Computational PDEs

Andreas Dedner

April 10, 2011

Contents

Acknowledgements

I would like to thank Andrew Stuart who provided me with his script on Computational PDEs. The following script is in major parts taken directly from Andrew's script. I just shortened it a bit and added one or two things.

Andreas Dedner

The following is from Andrew's Acknowledgements and I would also like to say my thanks to everybody who helped with this great script:

I would like to thank the many students and TAs who, through patient questions, improved the content of these notes; and the teaching assistants who have worked with me to produce solutions to exercises; and the individuals who have helped produce the notes, questions, and solutions in printed form. I am especially grateful to Maureen Doyle, Joel Franklin, Chen Greif, Menelaos Karavelos, Jim Lambers, Ofer Levi, Yvo Pokern, Brian Rogoff and Xiaowei Zhan.

Andrew Stuart

Part I

Background

Chapter 1

Model Problems

1.1 Boundary Value Problems

We model the temperature distribution in a heat conducting material occupying the volume $\Omega \subset \mathbb{R}^3$ with boundary Γ. Let $u\colon \Omega \to \mathbb{R}$ denote the temperature at a position $x \in \Omega$ and $q\colon \Omega \to \mathbb{R}^d$ the heat flux in the direction of increasing x_i. The intensity of the heat source is given by $f\colon \Omega \to \mathbb{R}$ and is assumed to be known, while $a > 0$ denotes the heat capacity coefficient. The total heat flux through the boundary S of any given part $V \subset \Omega$ in the outward direction (given by the outward normal n) is equal to the heat produced in V : $\int_S q \cdot n \; ds = \int_V f \; dx$. By the divergence theorem, $\int_S q \cdot n \; ds = \int_V \mathrm{div} q \; dx$ and thus we arrive at $\int_V \mathrm{div} q \; dx = \int_V f \; dx$. This equation holds for every V so that we can conclude that $\mathrm{div} q(x) = f(x)$ pointwise in Ω. Another component of our mathematical model is Fouriers law which states that heat flows from warm to cold regions with the heat flow proportional to the temperature gradient: $q(x) = -a(x)\nabla u(x)$ which leads to the well know *elliptic partial differential equation* for u:

$$-\mathrm{div}\big(a(x)\nabla u(x)\big) \;\; = \;\; f(x) \quad x \in \Omega \; .$$

To obtain an unique solution to this problem conditions for u on the boundary Γ have to be prescribed, so that we obtain a *boundary value problem*. Typical boundary conditions are *Dirichlet* conditions where the temperature u is prescribed and *Neumann* boundary conditions where the heat flux $q \cdot n$ in normal

direction is given:

$$\begin{aligned} u(x) &= g_D(x) \quad x \in \Gamma_D \ , \\ a(x)\nabla u(x) \cdot n(x) &= g_N(x) \quad x \in \Gamma_N \ , \end{aligned}$$

where $\Gamma_D \cup \Gamma_N = \Gamma$, $\Gamma_N \cap \Gamma_D = \emptyset$ and $\Gamma_D \neq \emptyset$.

In this lecture we will study the following model problem which adds advection and reaction to the diffusion process described above: Let $\Omega \subset \mathbb{R}^d$ be a bounded open set. Let $p, q, f \in C(\bar{\Omega}, \mathbb{R})$ and $g \in C(\partial\Omega, \mathbb{R})$. We study the **elliptic equation**

$$\begin{aligned} -\Delta u + p.\nabla u + qu &= f, \quad x \in \Omega, \\ u &= g, \quad x \in \partial\Omega. \end{aligned}$$

In our study of finite difference methods we will study classical solutions of this problem, where all derivatives appearing in the equation exist everywhere in Ω and the equation holds pointwise.

Remark This equation arises as a simplified model in many different areas, for example in *linear elasticity*: here the displacement field $u\colon \Omega \to \mathbb{R}^3$ of an elastic material under some external force has to be found. The main equations are *Newton's second law*: $\partial_t^2 u = \nabla \cdot \sigma + F$, the *strain displacement* relation $\epsilon = \frac{1}{2}(\nabla u + (\nabla u)^T)$ and *Hooke's law*: $\sigma = C\colon \epsilon$. Neglecting any dependency on time, leads to a an equation which is similar to our elliptic model problem. This equation is now for a vector valued function but the methods derived in this lecture can be applied to each component u_1, u_2, u_3.

The model for elastic deformation can also be derived by studying the deformation field minimizing some bending energy. Thus u is given as the function with $E(u) \leq E(v)$ for all $v \in V$. With $E(u) = \frac{1}{2}\int_\Omega C\epsilon(u)\colon \epsilon(u) - F \cdot u$ we can again derive an elliptic equation for u using the variational principals stating that $j(\delta) := \frac{d}{d\delta}E(u + \delta v)$ has to satisfy $j(0) = 0$ for arbitrary $v \in V$, if u minimizes the energy.

1.2 Diffusion

The *Poisson* equation $-\triangle u = f$ models the stationary heat distribution in some material. If we are interested in the evolution of the heat in time subject to some time dependent heat source or boundary conditions, then we have to study the solution $u\colon \Omega \times (0,\infty) \to \mathbb{R}$ of a *parabolic partial differential equation*.

A suitable model problem is the *heat equation* with given data $f, g \in C(\bar{\Omega}, \mathbb{R})$. The equation for u on an open bounded set $\Omega \subset \mathbb{R}^d$ is:

$$\begin{array}{rcll} \frac{\partial u}{\partial t} & = & \Delta u + f & (x,t) \in \Omega \times (0,\infty), \\ u & = & 0 & (x,t) \in \partial\Omega \times (0,\infty), \\ u & = & g & (x,t) \in \bar{\Omega} \times \{0\}. \end{array} \tag{1.1}$$

In our study of finite difference methods we will study classical solutions of this problem, where all derivatives appearing in the equation exist everywhere in $\Omega \times (0,\infty)$ and the equation holds pointwise.

1.3 Wave motion

If we write down the model which we derived for linear elasticity in one space dimension, we arrive at the model problem for a second order *hyperbolic pde*:

$$\partial_t^2 u(x,t) - c^2 \partial_x^2 u(x,t) = f(x,t) \quad x \in (a,b) \times (0,\infty) \ .$$

Again we need to prescribe initial and boundary conditions for u. The *wave equation* models the propagation of waves through some material, e.g., sound waves in air. We will not be studying this equation in this lecture but the methods derived for the other problems can also be applied to model the evolution of waves.

1.4 Schrödinger equation

A further model problem is given by the *periodic Schrödinger equation*: For a $2-$periodic function $g \in C([-1,1],\mathbb{R})$ the **periodic Schrödinger equation**

on $(-1,1)$ is given by:

$$\begin{array}{rcll} \frac{\partial u}{\partial t} & = & i\frac{\partial^2 u}{\partial x^2} & (x,t) \in (-1,1) \times (0,\infty), \\ u(-1,t) & = & u(1,t) & t \in (0,\infty), \\ \frac{\partial u}{\partial x}(-1,t) & = & \frac{\partial u}{\partial x}(1,t) & t \in (0,\infty), \\ u & = & g & (x,t) \in [-1,1] \times \{0\}. \end{array} \tag{1.2}$$

In our study of finite difference methods we will study classical solutions of this problem, where all derivatives appearing in the equation exist everywhere in $(-1,1) \times (0,\infty)$ and the equation holds pointwise.

1.5 Conservation Laws

Many physical laws are based on the conservation of some quantity, e.g., mass. The following equation states that the change of some quantity in some control volume V is determined by some flux over the boundary: $\frac{d}{dt}\int_V u = \int_{\partial V} q \cdot n$. If we can find some law connecting q to u, e.g., $q = f(u)$, we can derive a first order *hyperbolic equation*:

$$\partial_t u + \nabla \cdot f(u) = 0$$

Again we need to prescribe suitable initial and boundary condition to obtain a well posed problem. This simplest case is given by the *transport equation* with $f(u) = au$ for some given $a \in \mathbb{R}$. The **periodic transport problem** is then:

$$\begin{array}{rcll} \frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} & = & 0 & (x,t) \in (-1,1) \times (0,\infty), \\ u(-1,t) & = & u(1,t) & t \in (0,\infty), \\ u & = & g & (x,t) \in [-1,1] \times \{0\}. \end{array} \tag{1.3}$$

for a $2-$ periodic function $c \in C([-,1,1],\mathbb{R}^+)$.

1.6 Non-Linear problems

Note that in contrast to the model problems we looked at so far, the conservation law is a non-linear partial differential equation. The elliptic and parabolic

problems also often have to be studied in non-linear form. So for example the heat capacity a could depend not only on the position in the material but also on the value of the heat itself: $a = a(x, u(x))$. Then the equation for u becomes: $\nabla \cdot a(x, u(x))\nabla u(x) = f(x)$. Not only physics requires the solution of pdes. Many techniques used in computational imaging is based on solving pdes. Denoising of images can for example be based on solving a non-linear equation of the form: $\partial_t I = \nabla \cdot \left(\frac{a\nabla I}{a+\|\nabla I\|}\right)$.
Although the solution of non-linear problems presents many additional challenges, the schemes are often based on methods for the linear problems, so that the study of these forms the central part of any computational pde lecture.

1.7 Numerical Schemes

In most practical applications it is impossible to solve a given PDE analytically, instead some numerical approximation must be used. In general that means reducing the continuous model to a finite dimensional one. This can then be solved using some solver for systems of linear (or non-linear) equations. We will discuss this last step in detail but will focus in this lecture on reformulating a given analytical model as a finite dimensional problem.
There are a number of different approaches for numerically solving partial differential equations. The main ones are *finite difference* (FD) and *finite element* (FE) methods. The main difference between these methods is the form of the mathematical model they use, while the FD method is based on the classical pointwise formulation of the PDE (for example $-\triangle u(x) = f(x)$ for all $x \in \Omega$), the FE method is based on the variational formulation (in the example above that would correspond to $\int_\Omega \nabla u \cdot \nabla \phi = \int_\Omega f\phi$ for all $\forall \phi \in H^1(\Omega)$. The FD are consequently based on the classical spaces $C^m(\Omega)$ of continuous differentiable functions, while the FE methods make use of *Sobolev spaces* $H^m(\Omega)$ which are subsets of $L^(\Omega)$ [1]. For their construction the FD require a great deal of smoothness for both the data and the solutions, while for the construction of the FE mehtods only minimal smoothness is required - although for proofing convergence some additional smoothness is needed. We will start of with FD methods and will cover FE method more briefly in this lecture.

[1] we will not require any knowledge of their construction or properties for this lecture

Chapter 2

Mathematical Basics

2.1 Linear Algebra

2.1.1 Matrix Norms and Spectral Radius

Let $A \in \mathbb{R}^{n\times n}$ and let $\|\cdot\|$ denote a norm on $\mathbb{R}^n$. The expression

$$\|A\| = \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Ax\|}{\|x\|} = \sup_{x\in\mathbb{R}^n,\, \|x\|=1} \|Ax\|$$

defines the matrix norm induced by the vector norm $\|\cdot\|$. Note that, for induced norms,

$$\|Ax\| \leq \|A\|\|x\|.$$

If A is normal, i.e. $A^TA = AA^T$, then we can find an orthonormal basis of eigenvectors, $\{\phi^{(k)}\}_{k=1}^n$ for $\mathbb{R}^n$. We denote the eigenvalues by $\lambda^{(k)}$ so that we have

$$A\phi^{(k)} \;\; = \;\; \lambda^{(k)}\phi^{(k)}.$$

Then the matrix norm induced by the 2-norm can be expressed as the spectral radius:

$$\|A\|_2 = \rho(A) = \max_{k\in\{1,\ldots,n\}} |\lambda^{(k)}| \; .$$

Note that since $\mathbb{R}^{n\times n}$ is finite dimensional, all matrix norms are equivalent.

2.1.2 Eigenvalues of Toeplitz and Related Matrices

We state and prove results about the eigenvalues of various Toeplitz matrices (constant on diagonals) arising frequently in finite difference and finite element approximations of PDEs, as well as for a related tridiagonal matrix, arising in similar contexts.

Circulant Matrix

Let the **circulant** matrix A be defined as follows:

$$A = \begin{pmatrix} a & c & 0 & \dots & 0 & b \\ b & a & c & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & b & a & c \\ c & 0 & \dots & 0 & b & a \end{pmatrix} \in \mathbb{R}^{2J\times 2J} \tag{2.1}$$

Theorem 2.1.1. *Consider the eigenvalue problem*

$$A\phi^{(k)} = \lambda^{(k)}\phi^{(k)}$$

Let $\Delta x = J^{-1}$. *Then, writing* $\phi^{(k)} = (\phi^{(k)}_{-J}, \dots, \phi^{(k)}_{J-1}) \in \mathbb{C}^{2J}$ *we have, for* $k \in \{0, \dots, 2J-1\}$*:*

$$\begin{aligned} \phi^{(k)}_j &= \exp(ik\pi j\Delta x), \\ \lambda^{(k)} &= b\exp(-ik\pi\Delta x) + a + c\exp(ik\pi\Delta x). \end{aligned}$$

Proof. In index form, the eigenvalue problem is, for $\phi = \phi^{(k)}$, $\lambda = \lambda^{(k)}$:

$$b\phi_{j-1} + a\phi_j + c\phi_{j+1} = \lambda\phi_j, \quad j \in \{-J \dots, J-1\} \tag{2.2}$$

with the convention

$$\phi_{-J-1} = \phi_{J-1},\ \phi_J = \phi_{-J}.$$

It may be verified by substitution that choosing $\phi_j = \exp(ik\pi j\Delta x)$ with $\Delta x = J^{-1}$ solves the difference equation (2.2), provided k is an integer and

$$\lambda = b\exp(-ik\pi\Delta x) + a + c\exp(ik\pi\Delta x).$$

Since taking $k \in \{0, \ldots, 2J-1\}$ gives $2J$ distinct eigenvectors, the result follows. □

Remark Both $\lambda^{(k)}$ and $\phi^{(k)}$ are $2J$ periodic in k. Hence choosing $k \notin \{0, \ldots, 2J-1\}$ does not yield any new information. The eigenfunctions $\phi^{(k)}$, with $k \in \{0, \cdots, 2J-1\}$, form an orthogonal basis for $\mathbb{C}^{2J}$. Hence,

$$\|A\|_2 \;=\; \max_{k\in\{0,\ldots,2J-1\}} |\lambda^{(k)}|.$$

Tridiagonal Matrices

Let

$$\begin{pmatrix} a & b & 0 & \ldots & \ldots & 0 \\ b & \ddots & \ddots & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \ddots & \ldots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \ddots & \ddots & b \\ 0 & \ldots & \ldots & 0 & b & a \end{pmatrix} \in \mathbb{R}^{l\times l}.$$

Theorem 2.1.2. *Consider the eigenvalue problem*

$$A\phi^{(k)} = \lambda^{(k)}\phi^{(k)}$$

Let $\Delta x = (l+1)^{-1}$. Then, writing $\phi^{(k)} = (\phi_1^{(k)}, \ldots, \phi_l^{(k)}) \in \mathbb{R}^l$, we have for $k \in \{1, \ldots, l\}$:

$$\begin{aligned} \phi_j^{(k)} &= \sin(k\pi j\Delta x), \\ \lambda^{(k)} &= a + 2b\cos(k\pi\Delta x). \end{aligned}$$

Proof. In index form the eigenvalue problem is, for $\phi = \phi^{(k)}$, $\lambda = \lambda^{(k)}$:

$$b\phi_{j-1} + a\phi_j + b\phi_{j+1} = \lambda\phi_j, \quad j \in \{1, \ldots, l\},$$

with the convention $\phi_0 = \phi_{l+1} = 0$. It may be verified by substitution that choosing $\phi_j = \sin(k\pi j\Delta x)$ with $\Delta x = (l+1)^{-1}$ solves the difference equation provided that k is an integer and

$$\lambda = a + 2b\cos(k\pi\Delta x).$$

Since taking $k \in \{1, \ldots, l\}$ gives l distinct eigenvectors, the result follows. □

Remark Both $\lambda^{(k)}$ and $\phi^{(k)}$ are l periodic in k. Hence choosing $k \notin \{1, \ldots, l\}$ does not yield any new information. The eigenfunctions $\phi^{(k)}$, with $k \in \{1, \ldots, l\}$, form an orthogonal basis for $\mathbb{C}^{2J}$. The eigenfunctions $\phi^{(k)}$ form an orthogonal basis for $\mathbb{C}^l$. Hence,

$$\|A\|_2 \quad = \quad \max_{k\in\{1,\ldots,l\}} |\lambda^{(k)}|.$$

2.1.3 Rayleigh Coefficient

Given a square matrix A the Rayleigh coefficient is defined by

$$R(x) = \frac{(x, Ax)}{(x, x)}.$$

This function plays an important role in the analysis of eigenvalue problems. Note that if x is an eigenvector of A for the eigenvalue λ, then $R(x) = \lambda$. If A is hermitsch, i.e., $A = A^*$ then $(x, Ax) = (A^*x, x) = (Ax, x) = (x, Ax)$. In this case (x, Ax) is real and therefore $R(x)$ is also a real valued function.

If A has a complete set of orthonormal eigenvectors x_i with real eigenvalues λ_i, then for each x we have the representation $x = \sum_i \alpha_i x_i$ and $(x, Ax) = \sum_i \alpha_i(x, \lambda_i x_i) = \sum_{ij} \alpha_i\alpha_j\lambda_i(x_j, x_i) = \sum_i \lambda_i\alpha_i^2$. Furthermore $(x, x) = \sum_i \alpha_i^2$. Therefore we can conclude:

$$\max_i \lambda_i \geq R(x) \geq \min_i \lambda_i \ .$$

This result holds for example for symmetric matrices.

2.1.4 Square Root of Matrices

Let A be a matrix with positive real eigenvalues λ_i and a complete set of orthonormal eigenvectors x_i, then $A = U^{-1}DU$ with some regular matrix T and with $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ being a diagonal matrix with the eigenvalues on the diagonal. We can now define the square root of the matrix A using

$$A^{\frac{1}{2}} = U^{-1}D^{\frac{1}{2}}U$$

where the square root of a diagonal matrix is given by $D^{\frac{1}{2}} = \mathrm{diag}(\sqrt{(}\lambda_1), \ldots, \sqrt{(}\lambda_n))$. Note that with this definition we have $A^{\frac{1}{2}}A^{\frac{1}{2}} = A$. This definition can for example be applied to symmetric, positive definite matrices.

2.2 The Gronwall Lemma

In the numerical analysis of time discretizations the following Gronwall Lemma plays a central role:

Lemma 2.2.1. *Let $z_n \in \mathbb{R}^+$ satisfy*

$$z_{n+1} \leq Cz_n + D, \quad \forall n \geq 0$$

for some $C \geq 0$, $D \geq 0$ and $C \neq 1$. Then

$$z_n \leq D\frac{C^n - 1}{C - 1} + z_0C^n. \tag{2.3}$$

Proof. The proof proceeds by induction on n. Setting $n = 0$ in (2.3) yields

$$z_0 \leq z_0$$

which is obviously satisfied. We now assume that (2.3) holds for a fixed n and

prove that it is true for $n+1$. We have

$$\begin{aligned} z_{n+1} &\leq Cz_n + D \\ &\leq C\left[D\frac{C^n-1}{C-1} + z_0C^n\right] + D \\ &= D\frac{C^{n+1}-C}{C-1} + z_0C^{n+1} + D \\ &= D\left[\frac{C^{n+1}-C}{C-1} + \frac{C-1}{C-1}\right] + z_0C^{n+1} \\ &= D\frac{C^{n+1}-1}{C-1} + z_0C^{n+1} \end{aligned}$$

and the induction is complete. □

This lemma is a discrete analogue of the following continuous Gronwall lemma:

Lemma 2.2.2. *Let $z(t)$ satisfy*

$$z_t \leq az + b,\ z(0) = z_0,$$

for constants a, b. Then for $t \geq 0$

$$z(t) \leq e^{at}z_0 + \frac{b}{a}(e^{at}-1), \quad a \neq 0$$

and

$$z(t) \leq z_0 + bt, \quad a = 0.$$

Remark We will often use partial derivatives like $z_t, \partial_t z, u_x$, or $\partial_x u$ even for functions of only one variable, since the results are also to be applied to functions with more than one argument.

2.3 Three underlying ideas

If faced with some partial differential equation of the form $\mathcal{L}u = f$ (e.g., $\mathcal{L} = \partial_t - \triangle$) a number of different problems have to be solved:

- Find some function space V in which there exists a unique solution u to the problem. Often finding the right function space to use can be the most difficult part.

- Prove some properties of the solution u. Often these properties are motivated by physical considerations, e.g., if u is the concentration of some substance then any solution u to the problem should at least satisfy $u \geq 0$.

- Define some finite dimensional function space V_h and a discrete problem $\mathcal{L}_h u_h = f_h$ with a unique solution in V_h. The subscript h denotes some measure of the quality of the discretization where it is understood that the quality should increases if h is reduced - at the same time the dimension of V_h increases.

- Prove that u_h satisfies some suitable discrete version of the properties derived for u. So for example we could be satisfied with $u_h \geq -h$. Or we could only consider schemes with $u_h \geq 0$.

- Find a norm in which the difference between u_h and u goes to zero as the dimension of V_h goes to infinity (i.e. h tends to zero).

- Write a computer program to solve the finite dimension problem $\mathcal{L}_h u_h = f_h$. Here deriving efficient algorithms is a central task. This can involve quite a bit of additional mathematical analysis.

In the following we will describe three main concepts which are used to prove the convergence of a numerical scheme.

2.3.1 Consistency, Stability and Convergence

We outline an abstract framework in which many of the convergence proofs that we study can be placed. It may be summarized thus:

$$\text{Consistency} + \text{Stability} \Longrightarrow \text{Convergence}.$$

Given an original problem

$$\Phi(u) \;=\; 0$$

for which existence of a locally unique solution u is known, without necessarily being known explicitly, we approximate the problem by a (finite dimensional) problem:

$$\Phi_h(u_h) \;=\; 0. \tag{2.4}$$

Again h denotes some measure for the quality of the approximation.

