Numerical Analysis of PDE and SDE

Michael Pokojovy

Department of Mathematical Sciences, The University of Texas at El Paso

March 22, 2018

Outline

Operator Equations

Let X_0, X_1, Y_0 be complete metric spaces with $X_1 \hookrightarrow X_0$ and let $A\colon X_1 \to Y_0$ be a continuous mapping. Consider the operator equation

$$A(u) = f \text{ with } f \in Y_0.$$

Remark

We have the following obvious facts:

- *For the equation to be solvable, we need* $f \in \operatorname{im}(A)$.
- *For (locally) unique solvability,* A *needs to be (locally) injective.*
- *For continuous dependence on data, we moreover need* $A^{-1}\colon \operatorname{im}(A) \to X_1$ *to be continuous.*

Remark

We are not going to talk about operator inclusions (in particular, variational inequalities) and/or control and optimization problems. Also other kinds of inverse problems are outside of our scope.

Motivation

How do operator equations arise in applications?

- Let X be a Banach space. Minimize $J\colon X \to \mathbb{R}$. The first order optimality condition reads as: If x is an extremum of J, then $J'(x,\cdot) = 0$ in X'. Hence, $A\colon D(A) \subset X \to X'$, $A\colon x \mapsto J'(x,\cdot)$.
- Consider a system described by its state variable U and its flux V defined on a space or a space-time domain $\Omega \subset \mathbb{R}^d$. The conservation law suggests $\operatorname{div} U = 0$. Postulating a material law in the form $F(U,V) = 0$, we obtain a operator equation for (U,V).
- etc.

Typically, A is partial differential (or pseudo-differential) operator, sometimes with an additional integral, difference, algebraic or other structure.

Examples

Example (Nonlinear membrane)

For a domain $\Omega \subset \mathbb{R}^d$, consider the strain energy functional

$$J(u) = \int_\Omega \left(\sqrt{1 + |\nabla u|^2} - fu\right) \mathrm{d}x$$

mapping the Sobolev space $H_0^1(\Omega)$ into $[0, \infty)$. The associated operator equation (written as a partial differential equation) reads as

$$-\operatorname{div}\left(\frac{\nabla u}{\sqrt{1 + |\nabla u|^2}}\right) = f \text{ in } \Omega, u = 0 \text{ on } \partial\Omega.$$

Rodrigues, J-F. Obstacle problems in mathematical physics. Elsevier, 1987.

Examples – Cnt'd

Example (Lamé system)

Let $\Omega \subset \mathbb{R}^d$ be a domain and let $U\colon \Omega \to \mathbb{R}^d$ denote the displacement field of an elastic body occupying Ω. With $f\colon \Omega \to \mathbb{R}^d$ standing for the volumetric force acting on the body, the balance of momentum reads as

$$-\operatorname{div}\sigma = f,$$

where σ denotes the stress tensor. In this case, U is the basis field and σ is its flux. Assuming the body to be homogeneous and isotropic, selecting the geometric strain tensor

$$\varepsilon = \tfrac{1}{2}\left((\nabla U) + (\nabla U)^T + (\nabla U)^T(\nabla U)\right),$$

we obtain a hypoelastic material law

$$\sigma = \lambda \operatorname{tr}(\varepsilon) I_{d\times d} + \mu\varepsilon.$$

with λ, μ denoting the Lamé numbers. Hence, we arrive at the PDE

$$-\operatorname{div}\sigma(\nabla u) = f \text{ in } \Omega, \quad U = 0 \text{ on } \partial\Omega.$$

Here, we assumed the body to be clamped at the boundary.

Fu, Y. B., and Ogden, R. W. Nonlinear elasticity: theory and applications. Vol. 281. Cambridge University Press, 2001

Abstract Discretization Approach

Assuming we are given an operator equation $A(u) = f$ possessing a unique solution $u \in X_1$ for some $f \in Y_0$, we want to reduce it to a sequence of "simple" (typically, finitely dimensional) problems

$$A^h(u^h) = f^h \text{ for } h > 0.$$

For simplicity, we only consider the so called "conformal discretizations" with $A^h : X_1^h \hookrightarrow X_1 \to Y_0$.
The goal is to construct A^h and f^h such that the respective solution sequence $\{u^h\}_{h>0}$ converges to u in X_0 as $h \to 0$.

Consistency & Stability

From now on, we assume X_0 and Y_0 to be Banach spaces. Y_1 may remain though a metric space. Consider the numerical scheme

$$A^h(u^h) = f^h \text{ for } h > 0. \qquad (*)$$

There may exist a number $h_0 > 0$ such that Equation is uniquely solvable for any $h \in (0, h_0]$. Let X be a metric space with $X \hookrightarrow X_1$.

Definition

The numerical scheme $(*)$ is called X-consistent (of order $p > 0$) if

$$\|A^h(u^h) - f^h\|_{Y_0} \to 0 \quad \big(= O(h^p), \text{respectively} \big) \text{ as } h \to 0.$$

Definition

The numerical scheme $(*)$ is called stable if there exists a number $C > 0$ such that for all $h \in (0, h_0]$

$$\|u^h - v^h\|_{X_1} \leq C\|A^h(u^h) - A^h(v^h)\|_{Y_0} \text{ for any } u^h, v^h \in X_1^h.$$

Convergence and Lax' Principle

Definition

The numerical scheme $(*)$ is called X-convergent (of order $p > 0$) if

$$\|u^h - u\|_{X_0} \to 0 \quad \big(= O(h^p), \text{respectively} \big) \text{ as } h \to 0,$$

where $u \in X_1$ stands for the solution of the original operator equation.

Theorem (Lax' Principle)

An X-consistent (of order $p > 0$) stable scheme is X-convergent (of order p).

Proof.

Trivially, we have

$$\begin{aligned} \|u^h - u\|_{X_0} &\leq C\|A^h(u^h) - A^h(u)\|_{Y_0} = C\|A^h(u)\|_{Y_0} \\ &\to 0 \big(= O(h^p), \text{respectively} \big) \text{ as } h \to 0. \end{aligned}$$

Stochastic Operator Equations

Similar to deterministic problems, we can also consider stochastic operator equations. Let X_0, X_1, Y_0 be Polish spaces with $X_1 \hookrightarrow X_0$. Further, let A be a $C^0(X_1, Y_0)$-valued measurable variable with respect to a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. For an Y_0-valued random variable f with $f \in \mathrm{im}(A)$ P-a.s., we call the equation

$$A(u) = f$$

a stochastic operator equation.
For stochastic operator equations, probabilistic consistency, stability and convergence notions can be adopted.

Stationary PDE

The so-called "stationary" or "elliptic PDE" form a rather broad class of PDE. With $G \subset \mathbb{R}^d$ denoting a "space" domain, one can consider a nonlinear differential operator A, defined, say on some Sobolev space $W^{s,p}(G)$ given as

$$A(u) := a\big((\nabla^\alpha u)_{|\alpha| \leq s}\big)$$

for some tensor-valued function a (here, we employ the multi-index notation). Similarly, we consider the boundary operator B defined on some fractional Sobolev(-Slobodeckii) space $W^{\tilde{s},\tilde{p}}(\partial G)$ given as B

$$B(u) := b\big((\nabla^\alpha u)_{|\alpha| \leq \tilde{s}}\big)$$

for some other tensor-valued function b. A general boundary value problem for "elliptic" PDE reads as

$$A(u) = f \text{ in } G, \quad B(u) = g \text{ on } \partial G. \tag{$*$}$$

Equations of type $(*)$ have not fully been solved yet! Even in the linear case, the problem is often very challenging!

