Chapter 5: Numerical Solution of Parabolic Initial-Boundary Value Problems

Given functions $f = f(x,t)$, $(x,t) \in Q := (-a,+a) \times (0,T)$, $u_0 = u_0(x)$, $x \in [-a,+a]$ and $u_\ell = u_\ell(t)$, $u_r = u_r(t)$, $t \in [0,T]$, as well as $\kappa > 0$, consider the following initial-boundary value problem for the heat equation

$$
\begin{array}{lrcll}
(\star)_1 & \dfrac{\partial u}{\partial t} - \kappa \dfrac{\partial^2 u}{\partial x^2} & = f & \in & Q \ , \\
(\star)_2 & u(-a,t) & = & u_\ell(t) & t \in [0,T] \ , \\
(\star)_3 & u(-a,t) & = & u_r(t) & t \in [0,T] \ , \\
(\star)_4 & u(x,0) & = & u_0(x) & x \in [-a,+a] \ .
\end{array}
$$

A function $u \in C^1([0,T]; C^2([-a,+a]))$, which satisfies $(\star)_1 - (\star)_4$ pointwise, is said to be a classical solution of the initial boundary value problem.

5.1 Finite Difference Methods

We partition the computational domain $Q := [-a,+a] \times [0,T]$ into a grid

$$Q_{h,k} \; := \; \{(x_i,t_j) \in Q \mid x_i = -a + ih \;,\; 0 \leq i \leq N \;,\; t_j = jk \;,\; 0 \leq j \leq M\}$$

with step sizes $h := 2a/N \;,\; N \in \mathbb{N}$, and $k := T/M \;,\; M \in \mathbb{N}$.

We approximate the time derivative $\partial u/\partial t$ either by the forward difference quotient

$$(+)_1 \qquad \frac{\partial u}{\partial t}(x,t) \;\approx\; \frac{u(x,t+k) - u(x,t)}{k} \quad , \quad (x,t),(x,t+k) \in Q_{h,k} \;,$$

or by the backward difference quotient

$$(+)_2 \qquad \frac{\partial u}{\partial t}(x,t) \;\approx\; \frac{u(x,t) - u(x,t-k)}{k} \quad , \quad (x,t-k),(x,t) \in Q_{h,k} \;.$$

Moreover, we approximate the spatial derivative $\partial^2 u/\partial x^2$ by the central difference quotient

$$(\circ) \qquad \frac{\partial^2 u}{\partial x^2}(x,t) \;\approx\; \frac{u(x-h,t) - 2u(x,t) + u(x+h,t)}{h^2} \quad , \quad (x \pm h,t),(x,t) \in Q_{h,k} \;.$$

5.1.1 Explicit Difference Scheme

Denoting by $u_{h,k} : Q_{h,k} \to \mathbb{R}$ a grid function on $Q_{h,k}$ and using $(+)_1$ and $(\circ)$, we arrive at the explicit difference scheme

$$(\dagger)_1 \quad \frac{u_{h,k}(x_i,t_{j+1}) - u_{h,k}(x_i,t_j)}{k} - \kappa \frac{u_{h,k}(x_{i-1},t_j) - 2u_{h,k}(x_i,t_j) + u_{h,k}(x_{i+1},t_j)}{h^2} = f(x_i,t_j) ,$$

$$(\dagger)_2 \quad u_{h,k}(x_0,t_j) = u_\ell(t_j) \quad , \quad u_{h,k}(x_N,t_j) = u_r(t_j) \quad , \quad u(x_i,0) = u_0(x_i) .$$

The equation $(\dagger)_1$ can be solved for $u_{h,k}(x_i,t_{j+1})$ resulting in

$$(\dagger)_1' \quad u_{h,k}(x_i,t_{j+1}) = u_{h,k}(x_i,t_j) - \kappa \frac{k}{h^2} \left(-u_{h,k}(x_{i-1},t_j) + 2u_{h,k}(x_i,t_j) - u_{h,k}(x_{i+1},t_j)\right) + k\, f(x_i,t_j) .$$

