 3$.

At the end, a final Euler step is performed with stepsize h and a slope which is the weighted average of the four slopes k_i calculated before.

In averaging the four increments, greater weight is given to the increments at the midpoint. If f is independent of x, then the differential equation is equivalent to a simple integral and the classical RK4 method reduces to Kepler's rule. □

Order conditions

After performing the Taylor expansion the coefficients of the Taylor series of $\phi(t, x, h)$ and $\Delta(t, x, h)$ are compared. The goal is to choose the free parameters such that as many h-terms as possible from the expansion of $\phi(t, x, h)$ match those from $\Delta(t, x, h)$. An example was given in chap. 3.5.1. For the RK methods, this approach leads to the following order conditions

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

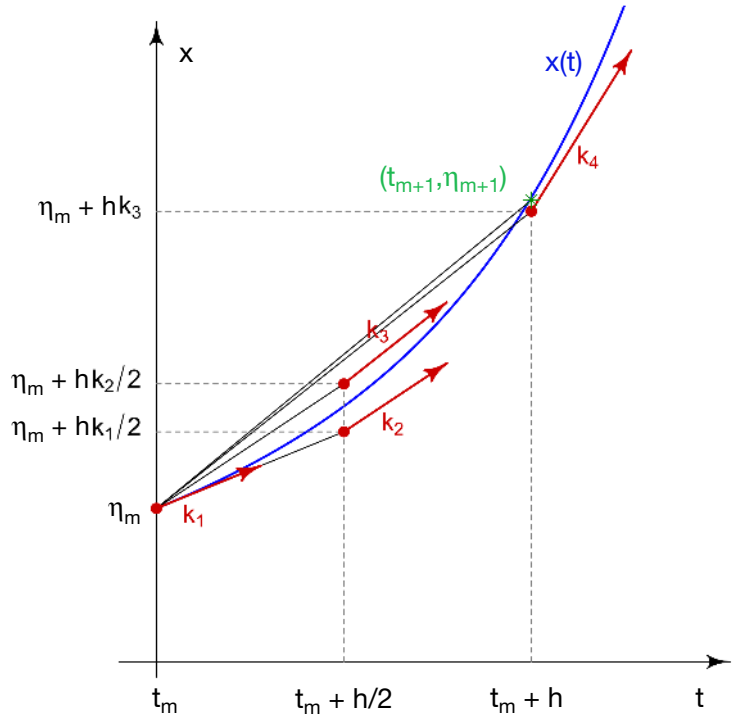


Figure 6: Classical RK4 method generating a sequence of slopes.

order p	order conditions
1	$\sum_i \quad b_i = 1$
2	$\sum_i \quad b_i c_i = 1/2$
3	$\sum_i \quad b_i c_i^2 = 1/3$
	$\sum_{i,j} \quad b_i \alpha_{ij} c_j = 1/6$
4	$\sum_i \quad b_i c_i^3 = 1/4$
	$\sum_{i,j} \quad b_i c_i \alpha_{ij} c_j = 1/8$
	$\sum_{i,j} \quad b_i \alpha_{ij} c_j^2 = 1/12$
	$\sum_{i,j,k} \quad b_i \alpha_{ij} \alpha_{jk} c_k = 1/24$
...	...

Remark

The first condition guarantees that the method is consistent at all: Because of $\phi(t,x,h) = \sum_{k=1}^{s} b_k \cdot f_k(t,x,h)$ we get

$$h \to 0 \;\Rightarrow\; \eta(t+h) \to \eta(x) \;\Rightarrow\; \left(\phi(t,\eta(t),h) = f(t,\eta(t)) \;\; \forall f \iff \sum_{i=1}^{s} b_i = 1 \right)$$

□

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Remark

In addition we often want that the *node condition* is satisfied:

$$c_i = \sum_{j=1}^{i-1} \alpha_{ij}$$

This condition guarantees, that we get the same numerical results no matter if the RK method is applied to a non-autonomous IVP or to the same problem after transformation into autonomous form. □

The numerical effort – an overview

The following table lists the minimum number of stages s_{min} (and with it in most cases the number of function evaluations) necessary to construct a RK method of order p. N denotes the number of order conditions that have to be fulfilled.

p	1	2	3	4	5	6	7	8
N	1	2	4	8	17	37	85	200
s_{min}	1	2	3	4	6	7	9	11

RK methods of order $p = 2, 4, 5, 8$ are often used in practical applications.

■ Example

3/8-rule of Kutta: RK method of order $p = 4$

$$\begin{array}{c|cccc} 0 & & & & \\ 1/3 & 1/3 & & & \\ 2/3 & -1/3 & 1 & & \\ 1 & 1 & -1 & 1 & \\ \hline & 1/8 & 3/8 & 3/8 & 1/8 \end{array}$$

Classical RK method: RK method of order $p = 4$

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

The method of Heun – that we already have discussed – is a RK method of order $p = 2$. □

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

3.6 Stepsize Control for One-Step Methods

3.6.1 Basic problem and Solution Strategy

Basic problem

A one-step method calculates in one integration step the solution $\eta(t_{m+1})$ at $t = t_{m+1}$ using $(t_m, \eta(t_m))$ only

$$t_m \;\rightarrow\; t_{m+1} = t_m + h_m\,, \quad t_m, t_{m+1} \in]t_0, t_f[$$

The selection of the correct stepsize h_m plays a crucial role.

If h_m is too small, then the
– discretization error ($\rightarrow$ truncated Taylor expansion) $\sim \mathcal{O}(h^{p+1})$: very small
– computational effort for the interval $[t_0, t_f]$ is high because of many steps
– rounding error is high because of many steps

If h_m is too large, then the inverse statements are true.

If h_m is chosen automatically, a compromise is required between the total error of the result and the computational effort.

We have almost no access to the rounding errors. We do not know the total discretization error that describes the difference between the true solution and its numerical approximation (convergence!) at the grid points t_i.

We only can get an estimate of the local discretization error (= error per step, consistency). That is not much, but unfortunately that mostly is all we have!

So we choose a rough estimate for the local discretization error (our tolerance tol); subject to that constraint we try to maximize h_m ($\rightarrow$ rounding error and computational effort decrease). ”Rough estimate” is meant literally: Often the required tolerance tol is only poorly approximated.

Instead of reaching $\|T(t_m, x(t_m), h_m)\| < tol$ we alternatively try to control the local discretization error per step length

$$\frac{\|T(t_m, x(t_m), h_m)\|}{h_m} < tol_2$$

Solution strategy

To obtain the local discretization error we have to compare after each step $t_m \rightarrow t_{m+1}$ the numerical result $\eta(t_{m+1})$ with the exact solution of the IVP for the same initial value $\eta(t_m)$. The exact solution unfortunately is unknown.

To overcome this difficulty, the following workaround is used: We calculate the step $t_m \rightarrow t_{m+1}$ twice with different accuracy and then we use the numerical approximation of higher precision as a substitute for the exact but unknown solution.

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

There are two standard ways to calculate these two approximations with different accuracy:

Either to take one method only and calculate the solution with the two stepsizes h_m (1 integration step) and $h_m/2$ (two succeeding integration steps)

or to choose two different methods with different orders of consistency and to perform one integration step each with the same stepsize h_m.

The mathematical justification of these strategies is given by the theorem on the asymptotic expansion of the global discretization error for one-step methods (Gragg's theorem).

Gragg's theorem

Assumption:

- Let be $f \in \mathcal{C}^{N+1}([a,b] \times \mathbb{R}^n, \mathbb{R}^n)$, $t_0 \in [a,b]$, $x' = f(t,x)$, $x(t_0) = x_0$.
- Let denote ϕ the increment function for a one-step method of order p.
- The stepsize is assumed to be constant: $h = h_m \ \forall m \ \Rightarrow \ h = \dfrac{t_m - t_0}{m}$

Claim:

$$\eta(t_m, h) = x(t_m) + \sum_{i=p}^{N} h^i e_i(t_m) + h^{N+1} E_{N+1}(t_m, h) \quad \forall m$$

and we get in addition:

$e_i(t_0) = 0$, $i = p, \ldots, N$, the residual term $E_{N+1}(t_m, h)$ is bounded $\forall h \leq H$ with H properly chosen and the $e_i(t_m)$ are independent of h!

Remark

The most important feature is: $e_i(x_m)$ is independent of h!!
Gragg's theorem guarantees the existence of an asymptotic expansion of the global discretization error ($\rightarrow$ convergence). □

3.6.2 One Method, Two Different Stepsizes

The local discretization error (error per step) should be smaller than a given tolerance tol. We investigate one step; instead of t_0 we might also write t_m.

We apply Gragg's theorem to a method which is convergent of order p; instead of h we write h_{old}, because here we insert our (old) estimate for the proper stepsize h (and want to find out, whether this estimate is good enough or not):

$$\begin{aligned} \eta(t_0 + h_{old}, h_{old}) &\doteq x(t_0 + h_{old}) + {h_{old}}^p e_p(t_0 + h_{old}) \\ \eta(t_0 + h_{old}, h_{old}/2) &\doteq x(t_0 + h_{old}) + \left(\frac{h_{old}}{2}\right)^p e_p(t_0 + h_{old}) \end{aligned}$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

We subtract the two equations, make a Taylor expansion and use $e_p(t_0) = 0$; then the following approximation of $e_p'(t_0)$ can be calculated numerically

$$\frac{\eta(t_0+h_{old},h_{old}) - \eta(t_0+h_{old},h_{old}/2)}{{h_{old}}^p\left(1-\frac{1}{2^p}\right)} = e_p(t_0+h_{old}) \doteq \underbrace{e_p(t_0)}_{=0} + e_p'(t_0)\cdot h_{old} \qquad (*)$$

The term $e_p'(t_0)$ is independent of h and h_{old} respectively!

Now we apply the Taylor expansion once more and directly to Gragg's theorem with the new and improved stepsize h_{new}:

$$\begin{aligned}
\eta(t_0+h_{new},h_{new}) &= x(t_0+h_{new}) + {h_{new}}^p e_p(t_0+h_{new}) \\
&+ {h_{new}}^{p+1} e_{p+1}(t_0+h_{new}) + \mathcal{O}({h_{new}}^{p+2}) \\
&\overset{Taylor}{=} x(t_0+h_{new}) + {h_{new}}^p(e_p(t_0) + h_{new}e_p'(t_0)) \\
&+ {h_{new}}^{p+1} e_{p+1}(t_0) + \mathcal{O}({h_{new}}^{p+2}) \\
&\doteq x(t_0+h_{new}) + {h_{new}}^{p+1} e_p'(t_0)
\end{aligned}$$

For the local discretization error we then obtain

$$\|\eta(t_0+h_{new},h_{new}) - x(t_0+h_{new})\| \doteq {h_{new}}^{p+1}\|e_p'(t_0)\| \overset{!}{\leq} tol \qquad (**)$$

Algorithm 4 (next page) gives a stepsize control based on $(*), (**)$. □

Remark

- Important feature: $e_p'(t_0)$ is independent of h_{old} and h_{new}; that is the only reason why a relation between $(*)$ and $(**)$ can be established.
- The stepsize control with two stepsizes can be easily understood, but it is not often used: too many function evaluations. □

3.6.3 Two Different Methods, One Stepsize

We use two methods of different orders of convergence $p, p+1$ and one common stepsize h. The mathematical derivation is similar to the case above and also uses Taylor expansions. This method can be efficiently programmed: Often only one additional function evaluation necessary.

And here are the details:

$$\begin{aligned}
\eta(t_0+h) &= x(t_0+h) + h^p e_p(t_0+h) + h^{p+1} e_{p+1}(t_0+h) + \mathcal{O}(h^{p+2}) \\
\hat{\eta}(t_0+h) &= x(t_0+h) + h^{p+1}\hat{e}_{p+1}(t_0+h) + \mathcal{O}(h^{p+2})
\end{aligned}$$

For the difference we get

$$\|\eta(t_0+h) - \hat{\eta}(t_0+h)\| = h^p\|e_p(t_0+h)\| + \ldots \doteq h^p\cdot h\cdot\|e_p'(x_0)\| \overset{!}{\leq} tol$$

Setting an "="-sign instead of the "$\leq$"-sign we finally obtain

$$h_{new} = \alpha\cdot h_{old}\sqrt[p+1]{\frac{tol}{\|\eta(t_0+h_{old}) - \hat{\eta}(t_0+h_{old})\|}}$$

For the safety factor we choose e.g. $\alpha = 0.9$. □

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Algorithm 4: Stepsize control with two different stepsizes

■ Start: Choose $h = h_{old}$, e.g. from the preceeding step, i.e.

$$t_{m-1} \to t_m = t_{m-1} + h_{old}$$

for $m = 0, 1, \ldots N-1$ **do**

Stepsize control for the step: $t_m \to t_{m+1} > t_m$

(A) Step $t_m \to t_{m+1} := t_m + h_{old}$: calculate $\eta(t_{m+1}, h_{old})$ and $\eta(t_{m+1}, h_{old}/2)$ from $\eta(t_m)$

(B) Calculate approximation for $e_p'(t_m)$ as in $(*)$

(C) Insert result into $(**)$ and calculate h_{new}:

$$h_{new}^{p+1} \leq \frac{tol}{\|e_p'(t_m)\|} \doteq tol \Big/ \frac{\|\eta(t_m + h_{old}, h_{old}) - \eta(t_m + h_{old}, h_{old}/2)\|}{h_{old}^{p+1}\left(1 - \frac{1}{2^p}\right)}$$

We calculate the $(p+1)$-th root and get h_{new}. Now we can check *a posteriori* whether the original stepsize selection was correct or not.

if $h_{old} > 2h_{new}$ **then**

Stepsize estimate was wrong:
Define $h_{old} := 1.5 \cdot h_{new}$, goto (A), repeat integration step

else

$$\begin{aligned} t_{m+1} &:= t_m + h_{old} \\ \eta(t_{m+1}) &:= \eta\left(t_m + h_{old}, h_{old}/2\right) \\ h_{old} &:= \min\left\{h_{new}, t_f - t_{m+1}\right\} \end{aligned}$$

Remark

This type of stepsize control (2 methods, 1 stepsize) is often used in RK methods (idea of Fehlberg). The methods are then denoted e.g. by RKF 4(5) and RKF 8(7). The method RKF 4(5) is consistent of order 4 with an embedded error estimator of order 5 ($\to$ order of the error estimator written in the bracket).

$$\begin{array}{c|ccccc} 0 & & & & & \\ c_2 & \alpha_{21} & & & & \\ c_3 & \alpha_{31} & \alpha_{32} & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ c_s & \alpha_{s1} & \alpha_{s2} & \cdots & \alpha_{s,s-1} & \\ \hline & b_1 & b_2 & \cdots & b_{s-1} & b_s \\ \hline & \hat{b}_1 & \hat{b}_2 & \cdots & \hat{b}_{s-1} & \hat{b}_s \end{array}$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Both methods are constructed simultaneously by an extended Butcher tableau. To obtain a sufficient number of free parameters for both methods, often one additional stage in the tableau is enough. □

Remark

In deviation from the strict theory, often the result of the better method is used as the initial value for the next integration step (e.g. RKF 8(7)). By this a small gain in precision is obtained without additional effort. □

■ Example

Three-body problem: Simulation of the motion of a rocket in the gravitational fields of Earth and Moon. Start from Earth orbit, flight to Moon, one revolution around the Moon, flight back to Earth and arrival in Earth orbit again (Apollo 13 type mission). The final accuracy describes the deviation at the end point.

Method	No. of steps	Final accuracy
Euler	24000	$> 10^0$
RK4	6000	$\approx 3 \cdot 10^{-1}$
RKF 5(4)	98	10^{-3}
RKF 8(5,3)	102	10^{-6}

3.7 Relation Between Convergence and Consistency

For explicit one-step methods the following property holds:

order of consistency = order of convergence

Let us investigate that in detail. The direction "order of convergence $p \Rightarrow$ order of consistency p" is directly contained in the theorem of Gragg. The opposite direction is still missing.

Theorem (consistency $\Rightarrow$ convergence)

Assumptions:

Let $x(t)$ solve the IVP $x' = f(t,x), x(t_0) = x_0$.
For the numerical solution let us use the general following one-step method

$$\begin{aligned} \eta_0 &:= x_0 \\ \eta_k &:= \eta_{k-1} + h_{k-1}\phi(t_{k-1}, \eta_{k-1}, h_{k-1}, f) \\ t_k &:= t_{k-1} + h_{k-1}, \quad k = 1, \ldots, N \end{aligned}$$

Furthermore, let f fulfill a global Lipschitz condition in x

$$\|f(t,x) - f(t,z)\| \leq L\|x - z\| \quad \forall x, z \in \mathbb{R}^n, \ \forall t \in I$$

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Claim:

$$\frac{\|T(t,x,h)\|}{h} \leq \sigma(h) \;\; \forall x, \forall t \in I, \forall h \in]0,H[\quad \wedge \quad \lim_{h\to 0} \sigma(h) = 0$$

$$\Longrightarrow \quad \|x(t_k) - \eta_k\| \leq \sigma(h_{max}) \cdot \frac{e^{L|t_k - t_0|} - 1}{L} \qquad \text{for } k = 1, \ldots, N$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Proof: We assume w.l.o.g. $t_0 < t_k, h_k > 0 \ \forall k$ and consider exact solutions $z_k(t)$ for the IVP with modified initial values

$$z_k(t_k) = \eta_k\,, \quad z_k' = f(t, z_k(t)) \quad \Rightarrow \quad z_0(t) = x(t)$$

Using the triangle inequality and the theorem on the continuous dependency of a solution from the initial values we get

$$\begin{aligned}
\|x(t_n) - \eta_n\| &= \|z_0(t_n) - z_n(t_n)\| \ \leq \ \sum_{k=1}^{n} \|z_{k-1}(t_n) - z_k(t_n)\| \\
&\leq \sum_{k=1}^{n} \|z_{k-1}(t_k) - \underbrace{z_k(t_k)}_{=\eta_k}\| e^{L|t_n - t_k|} \\
&= \sum_{k=1}^{n} \|z_{k-1}(t_k) - (\eta_{k-1} + h_{k-1}\phi(t_{k-1}, \eta_{k-1}, h_{k-1}, f))\| \, e^{L|t_n - t_k|} \\
&= \sum_{k=1}^{n} \left\| h_{k-1} \frac{z_{k-1}(t_k) - z_{k-1}(t_{k-1}) - h_{k-1}\phi(t_{k-1}, \eta_{k-1}, h_{k-1}, f)}{h_{k-1}} \right\| e^{L|t_n - t_k|} \\
&\leq \sum_{k=1}^{n} h_{k-1}\sigma(h_{k-1}) e^{L|t_n - t_k|} \ \leq \ \sigma(h_{max}) \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} e^{L(t_n - \xi)} \, \mathrm{d}\xi \\
&= \sigma(h_{max}) \int_{t_0}^{t_n} e^{L(t_n - \xi)} \, \mathrm{d}\xi \quad \Rightarrow \ \text{claim}
\end{aligned}$$

We estimated $h_{k-1} e^{L|t_n - t_k|}$ by the integral, for $e^{L|t_n - t_k|} \leq e^{L|t_n - \xi|} \ \forall \xi \in [t_{k-1}, t_k]$. □

3.8 Stiff ODEs

■ Introductory Example

$$\underbrace{\begin{pmatrix} x_1' \\ x_2' \end{pmatrix}}_{=:x'} = \underbrace{\begin{pmatrix} \frac{\lambda_1 + \lambda_2}{2} & \frac{\lambda_1 - \lambda_2}{2} \\ \frac{\lambda_1 - \lambda_2}{2} & \frac{\lambda_1 + \lambda_2}{2} \end{pmatrix}}_{=:A} \underbrace{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}}_{=:x}, \quad f(t,x) := Ax\,, \quad \lambda_1, \lambda_2 < 0$$

General analytic solution:

$$\begin{aligned}
x_1(t) &= c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t} \\
x_2(t) &= c_1 e^{\lambda_1 t} - c_2 e^{\lambda_2 t}
\end{aligned}$$

First we choose e.g. the *explicit Euler* as our solution method and obtain

$$\eta_{i+1} = \eta_i + h_i f(t_i, \eta_i) = \eta_i + h_i A \eta_i = (I + h_i A)\eta_i, \,, \quad \eta_i \in \mathbb{R}^2$$

and obtain in case of $h = h_i \ \forall i$ the numerical solution (proof by induction)

$$\begin{aligned}
\eta_{1,i} &= c_1(1 + h\lambda_1)^i + c_2(1 + h\lambda_2)^i \\
\eta_{2,i} &= c_1(1 + h\lambda_1)^i - c_2(1 + h\lambda_2)^i
\end{aligned}$$

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

The numerical approximation converges to 0 for $i \to \infty$ only if $|1+h\lambda_1| < 1 \ \wedge \ |1+h\lambda_2| < 1$; only in this case the numerical approximation matches the analytic solution in the limit.

We will see later on: The analytic solution is asymptotically stable, therefore the ODE problem is called stiff. The numerical solution "explodes", if the stepsize is too large.

Let us now choose $\lambda_2 \ll \lambda_1$ (e.g. $\lambda_1 = -1$ and $\lambda_2 = -1000$). Then for e.g. $t \geq 0.1$ the component $e^{\lambda_2 t}$ does "not" contribute to the numerical solution ($e^{-100} \approx 3 \cdot 10^{-44}$), but nevertheless it determines and reduces the stepsize of the integrator:

$$|1 - 1000h| < 1 \quad \Rightarrow \quad h < 0.002$$

If we use the *implicit Euler* instead

$$\eta_{i+1} = \eta_i + h_i f(t_{i+1}, \eta_{i+1}) = \eta_i + h_i A \eta_{i+1} \,,$$

then for $h = h_i \ \forall i$ we get the numerical solution (proof by induction)

$$\begin{aligned} \eta_{1,i} &= \frac{c_1}{(1-h\lambda_1)^i} + \frac{c_2}{(1-h\lambda_2)^i} \\ \eta_{2,i} &= \frac{c_1}{(1-h\lambda_1)^i} - \frac{c_2}{(1-h\lambda_2)^i} \end{aligned}$$

We get $\eta_i \to 0$ for $i \to \infty \ \forall h > 0$, for always $|1 - \lambda_i h| > 1$ is true because of the assumption $\lambda_i < 0, \ i = 1, 2$. There is no stepsize restriction for the implicit method. □

■ Repetition: stability of linear ODEs

Consider the linear system with constant coefficients

$$x'(t) = Ax(t) \,, \quad x(t_0) = x_0 \,, \quad A \in \mathbb{R}^{n \times n}$$

The system is stable, if for all EWs λ_i of A: $Re(\lambda_i) \leq 0$ and in case of $Re(\lambda_i) = 0$ the algebraic and the geometric multiplicity of λ_i are equal (i.e. the EW λ_i has multiplicity k_i and k_i linearly independent EVs).

The system is exponentially and thus also asymptotically stable, if $Re(\lambda_i) < 0$ is true for all EWs λ_i of A. □

■ How to characterize stiff ODEs?

Stiff ODE systems contain at least one asymptotically stable component. Here perturbations in the initial conditions are rapidly damped.

From the theorem on the continuous dependency of a solution from the initial values we get the (slightly simplified) criterion

$$(t_f - t_0) \| f_x(t, x) \| \leq L(t_f - t_0) \ll 1$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

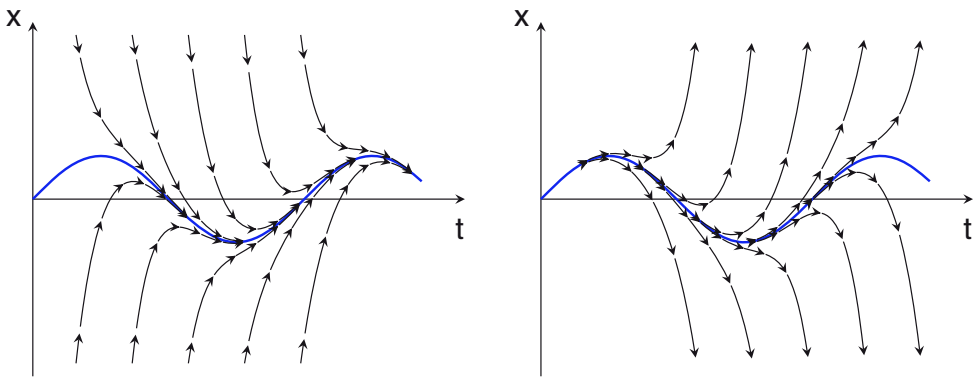


Figure 7: Examples of asymptotically stable (le., with large deviations in the initial values) and instable (ri., after small perturbations from the nominal trajectory) behaviour of the solutions of an ODE, nominal solution is marked in blue.

■ How to test solution methods for stiff ODEs?

Starting with the above characterization of stiff ODEs, we want to develop very simple ODEs which may serve us to decide whether a numerical method is suited for the solution of stiff ODEs.

Let be $x' = f(x)$ an (autonomous) ODE system with the exact solution $x(t)$ and the initial value $x(t_0) = x_0$. Let be $v(t)$ another solution of the same ODE, but for slightly modified initial values. By Taylor expansion we get

$$v'(t) = f(v(t)) \doteq f(x(t)) + f_x(x(t)) \cdot (v(t) - x(t))$$

Simplifying assumption 1:

$f_x(x(t))$ is only changing slowly , i.e. $f_x(x(t)) \approx const \approx J$. For $e(t) := v(t) - x(t)$ we get the new ODE $e'(t) = Je(t)$. We investigate the difference $e(t)$ to obtain the asymptotic behaviour sketched in Fig. 7.

$\Rightarrow$ 1^{st} test ODE: $\quad x'(t) = Ax(t)\,,\ \ x(0) = x_0\,,\ \ A \in \mathbb{R}^{n\times n}$

Simplifying assumption 2:

By a similarity transformation, in special cases J can be transformed to diagonal form: $\exists Q\ \ni\ Q^{-1}JQ = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. We define $p(t)$ by $p(t) := Q^{-1}e(t)$.

Then the 1^{st} test ODE decomposes into the following scalar ODEs

$$Q^{-1}e'(t) = Q^{-1}JQ \cdot Q^{-1}e(t) \ \ \Rightarrow\ \ p_i'(t) = \lambda_i p_i(t)\,,\ \ i = 1, \ldots, n$$

If J can be transformed to diagonal form, then the $\lambda_i \in \mathbb{C}$ are the EWs. Because a stiff system is characterized by asymptotic stability, we choose $Re(\lambda) < 0$

$\Rightarrow$ 2^{nd} test ODE: $\quad x'(t) = \lambda x(t)\,,\ \ x(0) = x_0\,,\ \ Re(\lambda) < 0$ (Dahlquist 1963)

The two simplifying assumptions preserve the fundamental properties of the ODE system!

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

There might exist stiff ODEs that do not satisfy the simplifying assumptions. A method which has been tested to work for Dahlquist's ODE not necessarily works well for these ODEs.

Requirements for a good numerical method:

The exact solution of the scalar system is $x(t_i+h) = e^{h\lambda}x(t_i)$.

The numerical solution should coincide as good as possible – but at least qualitatively – with the true and exact solution. The minimum requirement is

$$\begin{aligned} |x(t_i+h)| &\leq |x(t_i)| \quad \forall h \\ \lim_{h\cdot Re(\lambda)\to-\infty} x(t_i+h) &= 0 \end{aligned}$$

Definition (stability function $R(z)$)

The stability function is the function that allows an equivalent formulation of the numerical one-step method under consideration applied to the 2^{nd} test ODE

$$\eta_{i+1} = R(h\lambda)\eta_i$$

Here we use the special argument $z = h\lambda$ in $R(z)$.

Thus the stability function $R(z)$ is defined via the numerical solution after one step for the 2^{nd} test ODE

$$x'(t) = \lambda x(t)\,, \;\; x(0) = \eta_i\,, \;\; z = h\lambda$$

Alternatively, we can use the 1^{st} test ODE and analogously define $R(z)$ by

$$\eta_{i+1} = R(hA)\eta_i$$

■ Example

Explicit Euler: $R(z) = 1+z$

$$\eta_{i+1} = \eta_i + hf(t,\eta_i) = \eta_i + hA\eta_i = (I+hA)\eta_i$$

We substitute $hA \to z$, interpret the identity matrix as "1" and obtain $R(z) = 1+z$. The same formula we would obtain if we directly insert the 2^{nd} test ODE.

Implicit Euler: $R(z) = \dfrac{1}{1-z}$

For the implicit Euler applied to the 1^{st} test ODE we obtain

$$\eta_{i+1} = \eta_i + hA\eta_{i+1} \;\Leftrightarrow\; (I-hA)\eta_{i+1} = \eta_i \;\Leftrightarrow\; \eta_{i+1} = (I-hA)^{-1}\eta_i$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Trapezoidal rule: $R(z) = \dfrac{2+z}{2-z}$

From the original definition of the trapezoidal rule applied to the 1^{st} test ODE, we obtain the stability function $R(z)$ analogously to the implicit Euler

$$\begin{aligned} \eta_{i+1} &= \eta_i + \frac{h}{2}\left(f(t_i,\eta_i) + f(t_{i+1},\eta_{i+1})\right) = \eta_i + \frac{h}{2}\left(A\eta_i + A\eta_{i+1}\right) \\ &\Rightarrow \left(1 - \frac{h}{2}A\right)\eta_{i+1} = \left(1 + \frac{h}{2}A\right)\eta_i \\ &\Rightarrow \eta_{i+1} = (2-hA)^{-1}(2+hA)\,\eta_i \end{aligned}$$

□

■ Observation

For (almost) all methods we get:

If the method is explicit, then $R(z)$ is a polynomial;
if the method is implicit, then $R(z)$ is a rational function. □

Theorem

Let be $A \in \mathbb{R}^{n\times n}$ diagonalizable: $Q^{-1}AQ = D = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$. Let us define a numerical method by $\eta_{i+1} = R(hA)\eta_i$ with R rational function and assume that $Re(\lambda_i) < 0 \ \ \forall i$, i.e. for all EWs of A.

Then $\xi_i := Q^{-1}\eta_i$ satisfies the recursion $\xi_{i+1} = R(hD)\xi_i$; in addition, for $h > 0$ the so-defined numerical method converges as required

$$\eta_j \to 0 \text{ for } j \to \infty \quad \Longleftrightarrow \quad |R(h\lambda_i)| < 1 \ \ \forall i = 1\ldots n$$

Definition (stability)

A numerical method defined by $\xi_{i+1} = R(hA)\xi_i$ is called

$$\begin{aligned} \textit{absolutly stable} &:\Leftrightarrow \ |R(z)| < 1 \ \ \forall z \text{ with } Re(z) < 0 \\ \textit{A-stable} &:\Leftrightarrow \ |R(z)| \le 1 \ \ \forall z \text{ with } Re(z) \le 0 \\ \textit{L-stable} &:\Leftrightarrow \ \text{A-stable and in addition } \lim_{Re(z)\to-\infty} R(z) = 0 \end{aligned}$$

The set $\mathcal{S}_R := \{z \in \mathbb{C} \,||R(z)| \le 1\}$ is called *stability domain* and
$\mathcal{M}_R := \{z \in \mathbb{C} \,||R(z)| < 1\}$ is called *domain of absolute stability* of the method.

Remark

The larger the set $\mathcal{M}_R \cap \mathbb{C}_-$, the better a method is suited for the treatment of stiff ODEs.

For $\mathcal{M}_R \supseteq \mathbb{C}_- = \{z \in \mathbb{C} \,|\, Re(z) < 0\}$ the method is absolutly stable.

If $|R(z)| < 1$ for $z = h\lambda$, then if the neg. real part of λ increases, the stepsize h has to decrease to obtain the same value of the stability function $R(z)$. □

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

■ **Example** (stability domains)

Explicit Euler: $\mathcal{M}_{1+z} = \{z \in \mathbb{C} \,|\, |1+z| < 1\}$

Implicit Euler: $\mathcal{M}_{1/(1+z)} = \{z \in \mathbb{C} \,|\, |1-z| > 1\}$, i.e. the implicit Euler is absolute stable.