Gilbarg, D. and Trudinger, N. S. Elliptic Partial Differential Equations of Second Order, 2nd ed., Springer-Verlag, Berlin, 1984

Volevich, L. R. Solubility of boundary value problems for general elliptic systems, Mat. Sb. (N.S.), Volume 68(110), Number 3, 373416, 1965

A Model Problem

In this short lecture course, we look at the linear 2D Poisson equation with Dirichlet boundary conditions as a model problem:

$$-\triangle u = f \text{ in } G, \quad u = g \text{ in } \partial G,$$

where $G \subset \mathbb{R}^2$ is a Lipschitz domain.

Renardy, M., and Rogers, R. C. An introduction to partial differential equations. Vol. 13. Springer Science & Business Media, 2006

Most the techniques to be presented, can though be easily generalized to nonhomogeneous linear second order PDE with Dirichlet, Neumann or Robin boundary conditions or even PDE system (such as Lamé system, etc) in any space dimension.

In the following, we present two classical discretization approaches

- finite differences schemes
- finite elements schemes

Finite volumes, collocation methods, particle method, etc. are beyond our scope.

Poisson Equation

Let $G \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Further, let $f\colon \bar{G} \to \mathbb{R}$, $g\colon \partial G \to \mathbb{R}$. Consider the Poisson equation

$$-\triangle u = f \text{ in } G, \quad u = g \text{ in } \partial G,$$

The following existence and uniqueness results are known in the literature.

- **Weak L^p-solution:** Let $p, p' \in (1, \infty)$ such that $\frac{1}{p} + \frac{1}{p'} = 1$. Further, let $f \in L^{p'}(G)$, $g \in W^{1-1/p,p}(G)$. Then there exists a unique weak solution $u \in W^{1,p}(G)$.
- **Strong L^p-solution:** Let $p \in (1, \infty)$: If ∂G is piecewise $C^{1,1}$, with all cusps on the boundary are of angle being greater equal $\frac{\pi}{2}$, $f \in L^p(G), g \in W^{2-1/p,p}(\partial G)$, there exists a unique strong solution $u \in W^{2,p}(G)$.
- **Classical (Hölder) solution:** Let $\alpha \in (0, 1)$. If ∂G is $C^{2,\alpha}$, $f \in C^{0,\alpha}(\bar{G})$ and $g \in C^{2,\alpha}(\partial G)$, there exists a unique classical solution $u \in C^{2,\alpha}(G) \cap C^{0,\alpha}(\bar{G})$.

In the following, we assume the existence of a classical solution to Poisson equation.

Equidistant Lattices

For simplicity, we first let $G = (0,1)^2$ be the unit square and consider the equidistant lattice over G

$$G_h = \{(ih, jh) \,|\, i, j \in \{0, \ldots, M\}\}$$

with a space step $h = \frac{1}{M} > 0$ for some $M \in \mathbb{N}$.

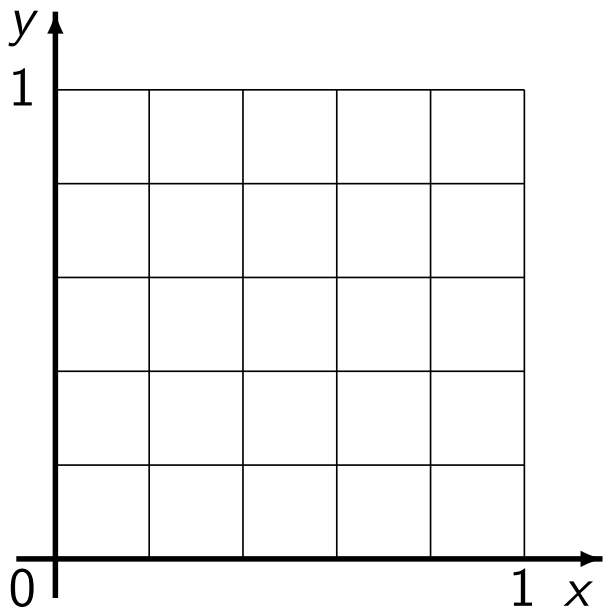


Figure: Lattice exaple with $M = 5$, $h = 0.2$

Finite Difference Approximation

The key idea of finite difference methods

Replace derivatives with diference quotients

For $v \in C^4([a,b],\mathbb{R})$, Taylor's theorem states

$$v(s \pm h) = v(s) \pm hv'(s) + \frac{h^2}{2}v''(s) \pm \frac{h^3}{6}v^{(3)}(s) + \frac{h^4}{24}v^{(4)}(\eta_\pm)$$

for $s, s \pm h \in [a,b]$ with $\eta_- \in [s-h, s]$, $\eta_+ \in [s, s+h]$. Therefore, for $v \in C^4([a,b])$,

$$\left|\frac{1}{h^2}(-v(s-h)+2v(s)-v(s+h)) + v''(s)\right| =$$

$$= \left|h^{-2}\left[-v(s)+hv'(s)-\frac{h^2}{2}v''(s)+\frac{h^3}{6}v^{(3)}(s)-\frac{h^4}{24}v^{(4)}(\eta_-)+2v(s)\right.\right.$$

$$\left.\left.-v(s)-hv'(s)-\frac{h^2}{2}v''(s)-\frac{h^3}{6}v^{(3)}(s)-\frac{h^4}{24}v^{(4)}(\eta_+)\right] + v''(s)\right|$$

$$= \frac{h^2}{24}\left|v^{(4)}(\eta_-)+v^{(4)}(\eta_+)\right| \le Ch^2 = O(h^2).$$

and, consequently,

$$-v''(s) = \frac{1}{h^2}\left(-v(s-h)+2v(s)-v(s+h)\right) + O(h^2).$$

Finite Difference Approximation – Cnt'd

Letting $u_{ij} = u(ih, jh)$, we get

$$-(\Delta u)_{ij} = \frac{1}{h^2}\big(- u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1} + 4u_{ij}\big) + O(h^2).$$

We define the discrete "interior" domain

$$\overset{\circ}{G}_h = \{(ih, jh) \in G_h \,|\, i, j \in \{1, \ldots, M-1\}\}$$

and replace the PDE with

$$h^{-2}\big(- u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1} + 4u_{i,j}\big) = f_{ij},$$

where $f_{ij} = f(ih, jh)$. For boundary mesh points $G_h^R = G \setminus \overset{\circ}{G}_h$, we set

$$u_{i\pm1,j} = \gamma_{i\pm1,j} \text{ or } u_{i,j\pm1} = \gamma_{i,j\pm1}$$

if $((i \pm 1)h, jh)$ or $(ih, (j \pm 1)h) \in \partial G$.
This leads to a linear algebraic system of dimension $(M-1)^2$ for the unknown numerical approximation u (the so-called lattice function)

$$Au = r \text{ with } A \in \mathbb{R}^{\overset{\circ}{G}_h, \overset{\circ}{G}_h}, u, r \in \mathbb{R}^{\overset{\circ}{G}_h}.$$

Finite Difference Approximation – Cnt'd

To explicitly represent A and r, we use the following enumeration for the lattice points: we start at the bottom left corner and proceed row by row from the left to the right. This leads to the following lattice function representation

$$u = (u^1, \ldots, u^{M-1}), \quad u^j = (u_{1,j}, u_{2,j}, \ldots, u_{M-1,j}) \in \mathbb{R}^{M-1},$$
$$f = (f^1, \ldots, f^{M-1}), \quad f^j = (f_{1,j}, f_{2,j}, \ldots, f_{M-1,j}) \in \mathbb{R}^{M-1}.$$