5.1.2 Implicit Difference Scheme

Using $(+)_2$ and $(\circ)$, we obtain the implicit difference scheme

$$(\ddagger)_1 \qquad \frac{u_{h,k}(x_i,t_j) - u_{h,k}(x_i,t_{j-1})}{k} \;-\; \kappa\, \frac{u_{h,k}(x_{i-1},t_j) - 2u_{h,k}(x_i,t_j) + u_{h,k}(x_{i+1},t_j)}{h^2} \;=\; f(x_i,t_j) \;,$$

$$(\ddagger)_2 \qquad u_{h,k}(x_0,t_j) \;=\; u_\ell(t_j) \quad , \quad u_{h,k}(x_N,t_j) \;=\; u_r(t_j) \quad , \quad u(x_i,0) \;=\; u_0(x_i) \;.$$

The equation $(\ddagger)_1$ represents a linear algebraic system in the unknowns $u_{h,k}(x_i,t_j)$

$$(\ddagger)_1^{'} \quad u_{h,k}(x_i,t_j) \;-\; \kappa\, \frac{k}{h^2}\, \big(u_{h,k}(x_{i-1},t_j) - 2u_{h,k}(x_i,t_j) + u_{h,k}(x_{i+1},t_j)\big) \;=\; u_{h,k}(x_i,t_{j-1}) \;+\; k\; f(x_i,t_j) \;.$$

5.1.2 The θ Method and the Crank-Nicolson Scheme

Given $0 \le \theta \le 1$, multiplying $(+)_2$ by θ (with t_j replaced by t_{j+1}) and $(+)_1$ by $1-\theta$ and adding the resulting equations, yields the so-called θ method

$$(\$)_1 \quad \frac{u_{h,k}(x_i,t_{j+1}) - u_{h,k}(x_i,t_j)}{k} - \kappa\,\theta\,\frac{u_{h,k}(x_{i-1},t_{j+1}) - 2u_{h,k}(x_i,t_{j+1}) + u_{h,k}(x_{i+1},t_{j+1})}{h^2} -$$

$$- \kappa\,(1-\theta)\,\frac{u_{h,k}(x_{i-1},t_j) - 2u_{h,k}(x_i,t_j) + u_{h,k}(x_{i+1},t_j)}{h^2} = \theta\, f(x_i,t_{j+1}) + (1-\theta)\, f(x_i,t_j) \;,$$

$$(\$)_2 \quad u_{h,k}(x_0,t_j) = u_\ell(t_j) \quad , \quad u_{h,k}(x_N,t_j) = u_r(t_j) \quad , \quad u(x_i,0) = u_0(x_i) \;.$$

For $\theta = 0$ we recover the explicit and for $\theta = 1$ the implicit method. The method for $\theta = 1/2$ is called the Crank-Nicolsen scheme.

The θ method can be written as the linear algebraic system

$$(*) \qquad (I_h + A_h(\theta))\, u_h^{j+1} \;=\; B_h(\theta)\, u_h^j \;+\; b_h^j \;.$$

Here, $u_h^j := (u_{h,k}(x_1, t_j), ..., u_{h,k}(x_{N-1}, t_j))^T \in \mathbb{R}^{N-1}$ represents the vector of unknowns, I_h is the $(N-1) \times (N-1)$ unity matrix, and the matrices $A_h(\theta), B_h(\theta) \in \mathbb{R}^{(N-1)\times(N-1)}$ are given by

$$A_h(\theta) \;=\; \begin{pmatrix} 2\kappa\theta\frac{k}{h^2} & -\kappa\theta\frac{k}{h^2} & 0 & \cdot & \cdot & 0 \\ -\kappa\theta\frac{k}{h^2} & 2\kappa\theta\frac{k}{h^2} & -\kappa\theta\frac{k}{h^2} & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & -\kappa\theta\frac{k}{h^2} & 2\kappa\theta\frac{k}{h^2} & -\kappa\theta\frac{k}{h^2} \\ 0 & \cdot & \cdot & 0 & -\kappa\theta\frac{k}{h^2} & 2\kappa\theta\frac{k}{h^2} \end{pmatrix} ,$$

$$B_h(\theta) \;=\; \begin{pmatrix} 1-2\kappa(1-\theta)\frac{k}{h^2} & \kappa(1-\theta)\frac{k}{h^2} & 0 & \cdot & \cdot & 0 \\ \kappa(1-\theta)\frac{k}{h^2} & 1-2\kappa(1-\theta)\frac{k}{h^2} & \kappa(1-\theta)\frac{k}{h^2} & 0 & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \kappa(1-\theta)\frac{k}{h^2} & 1-2\kappa(1-\theta)\frac{k}{h^2} & \kappa(1-\theta)\frac{k}{h^2} \\ 0 & \cdot & \cdot & 0 & \kappa(1-\theta)\frac{k}{h^2} & 1-2\kappa(1-\theta)\frac{k}{h^2} \end{pmatrix} .$$