Consistency refers to the idea that the true solution almost satisfies the approximate equations:

$$\|\Phi_h(u)\| \longrightarrow 0 \;\text{ as }\; h \longrightarrow 0.$$

Stability refers to h-indendent well-posedness: there exists M independent of h such that

$$\|v - w\| \leq M\|\Phi_h(v) - \Phi_h(w)\|$$

for all v, w from a set B of functions containing u and u^h. Note that stability implies uniqueness of the approximate solution, within B. Note also that, if (2.4) is a linear system $Au_h = f$, then stability implies $\|w\| \leq M\|Aw\|$. Hence A is invertible and $\|A^{-1}\|$ is bounded independently of h, in the induced matrix norm. In general if Φ_h is invertible than stability corresponds to Lipschitz continuity of Φ_h^{-1}.

Convergence of the approximate solution, u_h, to the solution of the original problem, u, requires

$$\|u - u_h\| \longrightarrow 0 \;\text{ as }\; h \longrightarrow 0$$

Consistency and stability together imply convergence:

$$\begin{aligned} \|u - u_h\| &\leq M\|\Phi_h(u) - \Phi_h(u_h)\| \\ &= M\|\Phi_h(u)\| \\ &\longrightarrow 0. \end{aligned}$$

Of course the preceeding discussion is not at all precise; we have simply tried to give a flavour of the proofs that will follow. In order to make such abstract

ideas precise the following issues need addressing: (i) we have not defined function spaces and appropriate norms; (ii) often the domain of Φ and Φ_h differ (infinite dimensional versus finite dimensional for example) and it is necessary to project between the two domains to make sense of certain expressions, such as the definition of consistency; (iii) we have said nothing about rates of convergence. Note that the right choice of norm in such analyzes is crucial and affects the analysis in significant ways.

2.3.2 Qualitative Properties and Stability

The method for establishing that the approximation inherits a qualitative property of the original problem is often the same as the method used to prove stability. Hence we will follow the format:

- write down the PDE and a qualitative property;
- provide the approximation and show it maintains the qualitative property;
- establish consistency and stability, implying convergence.

In the last two steps the methods of analysis will be closely related, and will imply a choice of norm for the stability estimates.

2.3.3 Cost and Error

Usually, there is a tradeoff between computational cost and the error incurred by approximation. The goal is to minimise cost per unit error.

One important aspect for the comparison of different numerical schemes is given by the *convergence rate* of the scheme. Often convergence proofs do not merely establish that $\|u - u_h\| \to 0$ but provide a rate of convergence: $\|u - u_h\| \leq Ch^p$ for h small enough. Here C is some constant which can depend on the data or even on the solution u but not on h. The higher p is, the faster is the rate of convergence. We speak of a *first order* scheme if

$p = 1$ of a *second order* scheme if $p = 2$ and so on. In general higher order schemes require more regularity of the data and of the solution u and require more computational time to obtain. Thus the convergence rate is only a first crude tool for comparing numerical schemes. Also the constant C is often not known and can differ greatly between numerical schemes so that a scheme with a small p but also a small C can perform better than a higher order scheme with large C for practically relevant values of h.

Part II

Finite Differences

Chapter 3

Introduction to Finite Difference Methods

3.1 Finite Differences

The basic idea of finite difference methods is to seek approximations to solutions of the PDE on a lattice. To approximate the derivatives appearing in the PDE, differences between lattice values at neighbouring points are used. We introduce the idea by considering functions of a single variable x.

Let $x_j = j\Delta x$, $\Delta x \ll 1$ and consider a smooth function $u : I \to \mathbb{R}$ for some open $I \subset \mathbb{R}$. We set $u_j = u(x_j)$. By Taylor expansion we have:

$$u_{j\pm1} = u_j \pm \Delta x \frac{\partial u}{\partial x}(x_j) + \frac{\Delta x^2}{2}\frac{\partial^2 u}{\partial x^2}(x_j) \pm \frac{\Delta x^3}{6}\frac{\partial^3 u}{\partial x^3}(x_j) + \mathcal{O}(\Delta x^4) \quad (3.1)$$

provided that $u \in C^4(I, \mathbb{R})$, $x_j \in I$ and Δx is sufficiently small. From this we see that

$$\frac{\partial^2 u}{\partial x^2}(x_j) = \frac{u_{j+1} - 2u_j + u_{j-1}}{\Delta x^2} + \mathcal{O}(\Delta x^2) \quad (3.2)$$

$$:= \frac{\delta^2 u_j}{\Delta x^2} + \mathcal{O}(\Delta x^2) \quad (3.3)$$

provided that $u \in C^4(I, \mathbb{R})$, $x_j \in I$ and Δx is sufficiently small.

We can stop the Taylor expansion at different powers of Δx and obtain similar approximations for the first derivatives. For example

$$\frac{\partial u}{\partial x}(x_j) = \frac{u_{j+1} - u_j}{\Delta x} + \mathcal{O}(\Delta x), \tag{3.4}$$

$$:= \frac{\Delta_+ u_j}{\Delta x} + \mathcal{O}(\Delta x) \tag{3.5}$$

and

$$\frac{\partial u}{\partial x}(x_j) = \frac{u_j - u_{j-1}}{\Delta x} + \mathcal{O}(\Delta x), \tag{3.6}$$

$$:= \frac{\Delta_- u_j}{\Delta x} + \mathcal{O}(\Delta x) \tag{3.7}$$

provided that $u \in C^2(I, \mathbb{R})$, $x_j \in I$ and Δx is sufficiently small. With the assumption that u is three times continuously differentiable we can find an improved approximation to the first derivative:

$$\frac{\partial u}{\partial x}(x_j) = \frac{u_{j+1} - u_{j-1}}{2\Delta x} + \mathcal{O}(\Delta x^2), \tag{3.8}$$

$$:= \frac{\Delta_0 u_j}{2\Delta x} + \mathcal{O}(\Delta x^2) \tag{3.9}$$

provided that $u \in C^3(I, \mathbb{R})$, $x_j \in I$ and Δx is sufficiently small.

3.2 Time-stepping

We illustrate the idea of finite difference methods through time-stepping methods for ODEs. Consider the equation

$$\frac{du}{dt} = f(u)$$

and let $U^n \approx u(n\Delta t)$ denote an approximation. Such an approximation can be computed by the following methods:

- $\frac{U^{n+1} - U^n}{\Delta t} = f(U^n)$ – Forward Euler;
- $\frac{U^{n+1} - U^n}{\Delta t} = f(U^{n+1})$ – Backward Euler;

- $\frac{U^{n+1}-U^n}{\Delta t} = \theta f(U^{n+1}) + (1-\theta) f(U^n)$ – θ-method;
- $\frac{U^{n+1}-U^n}{\Delta t} = f(\theta U^{n+1} + (1-\theta)U^n)$ – one-leg θ-method;
- $\frac{U^{n+1}-U^{n-1}}{2\Delta t} = f(U^n)$ – Leap-frog method.

3.3 Norms

Consider a function $u : \Omega \to \mathbb{R}$ with $\Omega \subset \mathbb{R}^d$. When we discretize in space we will obtain lattice approximations U_k where k is a multi-index ranging over a lattice Ω_Δ. We use U to denote the vector obtained from this indexed set of U_k. We use Δ to denote the mesh-spacing. For simplicity we will always use the same mesh-spacing in all dimensions.

When considering maximum principles our topology will be the supremum norm in space and we use the notation

$$\|u\|_\infty = \sup_{x\in\Omega} |u(x))|, \tag{3.10}$$

with Ω differing from example to example.

For the discretization we use the notation

$$\|U\|_\infty = \max_{k\in\Omega_\Delta} |U_k|, \tag{3.11}$$

with Ω_Δ differing from example to example. No confusion should arise from the dual use of the notation $\|\cdot\|_\infty$ since it will always be clear from the context whether a function or a vector is being measured.

When considering energy methods our topology will be the L^2 norm in space and we use the notation

$$\begin{array}{rcl} \langle u, v\rangle & = & \int_\Omega u(x)\bar{v}(x)dx, \\ \|u\|_2 & = & \left(\int_\Omega |u(x))|^2 dx\right)^{\frac{1}{2}}, \end{array} \tag{3.12}$$

with Ω differing from example to example, and the overbar denoting complex conjugate. We will work almost exclusively in the reals, and the complex conjugation will be redundant and then omitted. We will also use the notation

$$\|\cdot\|_{L^2} := \|\cdot\|_2, \ \|\cdot\|_{H^1} := \{\|\cdot\|_{L^2}^2 + \|\nabla\cdot\|_{L^2}^2\}^{\frac{1}{2}}.$$

For the discretization we use the notation

$$\begin{array}{rcl} \langle U, V \rangle & = & \sum_{k \in \Omega_\Delta} \Delta^d U_k \bar{V}_k, \\ \|U\|_2 & = & \left(\sum_{k \in \Omega_\Delta} \Delta^d |U_k|^2 \right)^{\frac{1}{2}}, \end{array} \tag{3.13}$$

with Ω_Δ differing from example to example, and the overbar denoting complex conjugate. As in the continuous case, we will work almost exclusively in the reals, and the complex conjugation will be redundant. No confusion should arise from the dual use of the notation for the inner-product $\langle \cdot, \cdot \rangle$ and the induced norm $\| \cdot \|_2$ since it will always be clear from the context whether a function or a vector is being measured. The scaling of the discrete L^2 norm is chosen so that it looks, formally, like an integral in the limit of small mesh spacing. This then implies that certain infinite dimensional norm inequalities carry over to the finite dimensional setting, with constants which are mesh independent. For example

$$\|U\|_2^2 \leq \max_{k \in \Omega_\Delta} |U_k|^2 \sum_{k \in \Omega_\Delta} \Delta^d.$$

Because of the scaling of the norm the constant $\sum_{k \in \Omega_\Delta} \Delta^d$ behaves likes the volume of Ω in the limit $\Delta \to 0$ and hence may be bounded independently of Δ. This leads to the bound

$$\|U\|_2^2 \leq C(\Omega) \|U\|_\infty^2. \tag{3.14}$$

Chapter 4

Boundary Value Problems and the Maximum Principle

In this chapter we study elliptic problems in dimensions 1 and 2 and their approximation by finite difference methods. We use the maximum principle, and hence the notation (3.10),(3.11).

4.1 Two Point BVPs

4.1.1 The Differential Equation

To introduce ideas in a simple context we study the Elliptic Problem (1.1) in dimension $d = 1$ with $\Omega = (a, b)$. We define

$$\mathcal{L}w(x) := -\frac{d^2w(x)}{dx^2} + p(x)\frac{dw(x)}{dx} + q(x)w(x) \tag{4.1}$$

and then study the two point boundary value problem

$$\begin{aligned} \mathcal{L}u &= f, \quad x \in (a, b), \\ u(a) - \alpha = u(b) - \beta &= 0. \end{aligned} \tag{4.2}$$

We use ideas related to the maximum principle in this chapter, and hence the norm (3.10) with $\Omega = (a, b)$.

We assume that p, q are continuous functions on the closed interval $[a, b]$ and hence that $\|p\|_\infty = P^* < \infty$. When discussing the consistency of the scheme, further regularity of p, q will implicitly be required.

Theorem 4.1.1. *Assume that $q(x) \geq Q_* > 0$. Then any solution of the equation*

$$\begin{aligned} \mathcal{L}w &= f, \quad x \in (a, b), \\ w(a) = w(b) &= 0, \end{aligned} \tag{4.3}$$

satisfies

$$\|w\|_\infty \leq \frac{1}{Q_*}\|f\|_\infty.$$

Proof. If $|w(x)|$ attains its maximum at $x = a$ or $x = b$ then the proof is complete. If it attains its maximum at $c \in (a, b)$ then, without loss of generality, we assume $w(c) > 0$ (otherwise consider $-w(x)$ and modify the following proof accordingly). Since $w(c)$ is an interior maximum we have $w'(c) = 0$ and $w''(c) \leq 0$. Hence,

$$\begin{aligned} Q_*\|w\|_\infty &= Q_* w(c) \\ &\leq q(c)w(c) \\ &= f(c) - p(c)w'(c) + w''(c) \\ &\leq f(c) \\ &\leq \|f\|_\infty. \end{aligned}$$

The result follows. □

Theorem 4.1.2. *Assume that $p, q, f \in C([a, b], \mathbb{R})$ and furthermore that q is strictly positive on $[a, b]$. Then $\exists P^*, Q_*, Q^* \in \mathbb{R}$ such that*

$$\begin{aligned} \|p\|_\infty &\leq P^*, \\ 0 < Q_* \leq q(x) &\leq Q^* \end{aligned}$$

and (4.2) has a unique solution.

Proof. Since the problem is elliptic, the Fredholm alternative shows that it suffices to show uniqueness of the solution, for arbitrary continuous f. Uniqueness follows by noting that, for two solutions $u_1(x)$, $u_2(x)$, the difference $w(x) = u_2(x) - u_1(x)$ solves (4.3) with $f = 0$. Hence Theorem 4.1.1 gives that $\|w\|_\infty = 0$. □

4.1.2 The Approximation

Let $U_j \approx u(x_j)$, $x_j = a + jh$, $Jh = b - a$. Set $p_j = p(x_j)$, $q_j = q(x_j)$ and define the operator $\mathcal{L}^h$ by

$$\mathcal{L}^h W_j = -\frac{\delta^2 W_j}{h^2} + p_j \frac{W_{j+1} - W_{j-1}}{2h} + q_j W_j. \tag{4.4}$$

The discrete analog of (4.2) is

$$\begin{aligned} \mathcal{L}^h U_j &= f(x_j), \quad j \in \{1, \ldots, J-1\}, \\ U_0 - \alpha = U_J - \beta &= 0. \end{aligned} \tag{4.5}$$

In matrix form we have

$$AU = F.$$

Here $U = (U_1, \cdots, U_{J-1})^T$ and

$$A = \frac{1}{h^2} \begin{pmatrix} b_1 & c_1 & & \\ a_2 & \ddots & \ddots & \\ & \ddots & \ddots & c_{J-2} \\ & & a_{J-1} & b_{J-1} \end{pmatrix},$$

where

$$\begin{aligned} a_j &= -[1 + hp(x_j)/2], \quad j = 1, \ldots, J-1, \\ b_j &= [2 + h^2 q(x_j)], \quad j = 1, \ldots, J-1, \\ c_j &= -[1 - hp(x_j)/2], \quad j = 1, \ldots, J-1, \\ F_j &= f(x_j), \ j = 2, ..., J-2 \\ F_1 &= f(x_1) - a_1 \alpha / h^2, F_{J-1} = f(x_{J-1}) - c_{J-1} \beta / h^2. \end{aligned} \tag{4.6}$$

We use the norm (3.11) with $\Omega_\Delta = \{1, \ldots, J-1\}$ and $\Delta = h$. We have the following existence result.

Theorem 4.1.3. *Under the same conditions as Theorem 4.1.2, (4.5) has a unique solution, provided $h \leq 2/P^*$.*

Proof. Since the problem is finite dimensional it suffices, by the Fredholm alternative, to show that, if $f(x) = 0$, $\alpha = \beta = 0$, then (4.5) has the unique

solution $U_j = 0$. We have, in this case,

$$\begin{aligned} a_j U_{j-1} + b_j U_j + c_j U_{j+1} = 0, \quad j \in \{1, \ldots, J-1\}, \\ U_0 = U_J = 0. \end{aligned}$$

Using $h \leq 2/P^*$ so that $1 \pm hp_j/2 \geq 0$ for $j \in \{1, \ldots, J-1\}$, we obtain

$$\begin{aligned} (2 + h^2 Q_*)|U_j| &\leq (2 + h^2 q_j)|U_j| \\ &= |b_j U_j| \\ &\leq |a_j U_{j-1}| + |c_j U_{j+1}| \\ &\leq [1 + hp_j/2]|U_{j-1}| + [1 - hp_j/2]|U_{j+1}|. \end{aligned}$$

Thus

$$(2 + h^2 Q_*)|U_j| \leq 2\|U\|_\infty, \quad j \in \{1, \ldots, J-1\}$$

and hence

$$\|U\|_\infty \leq \frac{2}{2 + h^2 Q_*} \|U\|_\infty$$

which implies $\|U\|_\infty = 0$ as required. □

4.1.3 Convergence

Definition 4.1.4. The difference scheme (4.5) is **stable** in the $\|\cdot\|_\infty$ norm if $\exists M \in \mathbb{R}$, independent of h and $h_c > 0$, such that, for all $w \in \mathbb{R}^{J-1}$

$$\|w\|_\infty \leq M\|Aw\|_\infty \quad \forall h \in (0, h_c).$$

Definition 4.1.5. The difference scheme (4.5) is **consistent** of order m in the $\|\cdot\|_\infty$-norm if $\exists C \in \mathbb{R}$, independent of h, and $h_c > 0$ such that:

$$\|Au - F\|_\infty \leq Ch^m \quad \forall h \in (0, h_c),$$

where

$$u = (u(x_1), \ldots, u(x_{J-1}))^T \in \mathbb{R}^{J-1}.$$

Remarks

1. Note the slight abuse of notation in that we use u to denote both the function $u \in C([a,b],\mathbb{R})$ and the vector $u \in \mathbb{R}^{J-1}$ found by sampling the function on the finite difference grid. This abuse is something we will repeat frequently in the following.

2. The quantity $T = Au - F$ is known as the **truncation error**, found by substituting the true solution into the approximation.

Lemma 4.1.6. *Under the assumptions of Theorem 4.1.3, the difference approximation is stable in the $\|\cdot\|_\infty$-norm, with $M = \frac{1}{Q_*}$.*

Proof. Let $Aw = z$. Then

$$\begin{aligned} a_j w_{j-1} + b_j w_j + c_j w_{j+1} &= h^2 z_j, \quad j \in \{1,\ldots,J-1\}, \\ w_0 = w_J &= 0. \end{aligned}$$

Thus, using $h \le 2/p^*$,

$$\begin{aligned} (2+h^2Q_*)|w_j| &\le (2+h^2 q_j)|w_j| \\ &= |b_j w_j| \\ &\le |a_j w_{j-1}| + |c_j w_{j+1}| + h^2|z_j| \\ &\le (1+hp_j/2)|w_{j-1}| + (1-hp_j/2)|w_{j+1}| + h^2|z_j| \\ &\le 2\|w\|_\infty + h^2\|z\|_\infty \end{aligned}$$

Hence

$$(2+h^2Q_*)\|w\|_\infty \le 2\|w\|_\infty + h^2\|z\|_\infty$$

and the result follows. □

Lemma 4.1.7. *If $u \in C^4([a,b],\mathbb{R})$ then (4.3) is consistent of order 2 in the $\|\cdot\|_\infty$-norm.*

Note that this result requires further regularity of p, q, over and above what has already been assumed, and can then be proved by Taylor series expansion, as explained in Chapter 3.

Theorem 4.1.8. *If $u \in C^4([a,b],\mathbb{R})$ then $\exists K$ independent of h, and $h_c > 0$ such that:*

$$\|U - u\|_\infty \quad \leq \quad Kh^2 \quad \forall h \in (0, h_c).$$

Proof. Let

$$Au \quad = \quad F + T$$

and note that $\|T\|_\infty \leq Ch^2$ by consistency. Also

$$AU \quad = \quad F.$$

Thus $e = u - U$ satisfies

$$Ae \quad = \quad T.$$

By stability we have

$$\|e\|_\infty \leq M\|Ae\|_\infty = M\|T\|_\infty \leq MCh^2$$

as required. □

Remarks

1. Subsequent convergence proofs will all be basically repeating the structure of this proof, although the stability estimate may not be written explicitly – for time-dependent problems it is often the least cubmerson notationally to prove convergence directly from consistency.

2. Note that, for linear problems, stability implies the existence of a unique solution to the finite differece equations, since it implies that A is invertible. But it does much more than this: it gives a mesh-independent bound on A^{-1} in the operator norm induced by the norm used in the definition of stability (see section 2.3.1.)

3. By extending the *stencil* of the method, i.e., including more neighboring values to discretize $\partial_{xx}u$, it is possible to extend the order of consistency of the scheme and thus the order of convergence. One possibility is a 5 point stencil: $\frac{1}{12h^2}(-u(x \pm 2h) + 16u(x \pm h) - 30u(x))$. This finite difference approximation is order four consistent but the different signs used for the neighboring values leads to computationally unfavourable properties so that it is not often used.

4.2 The Laplace Equation

4.2.1 The PDE

Now we study finite difference methods for the Elliptic Problem (1.1) in dimension $d = 2$. For simplicity we set p and q to zero, and consider homogeneous Dirichlet boundary conditions. We define

$$\mathcal{L}w = -\Delta w$$

and consider the Laplace equation

$$\begin{aligned} \mathcal{L}u &= f, \quad x \in \Omega, \\ u &= 0, \quad x \in \partial\Omega. \end{aligned} \tag{4.7}$$

We assume that $f \in C(\Omega, \mathbb{R})$.

Theorem 4.2.1. *If for $w \in C(\Omega, \mathbb{R})$ we have $\mathcal{L}w \leq 0 \quad \forall x \in \Omega$ then*

$$\max_{x\in\Omega} w(x) \leq \max_{x\in\partial\Omega} w(x).$$

Furthermore, if w attains its maximum M at $x \in \Omega$ then $w = M$ on the whole of Ω.

Remarks Taking u as the distribution of heat, the maximum principal states that if in the domain only heat sinks are present ($f \leq 0$) then the highest temperature value is found on the boundary of the domain.