Further let

$$B = h^{-2} \begin{pmatrix} 4 & -1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 4 & -1 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 4 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 4 & -1 & 0 \\ 0 & 0 & 0 & \ldots & -1 & 4 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -1 & 4 \end{pmatrix} \in \mathbb{R}^{M-1,M-1},$$

$$C = h^{-2} I \in \mathbb{R}^{M-1,M-1},$$
$$\tilde{g}^1 = (g_{1,0} + g_{0,1}, g_{2,0}, \ldots, g_{M-2,0}, g_{M-1,0} + g_{M,1}),$$
$$\tilde{g}^j = (g_{0,j}, 0, \ldots, 0, g_{M,j}), \quad j = 2, \ldots, M-2,$$
$$\tilde{g}^{M-1} = (g_{0,M-1} + g_{1,M}, g_{2,M}, \ldots, g_{M-2,M}, g_{M-1,M} + g_{M,M-1}).$$

Finite Difference Scheme

In the block form, we obtain the following financial difference scheme

$$\underbrace{\begin{pmatrix} B & -C & 0 & \dots & 0 & 0 & 0 \\ -C & B & -C & \dots & 0 & 0 & 0 \\ 0 & -C & B & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & B & -C & 0 \\ 0 & 0 & 0 & \dots & -C & B & -C \\ 0 & 0 & 0 & \dots & 0 & -C & B \end{pmatrix}}_{=:A} \begin{pmatrix} u^1 \\ u^2 \\ u^3 \\ \vdots \\ u^{M-3} \\ u^{M-2} \\ u^{M-1} \end{pmatrix}$$

$$= \underbrace{\begin{pmatrix} f^1 \\ f^2 \\ f^3 \\ \vdots \\ f^{M-3} \\ f^{M-2} \\ f^{M-1} \end{pmatrix} + h^{-2} \begin{pmatrix} \tilde{g}^1 \\ \tilde{g}^2 \\ \tilde{g}^3 \\ \vdots \\ \tilde{g}^{M-3} \\ \tilde{g}^{M-2} \\ \tilde{g}^{M-1} \end{pmatrix}}_{=:r}$$

Solving the Discretized Problem

Though the resulting problem is "merely" a system of linear algebraic equations, standards tools like Gauss & Jordan elimination are not applicable to explicitly compute the inverse matrix or study its properties. Therefore, an alternative approach is seminal here. In particular, the following two classical tools are available:

- L^2 approach (Fourier analysis, spectral or energy methods, etc.)
- L^∞ approach (inverse monotonicity, discrete maximum principle, etc.)

For time reasons, we will only look at the L^∞-approach. It should though be pointed out that the latter can be difficult to apply to equations with non-constant coefficients. The L^2-approach is discussed in the references below


Smith, G. D. Numerical solution of partial differential equations: finite difference methods. Oxford university press, 1985


Strikwerda, J. C. Finite difference schemes and partial differential equations. SIAM, 2004


http://www.ima.umn.edu/~arnold//8445.f11/notes.pdf

Foundations of the L^∞-Approach

Let $A \in \mathbb{R}^{N,N}$ be a matrix.

- A is called monotone (notation: $A \geq 0$) if $A_{ij} \geq 0$ for $i, j \in \{1, \ldots, N\}$, which is equivalent to
$$u \leq v \Rightarrow Au \leq Av \quad \forall u, v \in \mathbb{R}^N.$$
- We write $A \leq B$ iff $B - A \geq 0$
- A is called inverse monotone if it is invertible with $A^{-1} \geq 0$
- A is called L_0-matrix if $A_{ij} \leq 0$ for $i, j \in \{1, \ldots, N\}$ with $i \neq j$
- A is called M-matrix if it is an inverse monotone L_0-matrix.

Note that the inverse monotonicity for matrices is a discrete counterpart of the weak maximum principle for elliptic differential operators.
We will need the following important result.

Theorem (M-criterion)

An L_0-matrix $A \in \mathbb{R}^{N,N}$ is an M-matrix iff there exist a "majorizing element" $e \in \mathbb{R}^n$ with $e > 0$ such that $Ae \geq 0$ and the following "linkage" property is satisfied: For any $i_0 \in \{1, \ldots, N\}$ with $(Ae)_{i_0} = 0$ there exists a chain $i_0, i_1, \ldots, i_r \in \{1, \ldots, N\}$ such that $(Ae)_{i_r} > 0$ and $A_{i_{j-1}, i_j} \neq 0$ for any $j \in \{1, \ldots, r\}$.

What About the Discrete Poisson Equation?

Theorem

The matrix $A^h \in \mathbb{R}^{(M-1)^2,(M-1)^2}$ originating from the classical difference scheme is an M-matrix.

The proof is based on selecting $\mathbb{I} = (1, \ldots, 1)^T \in \mathbb{R}^{(M-1)^2}$ as a majorizing element.

This implies that A^h is invertible.

General Domains

Similar procedure can be applied to general curved domains. Letting

$$G_h = G \cap \mathbb{R}^2_h,$$

we associate any point $(x, y) \in G_h$ with its four neighbors $N_k = N_k(x, y, h) \in \bar{G},\ k = 1, 2, 3, 4.$

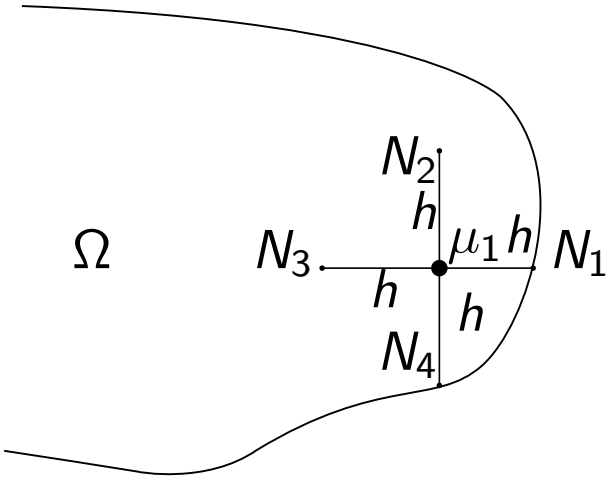


Figure: Neighbors

General Domains – Cnt'd

Let further

$$e^1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, e^2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, e^3 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, e^4 = \begin{pmatrix} 0 \\ -1 \end{pmatrix},$$

$$\mu_k = \mu_k(x, y, h) = \sup \left\{ \mu \in [0, 1] \,|\, (x, y) + the^k \in G, \forall t \in [0, \mu] \right\},$$

$$N_k = N_k(x, y, h) = (x, y) + \mu_k(x, y, h) he^k, \quad k = 1, 2, 3, 4.$$

The interior discrete domain is now

$$\overset{\circ}{G}_h = \{(x, y) \in \Omega_h \,|\, N_k(x, y, h) \in G, \quad k = 1, 2, 3, 4\}.$$

For any $(x, y) \in \overset{\circ}{G}_h$, we have $\mu_k(x, y, h) = 1$, $k = 1, 2, 3, 4$. For interior points, we use the same four-point discretization of Laplacian as before

$$h^{-2}(-u(x-h, y)-u(x+h, y)-u(x, y-h)-u(x, y+h)+4u(x, y)) = g(x, y).$$

General Domains – Cnt'd

The boundary lattice points are assinged to the set

$$G_h^R = G_h \setminus \overset{\circ}{G}_h = \{(x,y) \in G_h \,|\, N_k(x,y,h) \in \partial G \text{ for some } k = 1,2,3,4\}.$$

We need the following approximation result.