Moreover, $b_h^j \in \mathbb{R}^{N-1}$ is the vector

$$b_h^j = \begin{pmatrix} \kappa(1-\theta)\frac{k}{h^2}u_\ell(t_j) + \kappa\theta\frac{k}{h^2}u_\ell(t_{j+1}) + k(1-\theta)f(x_1,t_j) + k\theta f(x_1,t_{j+1}) \\ k(1-\theta)f(x_2,t_j) + k\theta f(x_2,t_{j+1}) \\ \cdot \\ \cdot \\ k(1-\theta)f(x_{N-2},t_j) + k\theta f(x_{N-2},t_{j+1}) \\ \kappa(1-\theta)\frac{k}{h^2}u_r(t_j) + \kappa\theta\frac{k}{h^2}u_r(t_{j+1}) + k(1-\theta)f(x_{N-1},t_j) + k\theta f(x_{N-1},t_{j+1}) \end{pmatrix} .$$

Definition 5.1 (Convergence of the θ Method)
The grid function $e_{h,k}(x,t) := u(x,t) - u_{h,k}(x,t)$, $(x,t) \in Q_{h,k}$ is called the global discretization error. The θ method is said to be convergent, if

$$\max_{(x,t)\in Q_{h,k}} |e_{h,k}(x,t)| \to 0 \quad (h,k \to 0) .$$

It is said to be convergent of order p_1 in t and of order p_2 in x, if

$$\max_{(x,t)\in Q_{h,k}} |e_{h,k}(x,t)| = O(k^{p_1} + h^{p_2}) \quad (h,k \to 0) .$$

5.2 Convergence of the θ-method

Definition 5.2 (Consistency and Order of Consistency)

Let $u(x,t)$, $(x,t) \in Q$ be the solution of the given initial-boundary value problem for the heat equation. The grid function $\tau_{h,k}(x,t)$, $(x,t) \in Q_{h,k}$ given by

$$\tau_{h,k}(x_i,t_{j+1}) := \frac{u(x_i,t_{j+1}) - u(x_i,t_j)}{k} - \kappa\,\theta\,\frac{u(x_{i-1},t_{j+1}) - 2u(x_i,t_{j+1}) + u(x_{i+1},t_{j+1})}{h^2} -$$

$$- \kappa\,(1-\theta)\,\frac{u(x_{i-1},t_j) - 2u(x_i,t_j) + u(x_{i+1},t_j)}{h^2} - (\theta\, f(x_i,t_{j+1}) + (1-\theta)\, f(x_i,t_j))$$

is called the local discretization error. The θ method is said to be consistent with the given initial-boundary value problem, if

$$\max\,\{|\tau_{h,k}(x_i,t_{j+1})| \mid 1 \le i \le N-1, 0 \le j \le M-1\} \;\to\; 0 \qquad (h,k \to 0)\ .$$

It is said to be consistent of order p_1 in t and p_2 in x, if

$$\max\,\{|\tau_{h,k}(x_i,t_{j+1})| \mid 1 \le i \le N-1, 0 \le j \le M-1\} \;=\; O(k^{p_1} + h^{p_2}) \qquad (h,k \to 0)\ .$$

Theorem 5.3 (Order of Consistency of the θ Method)

Assume $u_\ell, u_r \in C^2([0,T])$ and $u_0 \in C^4([-a,+a])$. Then, the solution of the given initial-boundary value problem satisfies $u \in C^2([0,T]; C^4([-a,+a]))$ and for $0 \le \theta \le 1$ the θ method is consistent of order $p_1 = 1$ and $p_2 = 2$, i.e.,

$$\max \{|\tau_{h,k}(x_i, t_{j+1})| \mid 1 \le i \le N-1, 0 \le j \le M-1\} = O(k + h^2) \qquad (h, k \to 0) .$$

If $u \in C^3([0,T]; C^4([-a,+a]))$, the Crank-Nicolsen scheme is consistent of order $p_1 = p_2 = 2$, i.e.,

$$\max \{|\tau_{h,k}(x_i, t_{j+1})| \mid 1 \le i \le N-1, 0 \le j \le M-1\} = O(k^2 + h^2) \qquad (h, k \to 0) .$$

Proof. The proof of the order of consistency follows easily by Taylor expansion.