Classical explicit Runge-Kutta method RK 4: $R(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24}$

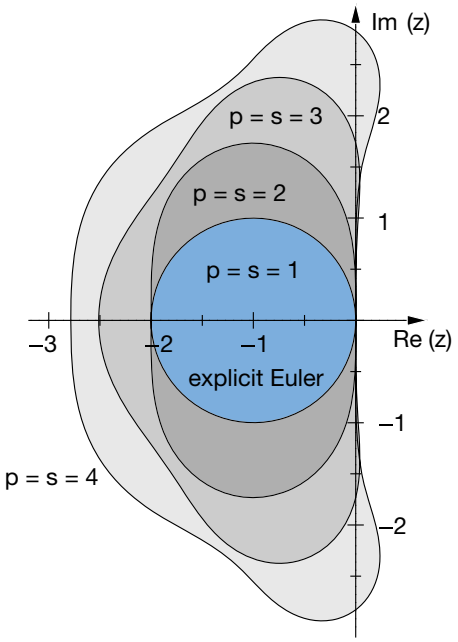


Figure 8: Stability domains of explicit RK methods of order p with s stages. $z = h\lambda \Rightarrow$ on the boundary of the stability domain: if $|\lambda|$ increases, then h has to decrease.

Implicit s-stage Runge-Kutta methods (IRK)

$$\phi(t,x,h) = \sum_{k=1}^{s} b_k \cdot f_k(t,x,h), \quad f_k(t,x,h) := f\left(t + c_k h, x + h \sum_{j=1}^{s} \alpha_{kj} f_j(t,x,h)\right)$$

These methods have excellent stability properties, but they are rather expensive numerically: In each integration step a system of nonlinear equations of dimension $(n \cdot s)$ has to be solved $\rightarrow \mathcal{O}(n^3 s^3)$ operations!

Example: A Radau-IIA method of order $p = 2s - 1$ is L-stable, e.g.

$1/3$	$5/12$	$-1/12$
1	$3/4$	$1/4$
	$3/4$	$1/4$

■ How to detect stiffness in an IVP for ODEs

Either analyze the local stability of the system after linearization or use a good explicit integrator first. If the stepsize $h \rightarrow 0$, switch to a stiff integrator. □

© Rainer Callies TUM 2020

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

4 Finite Differences

4.1 One-Dimensional Model Problem

4.1.1 Model problem

In an experiment elevation data along a mountain path are measured by GPS. Let the data be superimposed by heavy noise due to low signal strength. How to get a ”reasonable” altitude profile of the terrain structure?

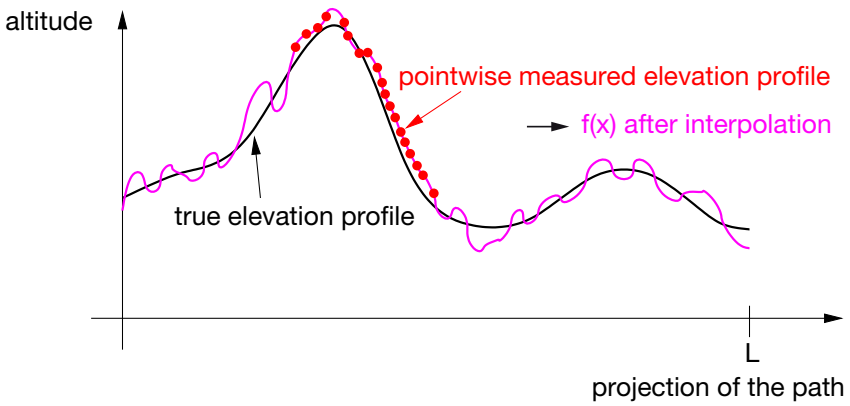


Figure 9: Measured (red) and true elevation profile of the path.

■ **Simple solution idea**

The measured elevation data (red) are interpolated by a spline $f : [0, L] \to \mathbb{R}$.

We want to get determine a smoothing curve $u : [0, L] \to \mathbb{R}$, which is close to f (i.e. $|u(x) - f(x)|$ is small $\forall x \in [0, L]$) but not noisy (i.e. $|u'(x)|$ small).

This leads us to the following objective function for a minimization problem

$$I(u) = \int_0^L (u(x) - f(x))^2 + \beta(u'(x))^2 \, \mathrm{d}x \stackrel{!}{=} \min$$

with $\beta > 0$ constant and chosen properly. It is no finite-dimensional problem of nonlinear optimization, because $u(x)$ has to be determined at infinitely many x-values. For sake of simplicity let the altitude at the initial and the final point ($u(0) = f(0)$ and $u(L) = f(L)$) be exactly known/measured.

■ **Mathematical realization of that idea**

For the solution we use the method of Lagrange: For an arbitrarily chosen function $\eta \in C^1([0, L], \mathbb{R})$ with $\eta(0) = 0 = \eta(L)$ $(*)$ we embed the optimal solution u into and compare it with the one-dimensional set of functions $v := u + \varepsilon\eta$ for $\varepsilon \in [-\varepsilon_0, \varepsilon_0]$. We have chosen $(*)$ because the values at the endpoints are prescribed and that has to be true also for all possible solution candidates.

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

If u is an optimum, then the following necessary condition holds

$$\frac{\mathrm{d}J(\varepsilon)}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} = 0 \quad \text{with} \quad J(\varepsilon) := I(u+\varepsilon\eta)$$

We differentiate $I(v)$ with respect to ε and get (using chain rule "chain" and integration by parts "p.I." on η')

$$\begin{aligned}
0 &= \frac{\mathrm{d}J(\varepsilon)}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} = \frac{\mathrm{d}I(u+\varepsilon\eta)}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\left(\int_0^L (v(x)-f(x))^2 + \beta(v'(x))^2\,\mathrm{d}x\right)\Big|_{\varepsilon=0} \\
&\overset{chain}{=} \int_0^L \big(2(u(x)-f(x))\cdot\eta(x) + 2\beta u'(x)\cdot\eta'(x)\big)\,\mathrm{d}x \\
&\overset{p.I.}{=} 2\int_0^L \big((u(x)-f(x))\cdot\eta(x) - \beta u''(x)\cdot\eta(x)\big)\,\mathrm{d}x + u'(x)\eta(x)\Big|_{x=0}^{x=L} \qquad (*)
\end{aligned}$$

The second term in $(*)$ vanishes, because $\eta(0) = 0 = \eta(L)$. We apply the Fundamental lemma (see below) to the integral and obtain, that a necessary condition for an optimum is that u solves the following boundary value problem (BVP)

Problem (P):

$$\begin{aligned}
-\beta u''(x) + u(x) &= f(x), \quad x \in]0, L[\\
u(0) &= f(0) \\
u(L) &= f(L)
\end{aligned}$$

The Fundamental lemma could be applied because the integral has to be zero for every choice of such a test function η:

Fundamental lemma

Let be $G \in \mathcal{C}^0([a,b],\mathbb{R})$, $\eta \in \mathcal{C}^1([a,b],\mathbb{R})$ with $\eta(a) = \eta(b) = 0$.

If $\forall\,\eta : \int_a^b \eta(x)G(x)\mathrm{d}x = 0$, then it follows: $G(x) \equiv 0$.

Remark

The boundary conditions in our example are $u(0) = f(0)$ and $u(L) = f(L)$, the function values are prescribed. That type of boundary condition is called Dirichlet condition.

If e.g. $u(L) = f(L)$ is omitted, then the new and special boundary condition $u'(L) = 0$ is necessary to fulfill $(*)$. If the derivative with respect to the exterior normal to the boundary (here in 1 D i.e. the ordinary derivative) is given, that type of boundary condition is called Neumann condition. □

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Remark

From problem (P) we see that solutions u have to be at least in $\mathcal{C}^2([0,L],\mathbb{R})$!
For f a continuous approximation by a polygon is sufficient. □

4.1.2 Numerical Approximation by Finite Differences

On $[0,L]$ we define a mesh with N gridpoints and a mesh size $h := L/(N-1)$

$$\Omega_h := \{x_i \,|\, x_i = (i-1)\cdot h, i = 1,\ldots,N\}$$

and want to approximate the exact values $u(x_i)$ by $U_i \approx u(x_i)$.

To obtain the U_i, the derivative $u''(x)$ is approximated by the difference quotient (Taylor expansion!)

$$u''(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + \mathcal{O}(h^2)$$

This Taylor expansion is possible only for $u \in \mathcal{C}^4([0,L],\mathbb{R})$!

Insertion into the BVP (P) gives the new discretized problem

Problem (P_h):

Consider the function $f \in \mathcal{C}^0([0,L],\mathbb{R})$ and choose $\beta > 0$. Let be $N \in \mathbb{N}$ and $h := L/(N-1)$.

Determine $U_i, i = 1,\ldots N$, such that

$$\begin{aligned} -\frac{\beta}{h^2}(U_{i+1} - 2U_i + U_{i-1}) + U_i &= f(x_i), \quad i = 2,\ldots,N-1, \\ U_1 &= f(x_1) \\ U_N &= f(x_N) \end{aligned}$$

□

In matrix notation (P_h) can be stated as a sparse linear system:

$$\left(-\frac{\beta}{h^2}\begin{pmatrix} 0 & & & & 0 \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ 0 & & & & 0 \end{pmatrix} + \begin{pmatrix} 1 & & & & 0 \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ 0 & & & & 1 \end{pmatrix}\right)\begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ \vdots \\ U_N \end{pmatrix} = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ \vdots \\ f(x_N) \end{pmatrix}$$

Remark (again)

We have obtained the difference approximation of $u''(x)$ using Taylor's theorem. From the first nonvanishing term of the error we see that the approximation quality $\mathcal{O}(h^2)$ stated there is valid only for $u \in \mathcal{C}^4([0,L],\mathbb{R})$!!

Therefore implicit smoothness assumptions are used ($\mathcal{C}^2$ from optimization, $\mathcal{C}^4$ or $\mathcal{C}^3$ from finite differences) which are not part of the original problem. □

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

4.1.3 Convergence of the Finite Difference Method

■ Introduction

The discretized problem (P_h) leads to a linear system. Two questions arise:

- Is the linear system uniquely solvable?
- Does the solution of (P_h) converge to the solution of (P) for $N \to \infty$ (or $h \to 0$)? If yes, with which convergence rate?

In chap. 4.1.3 we analyze these questions only for the one-dimensional model problem discussed in the previous chap. 4.1.1! □

Lemma (discrete maximum principle, special version)

Let $\{U_i, i = 1, \ldots, n\}$ solve the problem (P_h). Then

$$\begin{aligned} \max_{i=1,\ldots,N} U_i &\leq \max_{i=1,\ldots,N} f(x_i) \\ \min_{i=1,\ldots,N} U_i &\geq \min_{i=1,\ldots,N} f(x_i) \end{aligned}$$

Proof:

We only show the first property. Let U_j be the maximum of $\{U_i, i = 1, \ldots, n\}$.

If $j = 1$ (analogously for $j = N$), then

$$\max_{i=1,\ldots,N} U_i = U_1 = f(x_1) \leq \max_{i=1,\ldots,N} f(x_i)$$

If $j \in \{2, \ldots, N\}$, then because of the maximum property

$$U_j \geq U_{j-1} \quad \wedge \quad U_j \geq U_{j+1}$$

We insert this into the finite difference expression and obtain

$$\begin{aligned} (U_{j+1} - 2U_j + U_{j-1}) \leq 0 \quad &\Rightarrow \quad -\frac{\beta}{h^2}(U_{j+1} - 2U_j + U_{j-1}) \geq 0 \\ &\overset{(P_h)}{\Longrightarrow} \quad f(x_j) - U_j \geq 0 \quad \Rightarrow \quad U_j \leq f(x_j) \leq \max_{i=1,\ldots,N} f(x_i) \end{aligned}$$

■ Uniqueness □

We now address the first question.

The linear system is uniquely solvable, if the matrix is regular. For that we have to show that either the determinant of the matrix is non-zero – e.g. with the minor expansion formula (= Determinantenentwicklungssatz) – or that for the homogeneous system the only solution is $U := (U_1, \ldots, U_N) = 0$.

In our case the latter approach is simple if we use the discrete maximum principle:

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Consider the homogeneous system, i.e. $f(x_i) = 0 \;\; \forall i \in \{1, \ldots, N\}$.
Then $\min_{i=1,\ldots,N} f(x_i) = 0 = \max_{i=1,\ldots,N} f(x_i)$.

From the discrete maximum principle we get

$$0 \leq \min_{i=1,\ldots,N} U_i \leq \max_{i=1,\ldots,N} U_i \leq 0 \;\; \Rightarrow \;\; U_i = 0 \;\; \forall i \, .$$

Therefore the linear system is uniquely solvable.

■ Consistency, Stability, Convergence

Definition

Let denote $\Omega =]0, L[$ the domain, $h = L/(N-1)$ the discretization mesh size, $\Omega_h := \{x_i = (i-1) \cdot h, i = 1, \ldots, N\}$ the mesh of $N \in \mathbb{N}$ gridpoints; $u \in \mathcal{C}^4(\Omega, \mathbb{R})$.

Then the differential operator L^β is defined by

$$L^\beta := -\beta \frac{\mathrm{d}^2}{\mathrm{d}x^2} + 1 \quad \Rightarrow \quad L^\beta u(x) = -\beta u''(x) + u(x)$$

The discrete differential operator $L_h^\beta u : \Omega_h \to \mathbb{R}$ is defined by

$$\left(L_h^\beta u\right)(x_i) := -\frac{\beta}{h^2}\left(u(x_{i+1}) - 2u(x_i) + u(x_{i-1})\right) + u(x_i)$$

An operator always may be considered as a rule that tells us what to do e.g. with a function or a point. □

These definitions allow us a more compact and clear formulation of the following theorems.

Theorem (consistency, model problem)

For $u \in \mathcal{C}^4(\Omega, \mathbb{R})$ and our model problem we get

$$\max_{i=2,\ldots,N-1} \left| L^\beta u(x_i) - \left(L_h^\beta u\right)(x_i) \right| \leq C h^2$$

The constant C does not depend on h. The order of consistency is 2 because of the exponent in h^2.

Proof:

$$\left| L^\beta u(x_i) - \left(L_h^\beta u\right)(x_i) \right| = \left| -\beta u''(x_i) + u(x_i) - \left(L_h^\beta u\right)(x_i) \right| = \mathcal{O}(h^2)$$

because of the approximation of u'' by the Taylor-based finite difference formula in chap. 4.1.2. □

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Remark

As in case of the one-step methods for ODEs, consistency is a local characterization.

We insert the exact solution u at the grid points x_i into the homogeneous part of the exact differential equation and into its finite difference approximation and measure the maximum difference.

For a consistent method, the difference vanishes for $h \to 0$.

Attention: We do not compare the results U_i of the numerical solution of the ODE with the exact solution $u(x_i)$ here! □

Theorem (stability, continuous dependence on f, model problem)

Let be $U \in \mathbb{R}^N$ the solution of problem (P_h). Then

$$\max_{i=1,\ldots,N} |U_i| \leq \tilde{C} \max_{i=1,\ldots,N} |f(x_i)|$$

The constant $\tilde{C}$ does not depend on h. Such a method is called stable.

Proof:

We again apply the discrete maximum principle and directly obtain

$$\max_{i=1,\ldots,N} |U_i| \leq \max_{i=1,\ldots,N} |f(x_i)| \quad \Rightarrow \quad \tilde{C} = 1$$

Here we used that $\min U_i \geq \min f(x_i) \Rightarrow -\min U_i \leq -\min f(x_i)$ and in case of $\min f(x_i) < 0$ we get $-\min f(x_i) = \max(-f(x_i))$. □

Theorem (convergence, model problem)

Let be u the solution of (P) and $u \in \mathcal{C}^4(\Omega, \mathbb{R})$, let $U \in \mathbb{R}^N$ be solution of (P_h). Then

$$\max_{i=1,\ldots,N} |u(x_i) - U_i| \leq \hat{C} h^2$$

Such a method is called convergent of 2^{nd} order.

Proof:

We investigate the error $e : \Omega_h \to \mathbb{R}, e(x_i) := U_i - u(x_i)$ and prove that it solves problem (P_h) with a new right hand side $r(x_i)$ for $i = 2, \ldots, N$:

$$\begin{aligned} \left(L_h^\beta e\right)(x_i) &= -\frac{\beta}{h^2}(U_{i+1} - 2U_i + U_{i-1}) - \left(L_h^\beta u\right)(x_i) = f(x_i) - \left(L_h^\beta u\right)(x_i) \\ &= -\beta u''(x_i) + u(x_i) - \left(L_h^\beta u\right)(x_i) = L^\beta u(x_i) - \left(L_h^\beta u\right)(x_i) =: r(x_i) \end{aligned}$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

We now define $r(x_1) := 0, r(x_N) := 0$; then e solves the problem (P_h) with the new right hand side r instead of f.

Because we have proven that our method is stable:

$$\max_{i=1,\dots,N} |e(x_i)| \leq \tilde{C} \max_{i=1,\dots,N} |r(x_i)|$$

Because we have proven that our method is consistent:

$$|r(x_i)| \leq Ch^2 \quad \text{for } i = 2,\dots,N-1 \quad \wedge \quad r(x_1) = r(x_N) = 0$$

In total we get

$$\max_{i=1,\dots,N} |e(x_i)| \leq \hat{C} \quad \text{with } \hat{C} = C \cdot \tilde{C}$$

and $\hat{C}$ is independent on h. □

4.2 Quasilinear PDEs

Definition

A partial differential equation (PDE) for the scalar function $u(x_1,\dots,x_n)$ with the n independent varables $(x_1,\dots,x_n) \in D \subseteq \mathbb{R}^n$ is an equation of the form

$$F\left(x_1,\dots x_n,\ u, \frac{\partial u}{\partial x_1},\dots \frac{\partial u}{\partial x_n},\ \frac{\partial^2 u}{\partial x_1 \partial x_1},\dots \frac{\partial^2 u}{\partial x_1 \partial x_n},\ \dots\right) = 0\,.$$

If F is a linear function of u and its derivatives, then the PDE is called linear.

A PDE is called quasilinear, if F is at least linear in the highest order derivatives of u.

The order of the PDE is the highest order partial derivative of u in F. □

Notation

In PDEs, it is common to denote partial derivatives using subscripts. So e.g. for $u = u(x,y)$ we write:

$$u_x = \frac{\partial u}{\partial x},\quad u_{xx} = \frac{\partial^2 u}{\partial x^2},\quad u_{xy} = \frac{\partial^2 u}{\partial y\,\partial x} = \frac{\partial}{\partial y}\left(\frac{\partial u}{\partial x}\right).$$

Especially in physics, nabla (∇) is often used to denote spatial derivatives, and $\dot{u}, \ddot{u}$ for time derivatives. For example, the wave equation can be written as

$$\ddot{u} = c^2 \nabla^2 u = c^2 \Delta u$$

where Δ is the Laplace operator. □

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

■ Example

General scalar linear PDE of 2^{nd} order with 2 independent variables x, y:

$$a(x,y)u_{xx} + b(x,y)u_{xy} + c(x,y)u_{yy} + d(x,y)u_x + e(x,y)u_y + g(x,y)u = f(x,y)$$

with $a,b,c,d,e,f,g \in \mathcal{C}^0(\Omega,\mathbb{R}), \Omega \subset \mathbb{R}^2$ bounded domain and $|a| + |b| + |c| > 0 \ \forall (x,y) \in \Omega$.

General scalar quasilinear PDE of 2^{nd} order with 2 independent variables x, y:

$$\begin{aligned} &a(x,y,u,u_x,u_y)u_{xx} + b(x,y,u,u_x,u_y)u_{xy} + c(x,y,u,u_x,u_y)u_{yy} \\ &= f(x,y,u,u_x,u_y) \end{aligned}$$

with $a,b,c,f \in \mathcal{C}^0(D,\mathbb{R}), D \subset \mathbb{R}^5$ und $|a| + |b| + |c| > 0$.

If $f \equiv 0$, then the PDE is *homogeneous*, otherwise *inhomogeneous*. □

Classification of quasilinear and linear PDEs of second order

Consider the general quasilinear PDE with n independent variables $x \in \Omega$ and $\Omega \subseteq \mathbb{R}^n$ bounded domain (open, connected, bounded)

$$\begin{aligned} &\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(x,u,p)\frac{\partial^2 u}{\partial x_i \partial x_j}(x) = f(x,u,p) \\ &A = (a_{ij}) \ \text{ symmetric}, \ \ a_{ij}, f \in \mathcal{C}^0(Q,\mathbb{R}), Q \subset \Omega \times \mathbb{R} \times \mathbb{R}^n \\ &x := (x_1,\ldots,x_n), \ \ p := (p_1,\ldots,p_n), \ \ p_i := \frac{\partial u}{\partial x_i} = u_{x_i}, \ \ u(x) \in \mathbb{R}. \end{aligned}$$

In operator notation we write

$$\mathcal{L}u := \sum_{i,k=1}^{n} a_{ik}(x,u,p)\frac{\partial^2 u}{\partial x_i \partial x_j}, \ \ \mathcal{L}u(x) = f(x,u,p)$$

The PDE is in (x,u,p)

elliptic , if all EWs of A have the same sign (all are positive or negative)
parabolic , if exactly one EW is equal to zero and
all the other EWs of A have the same sign (are pos. or neg.)
hyperbolic , there is only one negative EW and all the rest are positive, or
there is only one positive EW and all the rest are negative.

A PDE is *elliptic/parabolic/hyperbolic*, if this property holds at every point of the domain.

Analogously we classify linear PDEs of second order

$$\mathcal{L} := -\sum_{i,k=1}^{n} a_{ik}(x)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{n} b_i(x)\frac{\partial}{\partial x_i} + c(x), \ \ \mathcal{L}u(x) = f(x)$$

□

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Well-posed PDE problems (Hadamard)

The mathematical term "well-posed problem" stems from a definition given by Hadamard. He believed that mathematical models of physical phenomena should have the properties that:

(1) a solution exists,
(2) the solution is unique,
(3) the solution's behavior changes continuously with the data (stability).

Examples of well-posed problems include the Dirichlet problem for Laplace's equation, and the heat equation with specified initial conditions.

Especially important in PDE applications is the correct determination of the initial data and the boundary values. Otherwise it might happen that no solution exists or that the solution changes dramatically even for a very small change in the data. □

4.3 Poisson Equation

We use the Poisson equation $-\Delta u = f$ as an example of an elliptic PDE. To obtain a numerical solution by Finite Difference methods, we proceed step by step as in the treatment of the one-dimensional model problem. We will use the Poisson equation again when we discuss the Finite Element method.

4.3.1 Derivation of the Poisson Equation

■ Problem

Let us describe the vertical displacement of a thin membrane (e.g. of a drum) caused by an external load f.

The resulting shape of the membrane can be obtained as the solution u of a variational problem from physics.

■ Notation

Let be $\Omega \in \mathbb{R}^2$ a bounded domain and $f \in \mathcal{C}^0(\Omega, \mathbb{R})$ a given function. Let denote $u : \Omega \to \mathbb{R}, \;\; (x,y) \mapsto u(x,y)$, the function that describes the vertical displacement of the membrane at every $(x,y) \in \Omega$.

Let the boundary $\partial\Omega$ of Ω consist of two parts Γ_D and Γ_N with

$$\Gamma_D \cup \Gamma_N = \partial\Omega \;\; \wedge \;\; \Gamma_D \cap \Gamma_N = \emptyset .$$

As boundary condition on Γ_D we assume a Dirichlet condition (function values prescribed)

$$u(x,y) := G(x,y) \;\; \text{for} \;\; (x,y) \in \Gamma_D$$

In addition, let $n(x,y)$ denote the exterior normal, i.e. the outward pointing unit normal vector on $\partial\Omega$; the directional derivative $\partial u/\partial n$ is calculated via the scalar product

$$\frac{\partial u}{\partial n}(x,y) = \langle n(x,y), \nabla u(x,y) \rangle_2$$

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

■ Physical model

From physics we get (without proof) for the potential energy of the deformed membrane

$$I(u) := \frac{1}{2}\int_\Omega \langle \nabla u, \nabla u\rangle_2 \,\mathrm{d}x\,\mathrm{d}y - \int_\Omega uf\,\mathrm{d}x\,\mathrm{d}y - \int_{\Gamma_N} uH\,\mathrm{d}s$$

The potential energy $I(u)$ consists of the strain energy (first integral, Verzerrungsenergie) minus the energy resulting from the external forces acting on the surface Ω and on the boundary Γ_N.

A physical system in equilibrium takes the state of minimum energy and therefore we get for u

$$I(u) \;\rightarrow\; \min! \quad\wedge\quad u = G \text{ on } \Gamma_D$$

■ Variational approach

We assume that a "classical solution" $u \in \mathcal{C}^2(\Omega,\mathbb{R}) \cap \mathcal{C}^1(\bar\Omega,\mathbb{R})$ with $\bar\Omega := \Omega \cup \partial\Omega$ exists and will obtain the Poisson equation as a necessary condition for a minimum.

For the solution we again (similar to chap. 4.1.1) use the method of Lagrange: For an arbitrarily chosen function $\eta \in \mathcal{C}^2(\Omega,\mathbb{R}) \cap \mathcal{C}^0(\bar\Omega,\mathbb{R})$ with $\eta\big|_{\Gamma_D} = 0$ we embed the optimal solution u into and compare it with the one-dimensional set of functions $v := u + \varepsilon\eta$ for $\varepsilon \in [-\varepsilon_0, \varepsilon_0]$. We have chosen that embedding because the values on the part Γ_D of the boundary are prescribed and that has to be true also for all possible solution candidates.

If u is an optimum, then the following necessary condition holds

$$\frac{\mathrm{d}J(\varepsilon)}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} = 0 \quad \text{with} \quad J(\varepsilon) := I(u+\varepsilon\eta)$$

Insertion yields

$$\begin{aligned} J(\varepsilon) &= \frac{1}{2}\int_\Omega \langle \nabla u + \varepsilon\nabla\eta, \nabla u + \varepsilon\nabla\eta\rangle_2 \,\mathrm{d}x\,\mathrm{d}y \\ &\quad - \int_\Omega (u+\varepsilon\eta) f\,\mathrm{d}x\,\mathrm{d}y - \int_{\Gamma_N} (u+\varepsilon\eta) H\,\mathrm{d}s \\ \Rightarrow \quad 0 = \frac{\mathrm{d}J(\varepsilon)}{\mathrm{d}\varepsilon}\Big|_{\varepsilon=0} &= \int_\Omega \langle \nabla u, \nabla\eta\rangle_2 \,\mathrm{d}x\,\mathrm{d}y - \int_\Omega \eta f\,\mathrm{d}x\,\mathrm{d}y - \int_{\Gamma_N} \eta H\,\mathrm{d}s \quad (*) \end{aligned}$$

Now we need something like "generalized integration by parts" in several dimensions. We use the following trick: We define the new vector field $F(x,y) := \eta \cdot \nabla u = (\eta \cdot u_x, \eta \cdot u_y) \in \mathbb{R}^2$ and apply Gauss's divergence theorem (Gaußscher Integralsatz)

$$\int_\Omega \text{div}\, F\,\mathrm{d}x\,\mathrm{d}y = \int_{\partial\Omega} \langle F, n\rangle_2 \,\mathrm{d}s$$

with div $F = \partial F_1/\partial x + \partial F_2/\partial y = u_{xx}\eta + u_x\eta_x + u_{yy}\eta + u_y\eta_y = \eta\Delta u + \langle \nabla u, \nabla\eta\rangle_2$.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Using this expression for the generalized integration by parts we get

$$\int_\Omega \langle \nabla u, \nabla \eta \rangle_2 \, \mathrm{d}x \, \mathrm{d}y = -\int_\Omega \eta \Delta u \, \mathrm{d}x \, \mathrm{d}y + \int_{\partial\Omega} \eta \langle \nabla u, n \rangle_2 \, \mathrm{d}s$$

We insert the last expression into $(*)$, use that $\eta\big|_{\Gamma_D} = 0$ and obtain

$$\begin{aligned} 0 &= \int_\Omega \eta \Delta u \, \mathrm{d}x \, \mathrm{d}y + \int_\Omega \eta f \, \mathrm{d}x \, \mathrm{d}y - \int_{\partial\Omega = \Gamma_N + \Gamma_D} \eta \langle \nabla u, n \rangle_2 \, \mathrm{d}s + \int_{\Gamma_N} \eta H \, \mathrm{d}s \\ &= \int_\Omega \eta (\Delta u + f) \, \mathrm{d}x \, \mathrm{d}y + \int_{\Gamma_N} \eta (H - \langle \nabla u, n \rangle_2) \, \mathrm{d}s \end{aligned}$$

This expression has to be valid for all $\eta \in \mathcal{C}^2(\Omega, \mathbb{R}) \cap \mathcal{C}^0(\bar{\Omega}, \mathbb{R})$ with $\eta\big|_{\Gamma_D} = 0$. Using the Fundamental lemma (in its generalized form) again, we get the following necessary condition for a minimum: u has to solve the

"Poisson equation" (PDE problem)

$$\begin{aligned} -\Delta u = -\left(\frac{\partial^2}{\partial x^2} u(x,y) + \frac{\partial^2}{\partial y^2} u(x,y) \right) &= f(x,y) \quad \text{on } \Omega \\ u(x,y) &= G(x,y) \quad \text{on } \Gamma_D \\ \frac{\partial u}{\partial n}(x,y) = \langle n(x,y), \nabla u(x,y) \rangle_2 &= H(x,y) \quad \text{on } \Gamma_N \\ u \in \mathcal{C}^2(\Omega, \mathbb{R}) \cap \mathcal{C}^1(\bar{\Omega}, \mathbb{R}) & \end{aligned}$$

On Γ_N a Neumann boundary condition has to be fulfilled (derivatives of the solution u prescribed). The Neumann boundary condition cannot be chosen freely in our example, but it is determined by the variational problem!

Remark

On Γ_D we do not need the $\mathcal{C}^1$-property of u, here $\mathcal{C}^0$ is sufficient. □

4.3.2 Poisson Equation and Properties of its Solution

Theorem (maximum principle)

Let be $u \in \mathcal{C}^2(\Omega, \mathbb{R}) \cap \mathcal{C}^0(\bar{\Omega}, \mathbb{R})$.

Maximum principle:
If $-\Delta u = f \leq 0$ in Ω, then u has its maximum on the boundary $\partial\Omega$.

Minimum principle:
If $-\Delta u = f \geq 0$ in Ω, then u has its minimum on the boundary $\partial\Omega$.

Comparison:
Let be $v \in \mathcal{C}^2(\Omega, \mathbb{R}) \cap \mathcal{C}^0(\bar{\Omega}, \mathbb{R})$ another function with $-\Delta u = f \leq -\Delta v = \tilde{f}$ in Ω and $u \leq v$ on $\partial\Omega$, then $u \leq v$ in Ω.

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Remark

A consequence of the maximum principle is that the solution changes continuously with the data on the boundary (in case of Dirichlet condition):

Consider $-\Delta u_1 = f$ and $-\Delta u_2 = f$ with $u_i(x) = G_i(x) \;\; \forall x \in \partial\Omega, i = 1,2$. We get $-\Delta w = 0$ for $w := u_1 - u_2$.