Lemma

Let $v \in C^3([-a,a],\mathbb{R})$ for some $a > 0$. Then, for any $\mu_0, \mu_1 \in (0,1]$ and any $h \le a$, there holds

$$\left| \frac{2}{\mu_0\mu_1(\mu_0+\mu_1)h^2} \Big\{ \mu_0 v(\mu_1 h) - (\mu_0+\mu_1)v(0) + \mu_1 v(-\mu_0 h) \Big\} - v''(0) \right| \le \frac{2}{3} h \cdot \max\{|v'''(x)| \,|\, |x| \le a\}.$$

General Domains – Cnt'd

With the result above, we obtain a discretization for boundary lattice points

$$f(x,y) = h^{-2}\Big\{ -\frac{2}{\mu_1(\mu_1+\mu_3)}u(N_1) - \frac{2}{\mu_3(\mu_1+\mu_3)}u(N_3) - \frac{2}{\mu_2(\mu_2+\mu_4)}u(N_2)$$
$$- \frac{2}{\mu_4(\mu_2+\mu_4)}u(N_4) + 2\Big(\frac{1}{\mu_1\mu_3} + \frac{1}{\mu_2\mu_4}\Big)u(x,y)\Big\}.$$

For each $N_k \in \partial G$, $k \in \{1,2,3,4\}$, we plug $u(N_k) = \gamma(N_k)$ due to $u = g$ on ∂G. This way, we obtain a finite difference scheme of the form

$$Au = r, \quad u \in \mathbb{R}^{G_h}.$$

Once again, for an explicit representation, the lattice points need to be enumerated.

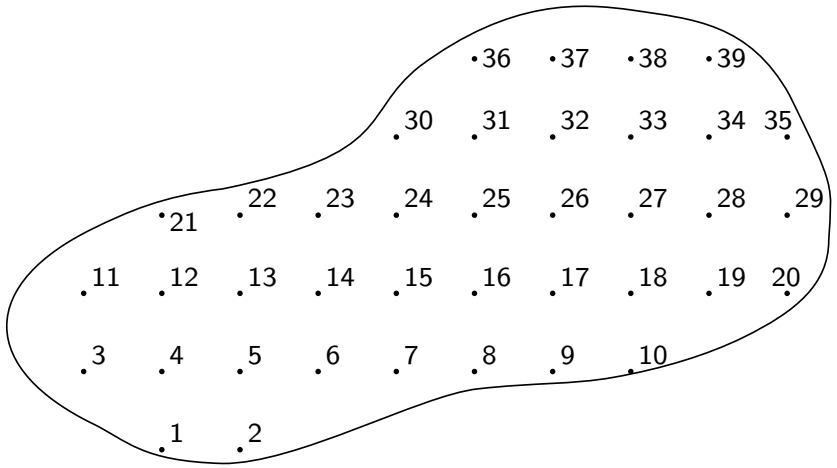


Figure: Lattice point enumeration example

Consistency

We adopt the consistency, stability & convergence notions from the Introduction. To this end, we let

$$A := \begin{pmatrix} -\triangle \\ (\cdot)|_{\partial G} \end{pmatrix}$$

a view it as mapping from $C^2(G) \cap C^0(\bar{G})$ to $C^0(G) \times C^0(\partial G)$.
This way, we have the following convergence result.

Theorem

The finite difference scheme is $C^4(G) \cap C^0(\bar{G})$-consistent of order 2. In particular, for any solution C^4-solution u of the original PDE we have

$$\|A^h \overline{u} - r^h\|_{L^\infty(G)} \leq \sup\{|(A^h \overline{u}_h)(x,y) - r^h(x,y)| \,|\, (x,y) \in G_h\} = O(h^2).$$

Remark

Despite of the $O(h)$-approximation around the boundary ∂G, we globally obtain the second order convergence by multiplying the equations corresponding to boundary points with h.

Stability

As for stability, we obtain

Theorem

The matrix A^h is an M. For any rectangle $(a,b)\times(c,d)\supset\bar{G}$, there exists a positive vector $e_h = e|_{G_h}\in\mathbb{R}^{G_h}$ with

$$e(x,y) = (x-a)(b-x)+(y-c)(d-y),\quad (x,y)\in\mathbb{R}^2,$$

which satisfies

$$A^h e_h \geq \rho e_h$$

with $\rho := \frac{16}{(b-a)^2+(d-c)^2} > 0$ if $h>0$ is sufficiently small.
A^h satisfies the stability estimate

$$\|u\|_{L^\infty(G_h)} \leq \frac{1}{\rho}\|A^h u\|_{L^\infty(G_h)} \qquad \text{for any } u\in\mathbb{R}^{G_h}$$

for sufficiently small $h>0$.

Semilinear Problems

Our techniques can easily be applied to a semilinear problem of the form

$$-\Delta u(x,y) = f(x,y,u(x,y)), \quad (x,y) \in G.$$

Discretizing the Laplacian as before, we obtain the following finite difference scheme

$$Au = G(u), \quad u \in \mathbb{R}^{G_h}$$

with the diagonal field

$$(G(u))(x,y) = g(x,y,u(x,y)) + \text{boundary terms}, \quad (x,y) \in G_h.$$

The resulting is usually solved with Newton's method.
For the nonlinear problem, one can also show the consistency as well as local statbility (in a neighborhood of the actual solution) implying the local convergence.

Weak Formulation

Again, as a model problem we consider the Poisson equation

$$-\triangle u = f \text{ in } G, \quad u = g \text{ in } \partial G,$$

In contrast to finite elements, we are now looking for a numerical approximation to the weak solution of our problem. For the sake of simplicity, we restrict ourselves to the Hilbert space situation. Consider the block operator

$$\begin{pmatrix} A \\ B \end{pmatrix} : H^1(G) \to H^{-1}(G) \times H^{1/2}(\partial G), \quad u \mapsto \begin{pmatrix} -\triangle u \\ u|_{\partial G} \end{pmatrix}.$$

(Note: $H^{-1}(G) = \left(H_0^1(G)\right)'$). With this notation, the Poisson equations rewrites as

$$Au = f, \quad Bu = g.$$

Weak Formulation – Cnt'd

For numerical method, the so-called variational formulation is usually preferable. Taking into account that $u \in H^1(G)$, for any test function $\varphi \in \mathcal{D}(G)$, we get in the sense of distributions

$$[-\triangle u](\varphi) = \int_G u \triangle \varphi \mathrm{d}x = -\int_G \nabla u \cdot \nabla \varphi \mathrm{d}x.$$

By a standard continuity and density argument, the latter is true for any $\varphi \in H_0^1(G)$.
Hence, the variational formulation of Poisson equation reads as: Find a function $u \in H^1(G)$ with $u|_G = g$ (note that the trace operator is well-defined) such that

$$\int_G \nabla u \cdot \nabla v \mathrm{d}x = \langle u, v \rangle_{H^{-1}(G);H_0^1(G)} \text{ for any } v \in H_0^1(G).$$

If g is trivial, the integral on the left-hand side can be viewed as a bilinear form on $H_0^1(G) \times H_0^1(G)$.

Variational Formulation

By the virtue of classical Sobolev extension theorems, $g \in H^{1/2}(G)$ can be continuously extended to an element of $H^1(G)$. This way, if u denotes the weak solution of Poisson equation, $u - g$ satisfies $H_0^1(G)$. The weak formulation further implies

$$\int_G \nabla u \cdot \nabla v \mathrm{d}x = \langle f, v \rangle_{H^{-1}(G);H_0^1(G)}$$

and, therefore,

$$a(u - g, v) = -a(g, v) + \langle f, v \rangle_{H^{-1}(G);H_0^1(G)}$$

with the bilinear form

$$a(u, v) := \int_G \nabla u \cdot \nabla v \mathrm{d}x.$$

With the bilinear form being coercive due to Poincaré inequality and the right-hand side being a continuous linear functional, $u - g$ is uniquely and continuously computable in terms of g. Therefore, u can also explicitly be computed.

Variational Formulation – Cnt'd

Both the weak and the variational formulations are equivalent. And we have the following result.