Definition 5.4 (Stability of Difference Schemes)

Consider the θ method in its algebraic formulation with perturbed data $\tilde{b}_h^j = b_h^j + \delta b_h^j, 0 \leq j \leq M-1,$ and $\tilde{u}_h^0 = u_h^0 + \delta u_h^0$ and denote by u_h^j and z_h^j the solutions of the unperturbed resp. the perturbed θ method. The θ method is called stable, if there exists $k_{max} > 0$ such that for any $\varepsilon > 0$ there is a $\delta = \delta(\varepsilon, k_{max}) > 0$ such that for all perturbations satisfying

$$\|\delta u_h^0\| \; + \; k \sum_{\ell=0}^{M-1} \|\delta b_h^\ell\| \; < \; \delta \; ,$$

for all $k \leq k_{max}$ there holds

$$\max_{0 \leq j \leq M} \; \|u_h^j \; - \; z_h^j\| \; < \; \varepsilon \; .$$

Lemma 5.5 (Stability estimate)

Using the same notations as in the previous definition, for $1 \leq j \leq M$ there holds

$$\|u_h^j - z_h^j\| \; \leq \; \|(I_h + A_h(\theta))^{-1} B_h(\theta)\|^j \; \|\delta u_h^0\| \; + \; \sum_{\ell=0}^{j-1} \|(I_h + A_h(\theta))^{-1} B_h(\theta)\|^{j-\ell-1} \; \|(I_h + A_h(\theta))^{-1}\| \; \|\delta b_h^\ell\| \; .$$

Proof. The proof follows by induction on j.

Definition 5.6 (Spectral radius of a matrix)

Let $A \in \mathbb{R}^{N \times N}$. Then, the spectral radius of A is given by

$$\rho(A) \; = \; \max \; \{|\lambda| \mid \lambda \text{ is eigenvalue of } A\} \; .$$

For the matrix norm $\|\cdot\|$ associated with the Euclidean norm we have $\rho(A^T A) = \|A\|^2$.

Lemma 5.7 (Eigenvalues of tridiagonal matrices)

Let $A \in \mathbb{R}^{N \times N}$ be a tridiagonal matrix of the form

$$A = \begin{pmatrix} a & b & \cdot & \cdot & \cdot & 0 \\ c & a & b & 0 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 0 & c & a & b \\ 0 & \cdot & \cdot & 0 & c & a \end{pmatrix} \; .$$

Then, the eigenvalues $\lambda_i(A)$, $1 \leq i \leq N$, are given by

$$\lambda_i(A) \; = \; a \; + \; 2\, b \sqrt{\frac{c}{b}} \cos(\frac{i\pi}{N+1}) \quad , \quad 1 \leq i \leq N \; .$$

Lemma 5.8 (Eigenvalues of $(I_h + A_h(\theta))^{-1} B_h(\theta)$)
The $(N-1, N-1)$-matrix $C_h(\theta) := (I_h + A_h(\theta))^{-1} B_h(\theta)$ has the eigenvalues

$$\lambda_i(C_h(\theta)) = \frac{1}{\theta} \frac{1}{1 + 4\kappa\theta(k/h^2)\sin^2(i\pi/(2N))} - \frac{1-\theta}{\theta} \quad , \quad 1 \le i \le N-1 .$$

Proof. We have $B_h(\theta) = \theta^{-1} I_h - \theta^{-1}(1-\theta)(I_h + A_h(\theta))$, and hence,

$$(\dagger) \qquad (I_h + A_h(\theta))^{-1} B_h(\theta) = \theta^{-1} (I_h + A_h(\theta))^{-1} - \theta^{-1} (1-\theta) I_h .$$

According to Lemma 5.7, the matrix $I_h + A_h(\theta)$ has the eigenvalues

$$\lambda_i(I_h + A_h(\theta)) = 1 + 2 \kappa \theta \frac{k}{h^2} (1 - \cos(\frac{i\pi}{N})) \quad , \quad 1 \le i \le N-1 .$$

Using the elementary trigonometric relation

$$1 - \cos(\alpha) = 2 \sin^2(\frac{\alpha}{2}) .$$

gives the assertion.