Corollary 4.2.2. *Equation (4.7) has a unique solution.*

Proof. By the preceeding theorem, the difference w of any two solutions satisfies

$$\max_{x\in\Omega} w(x) \leq \max_{x\in\partial\Omega} w(x) = 0.$$

Similarly,

$$\max_{x\in\Omega} -w(x) \leq 0$$

and hence $w = 0$ on the whole of Ω. □

4.2.2 The Approximation

For simplicity we take the domain to be a unit square: $\Omega = (0,1) \times (0,1)$. We set $x \longrightarrow (x,y) \in \mathbb{R}^2$, and $x_j = j\Delta x$, $y_j = j\Delta x$. Let $f_{j,k} = f(x_j, y_k)$. If $U_{j,k} \approx u(x_j, y_k)$ is desired, then a natural approximation to (4.7) is

$$\begin{array}{rcll} -\frac{1}{\Delta x^2}\delta_x^2 U_{j,k} - \frac{1}{\Delta x^2}\delta_y^2 U_{j,k} & = & f_{j,k}, & (j,k) \in \Omega_{\Delta x}, \\ U_{j,k} & = & 0, & (j,k) \in \partial\Omega_{\Delta x}, \end{array} \tag{4.8}$$

where

$$\begin{array}{rcl} \delta_x^2 W_{j,k} & = & W_{j+1,k} - 2W_{j,k} + W_{j-1,k}, \\ \delta_y^2 W_{j,k} & = & W_{j,k+1} - 2W_{j,k} + W_{j,k-1}, \\ \Omega_{\Delta x} & = & \{(j,k) \in \{1,\ldots,J-1\} \times \{1,\ldots,J-1\}\} \\ \overline{\Omega}_{\Delta x} & = & \{(j,k) \in \{0,\ldots,J\} \times \{0,\ldots,J\}\} \backslash \{(0,0),(J,J),(0,J),(J,0)\}, \\ \partial\Omega_{\Delta x} & = & \overline{\Omega}_{\Delta x} \backslash \Omega_{\Delta x}. \end{array} \tag{4.9}$$

Here, $J\Delta x = 1$. Defining

$$\mathcal{L}^{\Delta x} W_{j,k} = \frac{1}{\Delta x^2}[4W_{j,k} - W_{j-1,k} - W_{j+1,k} - W_{j,k-1} - W_{j,k+1}] \tag{4.10}$$

we obtain

$$\begin{array}{rcll} \mathcal{L}^{\Delta x} U_{j,k} & = & f_{j,k}, & (j,k) \in \Omega_{\Delta x}, \\ U_{j,k} & = & 0, & (j,k) \in \partial\Omega_{\Delta x}. \end{array} \tag{4.11}$$

We can write this as

$$AU = F \tag{4.12}$$

with the obvious definitions of U and F.

Theorem 4.2.3. *If* $\mathcal{L}^{\Delta x} W_{j,k} \leq 0, \quad (j,k) \in \Omega_{\Delta x}$ *then*

$$\max_{(j,k)\in\Omega_{\Delta x}} W_{j,k} \leq \max_{(j,k)\in\partial\Omega_{\Delta x}} W_{j,k}.$$

Furthermore, if $W_{j,k}$ *attains its maximum* M *at* $(j,k) \in \Omega_{\Delta x}$ *then* $W_{j,k} = M$ *on the whole of* $\overline{\Omega}_{\Delta x}$.

Proof. If the maximum is attained on the boundary and nowhere in the interior then we are done. If not, then we must prove that, if the maximum if attained

in the interior, $w_{j,k} \equiv M$ in $\overline{\Omega}_{\Delta x}$. Let $(j,k) \in \Omega_{\Delta x}$ and assume that the maximum M is attained here. Then

$$M = W_{j,k} \quad \leq \quad \frac{1}{4}\left[W_{j+1,k} + W_{j-1,k} + W_{j,k+1} + W_{j,k-1}\right].$$

Since M is a maximum this can only hold if W takes the same value M at the four neighbours of $j,k \in \Omega_{\Delta x}$, namely $(j \pm 1, k)$ and $(j, k \pm 1)$. Repeating this argument starting at each of these four neighbours, and recursing, shows that $W_{j,k} \equiv M$ in $\overline{\Omega}_{\Delta x}$. □

Corollary 4.2.4. *The problem (4.11) has a unique solution.*

Proof. See Exercise 4. □

4.2.3 Convergence

We use the norm (3.11) with $\Omega_\Delta = \Omega_{\Delta x}$.

Definition 4.2.5. The difference scheme (4.11) is **stable** in the $\|\cdot\|_\infty$ norm if $\exists M$, independent of Δx, and $\Delta_c > 0$, such that, for all $w \in \mathbb{R}^{(J-1)^2}$,

$$\|w\|_\infty \quad \leq \quad M\|Aw\|_\infty \quad \forall \Delta x \in (0, \Delta x_c).$$

Definition 4.2.6. The difference approximation is **consistent** of order m in the $\|\cdot\|_\infty$-norm if $\exists C$, independent of Δx, and $\Delta x_c > 0$ such that:

$$\|Au - F\|_\infty \leq C\Delta x^m \quad \forall \Delta x \in (0, \Delta x_c)$$

where $u = (\ldots, u(x_j, y_k), \ldots)^T \in \mathbb{R}^{(J-1)^2}$.

Lemma 4.2.7. *The difference approximation (4.11) is stable in the $\|\cdot\|_\infty$-norm.*

Proof. Let $\Phi_{j,k} = \frac{1}{4}\left[(x_j - \frac{1}{2})^2 + (y_k - \frac{1}{2})^2\right]$. Then

$$\begin{aligned} \Phi_{j,k} &\geq 0 \quad \forall (j,k) \in \overline{\Omega}_{\Delta x}; \\ -\mathcal{L}^{\Delta x}\Phi_{j,k} &= 1 \quad \forall (j,k) \in \Omega_{\Delta x}; \\ \Phi_{j,k} &\leq \frac{1}{8} \quad \forall (j,k) \in \partial\Omega_{\Delta x}. \end{aligned}$$

Let $AW = Z$, and $R = \|Z\|_\infty$. We will show that

$$\|W\|_\infty \leq \frac{1}{8}R,$$

which establishes stability.

Note that the linear system $AW = Z$ is equivalent to

$$\begin{array}{rcll} \mathcal{L}^{\Delta x} W_{j,k} & = & Z_{j,k}, & (j,k) \in \Omega_{\Delta x}, \\ W_{j,k} & = & 0, & (j,k) \in \partial\Omega_{\Delta x}. \end{array} \tag{4.13}$$

In writing this lattice version of the linear system we extend $W_{j,k}$ onto the boundary $\partial\Omega_{\Delta x}$ and set it to zero, in order to make sense of the application of $\mathcal{L}^{\Delta x}$ in the interior $\Omega_{\Delta x}$.

Now

$$\mathcal{L}^{\Delta x}(R\Phi_{j,k} \pm W_{j,k}) = -R \pm Z_{j,k} \leq 0.$$

Hence

$$\begin{array}{rcl} \max\limits_{(j,k)\in\Omega_{\Delta x}} (R\Phi_{j,k} \pm W_{j,k}) & \leq & \max\limits_{(j,k)\in\partial\Omega_{\Delta x}} (R\Phi_{j,k} \pm W_{j,k}) \\ & \leq & \dfrac{R}{8} \end{array}$$

But $R\Phi_{j,k} \geq 0$ in $\Omega_{\Delta x}$ and hence

$$\max_{(j,k)\in\Omega_{\Delta x}} (\pm W_{j,k}) \leq \frac{R}{8},$$

completing the proof. □

Define the truncation error

$$T_{j,k} = \mathcal{L}^{\Delta x} u_{j,k} - f_{j,k}$$

for $u_{j,k} = u(j\Delta x, k\Delta x)$.

Lemma 4.2.8. *If $u \in C^4(\overline{\Omega}, \mathbb{R})$ then, for some C independent of Δx,*

$$\max_{(j,k)\in\Omega_{\Delta x}} |T_{j,k}| \leq C\Delta x^2.$$

This can be proved by means of Taylor series exapnsion as explained in the previous chapter. If we define a vector $T = (\dots, T_{j,k}, \dots)^T$ then $T = AU - F$ and the result implies consistency of order 2 in the $\infty-$norm:

$$\|T\|_\infty \leq C\Delta x^2.$$

Theorem 4.2.9. *If $u \in C^4(\overline{\Omega}, \mathbb{R})$ then $\exists K$ independent of Δx, and $\Delta x_c > 0$, such that*

$$||u - U\|_\infty \quad \leq \quad K\frac{\Delta x^2}{8} \quad \forall \Delta x \in (0, \Delta x_c).$$

Proof. We have

$$Ae \quad = \quad T,$$

with $\|T\|_\infty \leq C\Delta x^2$. But the previous lemma gives stability and hence

$$\|e\|_\infty \leq \frac{1}{8}\|Ae\|_\infty = \frac{1}{8}\|T\|_\infty \leq \frac{C}{8}\Delta x^2.$$

□

Remarks

1. By extending the *stencil* of the method, i.e., including more neighboring values to discretize $\partial_{xx}u + \partial_{yy}u$, it is possible to extend the order of consistency of the scheme and thus the order of convergence. One possibility is a *compact* nine point stencil: $\frac{1}{60h^2}(4u(x \pm h, y) + 4u(x, y \pm h) + u(x \pm h, y \pm h) - 20u(x, y))$. In this form the scheme is still only consistent of order two but a minimal change in the evaluation of the right hand side leads to a scheme of order four. Instead of $f(x, y)$ one uses: $f_{j,k} = f(x_j, y_k) + \frac{1}{12}(f(x_{j\pm1}, y_k) + f(x_j, y_{k\pm1}) - 4f(x_j, y_k))$. The proof of this is left as exercise.

Chapter 5

Boundary Value problems and Energy Methods

We continue our study of the Elliptic Problem (1.1). We now use energy methods to study the PDE, its approximation and convergence of the approximation. We use inner-products and norms as given by (3.12), (3.13).

5.1 The Helmholtz Equation

5.1.1 The PDE

We consider (1.1) in the case $d = 2$, with $\Omega = (0,1) \times (0,1)$, with homogeneous Dirichlet boundary conditions. We consider the self-adjoint operator arising when $p \equiv 0$ and $q(x) = \mu$. a constant. This gives rise to the Helmholtz equation

$$\begin{aligned} -\Delta u + \mu u &= f \quad x \in \Omega, \\ u &= 0, \quad x \in \partial\Omega. \end{aligned} \tag{5.1}$$

We will make use of the **Poincaré inequality**:

$$\|\nabla u\|_2^2 \geq C_p^{-1} \|u\|_2^2. \tag{5.2}$$

which holds for functions u with zero boundary conditions. Here C_p^{-1} is the smallest eigenvalue of the Laplacian on Ω, subject to Dirichlet boundary conditions. For the particular domain we are considering we have $C_p^{-1} = 2\pi^2$. More details on this inequality will be given in the third part of the lecture.

Theorem 5.1.1. *Let $f \in L^2(\Omega)$ and $\mu > -C_p^{-1}$. Then (5.1) has a unique solution $u \in H^2(\Omega)$ satisfying*

$$\|u\|_2^2 \;\leq\; \frac{\|f\|_2}{\mu + C_p^{-1}}$$

Proof. Existence follows from the Lax-Milgram lemma. Taking the inner product of (5.1) with u yields:

$$\begin{aligned} \langle u, -\Delta u\rangle + \mu\langle u, u\rangle &= \langle f, u\rangle \\ \Longrightarrow \|\nabla u\|_2^2 + \mu\|u\|_2^2 &\leq \|f\|_2\|u\|_2. \end{aligned}$$

Hence

$$(C_p^{-1} + \mu)\|u\|_2^2 \;\leq\; \|f\|_2\|u\|_2$$

and the result follows. □

Remarks In proofs using the energy method, the main idea is often to multiply the pde with u or some derivative of u.

5.1.2 The Approximation

Using the notation (4.9) and (4.10) from the previous chapter, we consider the following approximation to (5.1):

$$\begin{aligned} \mathcal{L}^{\Delta x}U_{j,k} + \mu U_{j,k} &= f_{j,k}, \quad (j,k) \in \Omega_{\Delta x}, \\ U_{j,k} &= 0, \qquad (j,k) \in \partial\Omega_{\Delta x}. \end{aligned} \tag{5.3}$$

In matrix notation we have

$$(A + \mu I)U \;=\; F.$$

We use the inner product and norm on $\Omega_{\Delta x}$ defined by (3.13) with $d = 2$ and $\Delta = \Delta x$.

Exercise 3 shows that the eigenvalue problem

$$A\Phi^{(k,l)} = \lambda^{(k,l)}\Phi^{(k,l)},$$

has solution

$$\begin{aligned}\Phi^{(k,l)} &= \sin(k\pi x_m)\sin(l\pi y_n),\\ \lambda^{(k,l)} &= \frac{4}{\Delta x^2}\sin^2(k\pi\Delta x/2) + \frac{4}{\Delta x^2}\sin^2(l\pi\Delta x/2),\end{aligned}$$

for $(k,l) \in \{1,\ldots,J-1\} \times \{1,\ldots,J-1\}$. Thus A is positive-definite and has a unique positive-definite square root $A^{\frac{1}{2}}$. Hence

$$\langle u, Au\rangle = \|A^{\frac{1}{2}}u\|_2^2$$

so that

$$\frac{8}{\Delta x^2}\|u\|_2^2 \geq \|A^{\frac{1}{2}}u\|_2^2 \geq C_{p,\Delta x}^{-1}\|u\|_2^2 \tag{5.4}$$

where $C_{p,\Delta x}^{-1} = \lambda^{(1,1)} = \frac{8}{\Delta x^2}\sin^2(\pi\Delta x/2)$. Note that $C_{p,\Delta x} \longrightarrow C_p$ as $\Delta x \longrightarrow 0$ and that the lower inequality is a **discrete Poincaré inequality**; this is since, roughly, application of A corresponds to a discrete second derivative and hence application of $A^{\frac{1}{2}}$ to a discrete first derivative. The upper inequality is an example of an **inverse inequality**: something with no counterpart for continuous functions, and consequently with constants that blow-up with shrinking mesh.

Theorem 5.1.2. *Let $\mu > C_{p,\Delta x}^{-1}$. Then (5.3) has a unique solution satisfying*

$$\|U\|_2 \leq \frac{1}{\mu + C_{p,\Delta x}^{-1}}\|F\|_2$$

Proof. Multiplying the discrete equation with U, we have

$$\begin{aligned}\langle U, AU\rangle + \mu\langle U, U\rangle &= \langle F, U\rangle\\ \implies \|A^{\frac{1}{2}}U\|_2^2 + \mu\|U\|_2^2 &\leq \|F\|_2\|U\|_2.\end{aligned}$$

Hence

$$(C_{p,\Delta x}^{-1} + \mu)\|U\|_2^2 \leq \|F\|_2\|U\|_2.$$

The existence and uniqueness result follows because the bound implies

$$(C_{p,\Delta x}^{-1} + \mu)\|U\|_2 \leq \|F\|_2 = \|(A + \mu I)U\|_2. \tag{5.5}$$

This shows that $(A + \mu I)$ is injective and thus invertible. The same formula, in addition, gives the required bound on the solution. □

5.1.3 Convergence

We define the true solution restricted to the grid by

$$\begin{aligned} u_{j,k} &= u(j\Delta x, k\Delta x), \\ u &= (\ldots, u_{j,k}, \ldots)^T \in \mathbb{R}^{(J-1)^2} \end{aligned}$$

and the truncation error $T \in \mathbb{R}^{(J-1)^2}$ by

$$T = (A + \mu I)u - F.$$

From truncation error analysis we have:

Lemma 5.1.3. *If $u \in C^4(\bar{\Omega}, \mathbb{R})$ then, for some C independent of Δx,*

$$\max_{(j,k)\in\Omega_{\Delta x}} |T_{j,k}| \leq C\Delta x^2.$$

Thus, by (3.14), we have

$$\|T\|_2 \leq C(\Omega)^{\frac{1}{2}}\|T\|_\infty \leq C(\Omega)^{\frac{1}{2}} C\Delta x^2.$$

Remark

We prove convergence directly from this consistency result. Of course we are implicitly proving a stability result in the course of the proof; see Exercise 4.

Theorem 5.1.4. *If $u \in C^4(\bar{\Omega}, \mathbb{R})$ then*

$$\|u - U\|_2 \leq \frac{C\Delta x^2}{C_{p,\Delta x}^{-1} + \mu}$$

Proof. We have

$$\begin{aligned}(A+\mu I)u - F &= T,\\ (A+\mu I)U - F &= 0.\end{aligned}$$

Hence, $(A+\mu I)e = T$. By (5.5) we have, by taking $U = (A+\mu I)^{-1}w$,

$$\|(A+\mu I)^{-1}w\|_2 \leq \frac{\|w\|_2}{(C_{p,\Delta x}^{-1}+\mu)}.$$

Hence

$$\|(A+\mu I)^{-1}\|_2 \leq \frac{1}{(C_{p,\Delta x}^{-1}+\mu)}$$

and so

$$\begin{aligned}\|e\|_2 &= \|(A+\mu I)^{-1}T\|\\ &\leq \|(A+\mu I)^{-1}\|_2\|T\|_2\\ &\leq \frac{C(\Omega)^{\frac{1}{2}}C\Delta x^2}{C_{p,\Delta x}^{-1}+\mu}.\end{aligned}$$

□

Chapter 6

Initial Value Problems and Maximum Principles

In this chapter we study both the Transport Problem (1.3) and the Diffusion Problem (1.1) by means of finite difference approximations. Our primary tool of analysis is the maximum principle and we will use the norms (3.10) and (3.11).

6.1 The Transport Problem

6.1.1 The PDE

Here we study the Transport Problem with constant wave speed $c > 0$ (see (1.3) without the periodic assumption):

$$\begin{array}{rcll} \frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} & = & 0 & (x,t) \in (0,\infty) \times (0,\infty), \\ u & = & g & (x,t) \in [0,\infty) \times \{0\}, \\ u & = & h & (x,t) \in \{0\} \times (0,\infty). \end{array} \tag{6.1}$$

Assume for the moment that $g, h \in C([0,\infty), \mathbb{R})$ (further regularity conditions will be implicit in later developments). Furthermore we assume that they satisfy the compatibility condition $h(0) = g(0)$. By the method of characteristics,

there exists an exact solution:

$$u(x,t) \;=\; \begin{cases} h(t-\frac{x}{c}) & 0 \le x < ct \\ g(x-ct) & ct \le x < \infty \end{cases} \tag{6.2}$$

For computational purposes, the domain is restricted to $(x,t) \in [0,1] \times [0,T]$.

From the exact solution above, we infer:

Theorem 6.1.1.

$$\sup_{t\in[0,T]} \|u(\cdot,t)\|_\infty \;\le\; \max\left\{ \sup_{x\in[0,1]} |g(x)|,\ \sup_{t\in[0,T]} |h(t)| \right\}. \tag{6.3}$$

This well-posedness result can be used to establish uniqueness and continuous dependence of the solution on the data.

6.1.2 The Approximation

We introduce a spatial mesh through the points

$$\begin{aligned} x_j &= j\Delta x \\ J\Delta x &= 1,\ J \in \mathbb{N} \end{aligned}$$

where $\Delta x \ll 1$. We let $U_j(t)$ denote our approximation to $u(x_j,t)$. Thus

$$U_j(t) \;\approx\; u(x_j,t).$$

The spatially approximated transport problem then reads:

$$\begin{aligned} \tfrac{dU_j}{dt} + \tfrac{c}{\Delta x}[U_j - U_{j-1}] &= 0 \qquad (j,t) \in \{1,\dots,J\} \times (0,\infty), \\ U_j(t) &= g(x_j) \quad (j,t) \in \{0,\dots,J\} \times \{0\}, \\ U_j(t) &= h(t) \quad (j,t) \in \{0\} \times (0,\infty). \end{aligned} \tag{6.4}$$

If we denote the approximate solution by the vector $U = (U_1(t),\dots,U_J(t))^T$, and similarly $H(t) = \frac{c}{\Delta x}(h(t),0,\dots,0)^T \in \mathbb{R}^J$ and $G = (g(x_1),\dots,g(x_J))^T$ then the matrix form of the problem reads:

$$\begin{aligned} \frac{dU}{dt} + AU &= H \\ U(0) &= G \end{aligned}$$

where

$$A = \frac{c}{\Delta x}\begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ -1 & 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & -1 & 1 & 0 \\ 0 & \dots & \dots & -1 & 1 \end{pmatrix}.$$

Remarks We have so far derived a *semi discrete method* by only discretizing in space while still being continuous in time. This results in a system of ordinary differential equations for the unknown functions $U_j(t)$. This approach is often called *method of lines*.

Now we discretise time, introducting a temporal mesh, $t_n = n\Delta t$, $N\Delta t = T$, $N \in \mathbb{N}$, and apply the forward Euler method.

$$\begin{array}{rcll} \frac{U_j^{n+1}-U_j^n}{\Delta t} + \frac{c}{\Delta x}\left(U_j^n - U_{j-1}^n\right) & = & 0 & (j,n) \in \{1,\dots,J\} \times \{0,\dots N-1\}, \\ U_j^n & = & g(x_j) & (j,n) \in \{0,\dots,J\} \times \{0\}, \\ U_j^n & = & h(t_n) & (j,n) \in \{0\} \times \{1,\dots,N\}. \end{array}$$

From the matrix point of view this may be written as:

$$\begin{array}{rcc} U^{n+1} & = & (I - \Delta t A)U^n + \Delta t H \\ U^0 & = & G, \end{array} \tag{6.5}$$

where $U^n = (U_1^n, \dots, U_J^n)^T \in \mathbb{R}^J$. Here, we assume that $h(0) = g(0)$ so that, when the method requires knowledge of $h(0)$, it is equal to the value of $g(0)$.

Define the **Courant number** as follows:

$$\lambda = \frac{c\Delta t}{\Delta x}. \tag{6.6}$$

We define the norm (3.11) with $\Delta = \Delta x$ and $\Omega_\Delta = \{1,\dots,J\}$.