From the maximum principle we conclude

$$w(x) \leq \sup_{z\in\partial\Omega} w(z) \leq \sup_{z\in\partial\Omega} |w(z)|, \quad w(x) \geq \inf_{z\in\partial\Omega} w(z) \geq -\sup_{z\in\partial\Omega} |w(z)|$$

and with this

$$\sup_{x\in\Omega} |u_1(x) - u_2(x)| \leq \sup_{z\in\partial\Omega} |u_1(z) - u_2(z)| = \sup_{z\in\partial\Omega} |G_1(z) - G_2(z)|$$

From that we see that the Poisson equation with Dirichlet boundary conditions is well-posed in the sense of Hadamard (effect of changes in f not analyzed here). □

Remark

With the definition $\Delta u(x) := \sum_{i=1}^{n} u_{x_i x_i}(x)$ for $u \in \mathcal{C}^2(\Omega, \mathbb{R}) \cap \mathcal{C}^0(\bar{\Omega}, \mathbb{R})$ with $\Omega \in \mathbb{R}^n$ the Poisson equation can be generalized to $\mathbb{R}^n$. □

4.3.3 Grid, Difference Operators and Boundary Conditions

To use a finite difference method to approximate the solution to a problem, one must first discretize the problem's domain. Note that this means that finite-difference methods produce discrete numerical approximations to the derivatives.

A first introduction to that topic was given in the example "discretization of Laplace's equation" in chap. 2.

In this subchapter we consider the Poisson equation defined on a rectangular domain $\Omega \in \mathbb{R}^2$:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \;\; = \;\; f(x,y) \quad \text{for } (x,y) \in \Omega \subset \mathbb{R}^2$$

with $\Omega := \{(x,y)\,|\,x \in]0,(N+1)h[, y \in]0,(M+1)h[\}$ and define a uniform grid with meshsize h (refinement possible!).

The set of the gridpoints in the interior is defined by

$$\Omega_h := \{(x,y) \in \Omega\,|\,x = ih, y = jh \;\text{ for }\; i = 1,\ldots N, j = 1,\ldots M\}$$

and the set of the gridpoints on the boundary is defined by

$$\begin{aligned} \partial\Omega_h \;\; := \;\; & \{(x,y)\,|\,x = ih, y = 0 \vee y = (M+1)h \;\text{ for }\; i = 0,1,\ldots N, N+1\} \\ & \cup \; \{(x,y)\,|\,x = 0 \vee x = (N+1)h, y = jh \;\text{ for }\; j = 1,\ldots M\} \end{aligned}$$

We do not want to have the same grid point twice in the definition.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

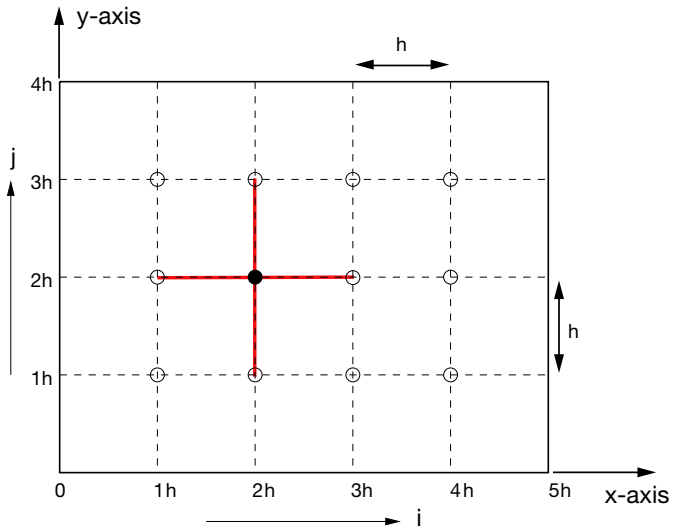


Figure 10: Rectangular domain Ω with $N = 4$ and $M = 3$, interior gridpoints (= elements of Ω_h) marked by circles together with one five-point stecil (red)

We make use of the uniform grid $x_i = i \cdot h, y_j = j \cdot h$ and define $u_{ij} := u(x_i, y_j)$. Our goal is not to calulate $u(x, y) \ \forall x, y \in \bar{\Omega}$, but only to get approximations $U_{i,j}$ of the exact solution $u_{i,j} := u(i \cdot h, j \cdot h)$ at the gridpoints.

Derivatives at the grid points are approximated by discrete difference operators which are derived by Taylor expansions.

■ Derivatives at the interior grid points $(x_i, y_j) \in \Omega_h$

First order partial derivatives:

$$\begin{aligned} u(x+h,y) &= u(x,y) + u_x(x,y) \cdot h + u_{xx}(x,y)\frac{h^2}{2} + \mathcal{O}(h^3) \\ u(x,y+h) &= u(x,y) + u_y(x,y) \cdot h + u_{yy}(x,y)\frac{h^2}{2} + \mathcal{O}(h^3) \end{aligned}$$

$$\begin{aligned} \text{Backward difference:} \quad u_x|_{i,j} &= \frac{1}{h}\Big(u_{i,j} - u_{i-1,j}\Big) && + \mathcal{O}(h) \\ u_y|_{i,j} &= \frac{1}{h}\Big(u_{i,j} - u_{i,j-1}\Big) && + \mathcal{O}(h) \\ \text{Forward difference:} \quad u_x|_{i,j} &= \frac{1}{h}\Big(u_{i+1,j} - u_{i,j}\Big) && + \mathcal{O}(h) \\ u_y|_{i,j} &= \frac{1}{h}\Big(u_{i,j+1} - u_{i,j}\Big) && + \mathcal{O}(h) \\ \text{Centered difference:} \quad u_x|_{i,j} &= \frac{1}{2h}\Big(u_{i+1,j} - u_{i-1,j}\Big) && + \mathcal{O}(h^2) \\ u_y|_{i,j} &= \frac{1}{2h}\Big(u_{i,j+1} - u_{i,j-1}\Big) && + \mathcal{O}(h^2) \end{aligned}$$

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

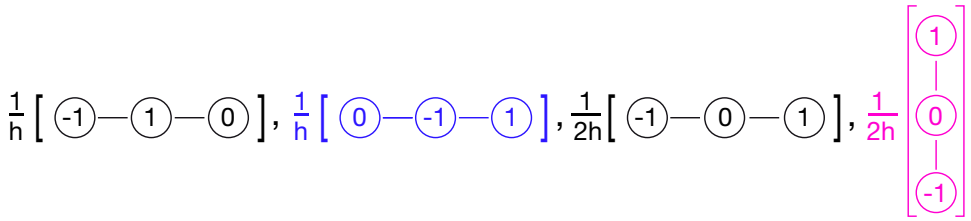


Figure 11: Computational molecules for backward, forward, centered difference approx. of $u_x|_{i,j}$ and centered difference approx. of $u_y|_{i,j}$ (from le. to ri.).

Some second order partial derivatives:

$$\begin{aligned} u_{xx}|_{i,j} &= \frac{u(x-h,y)-2u(x,y)+u(x+h,y)}{h^2}\Big|_{i,j} + \mathcal{O}(h^2) \\ &= \frac{u_{i-1,j}-2u_{i,j}+u_{i+1,j}}{h^2} + \mathcal{O}(h^2) \\ u_{xx}|_{i,j} &= \frac{1}{12h^2}\Big(u(x-2h,y)-16u(x-h,y)+30u(x,y) \\ &\qquad -16u(x+h,y)+u(x+2h,y)\Big)\Big|_{i,j} + \mathcal{O}(h^4) \end{aligned}$$

and analogous formulae for u_{yy}.

Based on those two formulae the Laplace operator Δu can be approximated by the 5-point stencil at (x_i, y_j)

$$\Delta u\Big|_{i,j} = \Delta_h u\Big|_{i,j} + \mathcal{O}(h^2), \quad \Delta_h u\Big|_{i,j} := \frac{1}{h^2}\left(u_{i+1,j}+u_{i-1,j}+u_{i,j+1}+u_{i,j-1}-4u_{i,j}\right)$$

or the non-compact 9-point stencil: $\Delta u\Big|_{i,j} = \Delta_h^{(9)} u\Big|_{i,j} + \mathcal{O}(h^4)$ (see fig. 12).

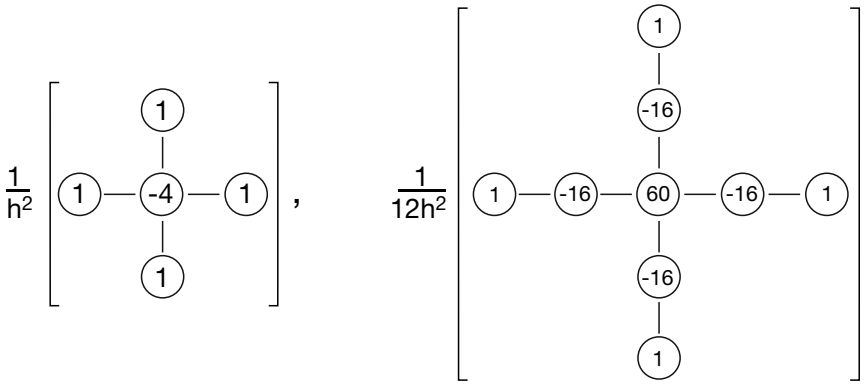


Figure 12: Computational molecules for the 5-point stencil Δ_h (le.) and the non-compact 9-point stencil $\Delta_h^{(9)}$ (ri.).

The approximation of the Laplace operator by the 5-point stencil at (x_i, y_j) leads to the equation for that gridpoint

$$U_{i-1,j}+U_{i+1,j}+U_{i,j-1}+U_{i,j+1}-4U_{i,j} = h^2 f(x_i, y_j) \tag{1}$$

© Rainer Callies TUM 2020

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Remark

The approximation of the Laplace operator by finite differences is possible only if u is sufficiently smooth (because of Taylor!). A much higher smoothness is necessary than for the analytical solution: $u \in \mathcal{C}^4(\Omega, \mathbb{R})$ in case of the 5-point stencil and $u \in \mathcal{C}^6(\Omega, \mathbb{R})$ in case of the non-compact 9-point stencil.

The discretization with the non-compact 9-point stencil includes values at points that are not closest neighbors. This leads to increased difficulties at points close to the boundary.

Because of these two drawbacks, discretizations of higher order are often not used. □

■ Boundary conditions at grid points $(x_i, y_j) \in \partial\Omega_h$

We consider the grid geometry in fig. 13 (section of the rectangular domain Ω with the set of gridpoints $\Omega_h \cup \partial\Omega$) and analyze how to treat the boundary conditions at (w.l.o.g) $(x_0, y_j) = (0, j \cdot h)$.

For the approximation of the Laplace operator Δ we have chosen the 5-point stencil Δ_h. The special stencil centered at (x_1, y_j) is also marked in fig. 13: This is the only stencil at an interior point in contact with our special grid point (x_0, y_j) on the discretized boundary $\partial\Omega_h$.

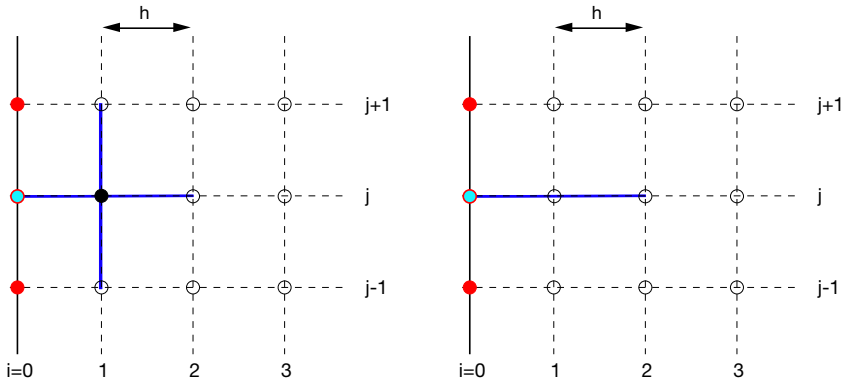


Figure 13: Grid geometry with boundary conditions at point (x_0, y_j) (blue/red): 5-point stencil centered at (x_1, y_j) (black) for Dirichlet boundary condition (le.) and extrapolation centered at (x_0, y_j) (blue/red) for Neumann boundary condition (ri.).

Dirichlet boundary condition:

$U_{0,j}$ is prescribed: $U_{0,j} = G(x_0, y_j)$. For the 5-point stencil this results in the following equation at (x_1, y_j):

$$U_{2,j} + U_{1,j+1} + U_{1,j-1} - 4U_{1,j} = h^2 f(x_1, y_j) - U_{0,j} = h^2 f(x_1, y_j) - G(x_0, y_j) \quad (2)$$

Neumann boundary condition:

The discretized Neumann condition

$$\frac{\partial u}{\partial n}(x, y) = H(x, y) \;\; \forall (x, y) \in \Gamma_{N,h} \subseteq \partial\Omega_h$$

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

at (x_0, y_j) in case of backward difference approximation of the derivative leads to

$$-\frac{\partial u(x,y)}{\partial x}\Big|_{0,j} = H(x_0, y_j) \quad \Rightarrow \quad \frac{u_{0,j} - u_{1,j}}{h} = H(x_0, y_j) + \mathcal{O}(h)$$

Unfortunately this poor approximation results in an additional local error (= consistency error) of order $\mathcal{O}(h)$, whereas the 5-point stencil has a consistency error of only $\mathcal{O}(h^2)$.

The approximation can be improved e.g. by extrapolation techniques. We demonstrate that for an arbitrary function $g \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$

$$\begin{aligned}
g(x+h) &= g(x) + hg'(x) + \frac{h^2}{2}g''(x) + \frac{h^3}{6}g'''(x) + \ldots \quad \Rightarrow \\
\frac{g(x+h)-g(x)}{h} &= g'(x) + \frac{h}{2}g''(x) + \frac{h^2}{6}g'''(x) + \ldots \\
\frac{g(x+2h)-g(x)}{2h} &= g'(x) + hg''(x) + \frac{2h^2}{3}g'''(x) + \ldots
\end{aligned}$$

By linear combination we get

$$2 \cdot \frac{g(x+h)-g(x)}{h} - \frac{g(x+2h)-g(x)}{2h} = g'(x) + \mathcal{O}(h^2)$$

We choose $g(x) = u(x,y)$ and obtain the following equation at (x_0, y_j):

$$\begin{aligned}
& 4u_{1,j} - 4u_{0,j} - u_{2,j} + u_{0,j} = -2h \cdot H(x_0, y_j) + \mathcal{O}(h^2) \\
\Rightarrow \quad & 4U_{1,j} - 3U_{0,j} - U_{2,j} = -2h \cdot H(x_0, y_j) \qquad (3)
\end{aligned}$$

with a consistency error of $\mathcal{O}(h^2)$.

■ Curvilinear boundary

If the domain Ω has a more complicated geometry, a modification of the discretization of the Laplace operator is necessary.

We consider the example in fig. 14.

On the intersections of the curvilinear boundary with the mesh we define additional points (red). The point A has the coordinates $(x_A, y_A) = (i \cdot h - h_A, j \cdot h)$ and the point B has the coordinates $(x_B, y_B) = (i \cdot h, j \cdot h + h_B)$ with $h_A, h_B > 0$. We modifiy the 5-point stencil centered at (x_i, y_j).

Using Taylor expansion again we get

$$\begin{aligned}
u_{xx}\Big|_{i,j} &= 2\left(\frac{u_{i+1,j}}{h(h+h_A)} - \frac{u_{i,j}}{h \cdot h_A} + \frac{u(x_A, y_A)}{h_A(h+h_A)}\right) + \mathcal{O}(h) \\
u_{yy}\Big|_{i,j} &= 2\left(\frac{u_{i,j-1}}{h(h+h_B)} - \frac{u_{i,j}}{h \cdot h_B} + \frac{u(x_B, y_B)}{h_B(h+h_B)}\right) + \mathcal{O}(h)
\end{aligned}$$

Addition completes the modified 5-point stencil:

$$u_{xx}\Big|_{i,j} + u_{yy}\Big|_{i,j} = \Delta_h^{(m)} u\Big|_{i,j} + \mathcal{O}(h)$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

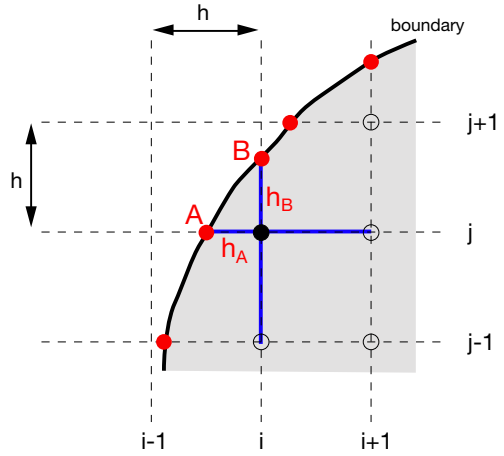


Figure 14: Curvilinear boundary and modified 5-point stencil.

For the difference equation at (x_i, y_j) we get

$$\alpha_{i,j}U_{i,j} + \alpha_{i+1,j}U_{i+1,j} + \alpha_{i,j-1}U_{i,j-1} + \alpha_A U_A + \alpha_B U_B = h^2 f(x_i, y_j) \qquad (4)$$

with $\alpha_{i,j}, \ldots, \alpha_B$ chosen according to the equations above. U_A, U_B are the approximations at the additional points A, B.

This scheme is called Shortley-Weller scheme. The order of consistency is only linear. In case of $h = h_A = h_B$ we get the usual 5-point stencil which is consistent of order 2.

Remark

This example shows that Finite Difference methods run into difficulties in case of more complicated geometries of Ω. □

4.3.4 Formulation of the Sparse Linear System

Each gridpoint (x_i, y_j) at which an approximation U_{ij} of the exact solution $u_{ij} = u(x_i, y_j)$ is required contributes one equation (see eq. (1)-(4) in the previous chap. 4.3.3) to the final linear system

$$A_h U = \tilde{f}_h \,.$$

We approximate the exact solution at all interior points (i.e. Ω_h) and at all boundary points with Neumann condition (i.e. $\Gamma_{N,h}$).

As an example consider $\bar{Q} := \{(x, y) | x \in [0, 5h], y \in [0, 4h]\}$ together with a grid with uniform mesh size h as in fig. 11. Let us assume Dirichlet boundary conditions with $r_{ij} := G(x_i, y_j) = u(x_i, y_j)$ on the boundary $\partial\Omega_h$ and let denote $f_{ij} = f(x_i, y_j)$. Then after ordering the unknowns U_{ij} into the vector $U \in \mathbb{R}^{12}$ in a proper way we obtain the following sparse linear system

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

$$
\begin{bmatrix}
4 & -1 & & & & -1 & & & & & & & & & \\
-1 & 4 & -1 & & & & -1 & & & & & & & & \\
 & -1 & 4 & -1 & & & & -1 & & & & & & & \\
 & & -1 & 4 & & & & & -1 & & & & & & \\
-1 & & & & 4 & -1 & & & & -1 & & & & & \\
 & -1 & & & -1 & 4 & -1 & & & & -1 & & & & \\
 & & -1 & & & -1 & 4 & -1 & & & & -1 & & & \\
 & & & -1 & & & -1 & 4 & & & & & -1 & & \\
 & & & & -1 & & & & 4 & -1 & & & & & \\
 & & & & & -1 & & & -1 & 4 & -1 & & & & \\
 & & & & & & -1 & & & -1 & 4 & -1 & & & \\
 & & & & & & & -1 & & & -1 & 4 & & & \\
 & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & & &
\end{bmatrix}
\begin{bmatrix}
U_{11} \\ U_{21} \\ U_{31} \\ U_{41} \\ U_{12} \\ U_{22} \\ U_{32} \\ U_{42} \\ U_{13} \\ U_{23} \\ U_{33} \\ U_{43}
\end{bmatrix}
= -
\begin{bmatrix}
h^2 f_{11} - r_{10} - r_{01} \\
h^2 f_{21} - r_{20} \\
h^2 f_{31} - r_{30} \\
h^2 f_{41} - r_{40} - r_{51} \\
h^2 f_{12} - r_{02} \\
h^2 f_{22} \\
h^2 f_{32} \\
h^2 f_{42} - r_{52} \\
h^2 f_{13} - r_{14} - r_{03} \\
h^2 f_{23} - r_{24} \\
h^2 f_{33} - r_{34} \\
h^2 f_{43} - r_{44} - r_{53}
\end{bmatrix}
$$

Figure 15: Sparse linear system $A_h U = \tilde{f}_h$ for the discretized Poisson problem with Dirichlet boundary conditions and the domain and grid as in fig. 11.

The structure of the matrix A_h depends on the chosen numbering of the grid points as can be seen in fig. 16 for a square domain with 25 interior points and Dirichlet boundary conditions. A number is assigned to each grid point; the number corresponds to the row of A_h that contains one of the equations (1)-(4) belonging to this grid point.

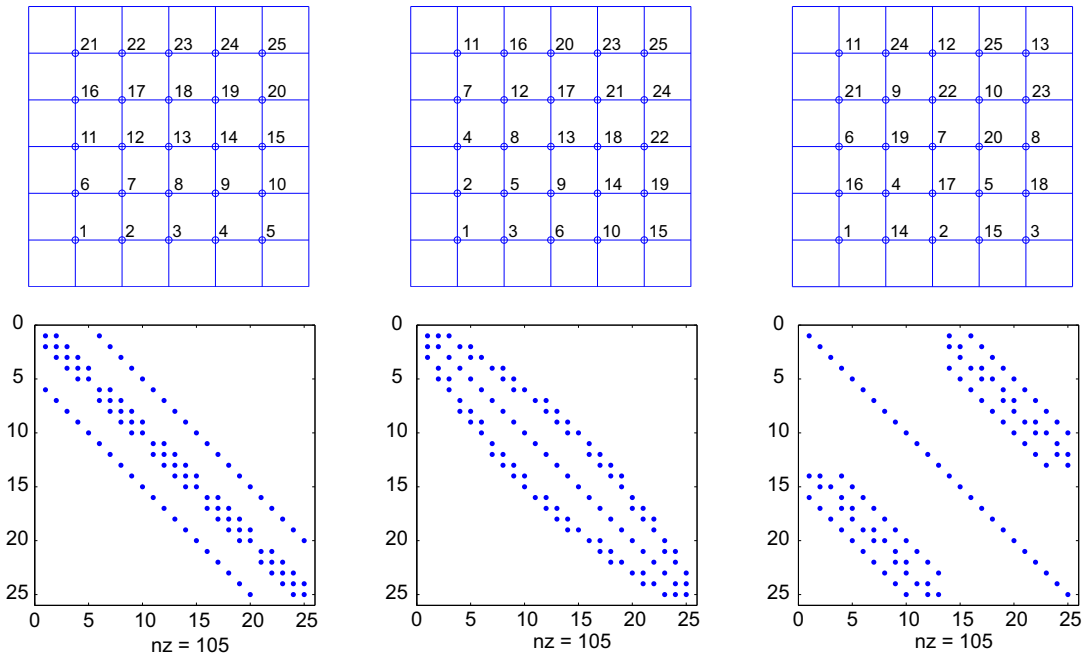


Figure 16: Structure of A_h depending on the numbering of the grid points. In all cases $A_h \in \mathbb{R}^{25\times 25}$ contains $nz = 105$ nonzero elements.

4.3.5 Analysis of the Finite Difference Discretization

We analyze the discretized Poisson problem in a similar way as the one-dimensional model problem in chap. 4.1.3.

We assume that the mesh size h is sufficiently small such that Ω_h is connected if Ω is connected.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

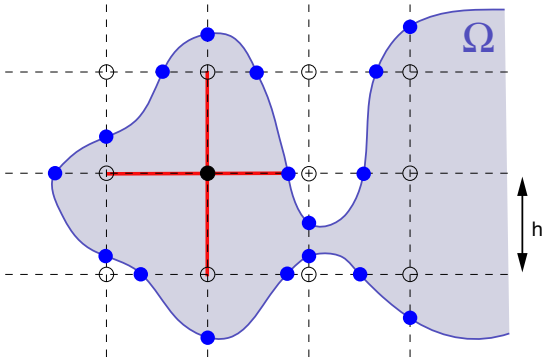


Figure 17: Mesh size h too large: Ω connected, Ω_h not connected.

Again a discretized maximum principle can be formulated:

Lemma (discrete maximum principle)

Consider the Poisson equation $-\Delta u = f$ with $f \leq 0$ in Ω and Dirichlet boundary conditions $u(x,y) = G(x,y)$ on $\partial\Omega$. Let denote Δ_h the discretization on a uniform grid using a 5-point stencil. Let us assume curvilinear boundaries (4). We obtain a sparse linear system $A_h U = \tilde{f}_h$.

If

$$\max_{(x_i,y_j)\in\Omega_h} U_{ij} \geq \max_{(x_i,y_j)\in\partial\Omega_h} G(x_i,y_j)$$

then $U_{ij} = const$ for all $(x_i,y_j) \in \Omega_h$.

Otherwise the discrete maximum is taken on the boundary $\partial\Omega_h$

$$\max_{(x_i,y_j)\in\Omega_h} U_{ij} \leq \max_{(x_i,y_j)\in\partial\Omega_h} G(x_i,y_j)$$

Remark

Analogously to the (non-discretized) Poisson equation in chap. 4.3.2 a discrete minimum principle and a discrete comparison principle can be formulated. The proof is similar to chap. 4.1.3 for the one-dimensional model problem. □

Theorem (uniqueness)

Consider the Poisson equation $-\Delta u = f$ and Dirichlet boundary conditions $u(x,y) = G(x,y)$ on $\partial\Omega$. Let denote Δ_h the discretization on a uniform grid using a 5-point stencil. Let us assume curvilinear boundaries (4). We obtain a sparse linear system $A_h U = \tilde{f}_h$.

Then A_h is non-singular and the sparse linear system is uniquely solvable.

Proof:

Similar to chap. 4.1.3 using the discrete maximum principle. □

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Question

How accurate is the Laplace operator approximated by the 5-point stencil? That is a local property.

Definition (consistency)

The difference scheme Δ_h is consistent with the Laplace operator Δ, if

$$\|\Delta_h v(x_i, y_j) - \Delta v(x_i, y_j)\| \leq \gamma(h) \quad \wedge \quad \lim_{h \to 0} \gamma(h) = 0 \quad \forall (x_i, y_j) \in \Omega_h$$

for all functions $v \in \mathcal{C}^2(\bar{\Omega}, \mathbb{R})$.

The scheme is consistent of order k, if for $v \in \mathcal{C}^{2+k}(\bar{\Omega}, \mathbb{R})$

$$\|\Delta_h v - \Delta v\| = \mathcal{O}(h^k) \quad \forall (x_i, y_j) \in \Omega_h \quad \text{and } h \to 0$$

□

Lemma

The 5-point stencil is consistent of order $k = 2$.

Proof: We directly get that from the Taylor expansion in case of constant h. □

Remark

Again, from consistency we cannot conclude convergence. We need stability in addition. □

Definition (global error)

The global error of the difference method Δ_h at $(x_i, y_j) \in \Omega_h$ is defined as the difference of the true and the approximated result

$$e(x_i, y_j) := U_{ij} - u(x_i, y_j) \quad \text{for } (x_i, y_j) \in \Omega_h$$

□

Remark

We investigate the global error and thus the convergence only on Ω_h (i.e. in the interior of the domain), not on the boundaries. For Dirichlet conditions, that is sufficient. □

Theorem

Consider the Poisson equation $-\Delta u = f$ and Dirichlet boundary conditions $u(x, y) = G(x, y)$ on $\partial\Omega$. Let denote Δ_h the discretization on a uniform grid with mesh size h using a 5-point stencil.

Consistency and stability $\Rightarrow$ convergence

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Proof and explanation:

- We define the (local) consistency error $r(x_i,y_j) := \Delta_h u(x_i,y_j) - \Delta u(x_i,y_j)$ for $(x_i,y_j) \in \Omega_h$ and obtain

$$\begin{aligned} \Delta_h e(x_i,y_j) &= \Delta_h U_{ij} - \Delta_h u(x_i,y_j) = -f(x_i,y_j) - \Delta_h u(x_i,y_j) \\ &= \Delta u(x_i,y_j) - \Delta_h u(x_i,y_j) = -r(x_i,y_j) \end{aligned}$$

Therefore, we get a new and discrete boundary value problem for the error

$$-\Delta_h e = r \quad \forall (x_i,y_j) \in \Omega_h \qquad \wedge \qquad e(x_i,y_j) = 0 \quad \forall (x_i,y_j) \in \partial\Omega_h$$

This is another discretized Poisson equation, which again can be written as a sparse linear system

$$A_h E = R \qquad \text{with} \quad E = \Big(e(x_i,y_j)\Big), R = \Big(r(x_i,y_j)\Big), \quad (x_i,y_j) \in \Omega_h$$

A_h is the same matrix as defined in chap. 4.3.4, E and R are vectors with the components $e(x_i,y_j), r(x_i,y_j)$ ordered in the same way as the components of U and $\tilde{f}_h$ in chap. 4.3.4. The boundary conditions for e already have been inserted.

- In the next step we show the stability of the system $A_h E = R$ using the discrete maximum principle.

We scale the errors

$$\tilde{e} := \frac{e}{\gamma(h)}, \quad \tilde{r} := \frac{r}{\gamma(h)} \quad \text{with} \quad \gamma(h) := \max_{(x_i,y_j)\in\Omega_h} |r(x_i,y_j)|$$

and consider the scaled system

$$-\Delta_h \tilde{e} = \tilde{r} \quad \wedge \quad |\tilde{r}(x_i,y_j)| \le 1 \;\; \forall (x_i,y_j) \in \Omega_h \quad \wedge \quad \tilde{e}(x_i,y_j) = 0 \;\; \forall (x_i,y_j) \in \partial\Omega_h$$

If Ω is bounded we can define a ball (= circle) $B_\varrho(0) := \{(x,y) \in \mathbb{R}^2 \,|\, x^2+y^2 < \varrho^2\}$ that completely contains Ω: $\Omega \subseteq B_\varrho(0)$.

The new defined function

$$w(x,y) := \frac{1}{4}\left(\varrho^2 - x^2 - y^2\right)$$

has the following properties

$$-\Delta w = -\Delta_h w = 1 \;\; \forall (x_i,y_j) \in \Omega_h \quad \wedge \quad w(x_i,y_j) \ge 0 \;\; \forall (x_i,y_j) \in \partial\Omega_h$$

From that we get

$$-\Delta_h \tilde{e} = \tilde{r} \le 1 = -\Delta_h w \;\; \forall (x_i,y_j) \in \Omega_h \quad \wedge \quad 0 = \tilde{e} \le w \;\; \forall (x_i,y_j) \in \partial\Omega_h$$

Now we can use the discrete comparison principle to obtain

$$\tilde{e} \le w \le \max(w) = \frac{\varrho^2}{4} \quad \forall (x_i,y_j) \in \Omega_h$$

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Analogously we perform the steps above for $\tilde{w} := -w$ and get $\tilde{e} \geq -\varrho^2/4$; we combine these two results and obtain (after multiplication with $\gamma(h)$) the stability condition

$$\max_{(x_i,y_j)\in\Omega_h} |\tilde{e}(x_i,y_j)| \leq \frac{\varrho^2}{4} \quad \Rightarrow \quad \max_{(x_i,y_j)\in\Omega_h} |e(x_i,y_j)| \leq \frac{\varrho^2}{4} \max_{(x_i,y_j)\in\Omega_h} |r(x_i,y_j)| = \frac{\varrho^2}{4}\gamma(h)$$

- Because the consistency condition $\lim_{h\to 0}\gamma(h) = 0$ holds, the total error e shows the same behaviour. This is convergence. □

Let us summarize our convergence results from the proof above:

Theorem (convergence of discretized Poisson)

Consider the Poisson equation $-\Delta u = f$ and Dirichlet boundary conditions $u(x,y) = G(x,y)$ on $\partial\Omega$. Let denote Δ_h the discretization using a 5-point stencil.