Theorem 5.9 (Stability of the θ-method)

For $\theta \geq \frac{1}{2}$, the θ-method is unconditionally stable. For $\theta < \frac{1}{2}$, the θ-method is stable, if

$$(\bullet) \qquad k \leq \frac{h^2}{2 - 4\theta} .$$

Proof. According to Definition 5.4 and Lemma 5.5, the θ-method is stable, if

$$(\ddagger) \qquad \rho\big((I_h + A_h(\theta))^{-1} B_h(\theta)\big) < 1 .$$

Taking Lemma 5.8 into account, $(\ddagger)$ is satisfied if and only if for all $1 \leq i \leq N-1$

$$|\frac{1}{\theta} \frac{1}{1 + 4\kappa\theta \frac{k}{h^2} \sin^2(\frac{i\pi}{2N})} - \frac{1-\theta}{\theta}| < 1 \Longleftrightarrow |1 - \frac{4\kappa \frac{k}{h^2} \sin^2(\frac{i\pi}{2N})}{1 + 4\kappa\theta \frac{k}{h^2} \sin^2(\frac{i\pi}{2N})}| < 1 \Longleftrightarrow$$

$$\frac{4\kappa \frac{k}{h^2} \sin^2(\frac{i\pi}{2N})}{1 + 4\kappa\theta \frac{k}{h^2} \sin^2(\frac{i\pi}{2N})} < 2 \Longleftrightarrow (\star) \quad (2 - 4\theta) \frac{k}{h^2} \sin^2(\frac{i\pi}{2N}) < 1 .$$

Now, $(\star)$ holds true for any $\theta \geq \frac{1}{2}$, whereas $(\bullet)$ implies $(\star)$ in case $\theta < \frac{1}{2}$.

Theorem 5.10 (Convergence of the θ-method)

Assume that for $\theta < \frac{1}{2}$ the condition $(\bullet)$ from Theorem 5.9 is satisfied. Then, for all $0 \leq \theta \leq 1$, the θ-method is convergent of order

$$\|u_{h,k}(x,t) - u(x,t)\| = O(k^p + h^2) \quad , \quad (x,t) \in Q_{h,k} \ ,$$

where $p = 2$, if $\theta = \frac{1}{2}$, and $p = 1$, otherwise.

Proof. We define $z_h^j := (u(x_1,t_j),...,u(x_{N-1},t_j))^T$. Then, z_h^j satisfies a perturbed θ-method with $\delta u_h^0 \equiv 0$ and $\delta b_h^j = k(\tau_{h,k}(x_1,t_j),...,\tau_{h,k}(x_{N-1},t_j))^T$. The consistency (Theorem 5.5) and the stability (Theorem 5.9) imply convergence. Moreover, the order of convergence is at least the same as the order of consistency.

Remark 5.11: The stability requirement $(\bullet)$ is a severe restriction on the step size k.

5.3 Implementation of the θ-Method

Except for $\theta = 0$, the θ-method is an implicit method which requires the numerical solution of the linear algebraic system

$$(\mathrm{I}_h + \mathrm{A}_h(\theta))\ \mathrm{u}_h^{j+1} \ = \ \mathrm{B}_h(\theta)\ \mathrm{u}_h^j \ + \ \mathrm{b}_h^j \ .$$

Since the coefficient matrix $\mathrm{I}_h + \mathrm{A}_h(\theta)$ is an $(\mathrm{N}-1, \mathrm{N}-1)$ tridiagonal matrix, for its numerical solution the direct solution by an LR-decomposition of $\mathrm{I}_h + \mathrm{A}_h(\theta)$ is the method of choice, since it only requires approximately $\mathrm{O}(8(\mathrm{N}-1))$ elementary operations.

Remark 5.12: The finite difference method can be generalized to the multidimensional Black-Scholes equation describing the price of basket options on d assets. However, for $d \geq 2$, we do not longer have a tridiagonal coefficient matrix, so that iterative methods such as the preconditioned cg-method become competitive. For $d \geq 4$, the required amount of computational work increases significantly so that other techniques such as sparse grid techniques should be taken into account.