Theorem 6.1.2. *If* $\lambda \in [0,1]$ *then*

$$\max_{n\in\{0,\dots,N\}} \|U^n\|_\infty \leq \max\left\{\sup_{x\in[0,1]} |g(x)|, \sup_{t\in[0,T]} |h(t)|\right\}. \tag{6.7}$$

Proof. First for all points which are not adjacent to the boundary we have:

$$U_j^{n+1} \;=\; (1-\lambda)U_j^n + \lambda U_{j-1}^n,\ (j,n) \in \{1,\ldots,J\} \times \{0,\ldots,N-1\}$$

Since $\lambda \in [0,1]$ we have $1-\lambda \in [0,1]$ too. Thus, for these (j,n),

$$\begin{aligned} |U_j^{n+1}| &\leq (1-\lambda)|U_j^n| + \lambda|U_{j-1}^n| \\ &\leq (1-\lambda) \max_{j\in\{0,\ldots,J\}} |U_j^n| + \lambda \max_{j\in\{0,\ldots,J\}} |U_j^n| \\ &= \max_{j\in\{0,\ldots,J\}} |U_j^n|. \end{aligned}$$

Now this turns into a statement about the maximum norms. Taking account of the boundary conditions we get:

$$\|U^{n+1}\|_\infty \;\leq\; \max\left\{ \|U^n\|_\infty,\ \sup_{t\in[0,T]} |h(t)| \right\}.$$

It follows by induction that, for $n \in \{0,\ldots,N\}$,

$$\|U^n\|_\infty \;\leq\; \max\left\{ \sup_{x\in[0,1]} |g(x)|,\ \sup_{t\in[0,T]} |h(t)| \right\}$$

since

$$\|U^0\|_\infty \;\leq\; \sup_{x\in[0,1]} |g(x)|.$$

□

6.1.3 Convergence

Let $u_j^n = u(x_j, t_n)$, noting that U_j^n is the computed approximation to u_j^n. Let

$$T_j^n = \left(\frac{u_j^{n+1} - u_j^n}{\Delta t} \right) + \frac{c}{\Delta x} \left(u_j^n - u_{j-1}^n \right) \quad (j,n) \in \{1,\ldots,J\} \times \{0,\ldots,N-1\}. \tag{6.8}$$

This is the truncation error found by substituting the true solution into the method. If we define $T^n = (T_1^n, \ldots, T_J^n) \in \mathbb{R}^J$ then we have, by the techniques of Chapter 3:

Lemma 6.1.3. *If* $u(x,t) \in C^2([0,1]\times[0,T],\mathbb{R})$ *then*

$$\max_{n\in\{0,\ldots,N-1\}} \|T^n\|_\infty \leq C(\Delta t + \Delta x)$$

for some constant C *independent of* Δt *and* Δx.

Remarks

1. By the exact solution (6.2), we see that the regularity assumption on u requires that $g, h \in C^2([0,T],\mathbb{R})$ as well as a C^2-continuity condition where h and g meet.
2. The previous lemma establishes consistency. The method used to prove the next theorem implicitly relies on a stability estimate. See Exercise 6.

Theorem 6.1.4. *If* $\lambda \in [0,1]$ *and* $u(x,t) \in C^2([0,1]\times[0,T],\mathbb{R})$ *then*

$$\max_{n\in\{0,\ldots,N\}} \|u^n - U^n\|_\infty \leq CT(\Delta t + \Delta x) \tag{6.9}$$

for the constant C *appearing in Lemma 6.1.3.*

Proof. Let $e_j^n = u_j^n - U_j^n$. Then, by linearity, we have:

$$\begin{array}{rcll} \frac{e_j^{n+1}-e_j^n}{\Delta t} + \frac{c}{\Delta x}\left(e_j^n - e_{j-1}^n\right) &=& T_j^n & (j,n) \in \{1,\ldots,J\}\times\{0,\ldots,N-1\}, \\ e_j^n &=& 0 & \{0,\ldots,J\}\times\{0\}s, \\ e_j^n &=& 0 & (j,n) \in \{0\}\times\{1,\ldots,N\}. \end{array}$$

Thus, for $(j,n) \in \{1,\ldots,J\}\times\{0,\ldots,N-1\}$, we have:

$$\begin{array}{rcl} e_j^{n+1} &=& (1-\lambda)e_j^n + \lambda e_{j-1}^n + \Delta t T_j^n \\ \implies |e_j^{n+1}| &\leq& (1-\lambda)|e_j^n| + \lambda|e_{j-1}^n| + \Delta t|T_j^n| \\ &\leq& (1-\lambda)\max\limits_{j\in\{0,\ldots,J\}}|e_j^n| + \lambda\max\limits_{j\in\{0,\ldots,J\}}|e_j^n| + \Delta t|T_j^n| \\ &=& \max\limits_{j\in\{0,\ldots,J\}}|e_j^n| + \Delta t|T_j^n| \end{array}$$

Since $e_0^n = 0$, $n \in \{0,\ldots,N\}$ we obtain, using Lemma 6.1.3:

$$\|e^{n+1}\|_\infty \leq \|e^n\|_\infty + C\Delta t(\Delta t + \Delta x)$$

By induction, using $e^0 = 0$, we obtain:

$$\|e^n\|_\infty \leq Cn\Delta t(\Delta t + \Delta x)$$

and hence that

$$\max_{n\in\{0,\ldots,N\}} \|e^n\|_\infty \leq CT(\Delta t + \Delta x)$$

□

6.2 The Heat Equation

6.2.1 The PDE

Here we study the Diffusion Model Problem (1.1) in dimension $d = 1$, with $\Omega = (0, 1)$ and with $f = 0$. Let $g \in C([0, 1], \mathbb{R})$ and consider the problem:

$$\begin{array}{rcll} \frac{\partial u}{\partial t} & = & \frac{\partial^2 u}{\partial x^2} & (x,t) \in (0,1) \times (0,\infty), \\ u & = & 0 & (x,t) \in \{0,1\} \times (0,\infty), \\ u & = & g & (x,t) \in [0,1] \times \{0\}. \end{array} \tag{6.10}$$

Theorem 6.2.1. *Let $t > s > 0$. Then:*

$$\min_{0\leq y\leq 1} u(y,s) \leq u(x,t) \leq \max_{0\leq y\leq 1} u(y,s) \quad \forall x \in [0,1] \tag{6.11}$$

This well-posedness result can be used to establish uniqueness and continuous dependence of the solution on the initial data.

6.2.2 The Approximation

Again, letting $x_j = j\Delta x$ and $J\Delta x = 1$, $J \in \mathbb{N}$, we introduce $U_j(t)$ as our approximation to $u(x_j, t)$ and consider the approximation

$$\begin{array}{rcll} \frac{dU_j}{dt} & = & \frac{1}{\Delta x^2}\delta^2 U_j & (j,t) \in \{1,\ldots,J-1\} \times [0,T], \\ U_j & = & 0, & (j,t) \in \{0,J\} \times [0,T], \\ U_j & = & g(x_j) & (j,t) \in \{0,J\} \times \{0\}. \end{array}$$

Adopting vector notation ($U = (U_1, \ldots, U_{J-1})^T$), we have:

$$\begin{aligned} \frac{dU}{dt} + AU &= 0, \\ U(0) &= G \end{aligned}$$

where

$$\begin{aligned} A &= \frac{-1}{\Delta x^2} \begin{pmatrix} -2 & 1 & & & \\ 1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & -2 \end{pmatrix}, \\ G &= (g(x_1), \ldots, g(x_{J-1}))^T. \end{aligned}$$

Letting $t_n = n\Delta t$ and $N\Delta t = T$, $N \in \mathbb{N}$, we now apply the theta method (with $\theta \in [0,1]$) in time to get:

$$\begin{array}{rcll} \frac{U_j^{n+1} - U_j^n}{\Delta t} &=& \frac{\theta}{\Delta x^2}\delta^2 U_j^{n+1} + \frac{1-\theta}{\Delta x^2}\delta^2 U_j^n & (j,n) \in \{1, \ldots, J-1\} \times \{0, \ldots, N-1\} \\ U_j^n &=& 0 & (j,n) \in \{0, J\} \times \{1, \ldots, N\} \\ U_j^n &=& g(x_j) & (j,n) \in \{0, \ldots, J\} \times \{0\} \end{array} \tag{6.12}$$

Vectorially, we have, for $U^n = (U_1^n, \ldots, U_{J-1}^n)^T$

$$\begin{aligned} (I + \Delta t\theta A)U^{n+1} &= (I - \Delta t(1-\theta)A)U^n \\ U^0 &= G \end{aligned} \tag{6.13}$$

Theorem 6.2.2. *Let $r = \frac{\Delta t}{\Delta x^2}$. If $r(1-\theta) \leq \frac{1}{2}$ then, for $j \in \{0, \ldots, J\}$,*

$$\min_{k \in \{0,\ldots,J\}} U_k^n \leq U_j^{n+1} \leq \max_{k \in \{0,\ldots,J\}} U_k^n.$$

Proof. We have, in the interior $(j,n) \in \{1, \ldots, J-1\} \times \{0, \ldots, N-1\}$,

$$(1 + 2r\theta)U_j^{n+1} = r\theta(U_{j-1}^{n+1} + U_{j+1}^{n+1}) + r(1-\theta)(U_{j-1}^n + U_{j+1}^n) + [1 - 2r(1-\theta)]U_j^n$$

Thus, define

$$U_{\max}^n = \max_{j \in \{0,\ldots,J\}} U_j^n.$$

The upper inequality simply states that

$$U^{n+1}_{\max} \leq U^n_{\max}$$

and it is this that we now prove.

Note that $1 - 2r(1-\theta) \geq 0$, $(1-\theta) \geq 0$, $\theta \geq 0$ and $r \geq 0$. For (j,n) in the interior,

$$\begin{aligned}(1+2r\theta)U^{n+1}_j &\leq 2r\theta U^{n+1}_{\max} + 2r(1-\theta)U^n_{\max} + [1-2r(1-\theta)]U^n_{\max} \\ &= 2r\theta U^{n+1}_{\max} + U^n_{\max}\end{aligned}$$

If $U^{n+1}_{\max}$ occurs for $j \in \{1,\ldots,J-1\}$ then

$$\begin{aligned}(1+2r\theta)U^{n+1}_{\max} &\leq 2r\theta U_{\max} + U^n_{\max} \\ \Longrightarrow U^{n+1}_{\max} &\leq U^n_{\max}\end{aligned}$$

If $U^{n+1}_{\max}$ occurs for $j \in \{0, J\}$ then

$$U^{n+1}_{\max} = 0 = U^n_0 \leq U^n_{\max}$$

Hence

$$U^{n+1}_{\max} \leq U^n_{\max} \quad n \in \{0,\ldots,N-1\}$$

and the upper inequality follows. The lower inequality is proved similarly, by considering

$$U^n_{\min} = \min_{j\in\{0,\ldots,J\}} U^n_j$$

□

6.2.3 Convergence

Let $u^n_j = u(x_j, t_n)$, noting that U^n_j is the computed approximation to u^n_j. Let

$$T^n_j = \frac{u^{n+1}_j - u^n_j}{\Delta t} - \frac{\theta}{\Delta x^2}\delta^2 U^{n+1}_j - \frac{(1-\theta)}{\Delta x^2}\delta^2 U^n_j \quad (j,n) \in \{1,\ldots,J-1\}\times\{0,\ldots,N-1\}$$

This is the truncation error. We define $T^n = (T^n_1,\ldots,T^n_{J-1}) \in \mathbb{R}^{J-1}$. By the techniques in Chapter 3, expanding about $\frac{t_{n+1}+t_n}{2}$, the following may be shown, using the norm (3.11) with $\Delta x = \Delta$ and $\Omega_{\Delta x} = \{1,\ldots,J-1\}$.

Lemma 6.2.3. *If* $u(x,t) \in C^{4\times 2}([0,1]\times[0,T],\mathbb{R})$ *then*

$$\max_{n\in\{0,\ldots,(N-1)\}} \|T^n\|_\infty \leq C[\Delta t + \Delta x^2]$$

for some constant C *independent of* Δt *and* Δx*. Also, if* $\theta = \frac{1}{2}$ *and* $u(x,t) \in C^{4\times 3}([0,1]\times[0,T],\mathbb{R})$ *then*

$$\max_{n\in\{0,\ldots,(N-1)\}} \|T^n\|_\infty \leq C[\Delta t^2 + \Delta x^2]$$

for some constant C *independent of* Δt *and* Δx*.*

Assumption For simplicity we assume that

$$\max_{n\in\{0,\ldots,(N-1)\}} \|T^n\|_\infty \leq C[\Delta t^2 + (1-2\theta)\Delta t + \Delta x^2]$$

in the following. The Lemma 6.2.3 gives conditions under which this holds.

Remark

We prove convergence directly from this consistency result. Of course we are implicitly proving a stability result in the course of the proof; see Exercise 7.

Theorem 6.2.4. *Let* $r(1-\theta) \leq \frac{1}{2}$ *and make the above assumption. Then*

$$\max_{n\in\{0,\ldots,N\}} \|u^n - U^n\|_\infty \leq CT\left[(1-2\theta)\Delta t + \Delta t^2 + \Delta x^2\right].$$

Proof. Let $e_j^n = u_j^n - U_j^n$. Then, by linearity,

$$\frac{e_j^{n+1}-e_j^n}{\Delta t} = \frac{\theta}{\Delta x^2}\delta^2 e_j^{n+1} + \frac{1-\theta}{\Delta x^2}\delta^2 e_j^n + T_j^n \quad (j,n)\in\{1,\ldots,J-1\}\times\{0,\ldots,N-1\}$$

Thus, for $(j,n)\in\{1,\ldots,J-1\}\times\{0,\ldots,N-1\}$ we get:

$$(1+2r\theta)e_j^{n+1} = r\theta(e_{j-1}^{n+1}+e_{j+1}^{n+1}) + r(1-\theta)(e_{j-1}^n+e_{j+1}^n) + [1-2r(1-\theta)]e_j^n + \Delta t T_j^n$$

with

$$\begin{aligned} e_0^n &= e_J^n = 0, \\ e_j^0 &= 0,\ j = 1,\ldots,J-1. \end{aligned}$$

We define

$$E^n = (e_1^n, \ldots, e_{J-1}^n)^T,\ T^n = (T_1^n, \ldots, T_{J-1}^n)^T.$$

Since $1 - 2r(1-\theta) \geq 0$, $(1-\theta) \geq 0$, $\theta \geq 0$ and $r \geq 0$ we have

$$(1+2r\theta)|e_j^{n+1}| \leq 2r\theta\|E^{n+1}\|_\infty + 2r(1-\theta)\|E^n\|_\infty + [1-2r(1-\theta)]\|E^n\|_\infty + \Delta t\|T^n\|_\infty$$

Hence, for $j \in \{1, \ldots, J-1\}$

$$(1+2r\theta)|e_j^{n+1}| \leq 2r\theta\|E^{n+1}\|_\infty + \|E^n\|_\infty + \Delta t\|T^n\|_\infty$$

which implies

$$\begin{aligned} (1+2r\theta)\|E^{n+1}\|_\infty &\leq 2r\theta\|E^{n+1}\|_\infty + \|E^n\|_\infty + \Delta t\|T^n\|_\infty \\ \implies \|E^{n+1}\|_\infty &\leq \|E^n\|_\infty + \Delta t\|T^n\|_\infty. \end{aligned}$$

By induction, using $\|E^0\|_\infty = 0$ and $0 \leq n\Delta t \leq T$, we obtain

$$\|E^n\|_\infty \leq n\Delta t\|T^n\|_\infty \leq T\|T^n\|_\infty$$

as required. □

Chapter 7

Initial Value Problems and Energy Methods

In this chapter we study the qualitative properties and then convergence of discretizations of PDEs by means of energy methods. When studying specific examples of PDEs and their approximations we use the notations (3.12) and (3.13).

To introduce the main ideas consider an ODE in a Hilbert space $\mathcal{H}$ with inner product $(\cdot,\cdot)$ and induced norm $\|\cdot\|$

$$\frac{du}{dt} \;=\; f(u) \tag{7.1}$$

where

$$(f(u),u)=0 \quad \forall u\in D(f)\subset\mathcal{H}. \tag{7.2}$$

Here, f is assumed to satisfy:

$$f : D(f) \;\longrightarrow\; \mathcal{H}$$

where $D(f)\subset\mathcal{H}$ is the domain of f.

Assuming that the solution lies in $D(f)$, this equation has the qualitative property that

$$\|u(t)\|^2 \;=\; \|u(0)\|^2 \tag{7.3}$$

since

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|u\|^2 &= \frac{1}{2}\frac{d}{dt}(u,u)\\ &= (\frac{du}{dt},u)\\ &= (f(u),u)\\ &= 0.\end{aligned}$$

If (7.1) is a PDE then it may be important to replicate the property (7.2) under spatial approximation; we address this on a case by case basis in subsequent sections. Analogously, when discretizing in time, it may also be important to preserve the property (7.3). Here we discuss briefly how time discretization affects (7.3), which is a consequence of (7.3).

Applying the one-leg variant of the θ-method to (7.1) gives

$$\frac{U^{n+1}-U^n}{\Delta t} = f(\theta U^{n+1}+(1-\theta)U^n). \tag{7.4}$$

Taking the inner product with

$$\theta U^{n+1}+(1-\theta)U^n = \frac{1}{2}\left(U^{n+1}+U^n\right)+(\theta-\frac{1}{2})(U^{n+1}-U^n) \tag{7.5}$$

gives, by (7.2):

$$\frac{1}{2\Delta t}\left\{\|U^{n+1}\|^2-\|U^n\|^2\right\}+\frac{1}{\Delta t}(\theta-\frac{1}{2})\|U^{n+1}-U^n\|^2 = 0 \tag{7.6}$$

Thus, $\|U^n\|^2 = \|U^0\|^2$ only when $\theta = \frac{1}{2}$. If $\theta \in [0,\frac{1}{2})$ then $\|U^n\|^2$ is an increasing sequence, and it is decreasing if $\theta \in (\frac{1}{2},1]$.

Similar methods of analysis apply to problems for which (7.2) is replaced by

$$\exists \alpha,\beta \geq 0:\ (f(u),u) \leq \alpha-\beta\|u\|^2 \quad \forall u \in D(f) \subset \mathcal{H}. \tag{7.7}$$

Under (7.7) with $\alpha = \beta = 0$ the norm of the true solution is non-increasing. It is of interest to find discretizations which preserve this property, without restriction on Δt; this requires $\theta \in [\frac{1}{2},1]$.

Under (7.7) with $\alpha, \beta > 0$ the true solution is ultimately bounded independent of initial data. It is of interest to find discretizations which preserve this property, without restriction on Δt; this also requires $\theta \in [\frac{1}{2}, 1]$.

For time-discrete problems we will frequently take the inner-product with (7.5) in the following; this will help us to analyse discretizations of PDEs satisfying (7.2) or (7.7).

7.1 The Transport Problem

7.1.1 The PDE

Here we study the Periodic Transport Problem (1.3) with constant wave speed $c > 0$:

$$\begin{array}{rcll} \frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} & = & 0 & (x,t) \in (-1,1) \times (0,\infty), \\ u(-1,t) & = & u(1,t) & t \in (0,\infty), \\ u(x,0) & = & g & (x,t) \in [-1,1] \times \{0\}. \end{array} \tag{7.8}$$

We use the inner-product and norm given by (3.12) with $\Omega = (-1,1)$.

Theorem 7.1.1.

$$\|u(t)\|_2^2 \;=\; \|u(0)\|_2^2 \quad \forall t > 0$$

Proof.

$$\begin{aligned} \langle u, \frac{\partial u}{\partial t} \rangle + c\langle u, \frac{\partial u}{\partial x} \rangle &= 0 \\ \implies \frac{1}{2}\frac{d}{dt}\|u\|_2^2 + \frac{c}{2}\int_{-1}^{1} \frac{\partial}{\partial x}\{u^2\}dx &= 0 \\ \implies \frac{1}{2}\frac{d}{dt}\|u\|_2^2 + \frac{c}{2}\left[u^2(1,t) - u^2(-1,t)\right] &= 0 \\ \implies \frac{1}{2}\frac{d}{dt}\|u\|_2^2 &= 0 \end{aligned}$$

and the result follows. □

7.1.2 A First Approximation

Now consider the upwind spatial finite difference approximation

$$\begin{aligned} \frac{dU_j}{dt} + \frac{c}{\Delta x}(U_j - U_{j-1}) &= 0 \quad (j,t) \in \{-J, \ldots, J-1\} \times (0,T], \\ U_{-J-1}(t) &= U_{J-1}(t) \quad t \in (0,T], \\ U_j(t) &= g(x_j) \quad (j,t) \in \{-J, \ldots, J-1\} \times \{0\}. \end{aligned}$$

We assume that g is $2-$periodic so that $g(x_{-J-1}) = g(x_{J-1})$ which is needed in the first time-step.

In matrix form we have:

$$\begin{aligned} \frac{dU}{dt} + AU &= 0 \\ U(0) &= G \end{aligned}$$

where $U = (U_{-J}, \ldots, U_{J-1})^T \in \mathbb{R}^{2J}$ and $G = (g(x_{-J}), \ldots, g(x_{J-1}))^T \in \mathbb{R}^{2J}$ and

$$A = \frac{c}{\Delta x} \begin{pmatrix} 1 & 0 & \ldots & \ldots & -1 \\ -1 & 1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \ldots & -1 & 1 & 0 \\ 0 & \ldots & \ldots & -1 & 1 \end{pmatrix}$$

Theorem 7.1.2.

$$\|U(t) - \bar{U}e\|_2^2 \longrightarrow 0 \text{ as } t \longrightarrow \infty,$$

where $\bar{U} = \frac{\Delta x}{2} \sum_{j=-J}^{J-1} U_j(0)$ *and* $e = (1, \ldots, 1)^T \in \mathbb{R}^{2J-1}$.