If $u \in \mathcal{C}^3(\bar{\Omega}, \mathbb{R})$, then the difference scheme converges to the exact solution and

$$\max_{(x_i,y_j)\in\Omega_h} |U_{ij} - u(x_i,y_j)| = \mathcal{O}(h) \quad \text{for } h \to 0$$

For $u \in \mathcal{C}^4(\bar{\Omega}, \mathbb{R})$ and a uniform grid with mesh size h, the difference scheme converges to the exact solution and

$$\max_{(x_i,y_j)\in\Omega_h} |U_{ij} - u(x_i,y_j)| = \mathcal{O}(h^2) \quad \text{for } h \to 0$$

Remarks

- For a uniform grid the discretization error $\gamma(h) = \mathcal{O}(h^2)$.
- With a refined numerical analysis we can show that for $\mathcal{O}(h^2)$-convergence a uniform grid is not necessary. □

4.4 1D Linear Advection Equation

4.4.1 Formulation of the PDE Problem

The linear advection equation (= transport equation) may be used in a model of various phenomena like the movement of pollutant in a river. In its simplest version and in one spatial dimension the linear advection equation is

$$u_t + v u_x = 0\,, \quad u = u(t,x)$$

with time $t > 0$ and space coordinate x. To complete the PDE problem let the initial condition for u be

$$u(0,x) = f(x)$$

For the moment we will ignore any boundary condition.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Remark

If we interpret the above defined PDE as a (partial) model of the transport of a soluble pollutant by a 1D river then $u(t,x)$ is pollutant concentration at time t and position x along the river and v is the (constant) velocity of the river. □

Exact solution of the liner advection equation

It can be easily shown that the exact solution is

$$u(t,x) = f(x - vt)$$

This means that $u(t,x)$ is just the initial concentration profile, $f(x)$, translated by vt along the x-axis. For $v > 0$, the translation is to the right and for $v < 0$, the translation is to the left. In either case the pollution moves downstream at the speed of the river. This model is unrealistic (because it e.g. neglects diffusion)

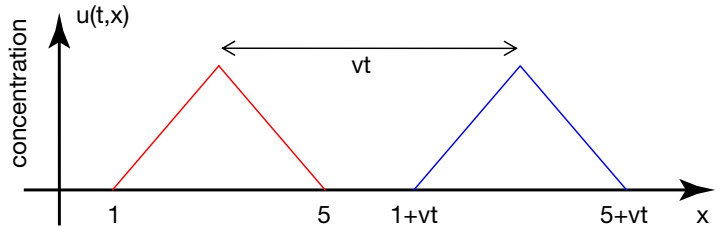


Figure 18: Concentration profile at initial time $t = 0$ (red) and after time t (blue), $v > 0$.

but is useful for learning purposes.

Example (simple explicit FD scheme applied to the 1D advection equation)

Consider an initial condition profile as in fig. 19 and the computational spatial domain $[0, 100]$ uniformly discretized with $x_0 = 0, \ldots, x_{100} = 100$ and $v = 0.5 > 0$. We use first order forward differences in both space and time to obtain the finite difference formulation.

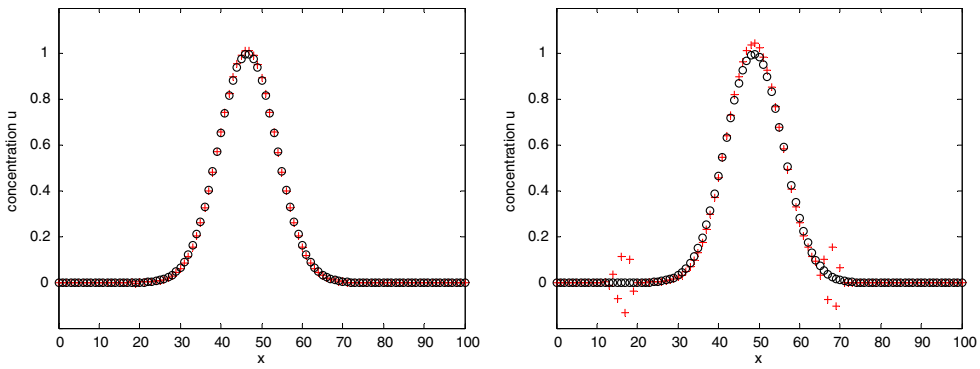


Figure 19: Comparison of numerical (+) and exact solutions (o) to the 1D linear advection equation using $\Delta t = 0.3$, 10 time steps (le.) and 25 time steps (ri.).

The numerical concentration peak has moved to the right place but is higher than the exact solution. More critical there is some noticeable divergence from

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

the exact solution in the numerical solution around $x = 15$ and $x = 67$. Something is wrong with this scheme! □

Question: Can we design other FD schemes to get more accurate results?

4.4.2 Explicit Schemes

In a first approach, we construct some explicit schemes, i.e. the data at the next time level is obtained from an explicit formula involving data from previous time levels only. By that, the solution is obtained row by row in the x-t-mesh.

To obtain such schemes we reformulate the PDE using

$$u_{tt} = \frac{\partial}{\partial t}\left(\frac{\partial}{\partial t}u\right) = \frac{\partial}{\partial t}\left(-v\frac{\partial}{\partial x}u\right) = -vu_{xt} = -vu_{tx} = -v\frac{\partial}{\partial x}\left(\frac{\partial}{\partial t}u\right) = v^2 u_{xx}$$

For fixed x, the Taylor expansion of $u(t+\Delta t, x)$ to order 3 gives,

$$\begin{aligned} u(t+\Delta t,x) &= u(t,x) + \Delta t \cdot u_t(t,x) + \frac{\Delta t^2}{2!}u_{tt}(x,t) + \mathcal{O}(\Delta t^3) \\ &= u(t,x) - v\Delta t \cdot u_x(t,x) + v^2\frac{\Delta t^2}{2!}u_{xx}(x,t) + \mathcal{O}(\Delta t^3) \end{aligned}$$

Now we only need information of the n-th time step to compute the new approximates for the $(n+1)$-th step, because only spatial derivates exist. Using the equivalent operator notation we get

$$L_x(\Delta t) := 1 - v\Delta t\frac{\partial}{\partial x} + v^2\frac{\Delta t^2}{2!}\frac{\partial^2}{\partial x^2}, \quad u(t+\Delta t,x) = L_x(\Delta t)u(t,x) + \mathcal{O}(\Delta t^3)$$

To design FD schemes we simply redefine $L_x(\Delta t)$ by replacing each continuous partial derivative by a finite difference approximation (denoted by δ_x, δ_{xx}).

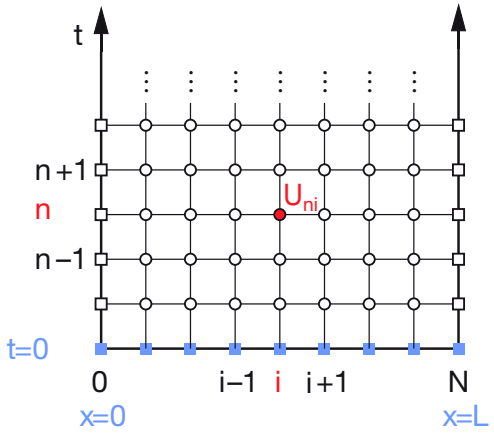


Figure 20: Mesh on a semi-infinite strip used for solution to the 1D linear advection equation. Solid blue squares indicate the location of the (known) initial values. Open squares indicate the location of the (known) boundary values. Open circles indicate the position of the interior points where the FD approximation is computed.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Consistent with the notation in the previous subchapters, $U_{n,i}$ is the approximation to the exact solution $u(t_n, x_i)$ at the n-th time step and the i-th spatial grid point. Some examples of FD schemes are now given.

■ Forward Time Central Space (FTCS) Scheme

$$\begin{aligned} \delta_x U_{n,i} &= \frac{U_{n,i+1} - U_{n,i-1}}{2\Delta x}, \quad \delta_{xx} U_{n,i} = 0 \\ \Rightarrow \quad U_{n+1,i} &= U_{n,i} - \frac{v\Delta t}{2\Delta x}(U_{n,i+1} - U_{n,i-1}) \end{aligned}$$

The scheme is first order in time and second order in space, i.e. it has a truncation error of $\mathcal{O}(\Delta t) + \mathcal{O}(\Delta x^2)$. Ghost values are required at both left and right ends of the computational domain.

■ First Order Upwind (FOU) Scheme (for $v > 0$)

$$\begin{aligned} \delta_x U_{n,i} &= \frac{U_{n,i} - U_{n,i-1}}{\Delta x}, \quad \delta_{xx} U_{n,i} = 0 \\ \Rightarrow \quad U_{n+1,i} &= U_{n,i} - \frac{v\Delta t}{\Delta x}(U_{n,i} - U_{n,i-1}) \end{aligned}$$

The scheme has a truncation error of $\mathcal{O}(\Delta t) + \mathcal{O}(\Delta x)$. A ghost value is required at the left end of the computational domain.

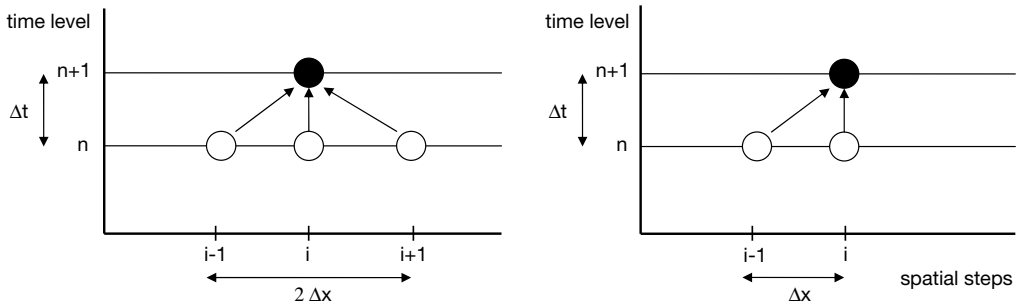


Figure 21: Stencil for the FTCS Scheme (le.) and the FOU Scheme (ri.).

■ Lax-Wendroff Scheme

$$\begin{aligned} \delta_x U_{n,i} &= \frac{U_{n,i+1} - U_{n,i-1}}{2\Delta x}, \quad \delta_{xx} U_{n,i} = \frac{U_{n,i+1} - 2Un,i + U_{n,i-1}}{\Delta x^2} \\ \Rightarrow \quad U_{n+1,i} &= U_{n,i} - \frac{v\Delta t}{2\Delta x}(U_{n,i+1} - U_{n,i-1}) + \frac{v^2\Delta t^2}{2\Delta x^2}(U_{n,i+1} - 2U_{n,i} + U_{n,i-1}) \end{aligned}$$

The scheme is second order in time and second order in space, i.e. the scheme has a truncation error of $\mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2)$. Ghost values are required at both left and right ends of the computational domain.

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

■ Lax-Friedrich Scheme (almost FTCS, $U_{n,i}$ replaced by mean value)

$$\begin{aligned}\delta_x U_{n,i} &= \frac{U_{n,i+1}-U_{n,i-1}}{2\Delta x}, \quad \delta_{xx} U_{n,i} = 0 \\ \Rightarrow \quad U_{n+1,i} &= \frac{U_{n,i+1}+U_{n,i-1}}{2} - \frac{v\Delta t}{2\Delta x}(U_{n,i+1}-U_{n,i-1})\end{aligned}$$

The scheme has a truncation error of $\mathcal{O}(\Delta t)+\mathcal{O}(\Delta x)$. Ghost values are required at both left and right ends of the computational domain. Although this scheme appears to be quite similar to the FTCS scheme its performance is very different.

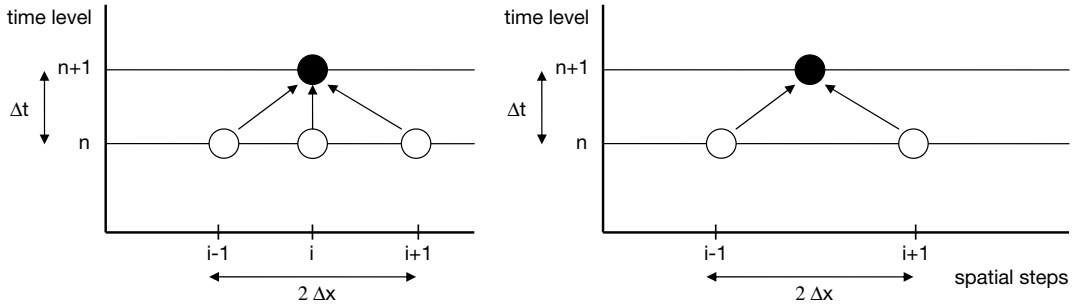


Figure 22: Stencil for the Lax-Wendroff (le.) and the Lax-Friedrich Scheme (ri.).

■ Multi-Level Schemes in General

So far all our schemes have been based on using data at the current time level (n) to advance to the next time level $(n+1)$. This approach can be extended to multi-level schemes by performing Taylor approximations at $t-\Delta t$ in addition:

$$u(t+\Delta t,x)-u(t-\Delta t,x) = 2\Delta t u_t(t,x)+\mathcal{O}(\Delta t^3) = -2v\Delta t u_x(t,x)+\mathcal{O}(\Delta t^3)$$

Dropping the error term, replacing the differential operator by the difference operator and using the usual discrete notation gives the general FD scheme in operator notation

$$U_{n+1,i} = U_{n-1,i} - 2v\Delta t \cdot \delta_x U_{n,i}$$

■ Leap-Frog Scheme (a special multi-level scheme)

$$\begin{aligned}\delta_x U_{n,i} &= \frac{U_{n,i+1}-U_{n,i-1}}{2\Delta x} \\ \Rightarrow \quad U_{n+1,i} &= U_{n-1,i} - \frac{v\Delta t}{\Delta x}(U_{n,i+1}-U_{n,i-1})\end{aligned}$$

The scheme has a truncation error of $\mathcal{O}(\Delta t^2)+\mathcal{O}(\Delta x^2)$. Ghost values are required at both left and right ends of the computational domain. Initial conditions are required at two time levels!

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

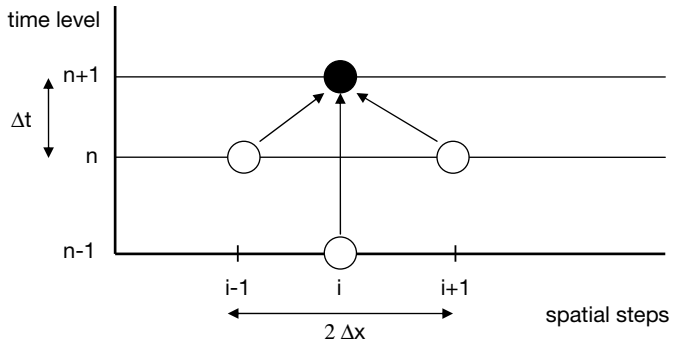


Figure 23: Stencil for the Leap-Frog Scheme.

4.4.3 Repetition: Discrete Fourier Transform (DFT)

We start the discussion of von Neumann stability analysis with a summary of the DFT (see last semester). It is an interpolation scheme well-suited to interpolate a periodic function f with some set of periodic basic functions instead of polynomials.

■ Example

Consider a 2π-periodic function f with $f(x) = f(x+2\pi), x \in \mathbb{R}$. Other periodicities can easily be transformed to 2π-periodicity. We want to interpolate f at the six equidistant nodes $x_k = 2\pi k/n$ for $k = 0, \dots 5$ by an interpolant $T(x)$ which is a linear combination of trigonometric basis functions:

$$T(x) = \sum_{j=-2}^{2} \gamma_j e^{ijx} = \sum_{j=-2}^{2} \gamma_j \left(e^{ix}\right)^j \quad \wedge \quad \gamma_k \text{ from } T(x_k) = f(x_k),\, k = 0, \dots, 4$$

Because of the special structure of the equidistant nodes

$$x_k = \frac{2\pi k}{5} \quad \Rightarrow \quad \left(e^{ix_k}\right)^l = \left(e^{i2\pi k/5}\right)^l = \omega^{kl}, \;\; \omega := e^{i2\pi/5}$$

the interpolation condition $T(x_k) = f(x_k)$ leads to the linear system

$$F\vec{\gamma} := \begin{pmatrix} \omega^0 & \omega^0 & \omega^0 & \omega^0 & \omega^0 \\ \omega^{-2} & \omega^{-1} & \omega^0 & \omega^1 & \omega^2 \\ \omega^{-4} & \omega^{-2} & \omega^0 & \omega^2 & \omega^4 \\ \omega^{-6} & \omega^{-3} & \omega^0 & \omega^3 & \omega^6 \\ \omega^{-8} & \omega^{-4} & \omega^0 & \omega^4 & \omega^8 \end{pmatrix} \cdot \begin{pmatrix} \gamma_{-2} \\ \gamma_{-1} \\ \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} f(x_0) \\ f(x_1) \\ f(x_2) \\ f(x_3) \\ f(x_4) \end{pmatrix} =: \vec{f}$$

We have seen that

$$F^H \cdot F = n \cdot I_n \quad \Rightarrow \quad F^{-1} = \frac{1}{n} F^H, \quad n = 5$$

and thus the matrix $\frac{1}{\sqrt{n}} F$ is a unitary matrix. We have computed the condition number of F: $\|F\|_2 \cdot \|F^{-1}\|_2 = 1$, the matrix is perfectly conditioned! □

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Remark

The solution parameters $\gamma_{-(n-1)/2}, \ldots, \gamma_{+(n-1)/2}$ (here: $n = 5$) are called *discrete Fourier coefficients* of the data stored in $\vec{f}$. □

Definition

Let be $f : [0, 2\pi] \to \mathbb{C}$ *piecewise continuous* (finite number of jumps of finite size in the real or imaginary part). Then the *Fourier series* of f is defined by

$$S_f(x) := \sum_{k=-\infty}^{\infty} c_k e^{ikx} \quad \text{with} \quad c_k := \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx} \, \mathrm{d}x \,, \quad k \in \mathbb{Z} \,.$$

Theorem

Let be $f \in \mathcal{C}_c^1([0, 2\pi], \mathbb{C})$ (function continuous everywhere, its first derivative piecewise continuous).

Then S_f converges uniformly to f. □

Important property!

Notice that γ_k is an approximation to this c_k or – after renaming the index – γ_l approximates c_l:

$$\begin{aligned} F\vec{\gamma} = \vec{f} \;\Rightarrow\; \gamma_l &= (F^{-1}\vec{f})_l = \left(\frac{1}{n} F^H \vec{f}\right)_k \\ &= \frac{1}{n} \sum_{k=0}^{n-1} f(x_k) \omega^{-lk} \;=\; \frac{1}{n} \sum_{k=0}^{n-1} f(x_k) e^{-ilx_k} \end{aligned}$$

With the periodicity condition $f(x_0)e^{-ilx_0} = f(x_n)e^{-ilx_n}$ we rewrite the last sum

$$\begin{aligned} 2\pi\gamma_l &= \frac{2\pi}{n} \left(\frac{1}{2} f(x_0) e^{-ilx_0} + \sum_{k=1}^{n-1} f(x_k) e^{-ilx_k} + \frac{1}{2} f(x_n) e^{-ilx_n} \right) \\ &\approx \int_0^{2\pi} f(x) e^{-ilx} \, \mathrm{d}x = 2\pi c_l \end{aligned}$$

The last sum can be viewed as a composite trapezoid rule approximation of the integral. □

Conclusion: The trigonometric interpolant

$$T_n(x) := \sum_{k=-(n-1)/2}^{(n-1)/2} \gamma_k e^{ikx} \qquad (n = 5 \text{ in our example})$$

is an approximation to the Fourier series obtained by (1) truncating the series, and (2) replacing the integral c_k with its approximation γ_k. □

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

4.4.4 Von Neumann Stability Analysis

The idea of a FD scheme is that $U_{n,i}$ approximates $u(t_n, x_i)$ and the approximation becomes better and better as Δx and Δt become smaller. With increasing mesh refinement round-off errors play an increasing role in the *difference* equation. On the other hand discretization errors are reduced.

Let the pointwise error (also called the 'global error'),

$$e_{n,i} := U_{n,i} - u(t_n, x_i)\,.$$

$e_{n,i}$ contains the accumulated (discretization and rounding) errors resulting from all previous steps and from the possible perturbation in the initial data.

Remark

Without perturbations on time level 0 the values $u(0, x_i)$ are known at all grid points and $U_{0,i}$ is taken to be $u(0, x_i)$ so $e_{0,i} = 0$ at all grid points. As iterations of the FD scheme introduce additional errors, in general $e_{n,i} \neq 0$. It may be that as iterations continue errors are compounded and amplified so that $e_{n,i}$ grows unboundedly making the FD scheme useless. □

Remark

Consistency is a condition on the structure of the formulation of the numerical algorithm.
The discretized PDE is compared with the true PDE and for finer and finer mesh the discretized problem (not the solution!) comes closer and closer to the true problem.

Stability is a condition on the solution of the numerical scheme.
Here the real numerical solution of the FD scheme is investigated and error propagation and amplification are analyzed.

Convergence is a condition on the solution of the numerical scheme.
The real numerical solution is compared to the exact solution of the PDE. □

Basic strategy in stability analysis

We assume that on time level n all errors $e_{n,i}$ are known. Let us investigate the next time step $n \to n+1$. The new errors made in this single step are neglected. Instead we investigate which influence do the accumulated errors $e_{n,i}$ obtained so far have on the results on the next time level: they propagate and cause errors – denoted by $\tilde{e}_{n+1,i}$ – on time level $(n+1)$.

A FD scheme is stable if and only if these propagated errors do not grow unboundedly with time, i.e.

$$|\tilde{e}_{n+1,i}| \leq |e_{n,i}| \qquad (*)$$

The analysis of stability due to von Neumann is based on that property $(*)$.

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

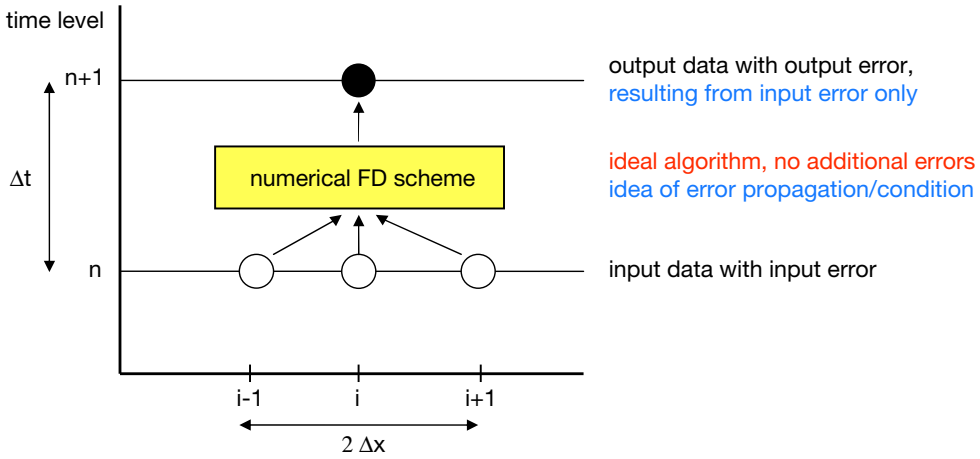


Figure 24: Neumann's stability analysis related to error propagation for 1 timestep.

Realization of that strategy

For a linear and consistent FD scheme and neglecting the (vanishing) truncation errors the exact solution $u(t_n, x_i)$ of the PDE satisfies the same difference scheme as the $U_{n,i}$ and so does the error $e_{n,i}$. Hence the errors do evolve over time in the same way as the numerical solution $U_{n,i}$ does.

■ **Example** (FOU scheme)

$$
\begin{aligned}
& U_{n+1,i} = U_{n,i} - c(U_{n,i} - U_{n,i-1}) \\
\Rightarrow \quad & [u_{n+1,i} + \tilde{e}_{n+1,i}] = [u_{n,i} + e_{n,i}] - c([u_{n,i} + e_{n,i}] - [u_{n,i-1} + e_{n,i-1}]) \\
\Rightarrow \quad & \tilde{e}_{n+1,i} = e_{n,i} - c(e_{n,i} - e_{n,i-1}) + \mathcal{O}(\Delta t) + \mathcal{O}(\Delta x)
\end{aligned}
$$

We neglect the errors in the time step $n \to n+1$ and therefore cancel especially the discretization error. We get

$$\tilde{e}_{n+1,i} = e_{n,i} - c(e_{n,i} - e_{n,i-1})$$

□

Next we will always assume that the boundary conditions are periodic.

The problem of stability for a linear problem with constant coefficients is well understood when the influence of boundaries can be neglected or removed. This is the case either for an infinite domain or for periodic boundary conditions on a finite domain.

In the latter case we consider that the computational domain on the x-axis of length L is repeated periodically and the non-periodic solution $u(t,x)$ on the finite interval $[0, L]$ is transformed into a periodic one.

In case of non-periodic Dirichlet boundary conditions the approach is also possible, because the *error values* are then zero at the boundaries (for details see below) and thus the errors are periodic even if the solution is not.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

■ Example

In our advection example periodicity is no restriction because at the beginning and at the end of a sufficiently long river (the spatial domain) the concentration $u(t,x)$ of the pollutant equals zero! □

We next interpolate the discrete error values $e_{n,i}$ at x_i by trigonometric interpolation to obtain a continuous error function $e_n(x)$ on time level n.

For that we assume a uniform spatial grid $x_0, \dots x_{2N+1}$ (even number of $2N+2$ nodes, $h = x_{j+1} - x_j$) with $U_{n,0} = U_{n,2N+1}$ and $u_{n,0} = u_{n,2N+1}$ and hence $e_{n,0} = e_{n,2N+1}$ because of periodicity.

Rescaling the spatial interval $[x_0, x_{2N+1}]$ to $[0, 2\pi]$ and applying the DFT, we may write,

$$e_n(x) = \sum_{k=-N}^{N} \gamma_{n,k} e^{jkx} \quad \Rightarrow \quad e_{n,i} = e_n(x_i) = \sum_{k=-N}^{N} \gamma_{n,k} e^{jkx_i}, \quad j := \sqrt{-1}$$

and $|\gamma_{n,k}|$ is an approximation of the amplitude of the k-th Fourier component. $e_n(x)$ can be regarded as the sum of $2N+1$ individual harmonic modes.

Analogously we apply the DFT to

$$\tilde{e}_{n+1,i} = \sum_{k=-N}^{N} \tilde{\gamma}_{n+1,k} e^{jkx_i}.$$

We insert the sums into the linear FD scheme for the errors, rearrange the coefficients and obtain

$$\sum_{k=-N}^{N} e^{jkx_i} \cdot \Big(\dots \Big) = 0, \quad i = 0, \dots, 2N$$

The resulting homogeneous linear system $Ax = 0$ has a nonsingular matrix $A = (e^{jkx_i})$. Therefore its solution is unique: The discretized equation (= FD scheme), which is satisfied by the error, must also be satisfied by each individual $\gamma_{n,k}$ and $\tilde{\gamma}_{n+1,k}$ respectively. Therefore, an arbitrary harmonic can be singled out and, when introduced into the scheme, stability requires that no harmonic mode should be allowed to increase in time without bound.

So we replace $e_{n,i}$ by $\gamma_{n,k}$ in $(*)$ to get the von Neumann condition for stability

$$G := \max_{-N \le k \le N} \left| \frac{\tilde{\gamma}_{n+1,k}}{\gamma_{n,k}} \right| \le 1.$$

G is called the amplification factor.

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Remark

For a better understanding, let us suppose that a special example set of errors exists such that only one harmonic mode with index k interpolates all errors at the x_i on time level n

$$e_n(x) = \gamma_{n,k} e^{jkx} \quad \Longleftrightarrow \quad e_{n,i} = \gamma_{n,k} e^{jkx_i}, \;\; i = 0, 1, \ldots, 2N$$

For a linear FD scheme the errors satisfy the same scheme (as we have seen), thus e.g. for the FOU scheme

$$\gamma_{n,k} e^{jkx_i} - c(\gamma_{n,k} e^{jkx_i} - \gamma_{n,k} e^{jk(x_i - \Delta x)} \;=\; \underbrace{(1 - c + ce^{-jk\cdot\Delta x})}_{=:\lambda} \cdot \gamma_{n,k} e^{jkx_i}$$

$$= \;\; \lambda \cdot \gamma_{n,k} e^{jkx_i} \;\; = \;\; \sum_{l=-N}^{N} \tilde{\gamma}_{n+1,l} e^{jlx_i}, \quad i = 0, 1, \ldots, 2N$$

Therefore we have $2N + 1$ conditions for the $2N + 1$ unknowns $\tilde{\gamma}_{n+1,l}$ and uniquely obtain $\tilde{\gamma}_{n+1,k} = \lambda \cdot \gamma_{n,k}$ and $\tilde{\gamma}_{n+1,l} = 0$ for $l \neq k$ (proof by insertion).

The effect of a single step of the numerical scheme in this special case is to multiply each error $\gamma_{n,k} e^{jkx_i}$ by the so-called magnification factor λ. In other words, $\gamma_{n,k} e^{jkx}$ assumes the role of an eigenfunction, with the magnification factor λ being the corresponding eigenvalue, of the linear operator governing each step of the numerical scheme. Continuing, we find that the effect of m further iterations of the scheme is to multiply the exponential by the mth power of the magnification factor:

$$e_{n+m}(x) = \lambda^m e_n(x) \quad \text{for} \quad e_n(x) = \gamma_{n,k} e^{jkx}$$

Thus, the stability of the scheme will be governed by the size of the magnification factor and it ist necessary that $|\lambda| \leq 1$. □

Remarks

Von Neumann stability analysis is applicable without further modifications only for linear PDEs with constant coefficients.

To apply this simple approach to multi-level schemes, additional considerations are necessary.

Stability analysis hasn't been worked out for most FD non-linear schemes, because it heavily relies on the theory of linear difference equations.

If $G \leq 1$ is always satisfied, the scheme is *stable*. That often occurs for implicit schemes.
If $G \leq 1$ can never be satisfied for $\Delta t > 0$ the scheme is *unconditionally unstable*.
Mostly for explicit schemes, $G \leq 1$ establishes a relation between the mesh sizes Δt and Δx. In that case the scheme is called *conditionally stable*. □

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

■ Example

We apply von Neumann stability analysis to the FOU scheme

$$U_{n+1,i} = U_{n,i} - c\cdot(U_{n,i} - U_{n,i-1}) = c\cdot U_{n,i-1} + (1-c)\cdot U_{n,i}\,, \quad c := \frac{v\Delta t}{\Delta x}$$

Step 1: Replace each instance of $U_{n,i}$ in the FD scheme by its corresponding single DFT component.

$$\tilde{\gamma}_{n+1,k} e^{jkx_i} = c\cdot\gamma_{n,k} e^{jkx_{i-1}} + (1-c)\gamma_{n,k} e^{jkx_i}\,, \quad x_{i-1} = x_i - \Delta x$$

Step 2: Rearrange to get G.

Dividing through by $\gamma_{n,k} e^{jkx_i}$ gives,

$$G = \left|\frac{\tilde{\gamma}_{n+1,k}}{\gamma_{n,k}}\right| = \left|(1-c) + c\cdot e^{-jk\Delta x}\right|$$

Step 3: Use the constraint $G \leq 1$ to obtain the condition for Δt (this step could be algebraically tricky).

Using the triangle inequality we estimate,

$$\left|(1-c) + c\cdot e^{-jk\Delta x}\right| \leq |(1-c)| + \left|c\cdot e^{-jk\Delta x}\right| = |1+c| + |c|$$

When $c \in [0,1]$ then $|1-c| = 1-c$ and $|c| = c$, therefore we get,

$$\left|(1-c) + c\cdot e^{-jk\Delta x}\right| \leq (1-c) + c = 1\,.$$

Hence the FOU scheme for the 1D linear advection equation is stable when $0 \leq c \leq 1$ which means that

$$\Delta t \leq \frac{\Delta x}{v}\,.$$

The FOU scheme is said to be conditionally stable. □

Remark

In a similar way we can prove (see (A41)) that for the 1D linear advection equation the FTCS scheme is unconditionally unstable and therefore useless even though it is consistent!