Theorem

Let $g \in H^{1/2}(G)$, $f \in H^{-1}(G)$. Then there exists a unique weak solution $u \in H^1(G)$ to Poisson equation.

The variational formulation is the commonly used in numerical literature. Usually, it is also prefered not to distinguish the weak and variational formulation. Besides, the problems arising from extending g onto G are subtly "overlooked."

Ritz & Galerkin Method

The Ritz & Galerkin method is based on the variation formulation of Poisson equation and the separability of underlying Hilbert spaces.
The basic idea is the following:

1. Select a finite dimensional subspace $V \subset H_0^1(G)$ represented by a basis $\{v_1, \ldots, v_n\}$
2. Approximate the numerical solution as
$$u = g + \sum_{k=j}^{n} c_j v_j,$$
where g is an H^1-extension of g onto G.
3. Now, computing u is equivalent to determining the c_i coefficients
4. Plugging the Galerkin ansatz into the weak formulation and taking into account that it suffices to require the variational equation only to hold on the basis, we get
$$\sum_{j=1}^{n} a(v_i, v_j)c_j = -a(g, v_i) + \langle f, v_i \rangle_{H^{-1}(G);H_0^1(G)} \text{ for } i = 1, \ldots, n,$$
which is now "just" a linear algebraic system for c_i's, i.e.,
$$Ac = r \quad (A \text{ stiffness matrix, } r \text{ load vector}).$$
5. Let V "converge" to H_0^1 in the sense that $\inf_{v \in V} \|u - v\| \to \infty$ for any $u \in H_0^1(G)$

Finite Elements

In contrast to analytical studies, where the basis functions v_j's are selected as the eigenfunctions of Laplacian, a fundamentally different approach is typically used in numerical analysis. Indeed, selecting v_j as an eigentfunction or any other function supported on the whole set of G, makes the stiffness matrix full or non-spare. Hence, it can be wise choice is to select v_j's in such a way that A is a sparse matrix (i.e., with lots of zeros). This also reflects the locality of underlying PDE. The set of v_j's is then referred to as a finite element space.

Remark

There are situations where making A sparse is not the optimal strategy. Within the framework of model reduction (cf. POD techniques), one prefers A to be low-dimensional, but typically not sparse.

Triangulation

We present a common approach to constructing finite element spaces. First, a triangulation G_{T_h} of the domain G is computed. For the triangulation to be admissible, the intersection between any two triangle needs either to be empty or consist of a vertex/node or an edge.

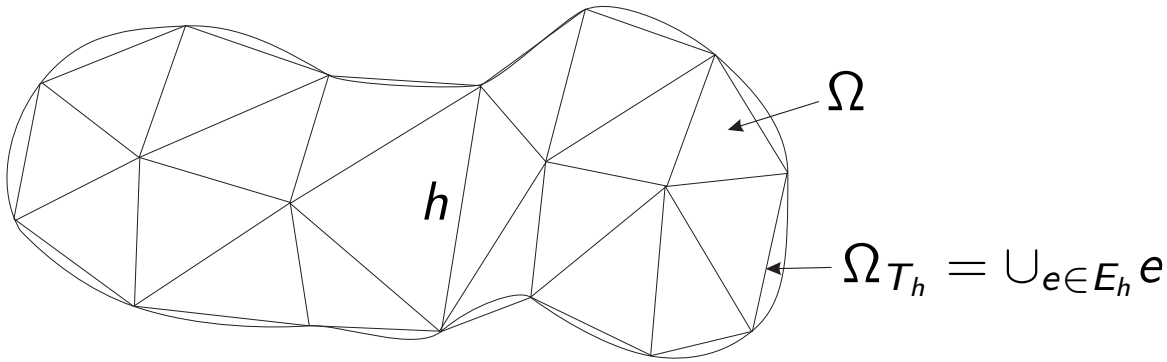
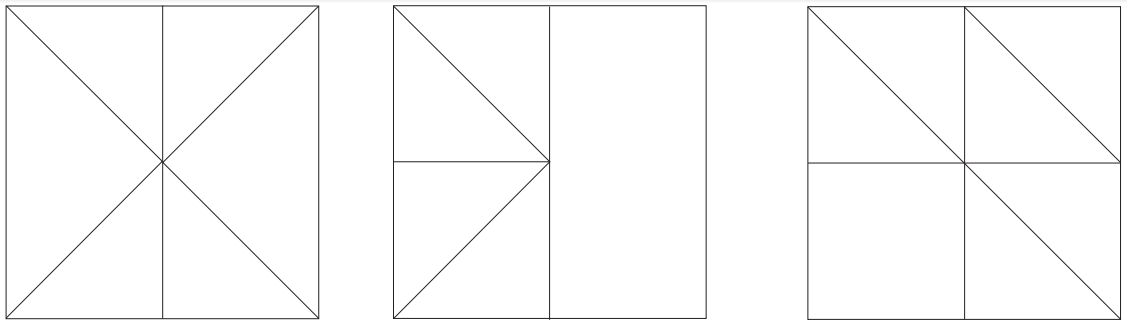


Figure: Triangulation

Triangulation – Cnt'd



admissible | not admissible | Friedrichs-Keller triangulation

Figure: Sample triangulations

The finite element space is then selected as

$$V_{T_h} = \{u \in C(G_{T_h}) \,|\, u \text{ is a polynomial of } \deg(u) \leq r \text{ in } x \text{ and } y \text{ on any triangle } e \in T_h \text{ and } u = 0 \text{ on } \partial G_{T_h}\}.$$

- $r = 1$: linear elements
- $r = 2$: quadratic elements
- etc.

We will only discuss linear finite elements.

Linear Finite Elements

Let P_i, $i = 1, \ldots, M$ are the numbered nodes of G_{T_h}, where the nodes that do not lie on ∂G are exaclty the P_i's, $i = 1, \ldots, m$ for $m < M$.

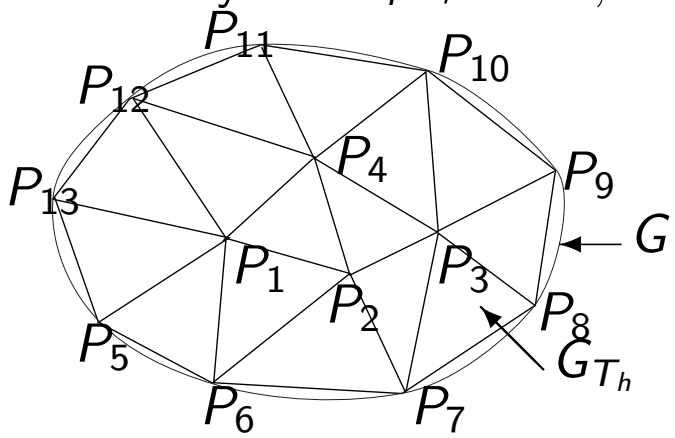


Figure: Node numeration, $M = 13$, $m = 4$

The functions u_i, $i = 1, \ldots, M$ are defined on G_{T_h} via

$$u_i(P_j) = \delta_{ij}, \quad 1 \leq i, j \leq M,$$

u is linear in x and y on each $e \in E_h$, $u \in C(G_{T_h})$.

One can easily show that this interpolation problem is uniquely solvable. The resulting functions u_i, $i = 1, \ldots, M$ satisfy $u_i \in H^1(G)$ and are referred to as form functions.