Proof. By Theorem 2.1.1 A has a set of orthogonal eigenfunctions $\phi^{(k)}$ and eigenvalues $\lambda^{(k)}$ given by

$$\begin{aligned} \phi^{(k)} &= (\phi_{-J}^{(k)}, \ldots, \phi_{J-1}^{(k)})^T \in \mathbb{C}^{2J} \\ \lambda^{(k)} &= [1 - \exp(-ik\pi\Delta x)]\frac{c}{\Delta x} \\ \phi^{(k)} &= \exp(ik\pi j \Delta x), \end{aligned}$$

for $k \in \{0, \cdots, 2J-1\}$. Notice that

$$\begin{aligned} \mathrm{Re}\,\{\lambda^{(k)}\} &= [1-\cos(k\pi\Delta x)]\tfrac{c}{\Delta x} > 0 \quad k \neq 0, \\ \lambda^{(0)} &= 0, \\ e &= \phi^{(0)}. \end{aligned} \tag{7.9}$$

The result follows by expanding U in the orthogonal basis $\{\phi^{(k)}\}_{k=0}^{2J-1}$. To see this let

$$U(t) = \sum_{k=0}^{2J-1} a_k(t)\phi^{(k)}$$

and observe that

$$\frac{da_k}{dt} + \lambda^{(k)} a_k = 0.$$

By (7.9) we deduce that

$$\begin{aligned} a_k(t) &\longrightarrow 0 \text{ as } t \longrightarrow 0 \quad \forall k \neq 0 \\ a_0(t) &= a_0(0) \quad \forall t \geq 0 \end{aligned}$$

But

$$a_0(0) = \frac{\langle U(0), \phi^{(0)} \rangle}{\|\phi^{(0)}\|_2^2} = \frac{\sum_{j=-J}^{J-1} \Delta x U_j(0)}{\sum_{j=-J}^{J-1} \Delta x} = \frac{1}{2J} \sum_{j=-J}^{J-1} U_j(0) = \frac{\Delta x}{2} \sum_{j=-J}^{J-1} U_j(0).$$

□

7.1.3 An Energy Conserving Approximation

The preceeding spatial approximation completely destroys the energy preserving property of the PDE. Here we describe a method which retains this property. The key is to discretize the spatial derivative in a symmetric fashion. Consider the scheme

$$\begin{aligned} \tfrac{dU_j}{dt} + \tfrac{c}{2\Delta x}(U_{j+1} - U_{j-1}) &= 0 && (j,t) \in \{-J, \ldots, J-1\} \times (0,T], \\ U_{-J-1}(t) &= U_{J-1}(t) && t \in (0,T], \\ U_J(t) &= U_{-J}(t) && t \in (0,T], \\ U_j(t) &= g(x_j) && (j,t) \in \{-J, \ldots, J-1\} \times \{0\}. \end{aligned} \tag{7.10}$$

In matrix form we have

$$\begin{array}{rcl} \frac{dU}{dt} + AU & = & 0 \\ U(0) & = & G \end{array} \tag{7.11}$$

with

$$A = \frac{c}{2\Delta x} \begin{pmatrix} 0 & 1 & \dots & \dots & -1 \\ -1 & 0 & 1 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & -1 & 0 & 1 \\ 1 & \dots & \dots & -1 & 0 \end{pmatrix} \tag{7.12}$$

In the following we use the norm and inner-product given by (3.13) with $\Delta = \Delta x$ and $\Omega_\Delta = \{-J, \dots, J-1\}$. With this inner-product A is skew-symmetric: since $A = -A^T$ we have, for all $v \in \mathbb{R}^{2J}$,

$$\langle v, Av \rangle = \langle A^T v, v \rangle = -\langle Av, v \rangle = -\langle v, Av \rangle$$

and so

$$\langle v, Av \rangle = 0 \quad \forall v \in \mathbb{R}^{2J}.$$

Theorem 7.1.3.

$$\|U(t)\|_2^2 = \|G\|_2^2 \quad \forall t > 0$$

Proof.

$$\begin{array}{rcl} \langle U, \frac{dU}{dt} \rangle + \langle U, AU \rangle & = & 0 \\ \implies \frac{1}{2} \frac{d}{dt} \|U\|_2^2 & = & 0 \end{array}$$

□

We now look at adding time discretization to this approximation. We employ the θ method to obtain, in matrix notation,

$$\begin{array}{rcl} \frac{U^{n+1} - U^n}{\Delta t} + A[\theta U^{n+1} + (1-\theta) U^n] & = & 0 \\ U^0 & = & G \end{array} \tag{7.13}$$

Theorem 7.1.4.

- *If* $\theta = \frac{1}{2}$ *then* $\|U^n\|_2^2 = \|U^0\|_2^2 \quad \forall n \in \{1, \ldots, N\}$.
- *If* $\theta \in [0, \frac{1}{2})$ *then* $\|U^n\|_2^2 \geq \|U^0\|_2^2 \quad \forall n \in \{1, \ldots N\}$.
- *If* $\theta \in (\frac{1}{2}, 1]$ *then* $\|U^n\|_2^2 \leq \|U^0\|_2^2 \quad \forall n \in \{1, \ldots, N\}$.

Proof. We are in the set-up (7.1) where f satisfies (7.2). Hence, taking the inner product with $\theta U^{n+1} + (1-\theta)U^n$ written as in (7.5) gives the result by (7.6). □

7.1.4 Convergence

To illustrate ideas we consider the semi-discrete scheme (7.10). We set $u_j(t) = u(x_j, t)$ and define the truncation error by

$$T_j(t) \quad = \quad \frac{du_j}{dt} + c\left\{\frac{u_{j+1}(t) - u_{j-1}(t)}{2\Delta x}\right\}.$$

Let $T(t) = (T_{-J}(t), \cdots, T_{J-1}(t))^T$. We assume that there is a constant C independent Δx and such that

$$\sup_{t\in[0,T]} \|T(t)\|_\infty \quad \leq \quad C\Delta x^2. \tag{7.14}$$

This bound on the truncation error is satisfied if $u \in C^3([-1,1] \times [0,T], \mathbb{R})$. Under this assumption we deduce from (3.14) (noting $C(\Omega) = 2$ here) that we have consistency in the sense that

$$\sup_{t\in[0,T]} \|T(t)\|_2 \quad \leq \quad \sqrt{2}C\Delta x^2.$$

Let $u(t) = (u_{-J}(t), \ldots, , u_{J-1}(t))^T \in \mathbb{R}^{2J}$. Then:

Theorem 7.1.5. *Assume that (7.14) holds. Then*

$$\sup_{t\in[0,T]} \|u - U\|_2 \quad \leq \quad \sqrt{2}CT\Delta x^2$$

Proof. If $e = u - U$ then, by linearity,

$$\frac{de}{dt} + Ae = T.$$

Taking the inner product with e gives, by the skew-symmetry of A,

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|e\|_2^2 &= \langle e, T\rangle \\ &\leq \|e\|_2\|T\|_2. \end{aligned}$$

Let $I \subseteq (0, T]$ denote the set where $e(t) \neq 0$. Then

$$\frac{d}{dt}\|e\|_2 \leq \|T\|_2 \quad \forall t \in I.$$

But for $t \notin I$ we have $\|e\| = 0$ and hence

$$\frac{d}{dt}\|e\|_2 \leq \|T\|_2 \quad \forall t \notin I.$$

Thus, for all $t \in (0, T]$,

$$\begin{aligned} \frac{d}{dt}\|e\|_2 &\leq \|T\|_2 \\ &\leq \sup_{t\in[0,T]} \|T(t)\|_2 \\ &\leq \sqrt{2}C\Delta x^2. \end{aligned}$$

It follows that

$$\|e(t)\|_2 \leq \sqrt{2}C\Delta x^2 \quad \forall t \in [0, T],$$

since $e(0) = 0$. □

Remark The preceeding proof implicitly uses a stability estimate. See Exercise 6.

7.2 The Heat Equation

7.2.1 The PDE

We again study the Diffusion Model Problem (1.1) in dimension $d = 1$, with $\Omega = (0, 1)$ and with $f = 0$, namely (6.10):

$$\begin{array}{rcll} \frac{\partial u}{\partial t} & = & \frac{\partial^2 u}{\partial x^2} & (x,t) \in (0,1) \times (0,\infty) \\ u & = & 0 & (x,t) \in \{0,1\} \times (0,\infty) \\ u & = & g & (x,t) \in [0,1] \times \{0\} \end{array}$$

We use the norms and inner-products defined by (3.12) with $\Omega = [0, 1]$.

Theorem 7.2.1.

$$\|u(x,t)\|_2^2 \leq e^{-2\pi^2 t}\|g\|_2^2 \quad \forall t \geq 0.$$

Proof.

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|u\|_2^2 &= \frac{1}{2}\langle u, \frac{\partial u}{\partial t}\rangle + \frac{1}{2}\langle \frac{\partial u}{\partial t}, u\rangle = \langle u, \frac{\partial u}{\partial t}\rangle \\ &= \langle u, \frac{\partial^2 u}{\partial x^2}\rangle = -\|\frac{\partial u}{\partial x}\|_2^2 \leq -\pi^2\|u\|_2^2. \end{aligned}$$

In the last line we used the Poincaré inequality. Integrating and using $u(x, 0) = g(x)$ gives the result. □

7.2.2 The Approximation

We use the norm (3.13) with $\Delta = \Delta x$ and $\Omega_\Delta = \{1, \ldots, J - 1\}$.

Define, for $U_j(t) \approx u(x_j, t)$, $x_j = j\Delta x$, $J\Delta x = 1$

$$A = \frac{1}{\Delta x^2}\begin{pmatrix} 2 & -1 & 0 & \ldots & \ldots & 0 \\ -1 & 2 & -1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & -1 & 2 & -1 \\ 0 & \ldots & \ldots & 0 & -1 & 2 \end{pmatrix},$$

and

$$U(t) = \begin{pmatrix} U_1(t) \\ U_2(t) \\ \vdots \\ \vdots \\ U_{J-1}(t) \end{pmatrix}, G = \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ \vdots \\ g_{J-1} \end{pmatrix}.$$

A natural semi-discrete approximation is then

$$\frac{dU}{dt} + AU = 0, \qquad U(0) = G. \tag{7.15}$$

By Theorem 2.1.2 we know that A has a complete set of orthogonal eigenvectors with eigenvalues

$$\lambda^{(k)} = 4\sin^2(k\pi\Delta x/2)/\Delta x^2, \qquad k = 1, \ldots J-1. \tag{7.16}$$

Note that $\lambda^{(1)} \longrightarrow \pi^2$ as $\Delta x \longrightarrow 0$ and that $\lambda^{J-1} \leq \frac{4}{\Delta x^2}$. Hence for the given inner product and norm, we have

$$\frac{4}{\Delta x^2}\|v\|_2^2 \geq \langle v, Av\rangle = \langle A^{1/2}v, A^{1/2}v\rangle \geq \lambda^{(1)}\|v\|_2^2. \tag{7.17}$$

The lower bound is a discrete Poincaré inequality, analogous to that appearing in (5.4). The upper bound is another example of an inverse inequality, also analogous to that appearing in (5.4).

Theorem 7.2.2. *If U solves (7.15) then*

$$\|U(t)\|_2^2 \leq e^{-2\lambda^{(1)}t}\|G\|_2^2 \quad \forall t \geq 0.$$

Proof. By (7.17) we have

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|U\|_2^2 &= \langle U, \frac{dU}{dt}\rangle \\ &= \langle U, -AU\rangle \\ &\leq -\lambda^{(1)}\|U\|_2^2. \end{aligned}$$

Hence $\|U\|_2^2 \leq e^{-2\lambda^{(1)}t}\|G\|_2^2$, this gives the desired result. □

7.2.3 Convergence

In the following it is useful to use the notation:

$$V^{n+\theta} \;=\; \theta V^{n+1} + (1-\theta)V^n$$

We now discretize (7.15) in time by the θ-method to obtain

$$\frac{U^{n+1}-U^n}{\Delta} + AU^{n+\theta} \;=\; 0,\, U^0 = G$$

for $U_j^n \approx u(x_j, t_n)$ and $U^n = (U_1^n, \ldots, U_{J-1}^n)^T \in \mathbb{R}^{J-1}$. We define the truncation error by

$$T^n \;=\; \frac{u^{n+1}-u^n}{\Delta t} + Au^{n+\theta}.$$

Fix $r = \frac{\Delta t}{\Delta x^2}$. We assume that there exists $C \in \mathbb{R}$ independent of Δt and Δx such that:

$$\max_{n\in\{0,\ldots,N-1\}} \|T^n\|_2 \;\leq\; C[(1-2\theta)\Delta t + \Delta t^2 + \Delta x^2]. \tag{7.18}$$

Lemma (6.2.3), combined with (3.14), gives conditions under which this holds.

Theorem 7.2.3. *If (7.18) holds then for $\delta, \varepsilon \in \mathbb{R}$*

- *if $\theta \in [\frac{1}{2}, 1]$ and $\delta^2 < 2\lambda^{(1)}$;*
  or
- *if $\theta \in [0, \frac{1}{2})$, $4(1-2\theta)r < 1-\epsilon$ and $\delta^2 < 2\epsilon\lambda^{(1)}$*

then

$$\sup_{0\leq n\Delta t\leq T} \|u^n - U^n\|_2 \;\leq\; C\sqrt{T}\left(\frac{1}{2} + \frac{1}{\delta^2}\right)^{\frac{1}{2}} \left[(1-2\theta)\Delta t + \Delta t^2 + \Delta x^2\right]$$

Proof. By linearity we have

$$e^{n+1} - e^n + \Delta t A e^{n+\theta} \;=\; \Delta t T^n.$$

Taking the inner product with $e^{n+\theta}$ and noting that

$$e^{n+\theta} = \frac{1}{2}(e^{n+1}+e^n)+(\theta-\frac{1}{2})(e^{n+1}-e^n)$$

we obtain

$$\frac{1}{2}\|e^{n+1}\|_2-\frac{1}{2}\|e^n\|_2^2+(\theta-\frac{1}{2})\|e^{n+1}-e^n\|_2^2 = -\Delta t\|A^{\frac{1}{2}}e^{n+\theta}\|_2^2+\Delta t\langle e^{n+\theta},T^n\rangle.$$

If $\theta\geq\frac{1}{2}$ then, by Cauchy-Schwarz and (7.17),

$$\|e^{n+1}\|_2^2 \leq \|e^n\|_2^2-2\Delta t\lambda^{(1)}\|e^{n+\theta}\|_2^2+\Delta t\delta^2\|e^{n+\theta}\|_2^2+\frac{\Delta t}{\delta^2}\|T^n\|_2^2$$

Since $2\lambda^{(1)}>\delta^2$ we get

$$\|e^{n+1}\|_2^2\leq\|e^n\|_2^2+\frac{\Delta t}{\delta^2}\|T^n\|_2^2.$$

Hence

$$\begin{aligned}\|e^n\|_2^2 &\leq \frac{n\Delta t}{\delta^2}\max_{n\in\{0,\ldots,N-1\}}\|T^n\|_2^2\\ &\leq \frac{T}{\delta^2}C^2\left[(1-2\theta)\Delta t+\Delta t^2+\Delta x^2\right]^2.\end{aligned}$$

The result follows.

If $\theta\in[0,\frac{1}{2})$ then

$$\begin{aligned}\|e^{n+1}\|_2^2 \leq\ & \|e^n\|_2^2+(1-2\theta)\Delta t^2\|Ae^{n+\theta}-T^n\|_2^2-2\Delta t\|A^{\frac{1}{2}}e^{n+\theta}\|_2^2\\ &+\Delta t\delta^2\|e^{n+\theta}\|_2^2+\frac{\Delta t}{\delta^2}\|T^n\|_2^2.\end{aligned}$$

But $\|a+b\|_2^2\leq 2\|a\|_2^2+2\|b\|_2^2$ and, by (7.17) for $w=A^{\frac{1}{2}}v$,

$$\begin{aligned}\|Av\|_2^2 &= \langle Aw,w\rangle\\ &\leq \frac{4}{\Delta x^2}\|w\|_2^2\\ &= \frac{4}{\Delta x^2}\|A^{\frac{1}{2}}v\|_2^2.\end{aligned}$$

Hence

$$\begin{aligned}\|e^{n+1}\|_2^2 &\leq \|e^n\|_2^2 + (1-2\theta)8r\Delta t\|A^{\frac{1}{2}}e^{n+\theta}\|_2^2 + (1-2\theta)2\Delta t^2\|T^n\|_2^2 \\ &\quad -2\Delta t\|A^{\frac{1}{2}}e^{n+\theta}\|_2^2 + \Delta t\delta^2\|e^{n+\theta}\|_2^2 + \frac{\Delta t}{\delta^2}\|T^n\|_2^2.\end{aligned}$$

But $(1-2\theta)4r < 1-\epsilon$ implies

$$\begin{aligned}\|e^{n+1}\|_2^2 &\leq \|e^n\|_2^2 - 2\Delta t\epsilon\|A^{\frac{1}{2}}e^{n+\theta}\|_2^2 + \Delta t\delta^2\|e^{n+\theta}\|_2^2 + \Delta t\left(\frac{1}{2}+\frac{1}{\delta^2}\right)\|T^n\|_2^2 \\ &\leq \|e^n\|_2^2 - \Delta t\left[2\epsilon\lambda^{(1)} - \delta^2\right]\|e^{n+\theta}\|_2^2 + \Delta t\left(\frac{1}{2}+\frac{1}{\delta^2}\right)\|T^n\|_2^2 \\ &\leq \|e^n\|_2^2 + \Delta t\left(\frac{1}{2}+\frac{1}{\delta^2}\right)\|T^n\|_2^2.\end{aligned}$$

The proof now follows as for $\theta \in [\frac{1}{2}, 1]$. □

Chapter 8

Underlying Principles

We have cosidered two types of problems.

Problem BVP: boundary value problems on a square domain $\Omega \subset \mathbb{R}^n$ with $n = 1, 2$. Find for f, h given, a function $u \in C^k(\Omega)$ satisfying

$$\begin{aligned} \mathcal{L}[u] &= f \quad \text{in } \Omega \ , \\ u &= h \quad \text{on } \partial\Omega \ , \end{aligned}$$

where the spatial operator $\mathcal{L}$ was of the form $\mathcal{L} = -\triangle u + p \cdot \nabla u + qu$ with given functions p, q from Ω to $\mathbb{R}$.

Problem IVP: initial value problems on a square domain $\Omega \times (0, \infty) \subset \mathbb{R}^n \times \mathbb{R}^+$ with $n = 1, 2$. Find for h, g a function $u \in C^{k \times l}(\Omega \times (0, \infty))$ satisfying

$$\begin{aligned} \partial_t u &= \mathcal{L}[u] \quad \text{in } \Omega \times (0, \infty) \ , \\ u &= h \quad \text{on } \partial\Omega \times (0, \infty) \ , \\ u &= g \quad \text{on } \Omega \times \{0\} \ . \end{aligned}$$

with a spatial operator of the form $\mathcal{L} = -\varepsilon\triangle u + p \cdot \nabla u + qu - f$ given functions p, q from Ω to $\mathbb{R}$.

In addition to Dirichlet boundary conditions $u = h$ we also considered periodic boundary conditions. In the following we will be considering Dirichlet boundary conditions if not mentioned otherwise.

To solve the **BVP** we set up a linear system of equation

$$AU = F$$

so that the solution $U = (U_j)_{j=1}^J \in \mathbb{R}^J$ is an approximation of $u(x_j)$ for J distinct points $x_j \in \Omega$. This is the method in **matrix form**.

We studied equidistant point sets which in one space dimension are of the form $a+jh$ with $(J+1)h = b-a$ where $\Omega = (a, b)$ thus x_0, x_{J+1} lie on the boundary of Ω and we can extend any vector U by defining $U_0 = h(a), U_{J+1} = h(b)$. In general to construct method in higher space dimension, methods from 1D are applied for each coordinate direction.

To solve the **IVP** we use the same technique to approximate $\mathcal{L}$ as used for the **BVP**. We thus obtained a **semi discrete** scheme for functions $U_j(t)$ which approximate $u(x_j, t)$. In matrix notation:

$$\frac{d}{dt}U(t) = AU(t) \ , \quad U(0) = G = (g(x_1), \ldots, g(x_J))^T \ .$$

To obtain a fully discrete scheme we applied a time stepping scheme to for ODEs. Again using matrix notation.

$$B^{n+1}U^{n+1} = A^n U^n \ , \quad U^0 = G$$

This is the method in **matrix form**. Here $U^n \approx U(t^n)$ with time steps $t^{n+1} = t^n + \Delta t^n$ with $\Delta t^n > 0$ and $t^0 = 0$. In general we used equidistant points $t^n = n\Delta t$.

In the following we discuss the different steps in more detail.

8.1 Time Stepping

To solve

$$\frac{d}{dt}U(t) = AU(t) \ , \quad U(0) = G \ ,$$

any solver for ODEs can be used. We only considered **one step** time discretization schemes, where U^{n+1} is computed using only the result from the

previous time step U^n. For example the *Leap frog scheme* mentioned in chapter 2 is not a one step method.

One step method are of the form

$$B^{n+1}U^{n+1} = A^nU^n \ , \quad U^0 = G$$

If $B^{n+1} = I$ we call the method **explicit** otherwise **implicit**.

We focused mainly on the θ-scheme, which is based on

$$\frac{U^{n+1} - U^n}{\Delta t^n} = \theta AU^{n+1} + (1-\theta)AU^n \ .$$

Rearranging terms gives us

$$(I - \theta\Delta t^n A)U^{n+1} = (I + (1-\theta)\Delta t^n A)U^n$$

so that we have

$$B^{n+1} = I - \theta\Delta t^n A \ , \quad A^n = I + (1-\theta)\Delta t^n A \ .$$

We always assumed that we have fixed time steps $\Delta t^n = \Delta t$ although the analysis of the method only becomes slightly more technical if varying time steps are allowed.

Convergence analysis for these schemes is mostly based on a truncation error estimate and some discrete form of Gronwall's lemma. Denote the norm of the error with $z_n = \|e^n\|$ then if we can show

$$z_{n+1} \leq Cz_n + \Delta t\|T^n\|$$

For $C > 0$ Gronwall's lemma provides a bound for the norm of the error:

$$z_n \leq \frac{C^n - 1}{C - 1}\Delta t \max_{k=0,\ldots,n-1} T_k$$

for $C \neq 1$; for $C = 1$ we simply have $z_n \leq n\Delta t \max_{k=0,\ldots,n-1} T_k$. We have used that $z_0 = \|e^0\| = 0$.

The factor T^n stems from the truncation error and should satisfy $\max_n \|T^n\| \leq \tilde{C}(\Delta t^p + \Delta x^q)$ and we need C small enough (see below). This requirement leads to some restriction on the time step.

The following two estimates for the exponential function play an important role

Lemma 8.1.1. *For $z \in \mathbb{R}$ the following holds:*

$$(1+z)^n \leq \exp(nz) \ ,$$

and

$$\frac{1}{(1-z)^n} \leq \exp(\frac{nz}{1-z}) \ .$$

Proof. The first estimate follows from Taylor series expansion: $\exp(z) = 1 + z + \frac{1}{2}\chi_z^2 \geq 1 + z$. $(1+z)^n \leq \exp(z)^n = \exp(nz)$.