The Lax-Friedrich scheme for the 1D advection equation is conditionally stable (see (A42)): The stability condition is fulfilled if the Courant number $C := v\Delta t/\Delta x$ satisfies $|C| < 1$ (Courant-Friedrichs-Lewy or CFL condition). □

Comment on the CFL condition

It is a fundamental stability condition of most explicit schemes for wave and convection equations and it expresses that the distance covered during the time interval Δt, by the disturbances propagating with speed v, should be lower than the minimum distance Δx between two mesh points. □

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

■ Example

The one-dimensional heat equation $u_t = a u_{xx}$ defined on the spatial interval $[0, L]$ can be discretized by the FTCS scheme as

$$U_{n+1,j} = U_{n,j} + r\left(U_{n,j+1} - 2U_{n,j} + U_{n,j-1}\right), \quad r = \frac{a\,\Delta t}{(\Delta x)^2}$$

Neumann stability analysis shows that

$$r = \frac{a\Delta t}{(\Delta x)^2} \leq \frac{1}{2}$$

is the stability requirement for the FTCS scheme as applied to the one-dimensional heat equation. In contrast to the advection equation, the FTCS scheme is applicable here!

A numerical example:
For copper, $a = 117 \cdot 10^{-6}$ m^2/s. If we choose a thin rod of length 1 m with a spatial resolution of 1 cm, then $\Delta x = 10^{-2}$. Stability restriction gives

$$\Delta t \leq \frac{\Delta x^2}{2a} = 10^{-4}/(2 \cdot 117 \cdot 10^{-6}) \approx 0.5\,[s]\,.$$

For $\Delta x = 10^{-3}\,[m]$ (i.e. 1 mm), we get $\Delta t \approx 0.005\,[s]$. □

4.4.5 Difference Equations

To extend the von Neumann analysis to multi-level schemes, we need some basic knowledge on difference equations. This chapter summarizes *P. Henrici, Elemente der numerischen Analysis 1, chap. 6, BI-HTB Nr. 551, 1964.*

Definition

A linear difference equation of order m with constant coefficients is defined by

$$x_n + a_1 x_{n-1} + a_2 x_{n-2} + \ldots + a_m x_{n-m} = b_n\,, \quad a_i \in \mathbb{C}\,,\, a_m \neq 0$$

with $\{b_n\} \subset \mathbb{C}$ a given sequence; the unknown sequence $\{x_n\} \subset \mathbb{C}$ has to be calculated. If $b_n = 0\ \forall n$, the difference equation is called homogeneous. A difference equation often is called recursion.

We further define $X := \{x_n\}$ and thus $(X)_n = x_n$ and $b := \{b_n\}$.

We introduce the additional sequence $\mathcal{L}X$ by the componentwise definition

$$(\mathcal{L}X)_n := x_n + a_1 x_{n-1} + a_2 x_{n-2} + \ldots + a_m x_{n-m}$$

□

Remark

X can be regarded as a generalization of a vector with countable, but infinitely many components. Multiplication by a scalar $(aX)_n = a(X)_n$ and vector addition $(X+Y)_n = (X)_n + (Y)_n$ are defined componentwise.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Then the operator $\mathcal{L}$ is linear:

$$\mathcal{L}(aX+bY) = a\mathcal{L}X + b\mathcal{L}Y$$

□

(A) Special solutions of homogeneous difference equations of order 2

We start the discussion with the simple case $m = 2$:

$$x_n + a_1 x_{n-1} + a_2 x_{n-2} = 0 \qquad (*)$$

This corresponds to homogeneous linear ODEs $x'' + a_1 x' + a_2 x = 0$, which always has (special) solutions $\exp(rt)$ with r properly chosen.

Motivated by this similarity and replacing the continuous variable t by the discrete variable n, we try to find special solutions of the difference equation with the structure $x_n = e^{r\cdot n} = z^n$, $z := e^r$.

Insertion yields

$$(\mathcal{L}X)_n = z^n + a_1 z^{n-1} + a_2 z^{n-2} = z^{n-2}(z^2 + a_1 z + a_2) = 0$$

For this we get either the trivial solution $z = 0$ or z is a root of the so-called characteristic polynomial $p(z) := z^2 + a_1 z + a_2$.

Theorem

Let be $z_1 \neq z_2$ two different (complex) solutions of the characteristic polynomial $p(z)$, then the two sequences

$$(X^{(1)})_n = z_1^n \quad \text{and} \quad (X^{(2)})_n = z_2^n$$

are solutions of $\mathcal{L}X = \vec{0}$.

If $z_1 = z_2$, then

$$(X^{(1)})_n = z_1^n \quad \text{and} \quad (X^{(2)})_n = n \cdot z_1^{n-1}$$

are solutions of $\mathcal{L}X = \vec{0}$.

Proof:

Case $z_1 \neq z_2$ is clear. For $z_1 = z_2$ we know that not only $p(z_1) = 0$, but also $p'(z_1) = 0$. Differentiation of $(\mathcal{L}X)_n = z^{n-2}p(z)$ yields

$$nz^{n-1} + a_1(n-1)z^{n-2} + a_2(n-2)z^{n-3} = (n-2)z^{n-3}p(z) + z^{n-2}p'(z)$$

For z_1 the right hand side is zero and thus nz_1^{n-1} is a solution of the difference equation $(*)$: With the substitution $x_n = nz_1^{n-1}$ into the left side we get again

$$x_n + a_1 x_{n-1} + a_2 x_{n-2} = 0$$

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

□

Example

$$x_n - 2x_{n-1} + x_{n-2} = 0 \;\Rightarrow\; p(z) = z^2 - 2z + 1 \;\Rightarrow\; z_1 = z_2 = 1$$

We get the solutions $(X^{(1)})_n = 1^n = 1$ and $(X^{(2)})_n = n \cdot 1^{n-1} = n$; insertion proves that they are solutions. □

Lemma

Let be $X^{(1)}, X^{(2)}$ two solutions of $\mathcal{L}X = \vec{0}$ and $c_1, c_2 \in \mathbb{C}$ arbitrary constants.

Then also $X := c_1 X^{(1)} + c_2 X^{(2)}$ is a solution (proof via linearity).

Remark

Situation very similar to linear ODEs. Let e.g. z_1 and $z_2 = \overline{z_1}$ complex roots of $p(z)$. Then $(X^{(1)})_n = z_1^n$ and $(X^{(2)})_n = (\overline{z_1})^n$ are complex solutions of $\mathcal{L}X = \vec{0}$.

With the corollary we get the real-valued solutions

$$(Y^{(1)})_n = \frac{1}{2}\left(z_1^n + (\overline{z_1})^n\right) = \operatorname{Re} z_1^n, \quad (Y^{(2)})_n = \operatorname{Im} z_1^n$$

By this e.g. from $(X^{(1)})_n = e^{in\varphi}$ we get $(Y^{(1)})_n = \cos(n\varphi)$ and $(Y^{(2)})_n = \sin(n\varphi)$.

Lemma

Let be b_n defined for $n \geq n_0$ and $N \geq n_0 + 1$.

Then the difference equation $\mathcal{L}X = b$ has exactly one solution, which takes preset values (vorgegebene Werte) for x_{N-1} and x_N.

Proof: (by contradiction)

Assume that $X^{(1)}$ and $X^{(2)}$ are 2 different solutions with $(X^{(1)})_{N-1} = (X^{(2)})_{N-1}$ and $(X^{(1)})_N = (X^{(2)})_N$.
Then the sequence $D = \{d_n\} := X^{(1)} - X^{(2)}$ solves $\mathcal{L}X = \vec{0}$ and $d_{N-1} = d_N = 0$.
$X^{(1)} \neq X^{(2)}$, so there exists (w.l.o.g.) a minimum index $n^* > N \;\ni\; d_{n^*} \neq 0$ (in case $n^* < N - 1$ we proceed analogously).
D is solution, i.e. $d_{n^*} + a_1 d_{n^*-1} + a_2 d_{n^*-2} = 0$ with $d_{n^*-1} = d_{n^*-2} = 0, d_{n^*} \neq 0$ ↯
□

Remark

From the lemma above we get: If X is solution of $\mathcal{L}X = \vec{0}$ with $x_{N-1} = x_N = 0$ (there exist two successive elements which are zero), then $X = \vec{0}$. □

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Theorem

Let be $X^{(1)}$ and $X^{(2)}$ two solutions of $\mathcal{L}X = \vec{0}$.

Then every solution X of $\mathcal{L}X = \vec{0}$ can be uniquely written as $X = c_1X^{(1)} + c_2X^{(2)}$ $\Longleftrightarrow$ the Wronski determinant

$$w_n := \begin{vmatrix} x_n^{(1)} & x_n^{(2)} \\ x_{n-1}^{(1)} & x_{n-1}^{(2)} \end{vmatrix}$$

is non-zero for at least one $n \in \mathbb{Z}$.

Proof:

$X = c_1X^{(1)} + c_2X^{(2)} \Rightarrow$ the system

$$\left\{ \begin{array}{rcrcl} c_1x_n^{(1)} & + & c_2x_n^{(2)} & = & x_n \\ c_1x_{n-1}^{(1)} & + & c_2x_{n-1}^{(2)} & = & x_{n-1} \end{array} \right\}$$

for the determination of (c_1, c_2) is uniquely solvable $\Leftrightarrow w_n \neq 0$.

"$\Leftarrow$": Lemma 1 $\rightarrow X$ and $c_1X^{(1)} + c_2X^{(2)}$ are solutions; Lemma 2 $\rightarrow$ because of the condition in brackets both solutions are identical. $\square$

Remark

We are now able to solve initial value problems for the difference equation $\mathcal{L}X = \vec{0}$. We have to find two special solutions $X^{(1)}, X^{(2)}$ with non-vanishing Wronski determinant. $c1, c_2$ are determined by the inital condition.

Possible initial conditions are $x_{-1} = 1 \wedge x_0 = 1$ (two components of the solution sequence X are given). $\square$

Example

Let the characteristic polynomial p of $\mathcal{L}X = \vec{0}$ have two different roots z_1, z_2. We determine the Wronski determinant of the corresponding special solutions $x_n^{(1)} = z_1^n$ and $x_n^{(2)} = z_2^n$

$$w_n := \begin{vmatrix} z_1^n & z_2^n \\ z_1^{n-1} & z_2^{n-1} \end{vmatrix} = (z_1z_2)^{n-1}(z_1 - z_2) \neq 0 \quad \forall n$$

$\Rightarrow$ the general solution of the difference equation is given by $c_1z_1^n + c_2z_2^n$. $\square$

Definition

Two sequences $X^{(1)}, X^{(2)}$ – which are not necessarily solutions of $\mathcal{L}X = \vec{0}$ – are called linearly dependent, if $\exists c_1, c_2 \in \mathbb{C} \ni c_1X^{(1)} + c_2X^{(2)} = \vec{0}$.

Otherwise they are linearly independent.

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Theorem

Let be $X^{(1)}$ and $X^{(2)}$ two solutions of $\mathcal{L}X = \vec{0}$ and $W = \{w_n\}$ the sequence of the Wronski determinants.

(1) If $X^{(1)}$ and $X^{(2)}$ linearly dependent, then $W = \vec{0}$, i.e. $w_n = 0 \ \forall n$.

(2) If $w_N = 0$ for at least one N, then $X^{(1)}$ and $X^{(2)}$ are linearly dependent.

(3) If $w_N \neq 0$ for one N, then $w_n \neq 0 \ \forall n$.

(B) Inhomogeneous difference equations of order 2

We consider the system $\mathcal{L}X = b$ or

$$x_n + a_1 x_{n-1} + a_2 x_{n-2} = b_n \qquad (**)$$

As in the homogeneous case, the results are similar to those for linear ODEs.

Theorem

Let be $X^{(1)}$ and $X^{(2)}$ two linearly independent solutions of $\mathcal{L}X = \vec{0}$ and Y a special ("partikuläre") solution of $\mathcal{L}Y = b$.

Then every solution X of $\mathcal{L}X = b$ can be written as

$$X = Y + c_1 X^{(1)} + c_2 X^{(2)}$$

with $c_1, c_2 \in \mathbb{C}$ properly chosen.

Proof:

For $Y = \{y_n\}$ we get by assumption: $y_n + a_1 y_{n-1} + a_2 y_{n-2} = b_n$.
Let be X an arbitrary solution of $(**)$ and let us define the difference $D := X - Y$. For that we obtain

$$d_n + a_1 d_{n-1} + a_2 d_{n-2} = 0$$

and thus D solves the homogeneous difference equation. Because of the linear independence, D can be written as $D = c_1 X^{(1)} + c_2 X^{(2)}$. With $X = Y + D$ we get the claim. □

Remark

The inhomogeneous problem is reduced to the determination of a special solution Y.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

As in the linear ODE case, we can either find a proper ansatz heuristically (e.g. if the b_n are polynomials, try for x_n polynomials of the same degree and determine the free constants) or use the general method of the variation of parameters ("Variation der Konstanten"). □

Theorem (variation of parameters)

Let be $X^{(1)}$ and $X^{(2)}$ two linearly independent solutions of $\mathcal{L}X = \vec{0}$ and $W = \{w_n\}$ the sequence of their Wronski determinants.

Then a special solution of $\mathcal{L}X = b$ is obtained from

$$x_n = \sum_{i=0}^{n} \begin{vmatrix} x_n^{(1)} & x_n^{(2)} \\ x_{i-1}^{(1)} & x_{i-1}^{(2)} \end{vmatrix} \cdot \frac{b_i}{w_i}$$

for $n \geq 0$.

Proof: by insertion. □

Example

Determine a special solution of

$$x_n - 2x_{n-1} + x_{n-2} = 1$$

We already treated the homogeneous problem in a previous example:

$$x_n - 2x_{n-1} + x_{n-2} = 1 \;\Rightarrow\; p(z) = z^2 - 2z + 1 \;\Rightarrow\; z_1 = z_2 = 1$$

and obtained the linear independent solutions $(X^{(1)})_n = 1^n = 1$ and $(X^{(2)})_n = n \cdot 1^{n-1} = n$ for the double root.

For the Wronski determinants in case of a double root we get

$$w_n = \begin{vmatrix} z_1^n & n z_1^{n-1} \\ z_1^{n-1} & (n-1) z_1^{n-2} \end{vmatrix} = -z_1^{2n-2} = -1$$

Using the formula from the theorem on the variation of parameters and inserting $b_n = 1 \;\; (n = 0, 1, 2, \ldots)$ we obtain

$$x_n = -\sum_{i=0}^{n} \begin{vmatrix} 1 & n \\ 1 & i-1 \end{vmatrix} = \sum_{i=1}^{n} (n+1+i) \overset{k:=n+1-i}{=} \sum_{k=1}^{n+1} k = \frac{(n+1)(n+2)}{2}$$

□

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

(C) Generalization to difference equations of order m

- Structure: $(\mathcal{L}X)_n = x_n + a_1 x_{n-1} + \ldots + a_m x_{n-m}$
- Every linear combination of two solutions of $\mathcal{L}X = \vec{0}$ again is a solution.
- Let be b_n defined for $n \geq n_0$ and $N \geq n_0 + m$. The difference equation $\mathcal{L}X = b$ has exactly one solution X which takes preset values for $x_N, x_{N-1}, \ldots, x_{N-m+1}$.

  Again $X \equiv \vec{0}$, if $x_n = 0$ for m successive values.
- Let be $X^{(1)}, \ldots, X^{(m)}$ solutions of $\mathcal{L}X = \vec{0}$. Then the Wronski determinant w_n is defined by

$$w_n := \begin{vmatrix} x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(m)} \\ x_{n-1}^{(1)} & x_{n-1}^{(2)} & \cdots & x_{n-1}^{(m)} \\ \vdots & \vdots & & \vdots \\ x_{n-m+1}^{(1)} & x_{n-m+1}^{(2)} & \cdots & x_{n-m+1}^{(m)} \end{vmatrix}$$

- Let be $X^{(1)}, \ldots, X^{(m)}$ solutions of $\mathcal{L}X = \vec{0}$ and w_n the Wronski determinant.

  If the $X^{(1)}, \ldots, X^{(m)}$ are linearly dependent, then $W = (w_n) = \vec{0}$.

  If one element of W is equal zero, then the $X^{(1)}, \ldots, X^{(m)}$ are linearly dependent.
- Let be $X^{(1)}, \ldots, X^{(m)}$ linearly independent solutions of $\mathcal{L}X = \vec{0}$ and Y a special ("partikuläre") solution of $\mathcal{L}Y = b$.

  Then every solution X of $\mathcal{L}X = b$ can be written as

$$X = Y + c_1 X^{(1)} + \ldots + c_m X^{(m)}$$

  with $c_1, \ldots, c_m \in \mathbb{C}$ properly chosen.
- A special solution Y again can be obtained via the strategy "variation of parameters".

  Let be $X^{(1)}, \ldots, X^{(m)}$ linearly independent solutions of $\mathcal{L}X = \vec{0}$ and $W = \{w_n\}$ the sequence of their Wronski determinants. Let be $B = (b_n)$ well-defined for $n \geq 0$.

  Then a special solution of $\mathcal{L}X = b$ is obtained from

$$x_n = \sum_{i=0}^{n} \begin{vmatrix} x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(m)} \\ x_{i-1}^{(1)} & x_{i-1}^{(2)} & \cdots & x_{i-1}^{(m)} \\ \vdots & \vdots & & \vdots \\ x_{i-m+1}^{(1)} & x_{i-m+1}^{(2)} & \cdots & x_{i-m+1}^{(m)} \end{vmatrix} \cdot \frac{b_i}{w_i}, \quad n \geq 0$$

The final question now is: How to obtain a set $\{X^{(1)}, \ldots, X^{(m)}\}$ of linearly independent solutions? The approach is similar to the case $m = 2$ and leads to the following theorem.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Theorem

Given is the linear homogeneous difference equation of order m

$$x_n + a_1 x_{n-1} + \ldots + a_m x_{n-m} = 0$$

Let be $p(z) = z^m + a_1 z^{m-1} + \ldots + a_{m-1} z^1 + a_m$ the characteristic polynomial of $\mathcal{L}X = \vec{0}$ and let be $z_1, \ldots, z_k \;\; (k \leq m)$ the k different roots of $p(z)$ with the multiplicity $l_i + 1$ ($\sum l_i = m - k$) of the i-th root z_i.

Then the m sequences $X^{(j)}$ with

$$x_n \, (= x_n^{(j)}) = \left\{ \begin{array}{ll} z_i^n & , \;\; p = 0 \\ n(n-1)\cdots(n-p+1) z_i^{n-p} & , \;\; p = 1, \ldots, l_i \end{array} \right\}, \;\; i = 1, 2, \ldots, k,$$

form a full set of m linearly independent solutions – i.e. a basis – of $\mathcal{L}X = \vec{0}$.

Alternatively a second set of m linearly independent sequences can be obtained by linear combinations of the elements of the just defined set. The new and simpler set reads as

$$x_n = n^p z_i^n, \quad p = 0, 1, \ldots, l_i, \;\; i = 1, 2, \ldots, k.$$

Example

Consider the difference equation

$$x_n - 2x_{n-2} + x_{n+4} = 0 \;\; \Rightarrow \;\; p(z) = z^4 - 2z^2 + 1 = (z+1)^2 (z-1)^2$$

with two double roots $z_1 = 1$ and $z_2 = -1$. From the theorem we get

$$x_n^{(1)} = 1, \;\; x_n^{(2)} = n, \;\; x_n^{(3)} = (-1)^n, \;\; x_n^{(4)} = n \cdot (-1)^n$$

□

Example

Assume that z is a fourfold root of $p(z)$. We want to prove that in this case $X = \{n^3 z^n\}$ is a solution of the difference equation (this illustrates the existence of an alternative set in the above theorem).

For that we try to represent X as a linear combination of the original solutions:

$$\begin{array}{rcl} n(n-1)(n-2) & = & n^3 - 3n^2 + 2n \\ n(n-1) & = & n^2 - n \\ n & = & n \end{array}$$

and $n^3 = n(n-1)(n-2) + 3n(n-1) + n$. From that we get

$$n^3 z^n = z^3 [n(n-1)(n-2) z^{n-3}] + 3z^2 [n(n-1) z^{n-2}] + z[n z^{n-1}]$$

This is the required linear combination of solutions of the first set in the above theorem to obtain a solution of the second set. □

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

4.4.6 Von Neumann Stability Analysis Extendend

Test example

Let us again considerate the (explicit) FOU scheme

$$\begin{aligned} U_{n+1,i} &= c\cdot U_{n,i-1} + (1-c)\cdot U_{n,i}\,, \quad \mu := e^{jk\Delta x} \\ \tilde{\gamma}_{n+1,k} &= c\gamma_{n,k}\frac{1}{\mu} + (1-c)\gamma_{n,k} \quad \Rightarrow \\ 0 &= \tilde{\gamma}_{n+1,k} - \left(\frac{c}{\mu} + (1-c)\right)\gamma_{n,k} \end{aligned}$$

This leads to the difference equation

$$x_{n+1} - \left(\frac{c}{\mu} + (1-c)\right) x_n = 0$$

For the characteristic polynomial and its root we obtain

$$p(z) = z - \left(\frac{c}{\mu} + (1-c)\right) \;\Rightarrow\; |z_1| = \left|\frac{c}{\mu} + (1-c)\right| \overset{!}{\leq} 1 \quad \text{for } c \in [0,1]\,,$$

because for those c we get $|1-c| = 1-c$ and $|c| = c$.

We obtain the basis solution $X^{(1)}$ with

$$(X^{(1)})_n = x_n = z_1^n$$

For a unique solution of our problem we still have to add the initial condition:

$$\begin{aligned} x_{n+p} &= \alpha z_1^{n+p}\,, \quad \alpha \in \mathbb{R} \\ p=0: \quad x_n &= \gamma_{n,k} = \alpha z_1^n \\ \Rightarrow\; x_{n+p} &= \tilde{\gamma}_{n+p,k} = \gamma_{n,k}\cdot z_1^p \end{aligned}$$

Starting with $x_n = \gamma_{n,k}$ we get the solution $x_{n+p} = \gamma_{n,k}\cdot z_1^p$ and thus the error component induced by $\gamma_{n,k}$ is damped for increasing p only for $|z_1| < 1$, which is the case for

$$0 < \Delta t < \frac{\Delta x}{v}\,.$$

□

Example (wave equation and implicit multi-level scheme)

Let us consider the wave equation $u_{tt} = a^2 u_{xx}$ on the (normalized) spatial interval $[0, 2\pi]$ with periodic boundary conditions.

We define a uniform spatial grid $0 =: x_0, \ldots x_{2N+1} := 2\pi$ (even number of $2N+2$ nodes, $\Delta x = x_{j+1} - x_j$) with $u_{n,0} = u_{n,2N+1}$ (because of periodicity) and uniform stepsize Δt in time.

For the implicit difference scheme we choose

$$\frac{U_{n+1,i} - 2U_{n,i} + U_{n-1,i}}{\Delta t^2} = a^2 \cdot \frac{U_{n+1,i+1} - 2U_{n+1,i} + U_{n+1,i-1}}{\Delta x^2}$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

and assume that $U_{n,0} = U_{n,2N+1}$ too. Hence for the error terms $e_{n,i} = U_{n,i} - u_{n,i}$ we obtain the same difference scheme because of linearity and we get $e_{n,0} = e_{n,2N+1}$.

We apply the DFT to interpolate the error values on the n-th level and write

$$e_n(x) = \sum_{k=-N}^{N} \gamma_{n,k} e^{jkx} \quad \Rightarrow \quad e_{n,i} = e_n(x_i) = \sum_{k=-N}^{N} \gamma_{n,k} e^{jkx_i}, \quad j := \sqrt{-1}$$

$|\gamma_{n,k}|$ is an approximation of the amplitude of the k-th Fourier component. $e_n(x)$ can be regarded as the sum of $2N+1$ individual harmonic modes.

We are now interested in the influence of errors on the levels $n-1$ and n on the error on level $n+1$ (error propagation). Due to the error interpolation by DFT all existing errors on the levels $n-1$ and n can affect the new error. Additional errors produced in the current step are neglected (as usual). It is required that all the amplitudes $|\gamma_{n,k}|$ of the single error modes are damped in the following time steps. We will write $\gamma_{n+1,k}$ instead of more precisely $\tilde{\gamma}_{n+1,k}$.

Insertion of the error terms into the difference scheme with $c := |a| \cdot \Delta t/\Delta x$ and $\mu_k := e^{jk\Delta x}$ yields

$$\sum_{k=-N}^{N} e^{jkx_i} \cdot \left(\left[\gamma_{n+1,k} - 2\gamma_{n,k} + \gamma_{n-1,k} \right] - c^2 \left[\gamma_{n+1,k} \cdot \mu_k - 2\gamma_{n+1,k} + \frac{\gamma_{n+1,k}}{\mu_k} \right] \right) = 0$$

for $i = 0, \ldots, 2N$. This again is a homogeneous linear system with zero as the unique solution; therefore the content in the brackets has to vanish and we get the recursion

$$\gamma_{n+1,k} \left(1 - c^2\mu_k + 2c^2 - \frac{c^2}{\mu_k} \right) - 2\gamma_{n,k} + 1 \cdot \gamma_{n-1,k} = 0$$

With

$$\mu_k - 2 + \frac{1}{\mu_k} = (-4) \cdot \left(\frac{e^{jk\Delta x/2} - e^{-jk\Delta x/2}}{2i} \right)^2 = -4\sin^2\left(\frac{k\Delta x}{2} \right)$$

and $s_k := \sin\left(\frac{k\Delta x}{2}\right)$ the characteristic polynomial associated to the difference equation reads

$$\left(1 + 4c^2 s_k^2\right) z^2 - 2z + 1 = 0$$

For the roots we get with $|z| = \sqrt{z \cdot \bar{z}}$

$$\begin{aligned} z_{1,2}^{(k)} &= \frac{1}{2(1+4c^2 s_k^2)} \left(2 \pm \sqrt{4 - 4(1+4c^2 s_k^2)} \right) \\ &= \frac{1 \pm \sqrt{-4c^2 s_k^2}}{1+4c^2 s_k^2} = \frac{1 \pm i \cdot 2c \cdot |s_k|}{1+4c^2 s_k^2} \quad \Rightarrow \\ \left| z_{1,2}^{(k)} \right| &= \frac{1}{\sqrt{1+4c^2 s_k^2}} < 1 \quad \forall c \end{aligned}$$

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Because the basic solution is $x_n = \left(z_i^{(k)}\right)^n$ for single roots, in that case the scheme is stable.

Double roots only exist for $s_k = 0$, which is impossible for $k \neq 0$, because

$$\Delta x = \frac{2\pi}{N+1} \quad \text{and} \quad k \in \{-N, \ldots, N\} \quad \text{and} \quad s_k := \sin\left(\frac{k\Delta x}{2}\right).$$

For stability considerations, we can stop here. If in addition we are interested in the unique solution compatible with the initial conditions, we make the ansatz of a linear combination of the basis solutions and insert the initial conditions

$$\begin{aligned} & & x_{n+p} &= \alpha\left(z_1^{(k)}\right)^{n+p} + \beta\left(z_2^{(k)}\right)^{n+p} \\ p=-1: & \quad \gamma_{n-1,k} = x_{n-1} &&= \alpha\left(z_1^{(k)}\right)^{n-1} + \beta\left(z_2^{(k)}\right)^{n-1} \\ p=0: & \quad \gamma_{n,k} = x_n &&= \alpha\left(z_1^{(k)}\right)^{n} + \beta\left(z_2^{(k)}\right)^{n} \end{aligned}$$

From these two equations we determine the unknown constants α and β. □

Summary

Von Neumann stability analysis can be extended to multi-level schemes, as can be seen in the example(s) above. Here difference equations for the amplitudes $\gamma_{n,k}$ of the error modes are formulated and the zeros $z_i^{(k)}$ of the associated characteristic polynomials are calculated. The linearly independent sequences that solve these difference equations should to be damped.

In case of single roots, the finite difference scheme is stable $\Leftrightarrow \max_{i,k}\left|z_i^{(k)}\right| \leq 1$. In case of "$< 1$", all error modes are damped.

A *sufficient* condition for instability is $\max_{i,k}\left|z_i^{(k)}\right| > 1$.

Because the approach is based on the theory of linear difference equations, nonlinear finite difference schemes cannot be analyzed by the von Neumann approach.

4.4.7 Implicit Schemes – Crank-Nicolson Scheme

FD schemes are called explicit if data at the next time level is obtained from an explicit formula involving data from previous time levels only. This normally leads to a (stability) restriction on the maximum allowable time step, Δt.

In implicit schemes data from the next time level occurs on both sides of the difference scheme that necessitates solving a system of linear equations. There is no stability restriction on the maximum time step Δt which may be much larger than in an explicit scheme for the same problem, as we have seen for the above example of the wave equation.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

We go back to the advection equation and choose

$$\delta_x U_{n,i} = \alpha \frac{U_{n,i+1} - U_{n,i-1}}{2\Delta x} + (1-\alpha)\frac{U_{n+1,i+1} - U_{n+1,i-1}}{2\Delta x}, \quad \alpha \in [0,1], \quad \delta_{xx} U_{n,i} = 0$$

This is a weighted average of central difference approximations to spatial derivatives at times levels n and $n+1$.

For $\alpha = 1/2$ we obtain the famous Crank-Nicolson scheme:

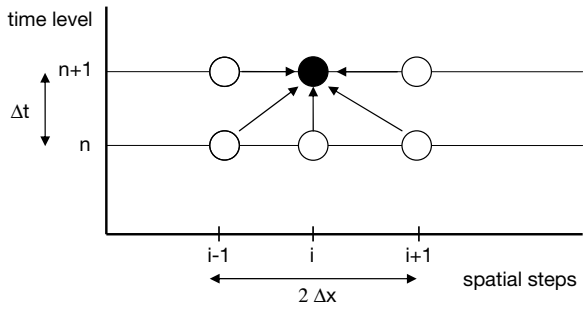


Figure 25: Stencil for the Crank-Nicolson Scheme.

$$U_{n+1,i} = U_{n,i} - \frac{c}{2}\left(\frac{U_{n,i+1} - U_{n,i-1}}{2} + \frac{U_{n+1,i+1} - U_{n+1,i-1}}{2}\right), \quad c = \frac{v\Delta t}{\Delta x}$$

The scheme has a truncation error of $\mathcal{O}(\Delta t) + \mathcal{O}(\Delta x^2)$. Ghost values are required at both left and right ends of the computational domain.

The scheme is implicit so values at time level $n+1$ are found by solving a tridiagonal system of linear equations: Rearranging so that data from the same time level is on the same side gives,

$$-cU_{n+1,i-1} + 4U_{n+1,i} + cU_{n+1,i+1} = cU_{n,i-1} + 4U_{n,i} - cU_{n,i+1} =: d_{n,i}$$

The definition of $d_{n,i}$ reflects that the data at time level n is assumed known. $U_{n+1,0}$ and $U_{n+1,N+1}$ on the left hand side are ghost values which may be known directly (or can be calculated in terms of neighbouring values depending on the type of boundary condition given in the problem).

This system is expressed as the matrix equation,

$$A_h U^{(n+1)} = \begin{pmatrix} 4 & c & 0 & & & \cdots & 0 \\ -c & 4 & c & 0 & & \cdots & 0 \\ 0 & -c & 4 & c & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & 0 & 0 \\ 0 & \cdots & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & & 0 & -c & 4 & c \\ 0 & \cdots & & & 0 & -c & 4 \end{pmatrix} \begin{pmatrix} U_{n+1,1} \\ U_{n+1,2} \\ \\ \vdots \\ \\ U_{n+1,N-1} \\ U_{n+1,N} \end{pmatrix} = d^{(n)}$$

This tridiagonal linear system is solved at each time step and the solution updated iteratively.