Finite Element Ansatz

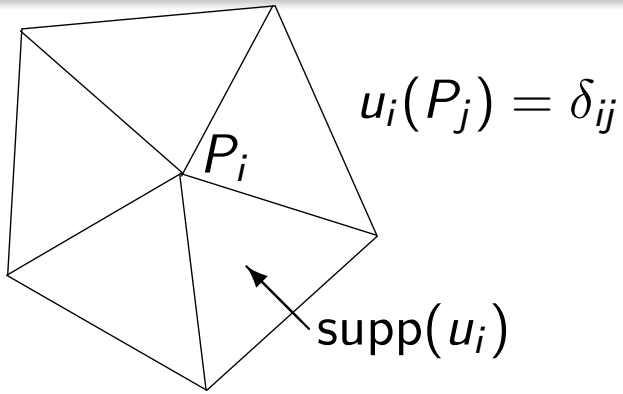


Figure: Support $\mathrm{supp}(u_i)$ of a form function u_i

We let

$$u_0 := \sum_{j=m+1}^{M} g(P_j) u_j,$$

where we recall $P_{m+1}, \ldots, P_M \in \partial\Omega$, and obtain

$$u_0(P_i) = \sum_{j=m+1}^{M} \gamma(P_j) \underbrace{u_j(P_i)}_{=\delta_{ij}} = \gamma(P_i), \quad i = m+1, \ldots, M.$$

Finite Element Ansatz – Cnt'd

The resulting ansatz reads as

$$\tilde{u} = u_0 + \sum_{j=1}^{m} c_j u_j = \sum_{j=1}^{M} c_j u_j$$

with $c_j = g(P_j)$ for $j = m+1, \ldots, M$.
The c_i coefficient corresponds then to the value of $\tilde{u}$ at the node P_i, $i = 1, \ldots, M$ since

$$\tilde{u}(P_i) = \sum_{j=1}^{M} c_j \underbrace{u_j(P_i)}_{=\delta_{ij}} = c_i, \quad i = 1, \ldots, M.$$

The unknowns are $c_1, \ldots, c_m$, while the coefficients $c_{m+1}, \ldots, c_M$ are known.

Remark

It should be pointed out that u_0 does **not** *belong to $H^1(G)$ in the limit as Ritz & Galerkin method requires. This problem is usually "overlooked" in most numerical literature.*

Assembling Procedure

Hence, we obtain the following system of linear algebraic equations

$$\begin{aligned} A_{ij} &= \int_G \nabla u_j \cdot \nabla u_i \, \mathrm{d}(x,y), \quad 1 \le i,j \le m, \\ r_i &= -\int_G (\nabla u_0 \cdot \nabla u_i - f u_i) \, \mathrm{d}(x,y), \quad 1 \le i \le m. \end{aligned}$$

Computing A and r is commonly referred to as assembling.
We replace G with G_{T_h} and f with

$$f_{T_h} = \sum_{j=1}^{m} f(P_j) u_j$$

and, taking into account $u_0 = \sum_{j=m+1}^{M} g(P_j) u_j$, obtain

$$\begin{aligned} A_{ij} &= \int_{G_{T_h}} \nabla u_i \cdot \nabla u_j \, \mathrm{d}(x,y), \quad 1 \le i,j \le m, \\ r_i &= \int_{G_{T_h}} f_{T_h} u_i - \nabla u_0 \cdot \nabla u_i \, \mathrm{d}(x,y) \\ &= \sum_{j=1}^{m} f(P_j) \int_{G_{T_h}} u_i u_j \, \mathrm{d}(x,y) - \sum_{j=m+1}^{M} g(P_j) \int_{G_{T_h}} \nabla u_j \cdot \nabla u_i \, \mathrm{d}(x,y), \quad i = 1, \ldots, m. \end{aligned}$$

Assembling Procedure – Cnt'd

Since the integrals over G_{T_h} can be writtten as a sum of integrals over respective triangles $e \in T_h$, i.e.,

$$\int_{\Omega_{T_h}} f(x,y)\,\mathrm{d}(x,y) = \sum_{e \in E_h} \int_e f(x,y)\,\mathrm{d}(x,y),$$

we only need to compute

$$\int_e \nabla u_i \cdot \nabla u_j\,\mathrm{d}(x,y), \quad \int_e u_i u_j\,\mathrm{d}(x,y), \quad 1 \le i,j \le M.$$

This can be done using the so-called isoparametric principle briefly described next.

Let $e \in T_h$ be an arbitrary triangle with edges P_{i1}, P_{i2}, P_{i3}.

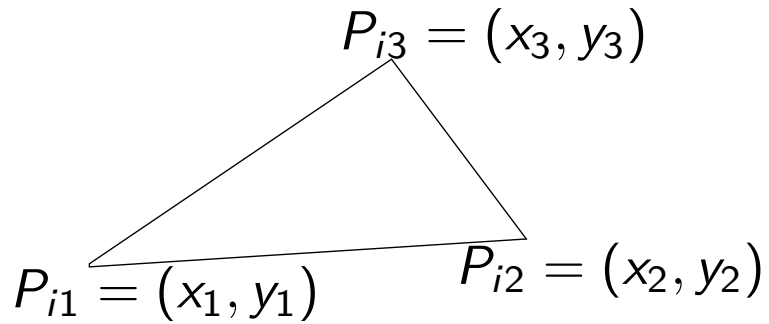


Figure: Triangle e along with its edges P_{i1}, P_{i2}, P_{i3}

Computing the Integrals

Only the following integrals are different from zero: $S_{jk} = \int_e \nabla u_{ij} \cdot \nabla u_{ik} \, \mathrm{d}(x, y)$, $k, j = 1, 2, 3$.
When assembling A, the symmetric matrix $S^e = (S_{jk})_{1 \leq j,k \leq 3} \in \mathbb{R}^{3,3}$ needs to be added up to the submatrix

$$\begin{pmatrix} A_{i1,i1} & A_{i1,i2} & A_{i1,i3} \\ A_{i2,i1} & A_{i2,i2} & A_{i2,i3} \\ A_{i3,i1} & A_{i3,i2} & A_{i3,i3} \end{pmatrix}.$$

We obtain

$$S^e = \frac{1}{2 \cdot F_e} C_e C_e^T \text{ with } C_e = \begin{pmatrix} y_2 - y_3 & x_3 - x_2 \\ y_3 - y_1 & x_1 - x_3 \\ y_1 - y_2 & x_2 - x_1 \end{pmatrix} \in \mathbb{R}^{3,2}$$

and $F_e = |e| = (x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1)$. Here, F_e is the area of triangle e with the edges P_{i1}, P_{i2}, P_{i3}.
The second integral is given via the matrix $M^e \in \mathbb{R}^{3,3}$ with

$$(M^e)_{jk} = \int_e u_{jk} \cdot u_{ik} \, d(x, y), \quad 1 \leq j, k \leq 3, \quad M^e = \frac{F_e}{24} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \in \mathbb{R}^{3,3}.$$

Convergence Analysis

For the sake of simplicity, we let $g \equiv 0$ and $f \in L^2(G)$. We consider the following numerical scheme: Find $u_h \in V_h$ such that

$$a(u_h, v) = f(v) \quad \forall v \in V_h \text{ with } f(v) = \langle f, v \rangle_{H^{-1}(G);H_0^1(G)}$$

for a finite dimensional subset $V_h \subset V$. Let $\bar{u}$ denote the solution to original PDE. Then $e := \bar{u} - u_h$ is the resulting error and we have the error equation

$$a(e, v) = 0 \quad \forall v \in V_h.$$

Lemma (Approximation Lemma)

Let $V_h \subset V$ be a subspace of V, $a(\cdot, \cdot)$ an inner product on V and $\|u\|_a = \sqrt{a(u,u)}$ the resulting norm. Then for any $u_h \in V_h$, we have

$$a(\bar{u} - u_h, v) = 0 \quad \forall v \in V_h,$$

which, in its turn, is equivalent with

$$\|\bar{u} - u_h\|_a = \inf\{\|\bar{u} - v\| \mid v \in V_h\}.$$

Convergence Analysis – Cnt'd

Lemma

The solution u_h to Galerkin equation $a(u,v) = b(v)$, $\forall v \in V_h$ is stable in the sense of estimate

$$\|u_h\|_{L^2(G)} \leq \frac{1}{\alpha}\|b\|_{V'}$$

for any $h > 0$, where

$$\|b\|_{V'} = \sup\left\{ \frac{|b(v)|}{\|v\|_{L^2(G)}} \,\middle|\, v \in V, v \neq 0 \right\}.$$