Noting that $1 + \frac{z}{1-z} = \frac{1}{1-z}$ provides the second estimate. □

Example:

For $C = 1$, i.e., $\|e^{n+1}\| \leq \|e^n\| + \Delta t\|T^n\|$ we can conclude

$$\max_{n, n\Delta t < T} z_n \leq C^n z_0 + n\Delta t T^n \leq T\tilde{C}(\Delta t^p + \Delta x^q) = TO(\Delta t^p + \Delta x^q) \ .$$

In some cases one has $C = 1 + c\Delta t$. For example when applying the **forward Euler** method to

$$\frac{d}{dt}u(t) = cu(t)$$

leads to $U^{n+1} = U^n + \Delta c U^n$ and thus to the **error equation** $e^{n+1} = (1 + c\Delta t)e^n + \Delta t T^n$ with $T^n = O(\Delta t)$. In this case Gronwall's lemma gives us

$$z_n \leq \Delta t\tilde{C}(\Delta t^p + \Delta x^q)\frac{C^n - 1}{C - 1} + z_0 C^n \leq \frac{\tilde{C}}{c}(\Delta t^p + \Delta x^q)((1 + c\Delta t)^n - 1) \ .$$

Using $(1+z)^n \leq \exp(nz)$

$$z_n \leq \frac{\tilde{C}}{c}(\Delta t^p + \Delta x^q)(\exp(c\Delta t n) - 1) \leq (\exp(cT) - 1)O(\Delta t^p + \Delta x^q) \ .$$

In both cases the method converges with the order of the truncation error.

The setting described above covers most explicit methods. For implicit schemes the argument is slightly different. We demonstrate this for the non-linear ODE

$$\frac{d}{dt}u(t) = f(u(t))$$

with Lipschitz continuous function f with Lipschitz constant L. Consider the implicit Euler method $U^{n+1} - \Delta t f(U^{n+1}) = U^n$. We know that the truncation error $T^n = \frac{u^{n+1}-u^n}{\Delta t} - f(u^{n+1})$ satisfies $\max_n |T^n| \leq \tilde{C}\Delta t$. The error satisfies $e^{n+1} = u^{n+1} - U^{n+1} = u^n + \Delta t f(u^{n+1}) - U^n - \Delta t f(U^{n+1}) + \Delta t T^n$. Taking the absolute value on both sides leads to

$$|e^{n+1}| \leq |e^n| + \Delta t |f(u^{n+1}) - f(U^{n+1})| + \Delta t T^n \leq |e^n| + \Delta t L |e^{n+1}| + \Delta t T^n.$$

Thus if $1 - L\Delta t > 0$ we have $|e^{n+1}| \leq \frac{1}{1-L\Delta t}|e^n| + \frac{\Delta t}{1-L\Delta t}|T^n|$. Defining $z_n = |e^n|$ and $C = \frac{1}{1-L\Delta t}, D = \frac{1}{1-L\Delta t}\max_n \|T^n\|$ we are in the situation covered by Gronwall's lemma if $C > 0$, i.e., $L\Delta t < 1$. Since $C - 1 = \frac{L\Delta t}{1-L\Delta t}$ and $C^n - 1 = ((1 - L\Delta t)^{-n} - 1)$ we have

$$z_n \leq \frac{1}{L} \max_n \|T^n\| ((1 - L\Delta t)^{-n} - 1) \ .$$

Using $(1 - L\Delta t)^{-n} \leq \exp(\frac{LT}{1-L\Delta t})$ we conclude that

$$\|e^n\| \leq \frac{1}{L} \max_n \|T^n\| \exp\big(\frac{LT}{1 - L\Delta t}\big)$$

if $\Delta t \leq \frac{1}{L}$.

8.2 Constructing Finite Difference Approximations

For both **BVP** and **IVP** the spatial discretizations we have studied were based on the finite difference approach. Hereby the spatial operator $\mathcal{L}[u]$ is evaluated in the points x_j and the derivatives are replaced by linear combinations of the values U_j. These approximations are obtained by Taylor expansion of u around the point x_j. The points used to discretize $\mathcal{L}[u](x_j)$ are called the **stencil** of the scheme. If we write down the scheme in matrix form then these give the non-zero entries in the jth row of the matrix A.

We focus of the one dimensional setting since this provides the basis also for higher dimensional discretizations.

We use the approach

$$\mathcal{L}[u](x_j) \approx \sum_{k=k_0}^{k_1} \alpha_{jk} U_{j+k} \ .$$

Where is general we have $k_0 \leq 0$ and $k_1 \geq 0$. The neighboring values $(U_{j+k})_{k=k_0}^{k_1}$ used in the discretization are the stencil of the method. The linear system now is

$$\sum_{k=k_0}^{k_1} \alpha_{jk} U_{j+k} = f(x_j)$$

for all $j \in \{1, \ldots, J\}$ with $j + k_0 > 0$ and $j + k_1 \leq J$. For j with $j + k_0 = 0$ or $j + k_1 = J$ the Dirichlet boundary conditions can be used, e.g., for $j + k_0 = 0$:

$$\sum_{k=k_0+1}^{k_1} \alpha_{jk} U_{j+k} = f(x_j) - \alpha_{jk_0} h(x_0) \ .$$

If $k_0 < -1$ some suitable discretization of derivatives in the point x_1 has to be used which is based on a stencil $U_0, \ldots, U_{\tilde{k}_1}$. Similar holds for the right boundary. For higher space dimension the approach is applied to each coordinate direction separately.

The main work is in obtaining the approximations of the derivatives. Let us assume that the points are equidistant, i.e., $x_{j+k} = x_j + kh$. In this case Taylor expansion of a function $u \in C^{p+1}([a, b])$ gives us:

$$u(x_{j+k}) = \sum_{l=0}^{p} \frac{1}{l!} (kh)^l \partial_x^l u(x_j) + C_k h^{p+1} \ .$$

Denote with $\delta_j^l = \partial_x^l u(x_j)$ then a linear combination of the $u(x_{j+k})$ values gives us:

$$\sum_{k=k_0}^{k_1} a_k u_{j+k} = \sum_{k=k_0}^{k_1} a_k \sum_{l=0}^{p} \frac{1}{l!} (kh)^l \delta_i^l + \sum_{k=k_0}^{k_1} a_k C_k h^{p+1} = \sum_{l=0}^{p} \delta_j^l \frac{h^l}{l!} \sum_{k=k_0}^{k_1} k^l a_k + \sum_{k=k_0}^{k_1} a_k C_k h^{p+1} \ .$$

Now we want to find an approximation $\delta_j^m = \sum_{k=k_0}^{k_1} a_k u_{j+k} + \sum_{k=k_0}^{k_1} a_k C_k h^{p+1}$. This lead to

$$\sum_{l=0, l \neq m}^{p} \delta_j^l \frac{h^l}{l!} \sum_{k=k_0}^{k_1} k^l a_k + \delta_j^m \left(\frac{h^m}{m!} \sum_{k=k_0}^{k_1} k^m a_k - 1 \right) = \sum_{k=k_0}^{k_1} a_k C_k h^{p+1} \ .$$

Assuming that the δ_i^l can attain arbitrary values this can only hold, if every term in the sum is equal to zero, thus for $l \in \{0, \ldots, p\}, l \neq m$ we have the equation

$$\frac{1}{l!} h^l \sum_{k=k_0}^{k_1} k^l a_k = 0$$

together with

$$\frac{1}{m!} h^m \sum_{k=k_0}^{k_1} k^m a_k = 1 \ .$$

This leads to a linear system of equations of the form

$$Da = b$$

where $D = (d_{lk})$ with $d_{lk} = k^l$ and $b_l = 0$ (for $l \neq m$) and $b_m = \frac{m!}{h^m}$ ($k = k_0, \ldots, k_1$). The approximation has the truncation error $C \sum_{k=k_0}^{k_1} a_k h^{p+1}$.

Example: Let us start with something well know. We want a stencil with $k_0 = -1$ and $k_1 = 1$ to approximate $\partial_x u(x_j)$. We take $p = 2$ (assuming u is smooth enough) which gives us the linear system of equation for a_{-1}, a_0, a_1:

$$\begin{aligned} a_{-1} + a_0 + a_1 &= 0 \\ -a_{-1} + a_1 &= \frac{1}{h} \\ a_{-1} + a_1 &= 0 \end{aligned}$$

We have three equation with three unknown which imply $a_{-1} = -a_1$ and $h(a_1 - a_{-1}) = 1$, so that we have $a_{-1} = -\frac{1}{2h}, a_1 = \frac{1}{2h}$ and $a_0 = -(a_{-1} + a_1) = 0$. Thus we arrive at the central finite difference approximation

$$\partial_x u(x_j) = \frac{1}{2h}(u_{x_{j+1}} - u_{x_{j-1}}) + h^3(-\frac{C_{-1}}{2h} + \frac{C_1}{2h}) = \frac{1}{2h}(u_{x_{j+1}} - u_{x_{j-1}}) + O(h^2) \ .$$

Note that we can not increase the order with this stencil since the next equation to solve would be $-a_{-1} - a_1 = 0$ which can not hold.

Lets us try $k_0 = 0$, $k_1 = 2$ again with $p = 3$, we arrive at

$$\begin{aligned} a_0 + a_1 + a_2 &= 0 \\ a_1 + 2a_2 &= \frac{1}{h} \\ a_1 + 4a_2 &= 0 \ . \end{aligned}$$

Thus $a_1 = -4a_2$ and $-2a_2 = \frac{1}{h}$. Thus $a_2 = -\frac{1}{2h}, a_1 = \frac{2}{h}$, and $a_0 = -a_1 - a_2 = -\frac{3}{2h}$. Again we get the order 2 and cannot increase the order because that would require $a_1 + 8a_2 = 0$.

One can also obtain finite difference approximations on points which are not equidistributed, e.g., for point $(x_j)_{j=0}^J$ with $x_0 < x_1 < \cdots < x_{J-1} < x_J$ with $\delta_i = x_i - x_{i-1} > 0$ for $j = 1, \ldots, J$ satisfying $ch \leq \delta_i \leq h$ for some $h > 0$ and $c < 1$ fixed. We cannot use the analysis presented above. But have to study the Taylor expansion directly.

Example: we again want to approximate $\partial_x u(x_j)$. First let us take the stencil x_j, x_{j+1}: $u(x_{j+1}) = u(x_j) + \delta_{j+1} u'(x_j) + O(h^2)$. We study the approximation $u'(x_j) = au(x_{j+1}) + bu(x_j) = (a+b)u(x_j) + a\delta_{j+1} u'(x_j) + O(h^2)$. This requires $a + b = 0$ and $a\delta_{j+1} = 1$, which is satisfied with $a = \frac{1}{\delta_{j+1}}, b = -\frac{1}{\delta_{j+1}}$. This gives us the approximation $u'(x_j) = \frac{1}{\delta_{j+1}}(u_{j+1} - u_j)$ which is the same as for equidistributed points. The second order scheme with error $O(h^2)$ is left as exercise.

Example: higher order derivatives are discretize using the same ideas. An example is given on the first exercise sheet for $\partial_x^4 u(x_i)$.

8.3 Stability in L^∞

The previous section shows how to obtain consistency in general. The second cruical part of the convergence proof is stability. In addition to proving a tool to prove convergence, the following also provides tools to show that the matrix A arrising in the finite difference method has an inverse, and thus that the discrete version for the BVP has a solution.

In most cases the matrices which appear in the finite difference methods have a very special structure. We have already made use of there Toeplitz form to obtain information on their eigenstructure and to prove L^2 norm convergence results.

A further observation is that they are *sparse*, i.e., only a few entries per line have non-zero entries corresponding to the stencil used.

In general they have a lot more structure on which the stability proofs in the maximum norm are based.

Consider the matrix arising from the central difference discretization of $-\partial_x u + qu = f$. With Dirichlet boundary conditions we have

$$A = \frac{1}{h^2} \begin{pmatrix} 2+h^2q & -1 & 0 & \dots & 0 & 0 \\ -1 & 2+h^2q & -1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -1 & 2+h^2q & -1 \\ 0 & 0 & \dots & 0 & -1 & 2+h^2q \end{pmatrix} \tag{8.1}$$

or with periodic boundary conditions:

$$A = \frac{1}{h^2} \begin{pmatrix} 2+h^2q & -1 & 0 & \dots & 0 & -1 \\ -1 & 2+h^2q & -1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -1 & 2+h^2q & -1 \\ -1 & 0 & \dots & 0 & -1 & 2+h^2q \end{pmatrix} \tag{8.2}$$

In both cases we have for $q > 0$

$$a_{ij} \leq 0 \text{ for } i \neq j, \quad a_{ii} > 0 \quad \text{and} \sum_j a_{ij} > 0$$

The last inequality can also be written as

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \ .$$

In the case of Dirichlet boundary conditions and $q = 0$ the last inequality only holds with strictly greater for the first and last row. For all other rows we have

$$|a_{ii}| = \sum_{j \neq i} |a_{ij}| \ .$$

But there is an additional property which we used for the proof of the maximum principle for the BVP in Chapter 3. That is that each interior point is connected to some point on the boundary. For the matrix that means that each row is "connected" with the first or last row in the following sense.

Definition 8.3.1. A square matrix $A = (a_{ij}) \in \mathbb{R}^{N\times N}$ is called **irreducible** if for all $i, j \in \{1, \ldots, N\}$ either $a_{ij} \neq 0$ or there exists a sequence $i_0, \ldots, i_s \in \{1, \ldots, N\}$ with $i_0 = i, i_s = j$ and $a_{i_{k-1}i_k} \neq 0$ for $k = 1, \ldots, s$.

A square matrix which is not irreducible is called **reducible**.

An example is

$$A = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & 0 \end{pmatrix} \tag{8.3}$$

Now taking for example $i = 5$ and $j = 2$ we can find a path through the matrix starting in row $i = 5$ and ending in column $j = 2$ containing only non-zero values: a_{54}, a_{43}, a_{32}. Similar argument hold for any zero entry in the matrix so that it is irreducible.

There are a few equivalent formulation for reducible matrices.

Lemma 8.3.2. *Let $A = (a_{ij}) \in \mathbb{R}^{N\times N}$ be a square matrix.*

A is reducible if and only if the indices $1, \ldots, N$ can be divided into two disjoint nonempty sets $i_1, i_2, \ldots, i_\mu$ and $j_1, \ldots, j_\nu$ with $\mu + \nu = N$ such that $a_{i_\alpha j_\beta} = 0$ for $\alpha = 1, \ldots, \mu$ and $\beta = 1, \ldots, \nu$.

A is reducible if and only if it can be placed into block upper-triangular form by simultaneous row/column permutations.

Next we define special classes of matrices:

Definition 8.3.3. Let $A = (a_{ij}) \in \mathbb{R}^{N\times N}$ be given. We say that A is

- an L_0 matrix if $a_{ij} \leq 0$ for $i \neq j$.
- an L matrix if it is an L_0 matrix with $a_{ii} > 0$.
- an M matrix if it is an L matrix, A^{-1} exists and $A^{-1} \geq 0$.

- strictly diagonal dominant if

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \ .$$

- irreducibly diagonal dominant if A is irreducible and

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|,$$

  and at least for one i we have

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \ .$$

Here we use the notation $A \geq 0$ or $u \geq 0$ if all entries of the matrix A or the vector u are non-negative.

Remark: thus the matrices from (8.1) and (8.2) are L matrices which are strictly diagonal dominant for $q > 0$ and (8.1) is irreducibly diagonal dominant for $q = 0$. We will show next that these are in fact M matrices. Note that for $q = 0$ the matrix (8.2) is an L matrix but neither strictly diagonal dominant nor irreducibly diagonal dominant. This reflects the fact that the matrix is a discretization for $-\partial_{xx} u = f$ and u is periodic which does not have a unique solution (thus it can not be an M matrix).

Theorem 8.3.4. *Let $A = (a_{ij}) \in \mathbb{R}^{N \times N}$ be given. Then the following are equivalent:*

1. *A^{-1} exists and $A^{-1} \geq 0$ (A is regular and inverse monoton).*
2. *$Ax \leq 0 \Longrightarrow x \leq 0$.*
3. *$Ay \leq Az \Longrightarrow y \leq z$.*

Proof. The equivalence of (2) and (3) is clear taking either $z = 0$ or $x = y - z$. So we will show the equivalence of (1) and (2).

(2)$\rightarrow$ (1): Let $x \in \mathbb{R}^N$ with $Ax = 0$. Then since $Ax \leq 0$ and $A(-x) \leq 0$ we have $x = 0$. Therefore A is injective and thus bijective so that A^{-1} exists.

Again using (2) we have that for $y \geq 0$ that $A^{-1}y \geq 0$ since from $0 \leq y = Ax$ follows $0 \leq x$ using $x = A^{-1}y$. Taking the unit vector $y = e_i \geq 0$ for a fixed $i = 1, \ldots, N$ we can conclude from $0 \leq A^{-1}$ and thus $0 \leq (A^{-1}y)_l = (A^{-1})_{li}$ which shows that $A^{-1} \geq 0$.

(1)$\rightarrow$ (2): For $y \geq 0$ we have $(A^{-1}y)_l = \sum_{j=1}^{N} (A^{-1})_{lj} y_j \geq 0$ for all $l = 1, \ldots, N$. For $x \in \mathbb{R}^N$ with $Ax \leq 0$ we have $y = -Ax \geq 0$ and thus $-x_l = (A^{-1}y)_l \geq 0$ which concludes the proof since $x \leq 0$. □

Some of the arguments used here can be found in the proof of Theorem **TODO**.

Theorem 8.3.5. *Let $A \in \mathbb{R}^{N \times N}$ be an L-matrix.*

If A is strictly diagonal dominant then A is an M matrix.

If A is irreducible diagonal dominant then A is an M matrix.

Proof. We will only prove the first part.

We will show that an L-matrix which is strictly diagonal dominant satisfies (2) from the previous theorem which proves that A is an M-matrix: Let for all $l = 1, \ldots, N$: $0 \geq (Ax)_l = \sum_{j=1}^{N} a_{lj} x_j$ hold. Assume that $x_l \geq x_j$ for all $j = 1, \ldots, L$. We have to show that $x_l \leq 0$. Assume that this is not the case, i.e., $x_l > 0$:

Sine $a_{ll} > 0$ we have Then $x_l \leq -\frac{1}{a_{ll}} \sum_{j \neq l} a_{lj} x_j = \frac{1}{|a_{ll}|} \sum_{j \neq l} |a_{lj}| x_j \leq x_l \frac{1}{|a_{ll}|} \sum_{j \neq l} |a_{lj}| < x_l$, the last inequality follows from $|a_{ll}| > \sum_{j \neq l} |a_{lj}|$ and $x_l > 0$. Therefore we arrive at the contradiction $x_l < x_l$ so that x_l must be non positive. That concludes the proof. □

Theorem 8.3.6. *Let $A \in \mathbb{R}^{N \times N}$ be an M-matrix and assume there exists a vector $\Phi \in$*
bbR^N with $\Phi > 0$ and $A\Phi > 0$. Then

$$\|x\|_\infty \leq C_\Phi \|Ax\|_\infty, \quad \text{with } C_\Phi = \frac{\max_i}{\min_i (A\Phi)_i} .$$

Proof. Let $Ax = y$. We have

$$\pm x_i = \sum_{j=1}^{N} (A^{-1})_{ij} (\pm y_j) \leq \|y\|_\infty \sum_{j=1}^{N} (A^{-1})_{ij}$$

taking into account that $(A^{-1})_{ij} > 0$ since A is an M matrix.

Define $c = \min_i (A\Phi)_i$. Thus $A\Phi \geq ce$ with $e = (1, \ldots, 1)^T \in \mathbb{R}^N$. With $z = cA^{-1}e$ we have $A\Phi \geq Az$ and thus using the previous Theorem we have $\Phi \geq z = cA^{-1}e$, i.e., $\Phi_i \geq c \sum_{j=1}^{N} (A^{-1})_{lj}$.

Combining both results we obtain:

$$\pm x_i \leq \|y\|_\infty \sum_{j=1}^{N} (A^{-1})_{ij} \leq \frac{\Phi_i}{c} \|y\|_\infty \leq \frac{\max_i \Phi_i}{c} \|Ax\|_\infty \ .$$

This concludes the proof. □

Remark The last theorem shows stability in the $\|\cdot\|_\infty$ norm for the scheme described by the matrix A. To fullfil the assumptions in the Theorem, the matrix A has to be an L matrix which is diagonally dominant. Furthermore one has to find a vector $\Phi \in \mathbb{R}^N$ with $\Phi > 0, A\Phi > 0$. In the case of a strictly diagonal dominant matrix $\Phi = e = (1, \ldots, 1)^T$ can be used. The theorem hold in this form also for irreducible diagonal dominant matrices and in this case finding Φ can be difficult. We have used this approach in Chapter 3 using a 2d version of $\Phi_j = \frac{1}{2}(x_j - \frac{1}{2})^2$.

Example 8.3.7. *Consider the BVP*

$$-\varepsilon \frac{\partial^2 u}{\partial x^2} + a \frac{\partial u}{\partial x} + pu = f$$

on $\Omega = [0, L]$ with periodic boundary conditions. Assume $\varepsilon, a, p > 0$. Consider a equidistant point set $\{x_i\}$ with spacing h on Ω, a standard 2. order consistent approximation for the second order derivative and consider for $\theta \in [0, 1]$:

$$\frac{\partial u}{\partial x}(x_i) \approx \frac{1}{2}\big((1 - 2\theta)u_i + (\theta - 1)u_{i-1} + \theta u_{i+1}\big) \ .$$

State the resulting finite difference scheme in matrix form and show that A is an M matrix under suitable restrictions on the spacing h. Prove convergence of the scheme in the $\|\cdot\|_\infty$ norm.

8.4 L^2 Convergence Analysis

The properties of the matrix A presented above provide existence of the inverse of A and thus prove that the discrete solution exists. Furthermore they provide tools for proving maximum principles and thus stability and convergence in L^∞. To obtain results using the energy method other tools are more useful.