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

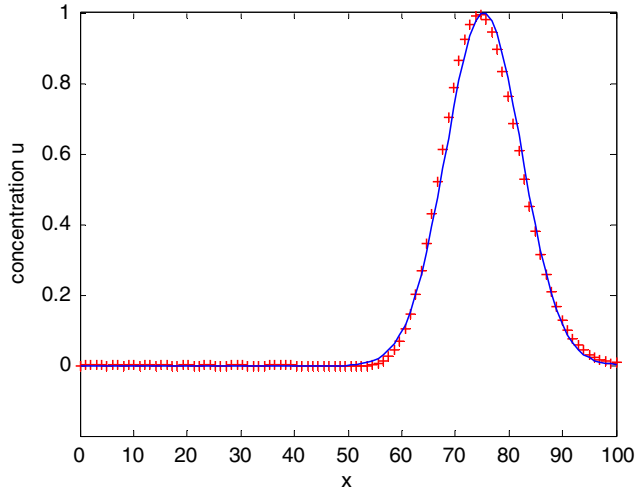


Figure 26: Comparison of numerical (+) and exact solutions (o) to the 1D linear advection equation using Crank-Nicolson scheme with $v = 0.5$, $c = 2.0$, 15 time steps.

Comment on advection and convection

Convection: a flow that combines diffusion and advection.
Diffusion: non-directional molecular transport of mass, heat, or momentum.
Advection: directional bulk transport of mass, heat, or momentum. □

■ Example

The advection-diffusion equation belongs to the class of parabolic PDEs. In 1 spatial dimension it is,

$$u_t + v u_x = K_x u_{xx}\,, \quad u = u(t,x)\,.$$

K_x is called the diffusion coefficient (in the x direction). If $K_x = 0$ then we get again the linear advection equation which we studied so far.

Using our previous interpretation of the linear advection equation in which $u = u(t,x)$ is a river pollutant concentration and v is the speed of the flow, we now get a more realistic description of pollutant transport. Not only does the initial pollutant move downstream with velocity v, the pollutant also diffuses into the surrounding water at rate K_x (the presence of second order spatial derivatives often indicates a diffusive process).

4.4.8 Matrix Stability Analysis

The linear FD scheme can be rewritten as a linear difference equation

$$AU^{(n+1)} = BU^{(n)} \quad \text{with} \quad A, B \in \mathbb{R}^{N\times N},\, U^{(n)} = (U_{n,1}, \ldots, U_{n,N})^T$$

For a consistent scheme and neglecting the (vanishing) truncation error the exact solution of the PDE satisfies the same scheme and so does the error vector

$$Au^{(n+1)} = Bu^{(n)} \quad \Rightarrow \quad Ae^{(n+1)} = Be^{(n)},\ e^{(n)} := U^{(n)} - u^{(n)}$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

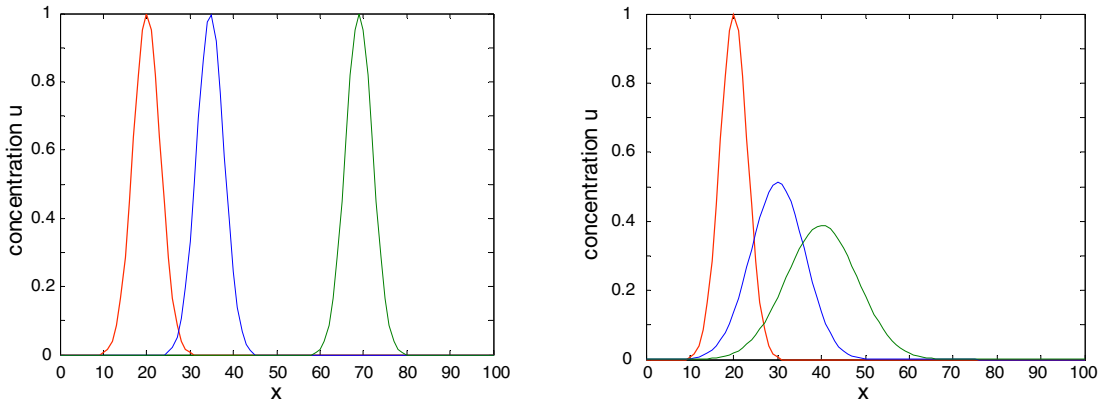


Figure 27: Time evolution of the exact solutions for pure advection (le.) and advection-diffusion (ri.). Each of the plots contains the initial profile (red) and two later solutions (blue and green).

From that we get

$$e^{(n+1)} = A^{-1}Be^{(n)} \quad \Rightarrow \quad \left\|e^{(n+1)}\right\| = \left\|A^{-1}Be^{(n)}\right\| \leq \left\|A^{-1}B\right\| \left\|e^{(n)}\right\|$$

Hence the FD scheme is stable if the error is not increasing (same idea as in von Neumann stability analysis) and this is true if

$$\left\|A^{-1}B\right\| \leq 1$$

The matrix norm used here is induced by the vector norm. Often the euclidian norm is used. Hence the stability of a (linear) FD scheme can be investigated by finding the norm of the matrix $A^{-1}B$.

This is the matrix method for stability and may be quite difficult to implement. It should be noted that there are many definitions of norms and a FD scheme may be stable in one norm but not in another.

The sharpest statements we obtain using the spectral radius, but this is possible only if the 2-norm and the spectral radius coincide for the matrix under investigation. □

■ Example

. . .

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

4.5 Multigrid Methods

To define and analyze basic properties of multigrid methods we use the FD approximation of a 1D Dirchlet boundary value problem as a simple example problem.

■ **Example 1** (cf. (A30) – smoothen the error)

We consider the BVP

$$\begin{aligned} -u_{xx} &= f && \text{for } \Omega =]0,1[\\ u(0) &= a,\ u(1) = b && \text{on } \partial\Omega \end{aligned}$$

On $[0,1]$ we define a uniform mesh Ω_h with $N+2$ gridpoints and a mesh size $h := 1/(N+1)$

$$\Omega_h := \{x_j \,|\, x_j = j \cdot h,\, j = 0,\ldots,N+1\}$$

For the $\mathcal{O}(h^2)$-approximation of u_{xx} we choose

$$u_{xx}(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + \mathcal{O}(h^2)$$

and we need $u \in \mathcal{C}^4([0,1],\mathbb{R})$ at minimum! With $f_j := f(x_j)$ and the numerical approximation $U_{j,h}$ of $u(x_j)$ we obtain the following tridiagonal system $A_h U_h = f_h$ of N equations for the calculation of the approximate solutions $U_{1,h},\ldots U_{N,h}$

$$\underbrace{-\begin{pmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & -2 & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & -2 \end{pmatrix}}_{A_h} \cdot \underbrace{\begin{pmatrix} U_{1,h} \\ U_{2,h} \\ U_{3,h} \\ \vdots \\ U_{N,h} \end{pmatrix}}_{U_h} = h^2 \cdot \underbrace{\begin{pmatrix} f_1 + a/h^2 \\ f_2 \\ f_3 \\ \vdots \\ f_N + b/h^2 \end{pmatrix}}_{f_h}$$

We denote the sparse matrix by A_h, because even if the h appears not directly in A_h, its size is affected by the choice of h!

We want to solve this problem iteratively by the damped Jacobi method

$$U_h^{(\nu+1)} = M_J(\omega,h) U_h^{(\nu)} + \omega D_h^{-1} f_h \quad \text{with} \quad M_J(\omega,h) := (I - \omega D^{-1} A_h),\ \ \omega \in]0,1]$$

Consider the eigenvalue problem for the damped Jacobi iteration

$$M_J(\omega,h) v_h^{(k)} = \lambda_h^{(k)} v_h^{(k)}$$

For the EWs $\lambda_h^{(k)}(\omega)$ and the EVs $v_h^{(k)}$ we get

$$\begin{aligned} \lambda_h^{(k)}(\omega) &= 1 - \omega \cdot \big(1 - \cos(k\pi h)\big), \\ v_h^{(k)} &= \big(\sin(k\pi h j)\big)_{j=1,N} = \begin{pmatrix} \sin(k\pi \cdot 1 \cdot h) \\ \vdots \\ \sin(k\pi \cdot N \cdot h) \end{pmatrix}, \quad k = 1,\ldots,N. \end{aligned}$$

$v_h^{(k)}$ can be regarded as the vector of function values obtained by evaluating a smooth function $g_k(x) := \sin(k\pi \cdot x)$ at the interior grid points $j \cdot h,\, j = 1,\ldots,N$.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Does that iterative solver converge to a fixed-point, i.e. $U_h^{(\nu)} \to U_h^*$ for $\nu \to \infty$?

For the spectral radius, we have that $\varrho(M_J(\omega,h)) < 1 \quad \forall \omega \in]0,1]$, because

$$\varrho(M_J(\omega,h)) = \max_{k,\omega} |\lambda_h^{(k)}(\omega)|$$

and thus convergence is guaranteed.

We see: Convergence depends on the method chosen (here: damped Jacobi) as well as – because of the structure and EWs of the matrix A_h – on the problem this method is applied to (here: 1D Laplace).

Please mind that due to the approximation of the derivatives by FD in the differential equation, U_h^* is not equal to the exact solution u at the grid points!

Which spectral radius do we get e.g. for $\omega = 1$ (classical Jacobi method)?

For $\omega = 1$ the spectral radius is for $h \to 0$

$$\varrho(M_J(1,h)) = \max_{k=1,\dots,N} |1 - 1 + \cos(k\pi h)| = \cos(\pi h) = 1 - \frac{1}{2}\pi h^2 + \mathcal{O}(h^4)$$

The convergence of our iterative solver deteriorates as $h \to 0$, because $\varrho \to 1$. The finer the grid is, the worse are the guaranteed convergence properties.

What do we obtain from a detailed analysis of the iteration error?

Let denote $\varepsilon_h^{(\nu)} := U_h^{(\nu)} - U_h^*$ the error after the ν-th iteration cycle. The EVs form a complete basis of the $\mathbb{R}^N$, therefore we can decompose the initial error at $\nu = 0$

$$\varepsilon_h^{(0)} = U_h^{(0)} - U_h^* = \sum_{k=1}^{N} e_{k,h}^{(0)} \cdot v_h^{(k)}$$

After one iteration cycle we get using the EW-/EV-property

$$\varepsilon_h^{(1)} = M_J(\omega,h)\varepsilon_h^{(0)} = \sum_{k=1}^{N} \lambda_h^{(k)}(\omega) \cdot e_{k,h}^{(0)} \cdot v_h^{(k)}$$

and after m iteration cycles we get (see p.15)

$$\varepsilon_h^{(m)} = M_J(\omega,h)^m \varepsilon_h^{(0)} = \sum_{k=1}^{N} \left(\lambda_h(\omega)^{(k)}\right)^m \cdot e_{k,h}^{(0)} \cdot v_h^{(k)}$$

We see: The smaller the k-th EW $\lambda_h^{(k)}(\omega)$ is, the faster the k-th component $e_{k,h}^{(0)}$ of the error $\varepsilon_h^{(0)}$ is damped. Damping is only weak for EWs close to 1.

How to damp at least one half of the error components efficiently for fine grids?

Let us divide the EWs into two groups: Group 1 contains the $\lambda_h^{(k)}$ with $1 \le k < N/2$ and group 2 contains the $\lambda_h^{(k)}$ with $N/2 \le k \le N$.

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

We want to choose the ω such that the error components which belong to $N/2 \leq k \leq N$ are damped as good as possible:

$$\begin{aligned} \mu &:= \max\left\{|\lambda_h^{(k)}|, N/2 \leq k \leq N\right\} \\ &< \max\{1-\omega, |1-\omega(1-\cos(\pi)|\} = \max\{1-\omega, |1-2\omega|\} \end{aligned}$$

Using this result we find that the optimal $\mu^* = 1/3$ is obtained using $\omega^* = 2/3$.

Each $g_k(x)$ that belongs to group 2 is more rapidly oscillating than each $g_k(x)$ that belongs to group 1. So for our example system by the choice of $\omega^* = 2/3$ we try to damp the so-called high-frequency components belonging to EWs from group 2 as good as possible, whereas the so-called low-frequency components belonging to EWs from group 1 are not included into the optimization procedure.

We also obeserve, that the worst-case situation is obtained for $k = 1$ which belongs to the $g_1(x)$ with the lowest frequency: For $N \gg 1$ we get $\lambda_h^{(1)}(\omega) \approx 1$ and extremely low damping of the respective error component $e_{1,h}^{(0)}$.

On the other hand after a few cycles m we get

$$\left|e_{k,h}^{(m)}\right| < \left(\frac{1}{3}\right)^m \left|e_{k,h}^{(0)}\right| \ll \left|e_{k,h}^{(0)}\right|$$

for all high-frequency components (group 2 with $N/2 \leq k \leq N$). For this reason, although the global error decreases slowly per iteration step, it is smoothed out very quickly – i.e. components which belong to EVs/$g_k(x)$ with high oscillations are damped – and this process does not depend on h! □

■ **Example 2** (two-grid method)

We consider again the BVP

$$\begin{aligned} -u_{xx} &= f && \text{for } \Omega =]0,1[\\ u(0) &= a,\ u(1) = b && \text{on } \partial\Omega \end{aligned}$$

and define two uniform meshes Ω_h (fine grid, $N+1$ gridpoints with N even) and Ω_H (coarse grid) with e.g. $H = 2h$. Using the same FD discretization as in the last example we get two discretized systems

$$A_h U_h = f_h \quad \text{and} \quad A_H U_H = f_H$$

of different dimension ($N-1$ and $(N/2-1$ in our example), but with the same (tridiagonal) structure of A_h and A_H.

Basic idea

The two-grid strategy combines two complementary schemes. The high-frequency components of the error are reduced by applying iterative methods like Jacobi or Gauss-Seidel schemes. For this reason these methods are called smoothers.

On the other hand, the low-frequency error components are effectively reduced by a coarse-grid correction procedure.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Realization of the two-grid idea in 7 steps

(1) Presmoothing steps on the fine grid:

Start with $U_h^{(0)}$ and try to solve the system $A_h U_h = f_h$ iteratively (e.g. by damped Jacobi), stop after m steps

$$U_h^{(0)} \to U_h^{(1)} \to \ldots \to U_h^{(m)}$$

The high-frequency components of the initial error $\varepsilon_h^{(0)} := U_h^{(0)} - U_h^*$ are efficiently damped in $\varepsilon_h^{(m)} := U_h^{(m)} - U_h^*$, i.e. only the smooth(er) functions $g_k(x)$ (see last example) contribute to the error $\varepsilon_h^{(m)}$ significantly.

(2) Calculation of the residual

The exact error $\varepsilon_h^{(m)}$ is the solution of the following equation which is equivalent to the original problem

$$A_h U_h^* = f_h \quad \Leftrightarrow \quad A_h \left(U_h^{(m)} - \varepsilon_h^{(m)} \right) = f_h \quad \Leftrightarrow \quad A_h \varepsilon_h^{(m)} = A_h U_h^{(m)} - f_h =: r_h^{(m)}$$

The residual $r_h^{(m)}$ can be simply calculated.

To analyze the residual, we also calculate the EWs $\mu_h^{(k)}$ and EVs $z_h^{(k)}$ of A_h. For that we use that $D = 2 \cdot I$ for our special matrix A_h in the damped Jacobi iteration:

$$\begin{aligned} M_J(\omega, h) v_h^{(k)} &= \lambda_h^{(k)} v_h^{(k)} \quad \Leftrightarrow \quad D^{-1} A_h v_h^{(k)} = \left(\frac{1 - \lambda_h^{(k)}}{\omega} \right) v_h^{(k)} \\ &\Leftrightarrow \quad A_h v_h^{(k)} = 2 \left(\frac{1 - \lambda_h^{(k)}}{2\omega} \right) v_h^{(k)} = 2(1 - \cos(k\pi h) \cdot v_h^{(k)} \end{aligned}$$

Therefore the EVs $z_h^{(k)} = v_h^{(k)}$ are the same as those of $M_j(\omega, h)$, only the EWs have to be transformed: $\mu_h^{(k)} = 2(1 - \cos(k\pi h)$.

Insertion into the residual gives

$$\begin{aligned} r_h^{(m)} &= A_h U_h^{(m)} - f_h = A_h U_h^* + A_h \varepsilon_h^{(m)} - f_h \\ &= \sum_{k=1}^{N-1} \left(\lambda_h(\omega)^{(k)} \right)^m \cdot e_{k,h}^{(0)} \cdot A_h v_h^{(k)} \\ &= \sum_{k=1}^{N-1} \left(\lambda_h(\omega)^{(k)} \right)^m \cdot \mu_h^{(k)} \cdot e_{k,h}^{(0)} \cdot v_h^{(k)} \end{aligned}$$

Also the residual is smooth(er) with a similar damping of the high-frequency error components ($0 < \mu_h^{(k)} < 4$); by the $\mu_h^{(k)}$ also the low-frequency components are damped.

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

(3) Restriction of the residual

If we inspect

$$A_h \varepsilon_h^{(m)} = r_h^{(m)} \qquad (*)$$

this again can be seen as a linear system that results from the FD approximation of the Poisson equation; here we know in addition that the new right hand side $r_h^{(m)}$ and the unknown solution $e_h^{(m)}$ are rel. smooth, i.e. varying not so rapidly.

That motivates the strategy to solve the Poisson equation for the new right hand side $r_h^{(m)}$ on a coarser grid: That is more efficient and possibly the accuracy is sufficient in that case. Later we have to prove (!) that our idea was good.

In the simplest approach, we cancel every second equation in $(*)$ and by that restrict our residual to the coarse grid with a mesh size $H = 2h$.

For this simple approach the restriction operator I_h^H is

$$I_h^H = \begin{pmatrix} 0 & 1 & 0 & & & & \\ & & 0 & 1 & 0 & & \\ & & & & \cdots & \cdots & \\ & & & & & 0 \quad 1 & 0 \end{pmatrix} \in \mathbb{R}^{(N/2-1)\times(N-1)}$$

and we obtain the $(N/2-1)$-dimensional linear system

$$A_H \varepsilon_H^{(m)} = I_h^H r_h^{(m)} =: r_H^{(m)}, \quad H = 2h$$

The restriction is no inverse operation (I_h^H is not a non-singular square matrix) and so we cannot avoid to loose information.

Here we use a better restriction based on the averaging operator

$$I_h^H = \frac{1}{4} \begin{pmatrix} 1 & 2 & 1 & & & & \\ & & 1 & 2 & 1 & & \\ & & & & \cdots & \cdots & \\ & & & & & 1 \quad 2 & 1 \end{pmatrix} \in \mathbb{R}^{(N/2-1)\times(N-1)}$$

Let be $g_h \in \mathbb{R}^{N-1}$ a vector – e.g. a discrete function on Ω_h –, then we obtain componentwise for the restricted vector $g_H \in \mathbb{R}^{N/2-1}$ – e.g. the (restricted) discrete function on Ω_H – from $g_H = I_h^H \cdot g_h$

$$(g_H)_j = \frac{1}{4}(g_h)_{2j-1} + \frac{2}{4}(g_h)_{2j} + \frac{1}{4}(g_h)_{2j+1}$$

(4) Solution of the coarse grid problem

We solve (directly or by another iterative scheme)

$$A_H \varepsilon_H^{(m)} = r_H^{(m)} \quad \Rightarrow \quad \varepsilon_H^{(m)} = \dots$$

For that solution we again can use a nested two-grid cycle!

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

(5) Coarse-grid correction

Because of smoothness one expects that $\varepsilon_H^{(m)}$ is an approximation to $\varepsilon_h^{(m)}$ on all grid points that Ω_h and Ω_h have in common, i.e. on all grid points $x_j \in \Omega_h \cap \Omega_H$.

To obtain an approximation of $\varepsilon_h^{(m)}$ for all the other grid points of the fine grid which are not grid points of the coarse grid too, we use interpolation. With the prolongation operator I_H^h we get an improved approximation on the fine grid

$$U_h^{(m+1)} = U_h^{(m)} - I_H^h \cdot \varepsilon_H^{(m)}$$

As an example we use linear interpolation with the prolongation operator

$$I_H^h := \frac{1}{2} \begin{pmatrix} 1 & & & \\ 2 & & & \\ 1 & 1 & & \\ & 2 & & \\ & 1 & \vdots & \\ & & \vdots & 1 \\ & & & 2 \\ & & & 1 \end{pmatrix}$$

If again $g_h \in \mathbb{R}^{N-1}$ is a discrete function on Ω_h (= vector) and $g_H \in \mathbb{R}^{N/2-1}$ on Ω_H, then we obtain componentwise from $g_h = I_H^h \cdot g_H$

$$(g_h)_j = \begin{cases} (g_H)_{j/2} & \text{if } j \text{ even} \\ \frac{1}{2}\left((g_H)_{(j-1)/2} + (g_H)_{(j+1)/2}\right) & \text{if } j \text{ odd} \end{cases}$$

(6) Postsmoothing steps on the fine grid:

Start with $U_h^{(m+1)}$ and try to solve the system $A_h U_h = f_h$ iteratively (e.g. by damped Jacobi), stop after $\tilde{m}$ steps

$$U_h^{(m+1)} \to U_h^{(m+2)} \to \ldots \to U_h^{(m+1+\tilde{m})}$$

(7) Loop

Continue with step (1) of the algorithm, if necessary. □

The fine-coarse-fine loop defined in steps (1)-(6) is called v-cycle, because from a fine grid we go down to a coarse grid and back again.

From an eigenvector analysis of the errors (see (A46)) we get the following essential result for the v-cycle in our example problem 2:

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Theorem

Consider the BVP problem

$$\begin{array}{rcll} -u_{xx} & = & f & \text{for } \Omega =]0,1[\\ u(0) & = & a,\ u(1) = b & \text{on } \partial\Omega \end{array}$$

Define the two-grid method (v-cycle) exactly as in example 2 and use the same notation. Choose $m = 2$ and $\omega = 2/3$. Then after the steps (1)-(5) (i.e. without postsmoothing) of one v-cycle we get

$$\|\varepsilon_h^{(m+1)}\|_2 = \|U_h^{(m+1)} - U_h^*\|_2 \leq 0.782 \cdot \|\varepsilon_h^{(0)}\|_2$$

Remarks

Each v-cycle reduces the error at least by a constant factor, and this is true also for $h \to 0$! That is an excellent result.

From example 2 we see that a perfect smoother followed by an exact solution at step 4 would leave no error. In reality, this will not happen.

Fortunately, a careful (but unfortunately not so simple) analysis shows that a v-cycle with good smoothing (better than by the damped Jacobi!) always reduces the error by a constant factor ϱ that is independent of h: A typical and good value is $\varrho = 0.1$; compare this e.g. with $\varrho = .99$ for Jacobi alone. We achieve a convergence factor ϱ that does not move up to 1 as $h \to 0$ and thus achieve a given relative accuracy in a fixed number of cycles. Since each step of each v-cycle requires only $\mathcal{O}(N)$ operations on sparse problems of size N, one loop (1-6) of the two-grid method is an $\mathcal{O}(N)$ algorithm. This does not change in higher dimensions. □

V-Cycles, W-Cycles and Full Multigrid

Clearly multigrid need not stop at two grids. Because of size, a direct solution of the problem in step (4) of example 2 often is not possible or efficient. And for an iterative solution, the lowest frequency is still low on the $H = 2h$ grid, and that part of the error does not decay quickly until we move to $4h$ or $8h$ or ... (or a very coarse $512h$).

The two-grid v-cycle extends in a natural way to more grids. It can go down to coarser grids (e.g. $2h$, $4h$, $8h$) and back up again to ($4h$, $2h$, h). This nested sequence of v-cycles is a V-cycle (capital V, see Fig. 28/le.). Because coarse grid iterations are much faster than fine grid iterations, a detailed mathematical analysis shows that time is well spent on the coarse grids. So the W-cycle that stays coarse longer is generally superior to a V-cycle.

The full multigrid cycle is asymptotically better than V or W. Fig. 28/ri. describes a typical multigrid scheme.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

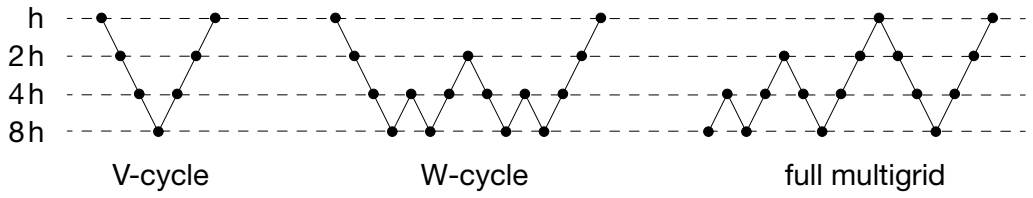


Figure 28: V-cycles, W-cycles and Full Multigrid use several grids several times.

Full multigrid starts on the coarsest grid. The solution on the $8h$ grid is interpolated to provide a good initial vector $U_{4h}^{(0)}$ on the $4h$ grid. A v-cycle between $4h$ and $8h$ improves it. Then interpolation predicts the solution on the $2h$ grid, and a deeper V-cycle makes it better (using $2h$, $4h$, $8h$). Interpolation of that improved solution onto the finest grid gives an excellent start to the last and deepest V-cycle.

The operation counts for a deep V-cycle and for full multigrid are certainly greater than for a two-grid v-cycle, but only by a constant factor. That is because the count is divided by a power of 2 every time we move to a coarser grid. For a differential equation in d space dimensions, we divide by 2^d. The cost of a V-cycle (as deep as we want) is less than a fixed multiple of the v-cycle cost:

$$\text{V-cycle cost} < \left(1 + \frac{1}{2^d} + \left(\frac{1}{2^d}\right)^2 + \dots\right) \cdot \text{v-cycle cost} = \frac{2^d}{2^d - 1} \cdot \text{v-cycle cost}$$

Full multigrid is nothing else than a series of inverted V-cycles, beginning on a very coarse mesh. Because of this and using the estimate just obtained we get

$$\text{full multigrid cost} < \frac{2^d}{2^d - 1} \cdot \text{V-cycle cost} < \left(\frac{2^d}{2^d - 1}\right)^2 \cdot \text{v-cycle cost}$$

The method works excellent in practice, if carefully programmed!

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

5 Finite Elements

The advantages of Finite Differences are that they are easy to implement and rather efficient.

On the other hand, complicated domains Ω are difficult to approximate especially for higher-dimensional domains and FD schemes only have a high order of convergence, if the continuous solution is very smooth (e.g. $u \in \mathcal{C}^4(\Omega, \mathbb{R})$ in case of the Poisson equation).

These problems are overcome by the Finite Element (FE) method.
In this chapter we only consider Dirichlet problems!

5.1 Linear Elliptic PDEs - Classical Solution

Definition (scalar product, Hilbert space)

Let X be a vector space over $\mathbb{R}$. A mapping

$$\langle \cdot, \cdot \rangle : X \times X \to \mathbb{R}$$

is called scalar product (or inner product), if $\forall x, y, z \in X, \forall \alpha \in \mathbb{R}$:

(1) $\langle x, x \rangle \geq 0 \ \wedge \ (x = 0 \Leftrightarrow \langle x, x \rangle = 0)$

(2) $\langle x + \alpha y, z \rangle = \langle x, z \rangle + \alpha \langle y, z \rangle$

(3) $\langle x, y \rangle = \overline{\langle y, x \rangle}$

By the definition $\|x\|_X := \sqrt{\langle x, x \rangle_X}$ a scalar product induces an associated norm.

If X equipped with this special norm is a Banach space ($\to$ completeness, i.e. every Cauchy sequence converges in X), then X is a Hilbert space. □

■ **Example**

$X = \mathbb{R}^n$ with $\langle x, y \rangle_2 := \sum_{i=1}^{n} x_i y_i$ for $x, y \in \mathbb{R}^n$ is a Hilbert space.

$X = \mathbb{Q}$ with $\langle x, y \rangle_2 := x \cdot y$ for $x, y \in \mathbb{Q}$ no Hilbert space, because $\exists \{x_k\} \subset \mathbb{Q}$: $x_k \to \sqrt{2} = 1.4142\ldots \notin \mathbb{Q}$. □

■ **Example** (L^p-space of p-integrable functions)

For $\Omega \subset \mathbb{R}^n$ and $1 \leq p < \infty$, the set of p-integrable functions

$$L^p(\Omega) := \left\{ f : \Omega \to \mathbb{R} \,|\, f \text{ measurable, Lebesgue integral } \int_\Omega |f(x)|^p \, \mathrm{d}\lambda < \infty \text{ exists} \right\}$$

in the p-norm

$$\|f\|_{L^p(\Omega)} := \left(\int_\Omega |f(x)|^p \, \mathrm{d}\lambda \right)^{1/p}$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

is a Banach space.

More precisely, it is a complete space with a seminorm, the zero function is not unique. Two function in $L^p(\Omega)$ which agree almost everywhere are identified.

$X = L^2$ is the only Hilbert space among L^p spaces. The inner product on L^2 is

$$\langle f, g \rangle_{L^2(\Omega)} := \int_\Omega f(x) g(x) \, \mathrm{d}\lambda$$

□

Riesz-Fréchet representation theorem

Let be H Hilbert space and $\phi : H \to \mathbb{R}$ a continuous linear functional, i.e.

$\to$ linearity:

$$\phi(\sum_{i=1}^{n} \alpha_i f_i) = \sum_{i=1}^{n} \alpha_i \phi(f_i) \quad \forall \alpha_i \in \mathbb{R}, \forall f_i \in H$$

$\to$ continuity:

$$\exists C < \infty : |\phi(f)| \leq C \|f\| \quad \forall f \in H \quad (\Rightarrow \quad \|\phi\| = \max_{\|f\|=1} |\phi(f)| < \infty)$$

Then there exists *exactly one* $\varphi \in H$ such that $\phi(f) = \langle f, \varphi \rangle \ \ \forall f \in H$.

Moreover $\|\varphi\|_H = \|\phi\|_{H^*}$, where H^* is the (dual) space, consisting of all continuous linear functionals from H into $\mathbb{R}$.

The reverse direction is valid too: Every $\varphi \in H$ generates a continuous linear functional on H by this formula.

The following theorem characterizes minimal solutions.

Characterization Theorem

Let be H Hilbert space, and suppose that $a : H \times H \to \mathbb{R}$ is a symmetric positive bilinear form, i. e.

– $a(u, v)$ is linear in u and v
– $a(u, u) > 0 \ \ \forall u \in H \setminus \{0\}$
– $a(u, v) = a(v, u) \ \ \forall u, v \in H$

In addition, let $\phi : H \to \mathbb{R}$ be a continuous linear functional.

Then $J(w) := 1/2 \cdot a(w, w) - \phi(w)$ attains its minimum over H at $u \in H$

$\iff \quad a(u, w) = \phi(w) \ \ \forall w \in H$.

There is at most one such minimal solution (if any).

This theorem is valid also for a linear space X instead of a Hilbert space H! The proof is the same.

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Proof:

This proof is meant as an exercise for the use of the new definitions.

Riesz theorem $\Rightarrow \quad \exists!\, \varphi \in H: \quad \phi(w) = \langle \varphi, w \rangle \quad \forall w \in H$. Let now be $u, w \in H, \varepsilon \in \mathbb{R}$

$$\begin{aligned} \Rightarrow \quad J(u+\varepsilon w) &= \frac{1}{2} a(u+\varepsilon w, u+\varepsilon w) - \langle \varphi, u+\varepsilon w \rangle \\ &= J(u) + \varepsilon \left[a(u,w) - \langle \varphi, w \rangle \right] + \frac{1}{2} \varepsilon^2 a(w,w)\,. \end{aligned}$$

"$\Leftarrow$ ": $u \in H$ satisfies $a(u,w) = \langle \varphi, w \rangle$, choose $\varepsilon = 1 \Rightarrow J(u+w) = J(u) + a(w,w)/2 > J(u) \quad \forall w \neq 0 \Rightarrow u$ unique minimal point (uniqueness proven by contradiction: assume $u_1 \neq u_2 \quad \ldots$)

"$\Rightarrow$ ": J has minimum at $u \quad \Rightarrow \quad \dfrac{\mathrm{d}}{\mathrm{d}\varepsilon} J(u+\varepsilon w)\Big|_{t\varepsilon=0} = 0 \quad \Rightarrow \quad a(u,w) - \langle \varphi, w \rangle = 0$

□

Definition (continuous bilinear form, H-elliptic)

Let be H Hilbert space and $a : H \times H \to \mathbb{R}$ bilinear form.

a is called *continuous*, if

$$\exists c > 0 \quad \ni \quad |a(u,v)| \leq c \|u\| \cdot \|v\| \quad \forall u, v \in H$$

A symmetric and continuous bilinear form a is called *H-elliptic* (= elliptic or coercive), if

$$\exists d > 0 \quad \ni \quad a(v,v) \geq d \|v\|^2 \quad \forall v \in H$$

The induced norm $\|v\|_a := \sqrt{a(v,v)}$ is called *energy norm*; it is equivalent to the norm of the Hilbert space. □

The important (!!) theorem of Lax-Milgram gives existence and uniqueness of the variational problem on convex sets.