Theorem (Cea's Lemma)

For the solution of Galerkin equation, we have

$$\|\overline{u} - u_h\|_{L^2(G)} \leq \frac{M}{\alpha} \cdot \inf\{\|\overline{u} - v\|_{L^2(G)} \,|\, v \in V_h\}.$$

Convergence Analysis – Cnt'd

By the virtue of Cea's lemma, the convergence is now only determined by the approximation quality $\inf\{\|\overline{u} - v\| \mid v \in V_h\}$.
For linear elements, we get

$$\|\overline{u} - u_h\|_{H^1(G)} \leq (1 + C^2) \cdot \inf\{\|\overline{u} - v\|_{H^1(G)} \mid v \in V_{T_h}\} \leq (1 + C^2)\|\overline{u} - w\|_{H^1(G)}$$

for any $w \in V_{T_h}$. Using interpolation, an appropriate w can be constructed. For the latter, one can show

$$\|w) - v\|_{H^1(G)} \leq C \cdot h \quad \forall v \in H^2(\Omega)$$

if $(T_h)_{0<h<h_1}$ is a triangulation family on G for which the maximal angle $\gamma_{h,\max}$ in each triangle $e \in T_h$ satisfies

$$\gamma_{h,\max} \leq \gamma_{\max} < \pi \text{ for } 0 < h \leq h_0.$$

The latter is called the maximal angle condition.

Theorem (Convergence)

Under the maximal angle condition, we have

$$\|\overline{u} - u_h\|_{L^2(G)} \leq \tilde{C} \cdot h \cdot \|u - u_h\|_{H^1(G)} \leq \tilde{C}\hat{C}h^2, \quad 0 < h \leq h_0,$$

if $\bar{u} \in H^2(G)$.

References

- Bathe, K.-J.. Finite element procedures. Klaus-Jurgen Bathe, 2006
- Brenner, S. C., and Scott, R.. The mathematical theory of finite element methods. Vol. 15. Springer Science & Business Media, 2008
- Ciarlet, P. G. The finite element method for elliptic problems. Vol. 40. SIAM, 2002
- Reddy, J. N. An Introduction to the Finite Element Method (3rd ed.). McGraw-Hill, 2006
- Strang, G., and Fix, G. F. An analysis of the finite element method. Vol. 212. Englewood Cliffs, NJ: Prentice-Hall, 1973
- Zienkiewicz, O. C., Taylor, R. L., and Zhu, J. Z. The Finite Element Method: Its Basis and Fundamentals (6th ed). Butterworth-Heinemann, 2005

Physical Model

Let $G \subset \mathbb{R}^d$ be a domain with a Lipschitz-boundary ∂G and $T > 0$ be a fixed number. Let a function $\theta\colon [0, T] \times \bar{G} \to \mathbb{R}$ denote the temperature measured with respect to a reference temperature θ_0 and let $q\colon [0, T] \times \bar{G} \to \mathbb{R}^d$ be the heat flux at a material point $x \in \bar{G}$ at time $t \in [0, T]$. With $\rho\colon \bar{G} \to (0, \infty)$ denoting the specific density and $c_\rho\colon \bar{G} \to (0, \infty)$ denoting the specific heat capacity, the energy conservation law reads as

$$\rho(x)c_\rho(x)\partial_t\theta(t, x) + \operatorname{div} q(t, x) = h(t, x) \text{ for } t \in (0, T),\ x \in G,$$

where h stands for the intensity of external heat sources.
To close this equation, a material law postulating a relation between the temperature and the heat flux is required. The classical way to do this consists in using Fourier's law of heat conduction stating

$$q(t, x) + \lambda(x)\nabla\theta(t, x) = 0 \text{ for } t \in (0, T),\ x \in G,$$

where $\lambda\colon \bar{\Omega} \to (0, \infty)$ denotes the heat conductivity being a material property. Finally, this leads to the classical parabolic heat equation

$$\rho(x)c_\rho(x)\partial_t\theta(t, x) - \operatorname{div}\big(\lambda(x)\nabla\theta(t, x)\big) = h(t, x) \text{ for } t \in (0, T),\ x \in G.$$

Khusainov, D., Pokojovy, M., and Racke, R. Strong and Mild Extrapolated L^2-Solutions to the Heat Equation with Constant Delay. SIAM Journal on Mathematical Analysis 47.1, 427-454, 2015

Model Problem

Consider the model problem

$$\begin{aligned} u_t &= u_{xx} + f(u, u_x, x, t) \text{ in } (0,1) \times (0,T), \\ u(x,0) &= u_0(x) \text{ for } 0 \leq x \leq 1, \\ u(0,t) &= g_0(t),\ u(1,t) = g_1(t) \text{ for } 0 \leq t \leq T. \end{aligned}$$

We use the following numerical principle for solving this time-dependent PDE.

Line method

First discretize with respect to space and then with respect to time.

Next, we illustrate the application of line method to our problem. Here, we use the finite differencee method for space discretization. It should be pointed out that the finite element method can be applied similarly.

Elliptic Problem

Carrying out a finite difference discretization for the respective elliptic problem

$$\begin{aligned} -w''(x) &= f(x, w(x), w'(x)) \text{ in } (0,1), \\ w(0) &= g_0,\ w(1) = g_1, \end{aligned}$$

we obtain

$$\begin{aligned} w(0) &= g_0, \\ \frac{1}{\Delta x^2}(-w(x-\Delta x) + 2w(x) - w(x+\Delta x)) &= \\ &= f\left(x, w(x), \frac{1}{2\Delta x}(w(x+\Delta x) - w(x-\Delta x))\right), \\ x &= j\Delta x, j = 1, \ldots, M-1, \\ w(1) &= g_1. \end{aligned}$$

Elliptic Problem – Cnt'd

Eliminating $w(0)$, $w(1)$, we get a nonlinear algebraic system for the vector $(w(\Delta x), \ldots, w(1-\Delta x))$ reading as

$$\frac{1}{\Delta x^2}\begin{pmatrix} 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 2 & -1 & 0 \\ 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{pmatrix}\begin{pmatrix} w(\Delta x) \\ w(2\Delta x) \\ w(3\Delta x) \\ \vdots \\ w(1-3\Delta x) \\ w(1-2\Delta x) \\ w(1-\Delta x) \end{pmatrix}$$

$$= \begin{pmatrix} f\left(\Delta x, w(\Delta x), \frac{1}{2\Delta x}(w(2\Delta x) - g_0)\right) + \frac{g_0}{\Delta x^2} \\ f\left(2\Delta x, w(2\Delta x), \frac{1}{2\Delta x}(w(3\Delta x) - w(\Delta x))\right) \\ \vdots \\ f\left(j\Delta x, w(j\Delta x), \frac{1}{2\Delta x}(w((j+1)\Delta x) - w((j-1)\Delta x))\right), j = 2, \ldots, M-2 \\ \vdots \\ f\left(1-2\Delta x, w(1-2\Delta x), \frac{1}{2\Delta x}(w(1-\Delta x) - w(1-3\Delta x))\right) \\ f\left(1-\Delta x, w(1-\Delta x), \frac{1}{2\Delta x}(g_1 - w(1-2\Delta x))\right) + \frac{g_1}{\Delta x^2} \end{pmatrix}.$$