8.4.1 The discrete Poincaré inequality

We first revisit the continuous Pointcaré inequality in 1D:

Theorem 8.4.1. *Let $\Omega = (a, b)$. For $u \in C^1(\Omega)$ with $u(a) = u(b) = 0$ the following inequality holds*

$$C_p \|\nabla u\|_2^2 \;\; \geq \;\; \|u\|_2^2. \tag{8.4}$$

Proof. Using $\int_a^x u'(x)\, dx = u(x) - u(a) = u(x)$ and the Hölder inequality we can conclude for $x \in (a, b)$

$$u^2(x) = \Big(\int_a^x u'(x)\, dx \Big)^2 \leq \int_a^x 1\, dx \int_a^x (u'(x))^2\, dx \leq (b-a) \int_a^b (u'(x))^2\, dx$$

So that

$$\int_a^b u^2(x)\, dx \leq (b-a)^2 \int_a^b (u'(x))^2\, dx$$

Therefore $\|u\|_2^2 \leq |\Omega|^2 \|\nabla u\|_2^2$ with $C_p = |\Omega|^2$. □

Remark: our estimate for the constant C_p is not optimal, i.e., it is too small. But in contrast to previous chapters we are here only interested in the qualitative behavior of the estimates.

We have previously used discrete versions of the Pointcaré inequality to prove estimates based on the energy methods.

Theorem 8.4.2. *Assume $J\Delta x = (b-a)$ and let $U = (U_j)_{j=1}^J \in \mathbb{R}$ be given and extend the vector by defining $U_0 = U_{J+1} = 0$. Define $V_j = \frac{\Delta^- U_j}{\Delta x}$ then*

$$C_p \|V|_2^2 \;\geq\; \|U\|_2^2 \tag{8.5}$$

with $C_p = (b-a)^2$.

The same estimate holds for the finite difference approximations $\frac{\Delta^- U_j}{\Delta x}$ and $\frac{\Delta_0 U_j}{\Delta x}$.

Proof. Using the properties of the telescope sum $\sum_{j=1}^k \Delta x V_j = \sum_{j=1}^k (U_j - U_{j-1}) = U_k - U_0 = U_k$ we obtain

$$\begin{aligned} U_k^2 &= \big(\sum_{j=1}^k \Delta x V_j\big)^2 = \big(\sum_{j=1}^k (\sqrt{\Delta x})(\sqrt{\Delta x} V_j)\big)^2 \\ &= \sum_{j=1}^k \Delta x \sum_{j=1}^k \Delta x V_j^2 \\ &\leq J\Delta x \sum_{j=1}^J \Delta x V_j^2 = (b-a)\|V\|_2^2 \,. \end{aligned}$$

So that

$$\sum_{k=1}^J U_k^2 \leq (b-a)^2 \|V\|_2^2 \,.$$

This concludes the proof for the backward difference. For the other finite difference approximation to proof follows similarly. □

Remark: Again we note that the bound is not optimal.

Not only the Pointcaré readily carries over from the continuous to the discrete; a further often used formula is integration by parts:

Theorem 8.4.3. *Assume $J\Delta x = (b-a)$ and let $U, V \in \mathbb{R}$ be given and extend the vectors by defining $U_0 = U_{J+1} = V_0 = V_{J+1} = 0$. Then*

$$((\frac{\delta^2 U_j}{\Delta x^2})_j, V) = -((\frac{\Delta^- U_j}{\Delta x})_j, (\frac{\Delta^- V_j}{\Delta x})_j) - U_J V_J \,.$$

The similar results holds using Δ^+ or Δ_0 if the additional term $U_J V_J$ is suitably replaced.

Proof.

$$\begin{aligned}
\Delta x \sum_{j=1}^{J} \Delta x \frac{\delta^2 U_j}{\Delta x^2} V_j &= \sum_{j=1}^{J} (U_{j-1} - 2U_j + U_{j+1}) V_j \\
&= \sum_{j=1}^{J} (U_{j-1} - U_j) V_j + \sum_{j=1}^{J} (U_{j+1} - U_j) V_j \\
&= \sum_{j=1}^{J} (U_{j-1} - U_j) V_j + \sum_{j=2}^{J} (U_j - U_{j-1}) V_{j-1} + U_{J+1} V_J - U_J V_J \\
&= \sum_{j=1}^{J} (U_{j-1} - U_j) V_j + \sum_{j=1}^{J} (U_j - U_{j-1}) V_{j-1} - U_J V_J
\end{aligned}$$

□

Remark: From the continuous case we would expect $((\frac{\delta^2 U_j}{\Delta x^2})_j, V) = -((\frac{\Delta^- U_j}{\Delta x})_j, (\frac{\Delta^- V_j}{\Delta x})_j)$. With a slight change in the definition of the scalar product and the extension of U, V for $j \notin \{1, \ldots, J\}$ we can even proof this result. To that end, extend for any $U \in \mathbb{R}^J$ the domain of U by defining: $U_j = 0$ for $j \in \mathbb{Z} \setminus \{1, \ldots, J\}$ and define a corresponding scalar product $(U, V) = \sum_{j \in \mathbb{Z}} U_j V_J$. With this definition the result of the Theorem can shown without the additional term $U_J V_J$.

Corollary 8.4.4. *Assume $J\Delta x = (b - a)$ and let $U \in \mathbb{R}$ be given and extend the vectors by defining $U_0 = U_{J+1} = 0$. Then*

$$-((\frac{\delta^2 U_j}{\Delta x^2})_j, V) \geq \|D^- U\|_2^2 \geq C_p \|U\|_2^2$$

The above also holds using Δ^+ or Δ_0.

Proof. Using the previous Theorem we have

$$-((\frac{\delta^2 U_j}{\Delta x^2})_j, U) = \|D^- U\|_2^2 + U_J^2 \geq \|D^- U\|_2^2$$

Using the discrete Poincaré inequality we obtain the result. □

Note that we have proven stability in L^2 (and H^1) for the discretization of our BVP. Consider for example the Helmholtz equation $\mathcal{L}[u] = -\partial_{xx}u + \lambda u$ with $\lambda > 0$.

Theorem 8.4.5. *Let* $F_j = -\frac{\delta^2 U_j}{\Delta x^2} + \lambda U_j$ *for* $U \in \mathbb{R}^J$ *defining* $U_0 = U_J = 0$. *Then*

$$\|U\|_2 \leq C\|F\|_2$$

. Therefore the finite difference approximation based on the central difference scheme in stable in the L^2 *norm.*

Proof. Taking the scalar product of the discrete scheme with U leads to

$$(F, U) = -((\frac{\delta^2 U_j}{\Delta x^2})_j + \lambda U, U) \geq C_p\|U\|_2^2 + \lambda(U, U)$$

using the previous results. Therefore

$$\|U\|_2^2 \leq \frac{1}{1+\lambda}(F, U) \leq \|F\|_2\|U\|_2 \ .$$

Dividing by $\|U\|_2$ prove the result. □

Remark: The result above can be easily extended to higher dimensions.

8.4.2 Von Neumann Stability Analysis

This approach can be applied for problems which are periodic in space. Thus we assume that our domain is $[0, 2\pi]$ and we have lattice consisting of equidistant points $x_j = jh$ $(j = 0, \ldots, J)$ with $Jh = 2\pi$. Our discrete solution vector $U^n \in \mathbb{R}^{J-1}$ and we assume that $U_J^n = U_0^n$. The van Neumann stability analysis uses a special form of perturbation based on Fourier modes to study the stability of numerical schemes. We thus study initial conditions of the form $H_l = \sum_{k=0}^{J-1} \alpha_k^0 \omega_l^k$ with Here we use $\omega_j = e^{ijh}$ where i is the imaginary number $\sqrt{-1}$. We define the vectors $\omega^k = (\omega_j^k)_{j=0}^{J-1}$ and use the abbreviation $\omega = \omega^1$. Note the following properties of the fourier modes ω_j:

1. $\bar{\omega}_j = e^{-ijh}$

2. $\omega_{j+l} = \omega_l \omega_j$.

3. $(\omega^k, \omega^l) = \begin{cases} J & l \equiv k \mod J \\ 0 & \text{otherwise} \end{cases}$

4. $\sum_{j=0}^{J-1} \omega_k^j \omega_l^j = (\omega^k, \omega^l)$

5. For a vector $(U_l)_{l=0}^{J-1} = (\sum_{k=0}^{J-1} \alpha_k \omega_l^k)_{l=0}^{J-1}$ we have $\|U\|_2 = \|\alpha\|_2$

6. If $(U_l)_{l=0}^{J-1} = (\sum_{k=0}^{J-1} \alpha_k \omega_l^k)_{l=0}^{J-1}$ then $\alpha_k = \frac{1}{J} \sum_{l=0}^{J-1} U_l \bar{\omega}_l^k$

This properties show that the vectors $(\omega^k)_{k=0}^J$ form a basis of $\mathbb{R}^J$. Thus the solution to the discrete evolution equation U^n can be written in the form

$$U_l^n = \sum_{k=0}^{J-1} \alpha_k^n \omega_l^k \ .$$

Example 8.4.6. *Consider the periodic transport equation on $[0, 2\pi] \times [0, T]$ with constant transport velocity $c > 0$:*

$$\begin{array}{rcll} \frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} & = & 0 & (x,t) \in (0, 2\pi) \times (0, T), \\ u(x,0) & = & g(x) & x \in [0, \pi] \\ u(\pi, t) & = & u(0,t) & t \in (0, T). \end{array} \tag{8.6}$$

We use an explicit in time, backward difference in space discretization:

$$\begin{array}{rcll} \frac{U_j^{n+1} - U_j^n}{\Delta t} + \frac{c}{h}\left(U_j^n - U_{j-1}^n\right) & = & 0 & (j,n) \in \{0, \ldots, J-1\} \times \{0, \ldots N-1\}, \\ U_j^0 & = & g(x_j) & j \in \{0, \ldots, J\} \\ U_J^n & = & U_0^n & n \in \{1, \ldots, N\}. \end{array}$$

*Inserting our ansatz (**??**) we arrive at*

$$\sum_{k=0}^{J-1} \left(\tfrac{1}{\Delta t}(\alpha_k^{n+1} - \alpha_k^n) + \tfrac{c}{h}(\alpha_k^n - \alpha_k^n \bar{\omega_1}^k) \right) \omega_j^k \ = \ 0 \qquad j \in \{0, \ldots, J-1\}.$$

Since $(\omega^1, \ldots, \omega^{J-1}$ are linear independent the above equation implies that each coefficint in the sum has to be zero:

$$\tfrac{1}{\Delta t}(\alpha_k^{n+1} - \alpha_k^n) + \tfrac{c}{h}(\alpha_k^n - \alpha_k^n \bar{\omega_1}^k) \ = \ 0 \qquad k \in \{0, \ldots, J-1\}.$$

As in the last chapter we define $r = \frac{c\Delta t}{h}$ *and defining* $M_k = 1 + r(\bar{\omega_1}^k - 1)$ *we arrive at*

$$\alpha_k^{n+1} = M_k \alpha_k^n \qquad k \in \{0, \ldots, J-1\}.$$

The factor M_k *describes the growth or decay of the* k*-th fourier mode during one time step. We can iterate the argument to obtain*

$$\alpha_k^N = (M_k)^n \alpha_N^0 \qquad k \in \{0, \ldots, J-1\}.$$

which describes how the each fourier mode of the perturbation in the initial data grow during the evolution from $t = 0$ *to* $t = T$*. A given fourier mode will grow if* $|M_k| > 1$ *and decay for* $|M_k| < 1$*. In this example will can compute* $|M_k|^2 = 1 - 4r(1-r)\sin^2(\frac{\pi k}{J})$ *where we have used Euler's formula* $e^{i\phi} = \cos(\phi) + i\sin(\phi)$ *and the half angle formula* $\sin^2(\frac{\phi}{2}) = \frac{1}{2}(1 - \cos(\phi))$*. In this case we see that no fourier mode will grow if* $r \in [0, 1]$ *- this was the Courant restriction we used to prove Theorem* **??***.*

We have computed $U_l^N = \sum_{k=0}^{J-1} (M_k)^N \alpha_k^0 \omega_l^k$ *and thus* $\|U^N\|_2^2 = J \sum_{k=0}^{J-1} |M_k|^{2N} |\alpha_k^0|^2$ *using the identity (***??***). Since* $|M_k| \leq 1$ *for all* k *if* $r \in [0, 1]$*, we conclude that*

$$\|U^N\|_2^2 \leq J \sum_{k=0}^{J-1} \alpha_k^0|^2 = \|U^0\|_2^2 \ .$$

Which corresponds to the result from Theorem **??***.*

Example 8.4.7. *Consider the periodic heat equation on* $[0, 2\pi] \times [0, T]$ *using an explicit in time, central difference in space discretization:*

$$\begin{array}{rcll} \frac{U_j^{n+1} - U_j^n}{\Delta t} - \frac{1}{h^2}\left(U_{j-1}^n - 2U_j^n + U_{j-1}^n\right) & = & 0 & (j,n) \in \{0, \ldots, J-1\} \times \{0, \ldots N-1\}, \\ U_j^0 & = & g(x_j) & j \in \{0, \ldots, J\} \\ U_J^n & = & U_0^n & n \in \{1, \ldots, N\}. \end{array}$$

*Inserting our ansatz (***??***) and repeating the arguments from the previous example we arrive at:*

$$\frac{1}{\Delta t}(\alpha_k^{n+1} - \alpha_k^n) - \frac{c}{h^2}(\alpha_k^n \omega_1^k - 2\alpha_k^n + \alpha_k^n \bar{\omega_1}^k) \ = \ 0 \qquad k \in \{0, \ldots, J-1\}.$$

As in the last chapter we define $r = \frac{\Delta t}{h^2}$ *and defining* $M_k = 1 + r(\omega_1^k - 2 + \bar{\omega_1}^k) = 1 + 2r(\cos(kh) - 1)$ *we arrive at*

$$\alpha_k^{n+1} = M_k \alpha_k^n \qquad k \in \{0, \ldots, J-1\}.$$

Note that $1 - 4r \leq M_k \leq 1$ *so that* $|M_k| \leq 1$ *if* $r \leq \frac{1}{2}$*. Thus we obtain* $\|U^N\|_2 \leq \|U^0\|_2$ *under the restrictions used in Theorem* **??***.*

When applying the von Neumann method for a one step time discretization scheme, we always arrive find a solution of the form

$$U_j^N = \sum_{k=0}^{J-1} (M_k)^N \alpha_k^0 \omega_j^k \ . \tag{8.7}$$

Initial perturbation grow if $|M_k| > 1$ and decay for $\|M_k\| < 1$. But note that methods can be stable even if $\|M_k\| > 1$ as the next example shows.

Example 8.4.8. *Study the forward in time, central in space finite difference discretization of the periodic transport equation:*

$$\begin{array}{rcll} \frac{U_j^{n+1}-U_j^n}{\Delta t} + \frac{c}{2h}\left(U_{j+1}^n - U_{j-1}^n\right) & = & 0 & (j,n) \in \{0,\ldots,J-1\} \times \{0,\ldots N-1\}, \\ U_j^0 & = & g(x_j) & j \in \{0,\ldots,J\} \\ U_J^n & = & U_0^n & n \in \{1,\ldots,N\}. \end{array}$$

This implies that

$$\frac{1}{\Delta t}(\alpha_k^{n+1} - \alpha_k^n) + \frac{c}{2h}(\alpha_k^n \omega_1^k - \alpha_k^n \bar{\omega_1}^k) \ = \ 0 \qquad k \in \{0,\ldots,J-1\}.$$

Thus $\alpha_k^{n+1} = (1 - \frac{1}{2} r(\omega_1^k - \bar{\omega_1}^k))\alpha_k^n = M_k \alpha_k^n$ *with* $r = \frac{c\Delta t}{h}$ *and* $M_k = 1 - ir\cos(kh)$. *As before we thus obtain the representation* $\alpha_k^N = (M_k)^N \alpha_k^0$. *A simple computation shows* $|M_k|^2 = 1 + r^2 \sin^2(kh)$. *Thus it is not possible to chose* r *so that* $|M_k|$ *remains bounded and thus all fourier modes will grow during each time step. Note that to reach a fixed end time* T *we must increase the number of time steps used if we decrease* Δt *so that this growth will become more dominant if we decrease* Δt. *Nevertheless the method is stable as we have shown in Theorem* **??** *under the condition* $r \leq \Delta t$. *Note that* $|M_k|^2 \leq 1 + r^2$.

If $|M_k| \leq 1$ is a desirable property or not depends on the problem under consideration. But is a limit to the amount that $|M_k|$ may exceed unity:

Definition 8.4.9. A finite difference scheme satisfies the *von Neumann condition* if there exists a positive constant $c > 0$ that is independent of $\Delta t, \Delta x$, and k, so that

$$|M_k| \leq 1 + c\Delta t \quad \forall \Delta t \leq \Delta t^*, h \leq h^* \ .$$

For a certain class of schemes, the von Neumann condition is a required and sufficient condition for stability as expressed in the following Theorem:

Theorem 8.4.10. *A constant coefficient scalar one level finite difference scheme is stable in the $\|\cdot\|_2$-norm if and only if it satisfies the von Neumann conditions.*

Proof. Suppose the von Neumann condition is satisfied. Using (**??**) we obtain

$$\|U^N\|_2^2 = J\sum_{k=0}^{J-1} |M_k|^{2N}|\alpha_k^0|^2 \leq (1+c\Delta t)^{2N}\|U^0\|_2^2 \leq e^{2cN\Delta t}\|U^0\|_2^2 \leq e^{2cT}\|U^0\|_2^2$$

where we used $(1+z) \leq e^z$. Thus we have $\|U^N\|_2 \leq C\|U^0\|_2^2$.

□

We conclude with an example using an implicit time discretization scheme.

Example 8.4.11. *We perform the von Neumann stability analysis for the heat equation with periodic boundary conditions using a backward Euler method in time and a central method in space. Inserting the Fourier expansion into the finite difference scheme and rearranging terms leads to $\alpha_k^N = (M_k)^N\alpha_k^0$ with*

$$M_k = 1 - \frac{2r(1-\cos(kh))}{1+2r(1-\cos(kh))} = 1 - \frac{4r\sin^2(\frac{1}{2}kh)}{1+4r\sin^2(\frac{1}{2}kh)} = \frac{1}{1+4r\sin^2(\frac{1}{2}kh)}\ .$$

Here we used $r = \frac{\Delta t}{h^2}$. Now if we require $|M_k| < 1$ (note that for the heat equations Fourier modes should decay), we get the stability condition $1 \leq |1+4r\sin^2(kh)|$ which is true for any choice of r. Therefore the method in unconditionally stable (compare with Theorem 7.2.3).

Example 8.4.12. *We conclude with an last example. Consider the transport equation with $c > 0$ and the forward Euler method in time. Consider the following discretization for the transport term*

$$\frac{1}{12h}(u_{i-2} - 8u_{i-1} + 8u_{i+1} - u_{i-2})$$

which is a fourthed order consistent approximation. The scheme is stable for $r = \frac{h}{\Delta t} \leq 1$.

One can show that $|M_k|^2 = 1 + r^2 m_k$ with $m_k = -\frac{1}{36}(8\sin(kx) - \sin(2kx))^2$. The results follows from $m_k \in (-2, 0]$.

Part III

Finite Element Methods

Chapter 9

Mathematical Background

When studying finite element methods we will study the weak or variational form of the partial differential equation. For our elliptic problem (1.1) the variation form is obtained by multiplying with some function $v \in C^1(\Omega)$ having compact support and then integrating over Ω. After using integration by parts one obtains in the case $g = 0$ the following problem:

$$\int_\Omega (\nabla u \cdot \nabla v + p \cdot \nabla uv + quv) = \int_\Omega fv$$

for all $v \in C_0^1(\Omega)$. If the solution u to this problem is in C^2 then the variational formulation is also the classical solution, i.e, satisfies the pde (1.1) pointwise. Note that in modeling a problem based on the concept of energy minimization (e.g. elasticity) the variation form is very natural.

It turns out that on general domains (1.1) does not have a classical solution but does have a solution in the variational sense. This solution is an element of the Sobolev space $H_0^1(\Omega)$ which is a subset of $L^2(\Omega)$ containing function which have a derivative and zero boundary conditions in a weak sense.

9.1 Sobolev Spaces

Note: the following only gives a brief overview of the Sobolev spaces $H^m(\Omega)$

and $H_0^m(\Omega)$.

We will use $L^2(\Omega) = \{u\colon \Omega \to \mathbb{R} : \|u\|_2 < \infty\}$ which is a Hilbert space with norm $\|u\|_2^2 = \int_\Omega u^2(x)\, dx$. The integral we use here is the Lebesque integral, which corresponds to the Riemann integral in the case of continuous functions but allows the integration of more general functions. For two function $u, v \in L^2$ we have $u = v$ if $u(x) = v(x)$ for almost all $x \in \Omega$, i.e., for all $x \in \Omega \setminus N$ where N is a set with measure zero. For example countable sets of points and lines in 2D have zero measure. The evaluation of $u \in L^2(\Omega)$ at some point $x \in \Omega$ or on the boundary of Ω is thus not meaningful. For the following we will not require a more formal definition of L^2.

Definition 9.1.1 (Weak derivatives)**.** A function $u \in L^2(\Omega)$ has weak derivatives if there are functions $w_i \in L^2(\Omega)$ for $i = 1, \ldots, d$ satisfying

$$\int_\Omega w_i \varphi = -\int_\Omega u \partial_i \varphi$$

for all $\varphi \in C_0^1(\Omega)$. In this case we use the notation $\partial_i u = w_i$. In a similar way we define the gradient of u or the divergence of a vector valued function, e.g., the weak divergence of a function $u \in [L^2(\Omega)]^d$, is a function $\nabla \cdot u \in L^2(\Omega)$ satisfying $\int_\Omega \nabla \cdot u\varphi = -\int_\Omega u \cdot \nabla\varphi$ for all $\varphi \in C_0^1(\Omega)$.

We now define the *Sobolev* space

$$H^1(\Omega) = \{u \in L^2(\Omega)\colon u \text{ has a weak gradient } \nabla u \in [L^2(\Omega)]^d\}.$$

Similar we can define weak derivatives of higher order and corresponding spaces $H^m(\Omega)$ with functions u having derivatives $\partial^\alpha u \in L^2(\Omega)$ where α is a multiindex with $|\alpha| \le m$.