Theorem (Lax-Milgram)

Let be H Hilbert space, $V \subset H$ closed and convex and $a : H \times H \to \mathbb{R}$ H-elliptic bilinear form.

For every continuous linear functional $\varphi : H \to \mathbb{R}$ the variational problem

$$J(v) := \frac{1}{2} a(v,v) - \langle \varphi, v \rangle \to \min!$$

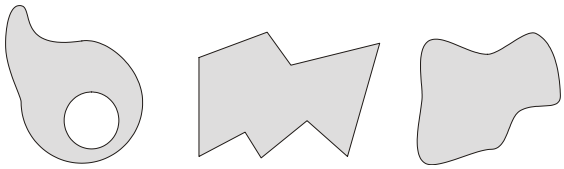
has a unique solution in V (i.e. existence and uniqueness).

Lipschitz continuous boundary, BL-domain

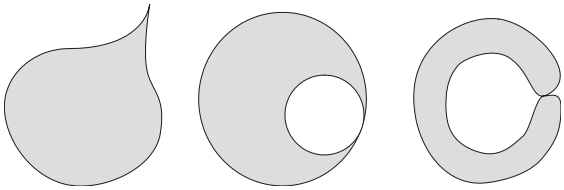
A domain with a Lipschitz continuous boundary is a domain whose boundary is piecewise smooth and "sufficiently regular".

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Instead of a precise definition of regularity, we illustrate that property by a few examples:



Lipschitz continuous boundaries



not Lipschitz continuous boundaries

Figure 29: Upper row: 3 examples of sets with Lipschitz continuous boundaries. Note that the smallest angle on the boundary is larger than 0.
Lower row: 3 examples of sets which boundaries are not Lipschitz continuous.

A bounded domain Ω with a Lipschitz continuous boundary $\partial\Omega$ is denoted by "Ω BL-domain" in these notes. □

Reduction to homogeneous boundary conditions

A certain class of quasilinear elliptic PDEs with Dirichlet conditions can be transformed to homogeneuos boundary conditions; this simplifies the further discussion.

Consider the Dirichlet boundary value problem

$$\begin{aligned} \mathcal{L}u &:= -\sum_{i,k=1}^{n} a_{ik}(x)\frac{\partial^2 u}{\partial x_i \partial x_j}\,, \\ \mathcal{L}u(x) &= f(x,u,p) \quad \forall x \in \Omega \subset \mathbb{R}^n\,,\ p_i := \frac{\partial u}{\partial x_i}\,, \\ u(x) &= g(x) \quad \forall x \in \partial\Omega\,. \end{aligned}$$

Let be Ω BL-domain. Suppose that $u_0 : \bar{\Omega} \to \mathbb{R}$ is an (arbitrary) function that satisfies $u_0|_{\partial\Omega} = g$; a possibility to obtain such an $u_0(x)$ is to choose $u_0|_{\partial\Omega} = g$ and then let u_0 linearly decay to zero. We define $w(x) := u(x) - u_0(x)$. Then

$$\mathcal{L}w = f_1(x,w,p) := f(x,u,p) - \mathcal{L}u_0 \ \ \forall x \in \Omega\,, \quad w(x) = 0 \ \ \forall x \in \partial\Omega\,.$$

For this type of problem we shall assume without simplification $g = 0$.

Definition (classical solution)

Let be $\Omega \subset \mathbb{R}^n$ BL-domain. Let be $u(x)$ a function that satisfies a quasilinear elliptic PDE of second order together with the boundary conditions.

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

u is called *classical solution* $:\Leftrightarrow$

$u \in \mathcal{C}^2(\Omega) \cap \mathcal{C}^0(\bar{\Omega})$ for Dirichlet boundary conditions $\left(u(x)|_{x\in\partial\Omega} \text{ given}\right)$

$u \in \mathcal{C}^2(\Omega) \cap \mathcal{C}^1(\bar{\Omega})$ for Neumann boundary conditions $\left(\left.\frac{\partial u(x)}{\partial n(x)}\right|_{x\in\partial\Omega} \text{ given}\right)$

and all these derivatives are bounded in Ω.

Important:
Higher order derivatives can be unbounded $\rightarrow$ convergence theory e.g. for FD methods then not applicable. □

Theorem (minimal property)

Let be $\Omega \subset \mathbb{R}^n$ BL-domain. Consider the linear elliptic PDE of 2^{nd} order

$$\begin{aligned} \mathcal{L}u(x) &:= -\sum_{i,k=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ik}(x)\frac{\partial u(x)}{\partial x_k}\right) + a_0(x)u(x) = f(x) \;\; \forall x \in \Omega \\ u(x) &= 0 \;\; \forall x \in \partial\Omega \end{aligned}$$

with a positive definite matrix (a_{ik}) and $a_0(x) > 0 \;\; \forall x \in \Omega$.

Every classical solution of this boundary value problem also solves the following minimization problem (= variational problem) and vice versa:

$$\begin{aligned} J(v) &\rightarrow \min! \quad \text{with} \\ J(v) &:= \int_\Omega \left[\frac{1}{2}\sum_{i,k=1}^{n} a_{ik}(x)\frac{\partial v(x)}{\partial x_i}\frac{\partial v(x)}{\partial x_k} + \frac{1}{2}a_0(x)v^2(x) - f(x)v(x)\right] \mathrm{d}x \end{aligned}$$

among all functions $v \in \mathcal{C}^2(\Omega) \cap \mathcal{C}^0(\bar{\Omega})$ with boundary condition $v(x)|_{x\in\partial\Omega} = 0$.

Proof:

With a slight modification of Green's first identity we get

$$\int_\Omega v(x)\frac{\partial}{\partial x_i}\left(a_{ik}(x)\frac{\partial}{\partial x_k}u(x)\right)\mathrm{d}x = -\int_\Omega a_{ik}(x)\frac{\partial v(x)}{\partial x_i}\frac{\partial u(x)}{\partial x_k}\mathrm{d}x \qquad (*)$$

We define

$$\begin{aligned} a(u,v) &:= \int_\Omega \left\{\sum_{i,k=1}^{n} a_{ik}(x)\frac{\partial u(x)}{\partial x_i}\frac{\partial v(x)}{\partial x_k} + a_0(x)u(x)v(x)\right\}\mathrm{d}x, \\ \langle f, v\rangle &:= \int_\Omega f(x)v(x)\,\mathrm{d}x \end{aligned}$$

and obtain using $(*)$: $a(u,v) - \langle f, v\rangle = \int_\Omega v(x)\left\{\mathcal{L}u(x) - f(x)\right\}\mathrm{d}x = 0$

Using the characterization theorem we directly obtain the minimal property. □

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Attention!

It is not granted, that for a given complicated boundary $\partial\Omega$ (e.g. with corners) or for not sufficiently smooth coefficients $a_{ik}(x)$ or ... a classical solution of the variational problem exists!

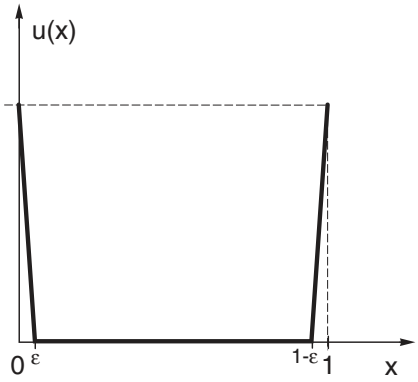
Therefore we have to introduce more general solution spaces!!

5.2 Function Spaces

■ Motivating example

Even for a bounded objective function $J(u)$ the minimum not necessarily is in the given function space. Consider e.g. the problem

$$J(u) := \int_0^1 u^2(t)\,\mathrm{d}t \to \min!, \quad u \in \mathcal{C}^0([0,1],\mathbb{R}) \ \cap \ u(0) = 1 = u(1).$$



The infimum of $J(u)$ is "0", but the corresponding u is not in $\mathcal{C}^0([0,1],\mathbb{R})$.

The difficulties can be overcome by solving the variational problem in a properly chosen Hilbert space and not e.g. in $\mathcal{C}^2([0,1],\mathbb{R})$. In the wrong space no solution exists! □

■ Example

Let be $\Omega \subset \mathbb{R}^n$ BL-domain.

Consider the space $\mathcal{C}^0(\Omega,\mathbb{R})$ (continuous functions) with

$$\langle u,v\rangle_0 := \int_\Omega u(x)v(x)\,dx\,, \quad \|u\|_0 := \sqrt{\langle u,u\rangle_0} \quad \forall\, u,v \in \mathcal{C}^0(\Omega,\mathbb{R})$$

We have seen that this space is not complete, i.e. not every Cauchy sequence of functions in $\mathcal{C}^0(\Omega,\mathbb{R})$ has a limit function that also is in $\mathcal{C}^0(\Omega,\mathbb{R})$.

Its completion – i.e. if we add all limit functions to this space – is the Hilbert space $L^2(\Omega,\mathbb{R})$ of the square-integrable functions with the same scalar product $\langle u,v\rangle_0 = \langle u,v\rangle_{L^2(\Omega)}$. □

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Definition

Let be $\Omega \subset \mathbb{R}^n$ BL-domain.

$\mathcal{C}^\infty(\Omega, \mathbb{R})$:= space of infinitely often differentiable functions;

$\mathcal{C}_0^\infty(\Omega, \mathbb{R})$:= subspace of $\mathcal{C}^\infty(\Omega, \mathbb{R})$, the functions $\varphi : \Omega \to \mathbb{R}$ of which have compact support in Ω, i.e.

$$\text{supp } \varphi := \overline{\{x \in \Omega \,|\, \varphi(x) \neq 0\}}$$

is bounded. φ is zero outside supp φ. Functions belonging to $\mathcal{C}_0^\infty(\Omega, \mathbb{R})$ are called test functions.

■ Example

Consider the function

$$\varphi : [-9, +7] \to \mathbb{R}, \quad \varphi(x) := \begin{cases} 0 & \text{for } |x| \geq 1 \\ \exp\left(-\dfrac{1}{1-x^2}\right) & \text{for } |x| < 1 \end{cases}$$

This function is infinitely often differentiable in the bounded domain $\Omega =]-9, 7[$ with supp $\varphi = [-1, 1] \subset \Omega$.

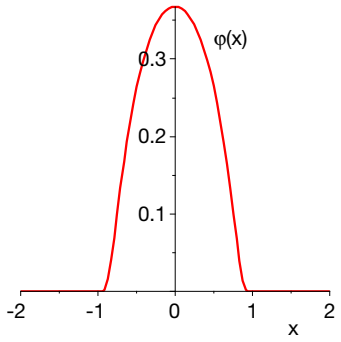


Figure 30: Example of a function $\varphi \in \mathcal{C}_0^\infty(\Omega, \mathbb{R})$ with supp $\varphi = [-1, 1] \subset \Omega$.

Definition

An n-dimensional multi-index is defined as an n-tuple $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ of non-negative integers with $|\alpha| = \alpha_1 + \alpha_2 + \ldots + \alpha_n$.

Let be α a multi index, then we define

$$D^\alpha u(x) := \frac{\partial^\alpha u(x)}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}}, \quad |\alpha| = \alpha_1 + \alpha_2 + \ldots + \alpha_n \,.$$

□

■ Example

$\alpha = 2$ and $u = u(x_1, x_2, x_3) = u(x, y, z)$, then

$$D^2 u(x, y, z) \in \left\{ u_{xx}, u_{yy}, u_{zz}, u_{xy} = \frac{\partial^2 u}{\partial x \partial y}, u_{xz}, u_{yz} \right\}$$

□

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Definition (weak derivative)

Let be $\Omega \subset \mathbb{R}^n$ BL-domain.

$u \in L^2(\Omega)$ has a *weak derivative* $v(x) := \partial^\alpha u(x)$

$$
\begin{aligned}
:\Leftrightarrow \quad & \exists v \in L^2(\Omega) \ni \langle \varphi, v\rangle_0 = (-1)^{|\alpha|}\langle \partial^\alpha \varphi, u\rangle_0 \\
\Leftrightarrow \quad & \int_\Omega v(x)\varphi(x)\,\mathrm{d}x = (-1)^{|\alpha|}\int_\Omega u(x) D^\alpha \varphi(x)\,\mathrm{d}x \quad \forall \varphi \in \mathcal{C}_0^\infty(\Omega, \mathbb{R})
\end{aligned}
$$

This definition is motivated by the technique of integration by parts. □

Remark

In a more general definition of the weak derivative, the space $L^1_{loc}(\Omega)$ is used instead of $L^2(\Omega)$. $L^1_{loc}(\Omega)$ contains all functions which are in $L^1(\Omega_1)$ for every compact sub-BL-domain $\Omega_1 \subset \Omega$. For us the definition with $L^2(\Omega)$ has certain advantages. □

Notation

In these lecture notes we alternatively will denote special weak derivatives by ∂ with index, e.g. for $u(x,y)$:

usual partial derivatives: $u_x = \dfrac{\partial}{\partial x}u = u_x\,, \quad \dfrac{\partial^2}{\partial x \partial y}u = u_{xy}\,, \quad \ldots$

weak derivatives: $\partial_x u\,, \quad \partial_{xy} u\,, \quad \ldots$

□

■ Example

Let be $\Omega =]a,b[\subset \mathbb{R}$ and $f \in \mathcal{C}^1(]a,b[, \mathbb{R})$ differentiable (in the classical sense), let be $\varphi \in C_0^\infty(I, \mathbb{R})$ test function, then by integration by parts we get

$$\int_\Omega f'(t)\varphi(t)\,\mathrm{d}t = -\int_\Omega f(t)\varphi'(t)\,\mathrm{d}t$$

Boundary terms vanish because $(\varphi(a) = 0, \varphi(b) = 0)$ for our test function.

If $f \in L^2(\Omega)$, then – even if f is not differentiable – a function $g \in L^2(\Omega)$ possibly exists with

$$\int_\Omega g(t)\varphi(t)\,\mathrm{d}t = -\int_\Omega f(t)\varphi'(t)\,\mathrm{d}t \quad \forall \varphi \in C_0^\infty(\Omega, \mathbb{R})$$

Such a g is called a weak derivative.

Example: $f(x) = |x|$, which is not (classically) differentiable at $x = 0$ and thus f not differentiable on $\Omega :=]a,b[\,,\ a < 0 < b$. For the weak derivative we get

$$\partial_x f(x) = \begin{cases} -1 & : \quad x < 0 \\ 0 & : \quad x = 0 \\ +1 & : \quad x > 0 \end{cases}$$

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

because for any test function $\varphi \in C_0^\infty(\Omega, \mathbb{R})$ integration by parts gives

$$\begin{aligned}\int_a^b \varphi'(x) \cdot f(x)\,\mathrm{d}x &= \int_a^0 -\varphi'(x)x\,\mathrm{d}x + \int_0^b \varphi'(x)x\,\mathrm{d}x \\ &= -\left(\int_a^0 \varphi(x)\cdot(-1)\,\mathrm{d}x + \int_0^b \varphi(x)\cdot 1\,\mathrm{d}x\right) \\ &= -\int_a^b \varphi(x)\cdot\partial_x f(x)\,\mathrm{d}x\end{aligned}$$

The weak derivative itself has no weak derivative, because it has a jump.

The definition $\partial_x f(0) = 0$ is arbitrarily and – because the weak derivative is defined via an integral – has no effect on the result.

Here you see the benefit of identifying functions in $L^2(\Omega)$ that agree almost everywhere. □

■ Example

Let us revisit the one-dimensional model problem in chap 4.1.1. Here the critical formula was

$$\begin{aligned}0 &= \left.\frac{\mathrm{d}J(\varepsilon)}{\mathrm{d}\varepsilon}\right|_{\varepsilon=0} = \left.\frac{\mathrm{d}I(u+\varepsilon\eta)}{\mathrm{d}\varepsilon}\right|_{\varepsilon=0} = \dots \\ &= 2\int_0^L \left[(u(x)-f(x))\cdot\eta(x) + \beta u'(x)\cdot\eta'(x)\right]\mathrm{d}x\end{aligned}$$

If we apply integration by parts in the next step we cannot avoid $u''(x)$ to appear.

If we use the weak derivative v instead, we get for all test functions $\eta \in \mathcal{C}^\infty([0,L],\mathbb{R})$ with $\eta(0) = 0 = \eta(L)$

$$\int_0^L v(x)\eta(x)\,\mathrm{d}x = -\int_0^L g(x)\eta'(x)\,\mathrm{d}x \quad \wedge \quad g := \beta\cdot u'$$

□

Remark

If a function is differentiable in the classical sense, then it is also weakly differentiable and both derivatives match (proof by insertion). □

Definition (Sobolev space)

Let be $\Omega \subset \mathbb{R}^n$ BL-domain.

The Sobolev space $H^m(\Omega) := H^m(\Omega, \mathbb{R})$ of order $m \in \mathbb{N}_0$ is defined by:

$$H^m(\Omega) := \left\{u \in L^2(\Omega)\,|\,u \text{ has (all) weak derivatives } \partial^\alpha u \;\; \forall |\alpha| \le m\right\}$$

The Sobolev space $H_0^m(\Omega)$ is the subspace of $H^m(\Omega)$, the functions $\varphi : \Omega \to \mathbb{R}$ of which have compact support in Ω, i.e.

$$\text{supp}\,\varphi := \overline{\{x \in \Omega\,|\,\varphi(x) \neq 0\}}$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

is bounded; φ is zero outside supp φ ($\rightarrow$ similar to $\mathcal{C}^\infty(\Omega,\mathbb{R}), \mathcal{C}_0^\infty(\Omega,\mathbb{R})$).

In $H^m(\Omega)$ we define a scalar product and an associated norm by

$$\langle u,v\rangle_m := \sum_{|\alpha|\le m} \langle \partial^\alpha u, \partial^\alpha v\rangle_0\,, \quad \|u\|_{m,\Omega} := \|u\|_m := \sqrt{\langle u,u\rangle_m}$$

Additionally we define the seminorm $|u|_m := \sqrt{\sum_{|\alpha|=m} \langle \partial^\alpha u, \partial^\alpha u\rangle_0}$.

A *seminorm* in a vector space X over $\mathbb{R}$ is a mapping $p\colon X \to \mathbb{R}_0^+$ with the properties $\forall x,y \in X, \lambda \in \mathbb{R}$

$$\begin{array}{rcll} p(\lambda x) & = & |\lambda| p(x) & \text{(absolute homogenity)} \\ p(x+y) & \le & p(x)+p(y) & \text{(subadditivity)} \end{array}$$

Compared to a norm, the uniqueness of the zero is missing. □

■ Example

The scalar product in $H^1(\Omega) = H^1(\Omega,\mathbb{R})$, $\Omega \subseteq \mathbb{R}^n$ is

$$\langle u,v\rangle_1 = \int_\Omega \left(u(x)v(x) + \sum_{i=1}^n \partial_{x_i} u(x) \cdot \partial_{x_i} v(x) \right) \mathrm{d}x\,.$$

Some important properties of Sobolev spaces

Let be $\Omega \subset \mathbb{R}^n$ BL-domain.

(i) $(H^m(\Omega), \|\cdot\|_m)$ is Hilbert space, $L^2(\Omega) = H^0(\Omega)$
$(H_0^m(\Omega), \|\cdot\|_m)$ is Hilbert space

(ii) $H^m(\Omega)$ is completion of $\mathcal{C}^\infty(\Omega,\mathbb{R}) \cap H^m(\Omega)$ w.r.t. $\|\cdot\|_m$,
$H_0^m(\Omega)$ is completion of $\mathcal{C}_0^\infty(\Omega,\mathbb{R}) \cap H_0^m(\Omega)$ w.r.t. $\|\cdot\|_m$

Attention (1):
Not every Cauchy sequence converges in $\mathcal{C}^\infty(\Omega,\mathbb{R})$ w.r.t. the $\|\cdot\|_m$-norm to an element from $\mathcal{C}^\infty(\Omega,\mathbb{R})$!

Attention (2):
"$\mathcal{C}^\infty(\Omega,\mathbb{R}) \cap H^m(\Omega)$" necessary because Ω is open, so a function $f \in \mathcal{C}^\infty(\Omega,\mathbb{R})$ may tend to ∞ close to the boundary such that f is not in L^2 (e.g. $f(x) = 1/x$ for $x \in \Omega =]0,1[$). We want to exclude that situation.

(iii) For $n = 1$ only, i.e. $\Omega \subset \mathbb{R}$: $\mathcal{C}^m([a,b],\mathbb{R}) \subset H^m([a,b]) \subset \mathcal{C}^{m-1}([a,b],\mathbb{R})$

(iv) For $n > 1$: $H^m(\Omega)$ can be continuously embedded into $\mathcal{C}^0(\bar\Omega,\mathbb{R})$ for $2m > n$, if the norm of $\mathcal{C}^0(\bar\Omega,\mathbb{R})$ is $\max_{x\in\bar\Omega} |y(x)|$, i.e.
$\exists c > 0 \ni \|u\|_{\mathcal{C}^0} \le c\|u\|_m \ \ \forall u \in H^m(\Omega)$.

(v) $H^0(\Omega) \supset H^1(\Omega) \supset H^2(\Omega) \supset \ldots$; $H_0^0(\Omega) \supset H_0^1(\Omega) \supset \ldots$;
$H^0(\Omega) = H_0^0(\Omega)$ and $H^i(\Omega) \supset H_0^i(\Omega), i \ge 1$.

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

The next theorem states, that all functions in $H^m(\Omega)$ can be approximated by smooth functions as good as we want.
That allows to transfer properties of classical functions (e.g. product rule is valid) to Sobolev functions.
The theorem tells us e.g. that $\mathcal{C}^\infty(\Omega,\mathbb{R})\cap H^1(\Omega)$ is dense in $H^1(\Omega)$; therefore, it is sufficient to prove an inequality like Poincaré-Friedrich (see below) only for $v\in\mathcal{C}^\infty(\Omega,\mathbb{R})$ and not for $v\in H^1(\Omega)$.

Theorem

Let be $\Omega\subset\mathbb{R}^n$ BL-domain and $m\geq 0$. Then $\mathcal{C}^\infty(\Omega,\mathbb{R})\cap H^m(\Omega)$ is dense in $H^m(\Omega)$, i.e.

for every $u\in H^m(\Omega)$ and every $\varepsilon>0$ there exists $\varphi\in\mathcal{C}^\infty(\Omega,\mathbb{R})\cap H^m(\Omega)$ with $\|u-\varphi\|_m\leq\varepsilon$.

The following inequalities show the equivalence of norm and seminorm. That is important, because often it is simpler to use the seminorm (e.g. in estimates).

Theorem (Poincaré-Friedrich inequality)

Let be Ω BL-domain and $W:=\{(x_1,\ldots,x_n)\,|\,x_i\in]0,s[\}$ an $n-$dimensional cube with edge length of s such that $\Omega\subset W$. Then

$$\|v\|_0\leq s|v|_1\quad\forall v\in H_0^1(\Omega)\,.$$

Corollary

Applying the Poincaré-Friedrich inequality also on the derivatives (induction) we get:

In $H_0^m(\Omega)$ the norms $\|\cdot\|_m$ and $|\cdot|_m$ are equivalent, more precisely

$$|v|_m\leq\|v\|_m\leq(1+s)^m|v|_m\quad\forall v\in H_0^m(\Omega)\,.$$

Definition (weak solution)

Let be Ω BL-domain. Consider the linear elliptic PDE of 2^{nd} order

$$\mathcal{L}u(x):=-\sum_{i,k=1}^{n}\frac{\partial}{\partial x_i}\left(a_{ik}(x)\frac{\partial u(x)}{\partial x_k}\right)+a_0(x)u(x)=f(x)\quad\forall x\in\Omega$$

with positive definite matrix (a_{ik}) and $a_0(x)\geq 0$ and Dirichlet boundary conditions $u(x)=0\ \forall x\in\partial\Omega$. Let be $f\in L^2(\Omega)$.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

The associated bilinear form and scalar product are defined by

$$\begin{aligned} a(u,v) &:= \int_\Omega \left\{ \sum_{i,k=1}^{n} a_{ik}(x)\partial_{x_i}u(x)\cdot\partial_{x_k}v(x) + a_0(x)u(x)v(x) \right\} \mathrm{d}x \\ \langle f,v\rangle_0 &:= \int_\Omega f(x)v(x)\,\mathrm{d}x \end{aligned}$$

A function $u \in H_0^1(\Omega)$ is called *weak solution* of the PDE above if

$$a(u,v) - \langle f,v\rangle_0 = 0 \qquad \forall v \in H_0^1(\Omega)\,.$$

Theorem (existence)

Assumptions same as in the definition of the weak solution above.

In addition let be $\mathcal{L}$ a *uniformly elliptic differential operator*, i.e.

$$\exists a^* > 0 \;\ni\; \xi^T A\xi = \sum_{k,j=1}^{n} \xi_k a_{kj}(x)\xi_j \geq a^*|\xi|^2 \quad \forall \xi \in \mathbb{R}^n, x \in \Omega$$

Then the Dirichlet problem always has a uniquely determined weak solution in $H_0^1(\Omega)$. This solution minimizes the variational problem

$$\frac{1}{2}a(v,v) - \langle f,v\rangle_0 \;\to\; \min! \qquad \forall v \in H_0^1(\Omega)\,.$$

Proof:

In short:
The proof uses the Cauchy-Schwarz inequality, the uniform elliptic differential operator, Poincaré-Friedrich's inequality, the H^1-elliptic bilinear form, Lax-Milgram and the characterization theorem and therefore also the assumptions made therein.

In detail:
Define $c_1 := \sup\left\{|a_{ik}(x)| \,\middle|\, x \in \Omega, 1 \leq i,k \leq n\right\}$; with the Cauchy-Schwarz inequality we get

$$\begin{aligned} & \left| \sum_{i,k=1}^{n} \int_\Omega a_{ik}\partial_{x_i}u\partial_{x_k}v\,\mathrm{d}x \right| \leq c_1 \cdot \sum_{i,k=1}^{n} \left| \int_\Omega \partial_{x_i}u\partial_{x_k}v\,\mathrm{d}x \right| \\ \overset{CSU}{\leq}\; & c_1 \cdot \sum_{i,k=1}^{n} \left[\int_\Omega |\partial_{x_i}u|^2\,\mathrm{d}x \cdot \int_\Omega |\partial_{x_k}v|^2\,\mathrm{d}x \right]^{1/2} \leq c_1 n^2 |u|_1 \cdot |v|_1 \\ \leq\; & c_1 n^2 \|u\|_1 \cdot \|v\|_1 \end{aligned}$$

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

With $c_2 := \sup\{|a_0(x)| \,|\, x \in \Omega\}$ we analogously get

$$\left|\int_\Omega a_0(x)\cdot uv\,\mathrm{d}x\right| \;\le\; c_2\left|\int_\Omega uv\,\mathrm{d}x\right| \;\le\; c_2\cdot\|u\|_0\|v\|_0 \;\le\; c_2\cdot s^2\|u\|_1\|v\|_1$$

In total: $a(u,v)\le C\cdot\|u\|_1\|v\|_1\,,\;\; C := c_1n^2+c_2s^2$.

For $\mathcal{C}^1$-functions (and thus especially for $\mathcal{C}^\infty$-functions) the uniform ellipticity implies the pointwise estimate

$$\sum_{i,k=1}^n a_{ik}(x)\partial_{x_i}v\partial_{x_k}v \ge a^*\sum_{i=1}^n(\partial_{x_i}v)^2$$

Integrating both sides and using $a_0\ge 0$ leads to:

$$a(v,v)\ge a^*\cdot\sum_{i=1}^n\int_\Omega(\partial_{x_i}v)^2\,\mathrm{d}x = a^*\,|v|_1 \quad \forall\, v\in H^1(\Omega,\mathbb{R})\,.$$

Because of the integration the estimate is valid for H^1 instead of only for $\mathcal{C}^\infty$ (because $\mathcal{C}^\infty$ is dense in H^1 and H^0 respectively).

Poincaré-Friedrich's inequality gives the equivalence of $|\cdot|_1$ and $\|\cdot\|_1$ on $H_0^1(\Omega)$. Therefore, a is a H^1-elliptic bilinear form on $H_0^1(\Omega)$. By Lax-Milgram there exists a unique (weak) solution of the variational problem, which because of the characterization theorem is a weak solution of the PDE. □

■ Example

Consider

$$-\Delta u(x) = f(x)\;\; \forall x\in\Omega\subset\mathbb{R}^2\,,\quad u(x)=0\;\; \forall x\in\partial\Omega$$

We write

$$\Delta = \frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2} \quad\Rightarrow\quad (a_{ik}) = \begin{pmatrix}1&0\\0&1\end{pmatrix},\;\; a_0=0$$

Therefore, for the weak solution $u(x)$ we get

$$a(u,v) = \int_\Omega\left(\sum_{i=1}^n\partial_{x_i}u(x)\cdot\partial_{x_i}v(x)\right)\mathrm{d}x = \int_\Omega f(x)v(x)\,\mathrm{d}x \quad \forall\, v\in H_0^1(\Omega)$$

□

Summary: Dirichlet problem for linear elliptic PDEs of second order

Classical solution:

Solution $u\in\mathcal{C}_0^2(\Omega)$, derivatives in $J(v)$ are the usual partial derivatives.

If it exists: Classical solution of the PDE equivalent to the solution of the variational problem $J(v)\to\min!$

Weak solution:

Solution $u\in H_0^1(\Omega)$, derivatives in $J(v)$ are weak derivatives.

Dirichlet problem always has a uniquely defined weak solution, which also is solution of the variational problem $J(v)\to\min!$

If a classical solution exists, weak and classical solution are identical. An approximation of the weak solution can be calculated numerically ($\to$ FE)! □

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

5.3 Ritz-Galerkin Method

5.3.1 Basic Principle of the Finite Element Method

$J(v)$ is not minimized $\forall v \in H_0^m(\Omega)$, but only w.r.t. a "suitably" chosen, finite-dimensional subspace $S_h \subset H_0^m(\Omega)$:

$$J(v) = \frac{1}{2}a(v,v) - \langle f, v\rangle_0 \to \min! \quad \forall v \in S_h\,.$$

The quality of the solution largely depends on the choice of $S_h := S_h(\Omega, \mathbb{R})$!

Remark

Because $S_h \subset H_0^m(\Omega)$ (i.e. a true subset), the approach is called *conforming*.

Rayleigh - Ritz method: Find $u_h \in S_h$ such that $J(u_h) \le J(v)\ \ \forall v \in S_h$

Galerkin method: Find $u_h \in S_h$ such that $a(u_h, v) = \langle f, v\rangle_0\ \ \forall v \in S_h$

For an H-elliptic bilinear form a both methods coincide $\to$ Ritz-Galerkin. □

How is the **approximation in a finite-dimensional subspace** realized?

Definition

$u_h \in S_h$ is a solution if

$$a(u_h, v) - \langle f, v\rangle_0 = 0 \quad \forall v \in S_h\,.$$

Suppose $\dim(S_h) = N$ and $\{\psi_1, \ldots, \psi_N\}$ is a basis of S_h. Then the last condition is equivalent to

$$a(u_h, \psi_i) = \langle f, \psi_i\rangle_0\,,\ \ i = 1, \ldots, N\,, \quad u_h = \sum_{k=1}^{N} z_k \psi_k$$

Using bilinearity we are led to the system of equations

$$\sum_{k=1}^{N} a(\psi_k, \psi_i) z_k = \langle f, \psi_i\rangle_0\,,\ \ i = 1, \ldots, N$$

which we can write in matrix notation

$$Az = b\,,\ A \in \mathbb{R}^{N\times N}\,, \quad A_{ki} := a(\psi_k, \psi_i)\,, \quad b_i = \langle f, \psi_i\rangle_0\,, \quad i,k = 1, \ldots, N\,.$$

A is called *stiffness matrix* or *system matrix*, b *load vector*.

The basis functions ψ_i are defined on the whole domain of Ω, but they should have a small and compact support $\operatorname{supp}\psi_i$ so that many components of A become zero!

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

5.3.2 One-dimensional Model Problem: $-u_{xx} = f$

We consider the one-dimensional model problem

$$\begin{aligned} -u_{xx}(x) &= f(x), \quad \Omega :=]0,1[\\ u(0) &= 0 \\ u(1) &= 0 \end{aligned}$$

Step 1: We use the definition of the weak derivative to get the weak form

$$-\int_0^1 v(x) \cdot \partial_{xx} u(x)\,\mathrm{d}x = \int_0^1 \partial_x v(x) \cdot \partial_x u(x)\,\mathrm{d}x = \int_0^1 f(x) \cdot v(x)\,\mathrm{d}x \quad \forall v \in H_0^1(\Omega)$$

Integration by parts causes only first derivatives to appear, so $v \in H_0^1(\Omega)$ and $f \in L^2(\Omega)$ is sufficient. The derivatives are further smoothened by the integration procedure.

Step 2: More abstract notation

$$a(u,v) := \int_\Omega \partial_x v(x) \cdot \partial_x u(x)\,\mathrm{d}x, \quad \langle f, v\rangle := \int_\Omega f(x) \cdot v(x)\,\mathrm{d}x, \quad \Omega =]0,1[$$

Step 3: Galerkin projection, general approach

Let be $\{\psi_k\}$ a finite-dimensional basis of S_h and $A_{ki} := a(\psi_k, \psi_i)$

A is symmetric because

$$a(\psi_k, \psi_i) = \int_\Omega \partial_x \psi_k(x) \cdot \partial_x \psi_i(x)\,\mathrm{d}x = \int_\Omega \partial_x \psi_i(x) \cdot \partial_x \psi_k(x)\,\mathrm{d}x = a(\psi_i, \psi_k)$$

A is positive because with $v := \sum_{i=1}^N y_i \psi_i(x)$ we get for $y \in \mathbb{R}^N$

$$y^T A y = \sum_{i,j} A_{ij} y_i y_j = a(v,v) = \int_\Omega (\partial_x v)^2\,\mathrm{d}x \geq 0$$

and $y^T A y = 0$ for $v = 0$ only. In total, A is positive definite.

Because A positve definite, the linear system $Az = b$ is uniquely solvable.

Step 4: Linear finite elements and global approach

As basis functions ψ_i in S_h we here choose hat functions defined by

$$\psi_i(x) := \begin{cases} 1 - (x_i - x)/(x_i - x_{i-1}) & \text{for } x \in [x_{i-1}, x_i] \\ 1 - (x - x_i)/(x_{i+1} - x_i) & \text{for } x \in [x_i, x_{i+1}] \\ 0 & \text{else} \end{cases}$$

The $\psi_j,\ j = 1, \ldots, N$, are in $H_0^1(\Omega) = H_0^1(]0,1[)$, they are piecewise linear and they have compact support: $\psi_j(x) \neq 0$ for $x \in [x_{j-1}, x_{j+1}]$ only.

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

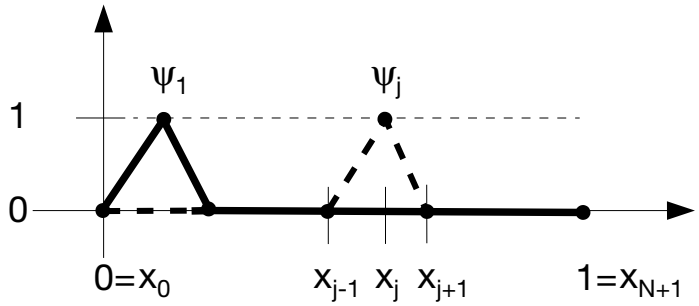


Figure 31: Hat functions. The discretization $0 < x_0 < x_1 < \ldots < x_{N+1} = 1$ is not necessarily equidistant.

The important property is

$$\psi_i(x_k) = \delta_{ik}\,, \quad i,k = 1,\ldots,N$$

and we get

$$u_h(x) = \sum_{k=1}^{N} z_k \psi_k(x) \quad \Rightarrow \quad u_h(x_i) = z_i$$

The approximative solution u_h is in $H_0^1(]0,1[)$ (and thus continuous) and interpolates the points (x_i, z_i). That is only an interpretation: The z_i cannot be chosen, but are obtained from the solution of the linear system $Az = b$.

We now calculate A for the chosen basis of hat functions row by row. We make a gobal approach, i.e. we consider the total domain $\Omega =]0,1[$ at once and do not restrict the analysis to one interval $[x_j, x_{j+1}]$.

The ψ_i have a very small support and therefore many elements of the matrix A are equal to zero. From $\operatorname{supp}\psi_i = [x_{i-1}, x_{i+1}]$ we immediately get

$$a(\psi_i, \psi_j) = 0 \quad \text{for} \quad |i-j| > 1$$

$\Rightarrow$ stiffness matrix A is tridiagonal $\Rightarrow$

the i-th row of A has 3 non-zero entries only: $A_{i,j} = a(\psi_i, \psi_j),\ j = i-1, i, i+1$.

Because of the simple structure of the ψ_i, the following integrals can be calculated analytically. With $h_i := x_i - x_{i-1}$ we get

$$A_{i,i-1} = a(\psi_i, \psi_{i-1}) = \int_0^1 \partial_x \psi_i(x) \cdot \partial_x \psi_{i-1}(x)\,\mathrm{d}x = \int_{x_{i-1}}^{x_i} \frac{1}{h_i} \cdot \frac{-1}{h_i}\,\mathrm{d}x = -\frac{1}{h_i}$$

and analogously $A_{i,i} = \dfrac{1}{h_i} + \dfrac{1}{h_{i+1}}$, $A_{i,i+1} = -\dfrac{1}{h_{i+1}}$.

The right hand side of the linear system is calculated in a similar way. If f is too complicated, numerical quadrature is used.
Here we use the simple example $f(x) = 2 \ \ \forall x \in]0,1[$ and get

$$\langle f, \psi_i \rangle = \int_0^1 2 \cdot \psi_i(x)\,\mathrm{d}x = 2 \cdot 0.5 \cdot (h_i + h_{i+1}) = h_i + h_{i+1}$$

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

The resulting linear system $Az = b$ for that special f is

$$\begin{pmatrix} 1/h_1+1/h_2 & -1/h_2 & & & \\ -1/h_2 & 1/h_2+1/h_3 & -1/h_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1/h_N \\ & & & -1/h_N & 1/h_N \end{pmatrix} z = \begin{pmatrix} h_1+h_2 \\ \\ \vdots \\ \\ h_N+h_{N+1} \end{pmatrix}$$

Remark

The structure of A and b as well as statements on the unique solvability of the linear system depend on the PDE and on the chosen basis functions. □

Step 5: Boundary conditions

The Dirichlet boundary conditions are satisfied by all basis functions, thus no additional steps are necessary.

Step 6: Short convergence analysis

We have proven that $Az = b$ has a unique solution. Now we are interested in the convergence properties of the method if we increase the number $N-1$ of interior mesh points, i.e. for $h_i \to 0$.

In the definition of H-elliptic bilinear forms, we have introduced the induced energy norm

$$\|v\|_a := \sqrt{a(v,v)} \quad \forall v \in H_0^1(\Omega)$$

Let $u \in H_0^1(\Omega)$ be the solution of the weak form $a(u,v) - \langle f, v\rangle = 0 \;\; \forall v \in H_0^1(\Omega)$ and $u_h \in S_h$ the Ritz-Galerkin solution $a(u_h,v) - \langle f, v\rangle = 0 \;\; \forall v \in S_h$. Subtraction gives

$$a(u-u_h, v) = 0 \quad \forall v \in S_h \quad \text{with } u \in H_0^1(\Omega),\, u_h \in S_h$$

$\Rightarrow$ fundamental property: W.r.t. the scalar product induced by a the error $u-u_h$ is orthogonal to the subspace S_h (Galerkin orthogonality).

For the energy norm – as for any other induced norm – the Cauchy-Schwarz inequality holds

$$|a(v,w)| \le \|v\|_a \|w\|_a \quad \forall v, w \in H_0^1(\Omega)$$

With that we get using the bilinearity

$$\begin{aligned} \|u-u_h\|_a^2 &= a(u-u_h, u-u_h) = a(u-u_h, u-v) + a(u-u_h, v-u_h) \\ &= a(u-u_h, u-v) \qquad (\text{for } a(u-u_h, v) = 0 \quad \forall v \in S_h) \\ &\le \|u-u_h\|_a \|u-v\|_a \\ \Rightarrow \quad &\|u-u_h\|_a \le \|u-v\|_a \quad \forall v \in S_h \end{aligned}$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

The approximation by the Galerkin method is the best possible approximation in the given subspace S_h w.r.t the special induced norm $\|\cdot\|_a$

$$\|u-u_h\|_a = \min\{\|u-v\|_a \,|\, \forall v \in S_h\}$$

That is nice, but not of great interest in applications. We are interested in the error w.r.t. a usual norm which is independent on the special problem a, e.g. in the L^2-norm

$$\|v\|_{L^2(\Omega)} = \sqrt{\langle v,v\rangle} = \left(\int_0^1 v(x)^2\,\mathrm{d}x\right)^{1/2}$$

Such estimates often are rather difficult to obtain and require a detailed analysis of the approximation space S_h. For our problem we get the error estimate

$$\begin{aligned}\|u-u_h\|_{L^2(\Omega)} &\leq Ch\|u-u_h\|_a \\ &\leq (Ch)^2\|u_{xx}\|_{L^2(\Omega)}, \quad \Omega=[0,1],\, h=\max_{1\leq i\leq N+1}(x_i-x_{i-1})\end{aligned}$$

Summary

Due to our theorem on p. 120 our one-dimensional example problem has a uniquely determined weak solution $u \in H_0^1(\Omega)$ that satisfies

$$a(u,v)-\langle f,v\rangle = 0 \quad \forall v \in H_0^1(\Omega)$$

and that minimizes the variational problem

$$J(v) := a(v,v)-\langle f,v\rangle \;\to\; \min \quad \forall v \in H_0^1(\Omega)$$

The Ritz-Galerkin solution $u_h \in S_h$ exists and is uniquely determined and computable and minimizes

$$J(v_h) := a(v_h,v_h)-\langle f,v_h\rangle \;\to\; \min \quad \forall v_h \in S_h \subset H_0^1(\Omega)$$

This result at that point tells us nothing about the desired solution u of the PDE.

It was not before step 6 that we have learned that in addition u_h is closest to u compared to all other $v \in S_h$ if we use the $\|\cdot\|_a$-norm. □

Remark

Here the energy norm is the L^2-norm for the derivative v'. If this norm has to be $<\infty$, it is a stronger condition as if the L^2-norm of the function v is $<\infty$. □

■ Numerical example

For the mixed boundary value problem

$$-u''(x) = \frac{\pi^2}{4}\sin\left(\frac{\pi}{2}x\right),\;\; u(0)=u'(1)=0 \;\Rightarrow\; u(x)=\sin\frac{\pi}{2}x$$

the results of numerical calculations with linear finite elements are summarized in the following graphs:

The method is convergent of order 2 in the L^2-norm and of order 1 in the energy norm. The Neumann boundary condition is automatically fulfilled, because it is indirectly included in the weak formulation. □

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

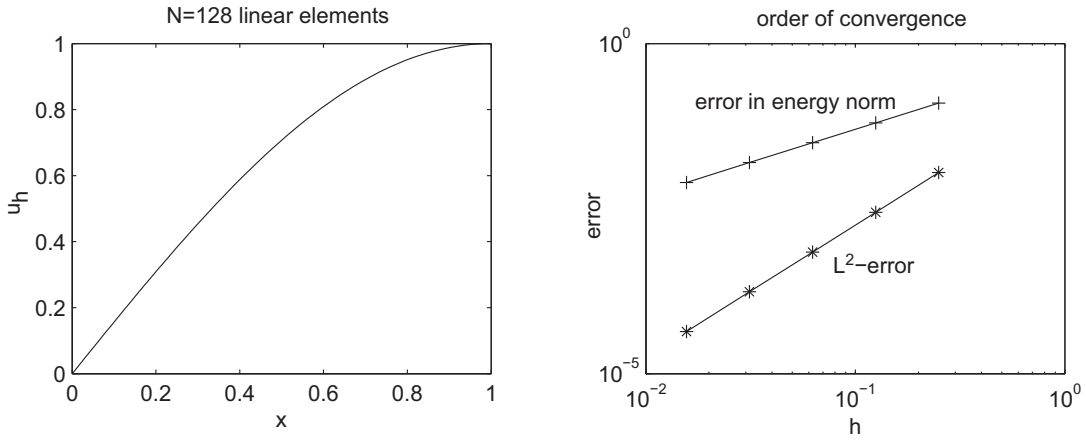


Figure 32: Solution using 128 linear finite elements (le.) and convergence properties depending on mesh size h (ri.).

5.3.3 Shape Functions and Local Approach

Let us consider the same problem as in chap. 5.3.2. Here we favour a local approach, that is more efficient to compute. The final results remain exactly the same.

We analyze only one single interval $[x_{i-1}, x_i]$, called element i. Because of the small and compact support, only 2 basis functions contribute to the solution on $[x_{i-1}, x_i]$:

$$u_h\Big|_{[x_{i-1},x_i]} = z_{i-1} \cdot \psi_{i-1}(x) + z_i \cdot \psi_i(x)$$

All other ψ_j vanish and thus do not contribute to the solution on that interval.

Next we transform $[x_{i-1}, x_i]$ to the unit interval $[0,1]$ via $\xi = (x - x_{i-1})/h_i\,,\ \ h_i := x_i - x_{i-1}$. We define the shape functions

$$N_1(\xi) := 1 - \xi\,, \quad N_2(\xi) := \xi\,, \quad 0 \le \xi \le 1$$

After transformation of $[x_{i-1}, x_i]$ to the reference interval $[0,1]$, each basis function (here: ψ_{i-1}, ψ_i) coincides with one of the normed shape functions. The shape functions are the same on each of the elements.

What is the contribution of the basis functions on element i to the total stiffness matrix. For that we analyze the following sub-matrix of A

$$\begin{aligned}
&\begin{pmatrix} a(\psi_{i-1},\psi_{i-1}) & a(\psi_{i-1},\psi_i) \\ a(\psi_i,\psi_{i-1}) & a(\psi_i,\psi_i) \end{pmatrix} \\
= &\begin{pmatrix} \int_{x_{i-2}}^{x_{i-1}} (\partial_x \psi_{i-1})^2 \,\mathrm{d}x & 0 \\ 0 & 0 \end{pmatrix} + \underbrace{\begin{pmatrix} \int_{x_{i-1}}^{x_i} (\partial_x \psi_{i-1})^2 \,\mathrm{d}x & \int_{x_{i-1}}^{x_i} \partial_x\psi_{i-1} \cdot \partial_x \psi_i \,\mathrm{d}x \\ \int_{x_{i-1}}^{x_i} \partial_x\psi_i \cdot \partial_x \psi_{i-1} \,\mathrm{d}x & \int_{x_{i-1}}^{x_i} (\partial_x \psi_i)^2 \,\mathrm{d}x \end{pmatrix}}_{=:A_i^{(e)}} \\
+ &\begin{pmatrix} 0 & 0 \\ 0 & \int_{x_i}^{x_{i+1}} (\partial_x \psi_i)^2 \,\mathrm{d}x \end{pmatrix}
\end{aligned}$$

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

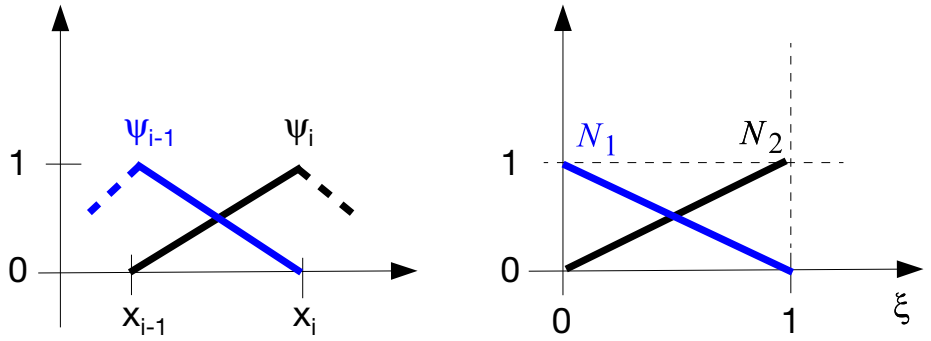


Figure 33: Linear basis functions (here: hat functions) ψ_i (le.) and corresponding shape functions (ri.)

The first summand contains contributions of element $i-1$, the third those of element $i+1$. We are interested in the second matrix which only contains contributions of element i: This matrix $A_i^{(e)}$ is called stiffness matrix on the element level.

For an efficient calculation of $A_i^{(e)}$ shape functions can be used. For the upper left component we e.g. get

$$\int_{x_{i-1}}^{x_i} \left(\partial_x \psi_{i-1}(x)\right)^2 \mathrm{d}x = h_i \int_0^1 \left(\frac{N_1'(\xi)}{h_i}\right)^2 \mathrm{d}\xi = \frac{1}{h_i} \int_0^1 \left(N_1'(\xi)\right)^2 \mathrm{d}\xi$$

With that we obtain for our model problem

$$A_i^{(e)} = \frac{1}{h_i} \begin{pmatrix} \int_0^1 \left(N_1'(\xi)\right)^2 \mathrm{d}\xi & \int_0^1 N_1'(\xi) \cdot N_2'(\xi)\, \mathrm{d}\xi \\ \int_0^1 N_2'(\xi) \cdot N_1'(\xi)\, \mathrm{d}\xi & \int_0^1 \left(N_2'(\xi)\right)^2 \mathrm{d}\xi \end{pmatrix} = \frac{1}{h_i} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

The quadrature is only related to the shape functions and thus independent on the special element i: It has to be done only once!

In the next step we place each stiffness matrix on the element level at its position in the (total) stiffness matrix A (assembly or compilation):

$$\begin{pmatrix} 1/h_1 & -1/h_1 & \\ -1/h_1 & 1/h_1 & \\ & & \end{pmatrix} + \begin{pmatrix} 0 & 0 & \\ 0 & 1/h_2 & -1/h_2 \\ & -1/h_2 & 1/h_2 \end{pmatrix} + \ \ldots \ = A$$

and obtain the same stiffness matrix A as in the global approach, but with only 4 quadratures (and many simple transformations).

5.3.4 Triangulation

Let us now generalize our approach to more than 1 dimension.

Before we can define S_h we have to mesh Ω. In the 1D example the mesh is the discretization $x_0 < x_1 < \ldots < x_{N+1}$ of an interval.

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Definition (admissible triangulation)

Let $\Omega \subset \mathbb{R}^d$ be a Lipschitz polytype ($d=2$ polygon, $d=3$ polyhedron).

A partition $\mathcal{T}_h = \{T_1, \ldots, T_N\}$ is called an admissible or regular triangulation, if all subdomains $T_i, i = 1, \ldots, N$, satisfy

- T_i is polygonal/polyhedral domain
- $\bar{\Omega} = \bigcup_{i=1}^{N} \bar{T}_i$ and $T_i \cap T_j = \emptyset$ for $i \neq j$
- $\partial T_i \cap \partial T_j, i \neq j$ is either empty or a common vertex, edge or face (in 3D).

We define h_T as the diameter of the smallest circle/ball that completely contains an element $T \in \mathcal{T}_h$. In the following, we omit the index i and write T instead of T_i.

A set $\{\mathcal{T}_h\}_{h \in \Theta}$ of admissible triangulations is called shape regular, if and only if there exists a constant $0 < c_s$ such that for each $T \in \mathcal{T}_h$

$$\varrho_T \geq c_s h_T \qquad \forall T \in \mathcal{T}_h, \ \forall h \in \Theta$$

where $2\varrho_T$ is the diameter of the largest inscribed ball in T and Θ a set of different values h (e.g. h characterizes the mesh size by $h := \max_{T \in \mathcal{T}_h} h_T$). □

Remark

We try to avoid very small values of c_s. □

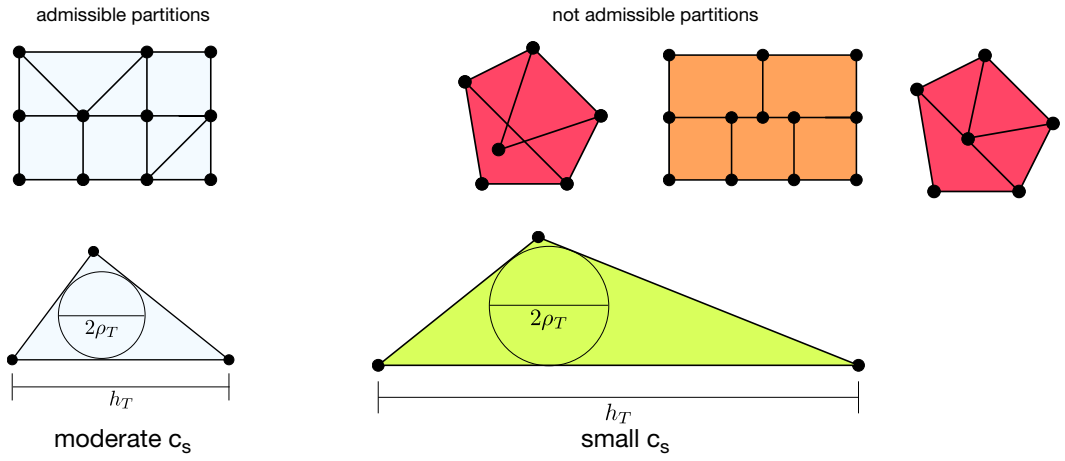


Figure 34: Examples for admissible and non admisible triangulations (upper row) and visualization of the definition of shape regularity (lower row).

For a fixed triangulation $\mathcal{T} := \mathcal{T}_h$ the following notation is used

set of vertices	$\mathcal{V} = \{V_i\}$	
set of edges	$\mathcal{E} = \{E_{ik}\}$	
set ot triangles	$\mathcal{T} = \{T_{ikl}\}$	(for $\mathbb{R}^2$)
set of faces	$\mathcal{F} = \{F_{ikl}\}$	(for $\mathbb{R}^3$)
set of tetrahedra	$\mathcal{T} = \{T_{iklm}\}$	(for $\mathbb{R}^3$)

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

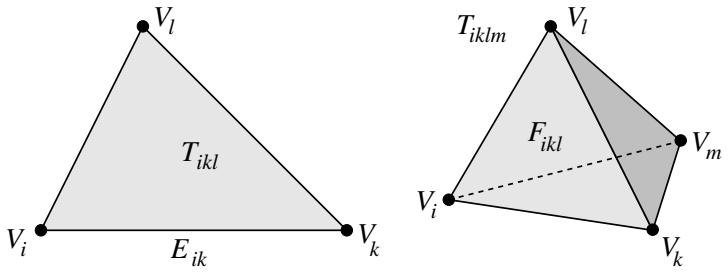


Figure 35: Elements of a fixed triangulation in $\mathbb{R}^2$ and $\mathbb{R}^3$.

5.3.5 Ansatz and Shape Functions in 1D

- Reconsider the 1D problem in chap. 5.3.2. Here we got for the hat functions $\psi_i(x) = 0$ for $x \in [0,1] \backslash]x_{i-1}, x_{i+1}[$ and $\psi_i(x_i) = 1$ (interpolation).

  Using the associated shape functions N_1, N_2 on $[0,1]$ in chap. 5.3.3, we can rewrite the interpolation property using a linear ansatz:
  Determine the linear polynomial $p(\xi) = U_1 \cdot N_1(\xi) + U_2 \cdot N_2(\xi)$ for $\xi \in [0,1]$ that satisfies $p(0) = U_1, p(1) = U_2$.

  For the shape functions we have obtained $N_1(\xi) = \xi, N_2(\xi) = 1 - \xi$.

- Instead of the hat functions we can use piecewise quadratic functions $\psi_i(x), \psi_{i-1/2}(x)$ defined on $\Omega =]0,1[$ such that

$$u_h(x) = \sum_{k=1}^{N} z_k \psi_k(x) + \sum_{i=1}^{N+1} z_{k-1/2} \psi_{k-1/2}(x)$$

  These functions fulfill the following conditions $\left(x_{k-1/2} := (x_{k-1} + x_k)/2\right)$:

$$\begin{aligned} \psi_i &= 0 \ \forall x \in \Omega \backslash]x_{i-1}, x_{i+1}[, \ \psi_i(x_k) = \delta_{ik}, \ \psi_i(x_{k-1/2}) = 0 \\ \psi_{i-1/2}(x) &= 0 \ \forall x \in \Omega \backslash]x_{i-1}, x_i[, \ \psi_{i-1/2}(x_k) = 0, \ \psi_{i-1/2}(x_{k-1/2}) = \delta_{ik} \end{aligned}$$

  Using the associated shape functions N_1, N_2, N_3 on $[0,1]$ in (A54), we can rewrite the interpolation property using a quadratic polynomial as an ansatz: $p(\xi) = U_0 \cdot N_1(\xi) + U_{1/2} \cdot N_2(\xi) + U_1 \cdot N_3(\xi)$ that satisfies the three conditions $p(0) = U_0, p(1/2) = U_{1/2}, p(1) = U_1$

  The shape functions are
  $N_1(\xi) = 2\xi^2 - 3\xi + 1, N_2(\xi) = -4\xi^2 + 4\xi, N_3(\xi) = 2\xi^2 - \xi$.

5.3.6 Linear Ansatz Functions in 2D

Definition

Consider a triangulation $\mathcal{T}_h$ of a BL-domain $\Omega \in \mathbb{R}^2$ by triangles. The space of linear finite elements on $\mathcal{T}_h$ is defined by

$$S_h^1 := \left\{ v \in \mathcal{C}^0(\bar{\Omega}) \,\middle|\, v\big|_T \in P_1(T), T \in \mathcal{T}_h \right\}$$

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

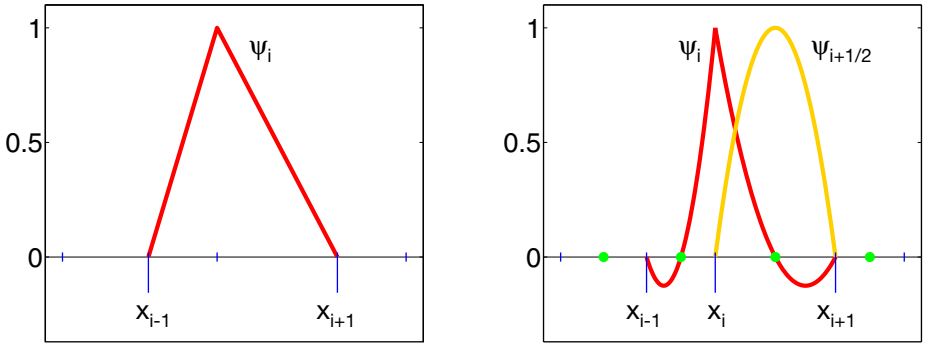


Figure 36: Hat function (le.) and piecewise quadratic functions (ri.)

with the space $P_1(T) := \{a_0 + a_1x_1 + a_2x_2 \,|\, x = (x_1, x_2) \in T\}$ of ansatz functions consisting of linear polynomials on each triangle $T \subset \mathbb{R}^2$. The space S_h^1 is a global ansatz space and is called conformal FE space of order 1.

Remark

An element $v\Big|_T \in P_1(T)$ is uniquely determined by the function values $v(p_i)$, where the $p_i, i = 1, 2, 3$, are the are the vertices of T. P_h is defined as the set of all vertices of $\mathcal{T}_h$.

In general it is not sufficient to check whether $\dim P_h$ is equal to the number of equations of condition! □

Theorem

The interpolation task:

$$\text{Determine } u_h \in S_h^1 \text{ with } u_h(p) = z_p, p \in P_h$$

has a unique solution $\forall\, z_p \in \mathbb{R}$ and $\forall\, p \in P_h$.

Definition

The functions $\psi_p \in S_h^1$ defined by

$$\psi_p(q) := \delta_{pq}\,, \quad q, p \in P_h$$

form a basis of the S_h^1. The ψ_p are again called hat functions and the basis is called *nodal basis*.

Remark

For $u_h \in S_h^1$ we get

$$u_h(x_1, x_2) = \sum_{p \in P_h} z_p \psi_p(x_1, x_2) \;\; \forall\, (x_1, x_2) \in \Omega \quad \text{and} \quad z_p = u_h(p)\,,$$

which solves the interpolation task. □

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

Let us now calculate the $\dim S_h^1$:

Let be n_T the number of triangles in $\mathcal{T}_h$, n_p the number of vertices and $n_T(P)$ the number of triangles, that contain the same point $P \in \mathbb{R}^2$. Because of the continuity we get

$$\begin{aligned} \dim S_h^1 &= 3n_T - \sum_P (n_T(P) - 1) = 3n_T - \sum_P n_T(P) + \sum_P 1 \\ &= 3n_T + n_p - 3n_T = n_P \end{aligned}$$

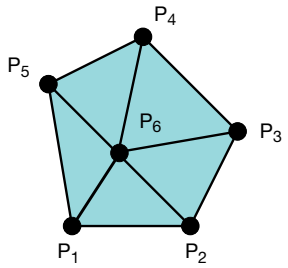


Figure 37: Example: $n_T = 5$, $n_T(P_i) = 2$ for $i = 1,\ldots,5$, $n_T(P_6) = 5$ and $n_p = 6$.

5.3.7 Finite Element

Definition

A Finite Element is defined as a triple (T, Π_T, Σ_T) with the following properties:

1. T is a polyhedron in $\mathbb{R}^d$. The parts of the surface ∂T are called faces.
2. The space of the ansatz functions Π_T is a subspace of $\mathcal{C}^0(T)$ with finite dimension s.
3. $\Sigma_T = \{\sigma_1, \ldots, \sigma_s\}$ is a set of s linearly independent functionals on Π_T, i.e. $\sigma_j : \Pi_T \to \mathbb{R}$.

   Usually the functionals involve point evaluation of a function or its derivatives at points in T.

The unique basis $\{\psi_1, \ldots, \psi_s\}$ of Π_T fulfilling

$$\sigma_i(\psi_k) = \delta_{ik}$$

is called nodal basis. □

■ Example

$T \subset \mathbb{R}^2$ is a (non-degenerate) triangle. $\Pi_T = P_1(T)$ with $\dim P_1(T) = 3$.

$\Sigma_T = \{\sigma_1, \sigma_2, \sigma_3\}$ with $\sigma_i(f) = f(p_i)$; p_i denotes a vertex of the triangle and $f \in \Pi_T$, $i = 1, 2, 3$.

© Rainer Callies TUM 2020

This document is available free of charge on StuDocu.com
Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)

If we choose for the f the hat functions ψ_k from the last chapter, then we get

$$\sigma_i(\psi_k) = \psi_k(p_i) = \delta_{ik}$$

and thus a nodal basis. Thus we see that Σ_T contains the (generalized) interpolation conditions.

Let us choose a reference element T with the vertices $(0,0),(1,0),(0,1)$. Then the local nodal basis is given by the functions

$$\psi_1(x,y) = 1-x-y\,,\;\; \psi_2(x,y) = x\,,\;\; \psi_3(x,y) = y$$

□

Heruntergeladen durch Option Some (shuhao.zhang.x@gmail.com)