The space $H^m(\Omega)$ is a Hilbert space with scalar product given by

$$(u, v)_{H^m} = \sum_{|\alpha| < m} \int_\Omega \partial^\alpha u \partial^\alpha v.$$

Note that $H^0(\Omega) = L^2(\Omega)$ with the same scalar product $(u, v)_{H^0} = (u, v)_2$.

Theorem 9.1.2. *For each $u \in H^1((a, b))$ there is a function $\bar{u} \in C^\infty((a, b))$ with $u = \bar{u}$ almost everywhere.*

For each $u \in H^2(\Omega)$ with $\Omega \subset \mathbb{R}^2$ there is a function $\bar{u} \in C^0(\Omega)$ with $u = \bar{u}$ almost everywhere if Ω is a bounded domain with Lipschitz boundary.

This Theorem shows that functions in H^m can be thought of as being continuous if m is high enough (depending on the space dimension). This type of result is refered to as *Sobolev embeddings*.

Theorem 9.1.3. *Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with Lipschitz boundary and let Ω_1, Ω_2 be bounded domain with Lipschitz boundary which form a partition of Ω, i.e., $\Omega_1 \cup \Omega_2 = \Omega$ and $\Omega_1 \cap \Omega_2 = \emptyset$. Then if $u \in C^0(\Omega)$ and $u|_{\Omega_i} \in C^1(\Omega_i)$ for $i = 1, 2$ then $u \in H^1(\Omega)$.*

The following Theorem provides a different view of function in H^m:

Theorem 9.1.4. *For each $u \in H^m(\Omega)$ there is a sequence $(v_i)_{i\in\mathbb{N}} \subset C^\infty(\Omega) \cap H^m(\Omega)$ with $\|u - v_i\|_{H^m} \to 0$ as $i \to \infty$. Thus $C^\infty(\Omega) \cap H^m(\Omega)$ is dense in $H^m(\Omega)$ with respect to the $\|\cdot\|_{H^m}$ norm.*

Finally we need to define functions which have zero boundary conditions in a weak sense.

Definition 9.1.5. We define $H_0^m(\Omega)$ to be the set of all functions for which there exists a sequence $(v_i)_{i\in\mathbb{N}} \subset C_0^\infty(\Omega) \cap H^m(\Omega)$ with $\|u - v_i\|_m \to 0$ as $i \to \infty$. Note that $H_0^m \subset H^m$.

Note that in general some interpretation of $u|_{\partial\Omega}$ is not meaningful for function in L^2. But in 2D it is possible to give $u \in H^1(\Omega)$ restricted to the boundary some meaning as function in $L^2(\Omega)$. In this sense we have for functions $u \in H_0^1(\Omega)$ that $u = 0$ almost everywhere on $\partial\Omega$. The corresponding result is often refered to as *trace theorem*.

We can summarize the relation between the Sobolev spaces with the following diagram:

$$\begin{array}{rcccccccc} L^(\Omega) & = & H^0(\Omega) & \supset & H^1(\Omega) & \supset & H^2(\Omega) & \supset & \cdots \\ & & \cup & & \cup & & \cup & & \\ & & H_0^0(\Omega) & \supset & H_0^1(\Omega) & \supset & H_0^2(\Omega) & \supset & \cdots \end{array}$$

We conclude with the Poincarre inequality which we already encountered in chapter 5:

Lemma 9.1.6. *If $v \in H_0^1(\Omega)$ then there is a constant $C_p > 0$ such that*

$$\|v\|_2^2 \leq C_p \|\nabla v\|_2^2.$$

9.2 Introduction to Finite Element Methods

In this chapter we will again study our elliptic problem (1.1) with homogeneous boundary conditions

$$\begin{aligned} -\Delta u + p.\nabla u + qu &= f, \quad x \in \Omega, \\ u &= 0, \quad x \in \partial\Omega. \end{aligned}$$

The finite element method is based on the weak or variational form of this problem: Let $\mathcal{V} = H_0^1(\Omega)$ and define the bilinear form and L^2 inner-product, respectively, by

$$\begin{aligned} a(u,v) &= \int_\Omega \left[\nabla u \cdot \nabla v + (p \cdot \nabla u)v + quv\right] dx, \\ \langle u, v\rangle &= \int_\Omega uv dx. \end{aligned} \tag{9.1}$$

Then the weak formulation of (1.1) is to find

$$u \in \mathcal{V} : \; a(u,v) = \langle f, v\rangle \quad \forall v \in \mathcal{V}. \tag{9.2}$$

Under certain regularity assumptions on f this variational formulation is equivalent to the original strong form of the problem for classical solutions.

9.3 Galerkin Method

Let $\mathcal{V}$ be a Hilbert space. Consider a weak formulation of a linear PDE specified via a bilinear form $a : \mathcal{V} \times \mathcal{V} \longrightarrow \mathbb{R}$, and a linear form $L : \mathcal{V} \longrightarrow \mathbb{R}$ to give the problem of finding

$$u \in \mathcal{V} : \; a(u,v) = L(v) \quad \forall v \in \mathcal{V}.$$

This problem can be approximated by specifying a finite dimensional subspace $\mathcal{V}^h \subset \mathcal{V}$ and seeking a solution in $\mathcal{V}^h$ instead. This leads to a finite dimensional problem to be solved for the approximation u^h:

$$u^h \in \mathcal{V}^h : a(u^h, v) = L(v) \quad \forall v \in \mathcal{V}^h.$$

This is known as the **Galerkin method**. For finite element methods, one uses $\mathcal{V}^h = \text{span}\{\phi_j\}_{j=1}^N$ where ϕ_j is locally supported on some mesh or grid. The idea extends to problems which are time-dependent, with a variational structure in space which is analogous to that outlined here.

Notice that for finite difference methods the domain of the unknown is approximated by a finite lattice. In contrast, finite element methods approximate the space where the unknown function lies, by restricting the solution to a finite dimensional subspace.

9.4 Norms

Consider a function $u : \Omega \to \mathbb{R}$ with $\Omega \subset \mathbb{R}^d$. When studying finite element methods we will use norms in the Sobolev spaces L^2, H^1 and H^2, as well as the corresponding semi-norms. We use the notation

$$\|\cdot\|_{L^2} := \|\cdot\|_2, \ \|\cdot\|_{H^1} := \{\|\cdot\|_{L^2}^2 + \|\nabla\cdot\|_{L^2}^2\}^{\frac{1}{2}}, \ \|\cdot\|_{H^2} := \{\|\cdot\|_{H^1}^2 + \|\nabla\nabla\cdot\|_{L^2}^2\}^{\frac{1}{2}} \tag{9.3}$$

for the norms, with the L^2 norm defined through the standard inner-product (3.12), and the notation

$$|\cdot|_{H^1} = \|\nabla\cdot\|_{L^2} \leq \|\cdot\|_{H^1}, \ |\cdot|_{H^2} := \|\nabla\nabla\cdot\|_{L^2} \leq \|\cdot\|_{H^2} \tag{9.4}$$

for the corresponding semi-norms. When the bilinear form $a(\cdot,\cdot)$ defines an inner-product then we use the notation

$$\|u\|_a = \{a(u,u)\}^{\frac{1}{2}}. \tag{9.5}$$

9.5 Consistency and Stability

For finite difference methods all our proofs proceeded, either explicitly or implicitly, via a consistency and a stability inequality. In finite element methods there is an analogous structure which we will exploit and we outline it here for time-independent problems.

The analogue of **stability** is the **Galerkin orthogonality property** which states that, for $e = u - u^h$,

$$a(e, v) = 0 \quad \forall v \in \mathcal{V}^h. \tag{9.6}$$

If the bilinear form $a(\cdot, \cdot)$ induces an inner-product, and then norm via (9.5), then the orthogonlaity property states that the error is always orthogonal to the subspace $\mathcal{V}^h$, in that inner-product. From this we obtain the following abstract result which underlies most proofs of convergence for finite element methods.

Theorem 9.5.1. *The approximation u^h is the optimal appropximation to u in $\mathcal{V}^h$ in the sense that*

$$\|u - u^h\|_a \leq \|u - v\|_a \quad \forall v \in \mathcal{V}^h. \tag{9.7}$$

Proof. For any $v \in \mathcal{V}^h$, the orthogonality property (9.6) gives

$$\begin{aligned} a(e, e) &= a(e, e + u^h - v) && (9.8) \\ &= a(e, u - v) && (9.9) \end{aligned}$$

Thus, by Cauchy-Schwarz,

$$\|e\|_a^2 \leq \|e\|_a \|u - v\|_a \quad \forall v \in \mathcal{V}^h \tag{9.10}$$

implying (9.7). □

Let P^h denote some projection from $\mathcal{V}$ to $\mathcal{V}^h$. The analogue of **consistency** is to bound

$$\|u - P^h u\|_a \leq Ch^r \tag{9.11}$$

where h is some measure of the fineness of the mesh underlying the finite element method.

Combining (9.7) with $v = P^h u$ and (9.11) gives the **convergence** result

$$\|u - u^h\|_a \leq \|u - P^h u\|_a \leq Ch^r. \tag{9.12}$$

Analogous ideas hold for time-dependent problems, but then the Galerkin orthogonality property (9.6) is generalized to a time-dependent statement about the error. Integrating this gives a stability estimate, and then a consistency estimate like (9.11) gives convergence.

Chapter 10

Boundary Value Problems

10.1 Introduction

In this chapter we study finite element methods and their application to the Elliptic Problem (1.1) in arbitrary dimension. Our setting will be the weak formulation of the problem (9.2). We use the norms and semi-norms given by (9.3)–(9.5), as well as the standard L^2 inner-product $\langle \cdot, \cdot \rangle$ given by (3.12).

10.2 Laplace's Equation

For simplicity we concentrate on Laplace's equation. We start the development in arbitrary dimension and with arbitrary apporixmation space, but specify to $d = 2$ and piecewise linear approximation space in later developments.

10.2.1 The PDE

We consider the Elliptic Problem (1.1) in dimension $d \leq 2$ with $\Omega \subset \mathbb{R}^d$ bounded and open. We also take $p = 0$ for simplicity, and assume that $q \geq 0$

in $\bar{\Omega}$. This gives

$$\begin{aligned} -\Delta u + qu &= f, \quad x \in \Omega, \\ u &= 0, \quad x \in \partial\Omega. \end{aligned} \tag{10.1}$$

The weak form of this problem is defined from (9.2) as follows. Let $\mathcal{V} = H_0^1(\Omega)$ and

$$a(u,v) = \int_\Omega \{\nabla u \cdot \nabla v + quv\} dx.$$

Then we have to find

$$u \in \mathcal{V} : \; a(u,v) = \langle f, v \rangle \quad \forall v \in \mathcal{V}. \tag{10.2}$$

This can be obtained by multiplying (10.1) by $\mathcal{V}$ and using Green's formula. Classical solutions of (10.1) solve (10.2), and the converse is also true under regularity conditions on f and Ω.

10.2.2 The Approximation

We let $\mathcal{V}^h \subset \mathcal{V}$ be a finite dimensional approximating space. This gives rise to the finite element method to find

$$u^h \in \mathcal{V}^h : \; a(u^h, v) = \langle f, v \rangle \quad \forall v \in \mathcal{V}^h \tag{10.3}$$

Since $\mathcal{V}^h$ is finite-dimensional we may assume that

$$\mathcal{V}^h = \text{span}\{\Psi_j, \, j \in \mathcal{J}\}.$$

Let $M = |\mathcal{J}|$, the cardinality of $\mathcal{J}$. Then (10.3) is equivalent to

$$u^h \in \mathcal{V}^h : \; a(u^h, \Psi_i) = \langle f, \Psi_i \rangle \quad \forall i \in \mathcal{J}.$$

Furthermore, an arbitrary element of $\mathcal{V}^h$ can be expanded in terms of the Ψ_i; in particular we may write

$$u^h(x) = \sum_{j \in \mathcal{J}} U_j \Psi_j(x),$$

noting that determination of $U = (\ldots, U_j, \ldots) \in \mathbb{R}^M$ is equivalent to determination of u^h. Thus we obtain the linear system

$$AU = F \tag{10.4}$$

where

$$\begin{aligned} U &= (U_1, \ldots, U_M)^T, \\ F &= (F_1, \ldots, F_M)^T, \\ F_i &= \frac{1}{h_i}(f, \Psi_i), \\ A_{ij} &= \frac{1}{h_i} a(\Psi_i, \Psi_j), \end{aligned}$$

and

$$h_i = \frac{1}{2}\text{Vol}\{\text{ssupp}(\Psi_i)\}.$$

The scaling of each equation by h_i is chosen to make the relationship with finite difference methods transparent.

One Dimensional Example

As a concrete example we consider (10.1) with $q(x) \equiv \mu \geq 0$ a constant. We approximate $\mathcal{V}$ by piecewise linear continuous functions on an equi-partition of $\Omega = (0, 1)$ so that

$$\mathcal{V}^h = \{v \in \mathcal{V} : \, v \, \text{is linear on} \, I_j, \, j = 1, \ldots, J\} \bigcap C^0(0, 1).$$

We set $I_j = [x_{j-1}, x_j]$ for $x_l = l\Delta x$, $J\Delta x = 1$.

If we define $\Psi_j(x)$ to be the unique piecewise linear function satisfying

$$\begin{aligned} \Psi_j(x_i) &= \delta_{ij}, \\ \Psi_j|_{I_i} &\text{ is } \text{linear}, \end{aligned}$$

then

$$\mathcal{V}^h = \text{span}\{\Psi_j, \, j = 1, \ldots, J-1\}.$$

The explicit form of the Ψ_j is given as:

$$\Psi_j(x) = \begin{cases} (x - x_{j-1})/\Delta x, & x \in I_j, \\ (x_{j+1} - x)/\Delta x & x \in I_{j+1}, \\ 0, & x \notin I_j \cup I_{j+1} \end{cases}$$

If $v(x) \in \mathcal{V}^h$ is written as

$$v(x) = \sum_{j=1}^{J-1} V_j \Psi_j$$

then $V_i = v(x_i)$. Thus an expansion in the basis defined by the Ψ_j will lead naturally to equations for nodal values of the approximation.

Now

$$\begin{aligned} \int_0^1 \Psi_j^2(x)dx &= 2\int_{x_{j-1}}^{x_j} \frac{(x-x_{j-1})^2}{\Delta x^2}dx \\ &= 2\int_0^1 y^2 \Delta x dy \\ &= \frac{2}{3}\Delta x \end{aligned}$$

and

$$\begin{aligned} \int_0^1 \Psi_j(x)\Psi_{j-1}(x)dx &= \int_{x_{j-1}}^{x_j} \frac{(x-x_{j-1})(x_j-x)}{\Delta x^2}dx \\ &= \int_0^1 y(1-y)\Delta x dy \\ &= \frac{\Delta x}{6}. \end{aligned}$$

Also

$$\begin{aligned} \int_0^1 \left(\frac{d\Psi_j}{dx}\right)^2 dx &= 2\int_{x_{j-1}}^{x_j} \frac{1}{\Delta x^2}dx \\ &= \frac{2}{\Delta x} \end{aligned}$$

and

$$\begin{aligned} \int_0^1 \frac{d\Psi_j}{dx}\frac{d\Psi_{j-1}}{dx} &= \int_{x_{j-1}}^{x_j} \frac{-1}{\Delta x^2} \\ &= \frac{-1}{\Delta x}. \end{aligned}$$

Finally

$$F_j \;=\; \frac{1}{\Delta x}\int_0^1 \Psi_j f dx.$$

and introducing

$$A_1 = \frac{1}{6}\begin{pmatrix} 4 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 4 \end{pmatrix}, \qquad A_0 = \frac{-1}{\Delta x^2}\begin{pmatrix} -2 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{pmatrix}$$

we obtain from (10.4) the linear system

$$(A_0 + \mu A_1)U = F. \tag{10.5}$$

Notice that the approximation (10.5) is very similar to a finite difference approximation of the heat equation: the matrix A_0 is identical to the standard approximation of the negative of the second derivative. However, the overall approximation differs in two ways: firstly, the term involving μ contains the matrix A_1 where the identity would naturally appear in a finite difference approximation (replacing A_1 by the identity is known as **mass-lumping**); secondly F is found by testing the function f against scaled basis functions, not against delta functions.

Two Dimensional Example

We now study the case where $\Omega \subset \mathbb{R}^2$ is bounded and polygonal, $q \equiv 0$ and we write

$$\Omega \;=\; \cup_{K\in T_h} K$$

where $T_h = \{K_1, \ldots, K_m\}$ is a set of triangles satisfying $K_i \cap K_j = \emptyset$. We consider the approximation space to be made up of piecewise linear functions, which are linear on each triangle and continuous across edges. This is encapsulated in the definition

$$\mathcal{V}^h \;=\; \{v \in \mathcal{V} :\; v(\cdot)|_K \,\text{is linear}\, \forall K \in T_h\} \bigcap C(\overline{\Omega}, \mathbb{R}).$$

Here, $v(\cdot)|_K$ denotes the restriction to K of $v(\cdot)$.

Any $v \in \mathcal{V}^h$ is then uniquely specified by the values at the nodes $\{N_i\}_{i=1}^M$ of the triangles. Define $\{\Psi_j(x)\}_{j=1}^M$, $\Psi_j \in \mathcal{V}^h$ by

$$\Psi_j(N_i) \quad = \quad \delta_{ij}$$

Then the Ψ_j span $\mathcal{V}^h$ and so, if

$$v(x) \quad = \quad \sum_{j=1}^{M} V_j \Psi_j(x) \quad \forall v \in \mathcal{V}^h,$$

then $V_i = v(N_i)$.

Writing

$$u^h(x) \quad = \quad \sum_{j=1}^{M} U_j \Psi_j(x)$$

we obtain the linear system (10.4) as above. For simple geometries Ω and uniform partitions T_h the matrix A is the same as that appearing in the finite difference approximation (4.12). The right hand side, however, differs; in the finite difference case it is found by integrating f against delta measures whereas for finite elements a scaled basis functions is used.

10.2.3 Convergence

For the convergence analysis we concentrate on the preceeding two examples if approximation by piecewise linear functions. In one or two dimensions we let $\Delta_h = \{$a family of triangulations T_h of $\Omega\}$. with "triangulation" refering to writing Ω as the union of intervals in one dimension.

Definition 10.2.1. Given $v \in C(\Omega, \mathbb{R})$, $P^h v$ denotes the piecewise linear interpolant of v on a given triangulation.

Thus

$$(P^h v)(x) = \sum_{j=1}^{M} v(N_j) \Psi_j(x).$$

In the following we use the notation:

- h_K = longest side of a triangle in $K \subset T_h$;
- ρ_K = diameter of the largest circle inscribing a triangle in $K \subset T_h$;
- $r_K = \rho_K / h_K$ for any triangle $K \subset T_h$;
- $h = \max_{K \subset T_h} h_K$.

In the one dimensional example we have $h_k = \rho_k = h = \Delta x$ and $r_K = 1$. Notice that, for bounded $\Omega \subset \mathbb{R}^d$, $d \leq 2$, $v \in H^2(\Omega)$ is continuous and thus we may apply P^h to it. The following consistency result will be useful in what follows.

Lemma 10.2.2. *Let $v \in H^2(\Omega)$. Assume that $\exists \beta$, $h_c : \ r_K > \beta$ for all $K \subset T_h$ and $h \in (0, h_c)$. Then $\exists C > 0$:*

$$\begin{aligned} \|v - P^h v\|_{L^2} &\leq Ch^2 |v|_{H^2}, \\ |v - P^h v|_{H^1} &\leq C\frac{h}{\beta} |v|_{H^2}. \end{aligned}$$

Theorem 10.2.3. *Consider u solving (10.2) and u^h solving (10.3). If $u \in H^2(\Omega)$ and $\exists \beta$, $h_c : \ r_K > \beta$ for all $K \subset T_h$ and $h \in (0, h_c)$ then $\exists C > 0$:*

$$\|u - u^h\|_{H^1} \leq Ch|h|_{H^2} \quad \forall h \in (0, h_c).$$

Proof. Let $e = u = u^h$. From Theorem 9.5.1 and the definition of a we have that

$$|e|_{H^1} \leq |u - v|_{H^1} \quad \forall v \in \mathcal{V}^h.$$

Choosing $v = P^h u$ and applying Lemma 10.2.2 gives, by (9.4),

$$\|\nabla e\|_{L^2} \leq \frac{C}{\beta} h |u|_{H^2}. \tag{10.6}$$

But, for $v \in \mathcal{V}$, (9.3) and the Poincaré inequality (5.2) gives

$$\|e\|_{H^1}^2 \leq (1 + C_p) \|\nabla e\|_{L^2}^2$$

and so the result follows. □

Theorem 10.2.4. *Under the assumptions of the previous theorem,* $\exists C > 0$*:*

$$\|u - u^h\|_{L^2} \leq Ch^2|u|_{H^2} \quad \forall h \in (0, h_c).$$

Proof. As in the proof of Theorem 9.5.1 we start from the orthogonality property for the error:

$$a(e, v) = 0 \quad \forall v \in \mathcal{V}^h.$$

Now let ϕ solve

$$a(\phi, v) = \langle e, v \rangle \quad \forall v \in \mathcal{V}.$$

By elliptic regularity we have

$$\|\phi\|_{H^2} \leq C_{\text{stab}}\|e\|_{L^2}.$$

Now

$$\begin{aligned} \langle e, e \rangle &= a(e, \phi) \\ &= a(e, \phi - P^h\phi). \end{aligned}$$

Hence, by (10.6) and Lemma 10.2.2,

$$\begin{aligned} \|e\|_{L^2}^2 &\leq |e|_{H^1}|\phi - P^h\phi|_{H^1} \\ &= \|\nabla e\|_{L^2}\|\nabla(\phi - P^h\phi)\|_{L^2} \\ &\leq \left(\frac{C}{\beta}h|u|_{H^2}\right)\left(C\frac{h}{\beta}|\phi|_{H^2}\right) \\ &\leq Kh^2|u|_{H^2}\|\phi\|_{H^2}. \end{aligned}$$

Thus

$$\|e\|_{L^2}^2 \leq KC_{\text{stab}}h^2\|e\|_{L^2}$$

and the result follows. □