Elliptic Problem – Cnt'd

The latter system can be written in a compact form as

$$A^{\Delta x} w = G^{\Delta x}(w), \quad w \in \mathbb{R}^{G_{\Delta x}}$$

with

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

$$G^{\Delta x}(w) = \begin{pmatrix} f\left(\Delta x, w(\Delta x), \frac{1}{2\Delta x}(w(2\Delta x) - g_0)\right) + \frac{g_0}{\Delta x^2} \\ f\left(j\Delta x, w(j\Delta x), \frac{1}{2\Delta x}(w((j+1)\Delta x) - w((j-1)\Delta x))\right), \\ j = 2, \dots, M-2 \\ f\left(1 - \Delta x, w(1 - \Delta x), \frac{1}{2\Delta x}(g_1 - w(1 - 2\Delta x))\right) + \frac{g_1}{\Delta x^2} \end{pmatrix}.$$

Semidiscretization in Space

We come back to the original time-dependent PDE and write it as an ODE for

$$\begin{aligned} v(t) &= (u(\Delta x, t), u(2\Delta x, t), \ldots, u(1-2\Delta x, t), u(1-\Delta x, t)) \\ &= (v_1(t), v_2(t), \ldots, v_{M-2}(t), v_{M-1}(t)) \in \mathbb{R}^{M-1}, \quad 0 \le t \le T \end{aligned}$$

reading as

$$\begin{aligned} v'(t) &= -A^{\Delta x} v(t) + H^{\Delta x}(v(t)) + r^{\Delta x}(t) \\ &=: F_{\Delta x}(v(t), t), \quad 0 \le t \le T, \\ v(0) &= v^0 = (u_0(\Delta x), \ldots, u_0((M-1)\Delta x)). \end{aligned}$$

Time Integration

The ODE we obtained above can essentially be discretized using any time integrator available for ODE. The latter should be done to achieve the desired consistency order.
For stability and therefore convergence, other aspects are important. Most difficulties arise from the fact that the ODE under consideration is a stiff one. Based on whether the time integrator is implicit or explicit, either conditional or unconditional stability follows. In the former case, the resulting full discretization is stable only for sufficiently small time steps related to space steps. Moreover, stability properties may and often do depend on the selection of topology.
In this lecture, we will use the classical ϑ-method as our time integrator. Taking particular values of ϑ, we get the following well-known schemes:

$\vartheta = 0$: explicit Euler scheme

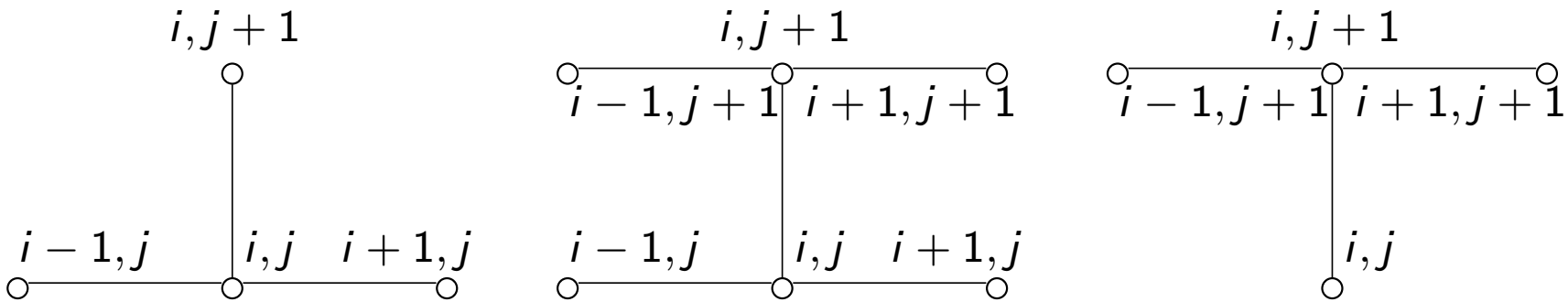
$\vartheta = 1/2$: Crank & Nicolson scheme or Simson/trapezoidal rule

$\vartheta = 1$: implicit Euler scheme

ϑ-Method

For given $\Delta t = \frac{T}{N} > 0$ and v^0, let v^j approximate $v(j\Delta t)$, $j = 0, \ldots, N$. Selecting $\vartheta \in [0, 1]$, the ϑ-method reads as

$$\begin{aligned} v_{j+1} &= v^j + \Delta t \left[\vartheta F_{\Delta x}(v^{j+1}, t_{j+1}) + (1-\vartheta) F_{\Delta x}(v^j, t_j)\right], \\ & \quad j = 0, \ldots, N-1, \\ v^0 &= (u_0(x_1), \ldots, u_0(x_{M-1})). \end{aligned}$$



$\vartheta = 0$ $\quad$ $\vartheta = \frac{1}{2}$ $\quad$ $\vartheta = 1$

Figure: Difference schemes for various ϑ values

Feasibility

For $\vartheta \neq 0$, implementing ϑ-method makes it necessary to solve a system of nonlinear algebraic equations at each time level. The following result holds true.

Theorem

Let $f \in C^1(\mathbb{R} \times [0,1] \times [0,T], \mathbb{R})$ *and*

$$\left| \frac{\partial f}{\partial u}(u,x,t) \right| < \mu$$

for $u \in \mathbb{R}$, $0 \leq x \leq 1$ *und* $0 \leq t \leq T$. *The implicit* ϑ*-scheme* ($\vartheta > 0$)

$$v^{j+1} = v^j + \Delta t \cdot \left[\vartheta F_{\Delta x}(v^{j+1}, t_{j+1}) + (1-\vartheta) F_{\Delta x}(v^j, t_j) \right], \; j = 0, \ldots, N-1$$

can be resolved for v^{j+1} *at any time level* $t_j = j\Delta t$ *for all* Δt *satisfying* $\vartheta \mu \Delta t < 1$.

Consistency

First, we write the fully discrete problem as

$$T^h(u) = 0, \quad u \in \mathbb{R}^{G_h}, \quad T^h : \mathbb{R}^{G_h} \longrightarrow \mathbb{R}^{G_h},$$

where we selected $h = (\Delta x, \Delta t)$ and $\Delta x = \frac{1}{M}$, $\Delta t = \frac{T}{N}$ and introduced (after slightly changing the notation) the space-time lattice

$$G_h = \{(x_i, t_j) = (i\Delta x, j\Delta t) \,|\, i = 1, \ldots, M-1,\ j = 0, \ldots, N\}.$$

For $u \in \mathbb{R}^{G_h}$, we have

$$\begin{aligned} u &= (u_1^0, \ldots, u_{M-1}^0, \ldots, u_1^1, \ldots, u_{M-1}^1, u_1^N, \ldots, u_{M-1}^N), \\ u_i^j &= u(x_i, t_j) = u(i\Delta x, j\Delta t), \quad i = 1, \ldots, M-1,\ j = 0, \ldots, N. \end{aligned}$$

Now, $T^h(u) = 0$ can explicitly be written as

$$(T^h(u))_i^j = \begin{cases} u_i^0 - u_0(x_i), & \begin{cases} j = 0, \\ i = 1, \ldots, M-1, \end{cases} \\ \frac{1}{\Delta t}\left(u_i^j - u_i^{j-1}\right) & \\ \quad -\vartheta\left[\frac{1}{\Delta x^2}\left(u_{i-1}^j - 2u_i^j + u_{i+1}^j\right) + f(u_i^j, x_i, t_j)\right] & \begin{cases} j = 1, \ldots, N, \\ i = 1, \ldots, M-1, \end{cases} \\ \quad -(1-\vartheta)\left[\frac{1}{\Delta x^2}\left(u_{i-1}^{j-1} - 2u_i^{j-1} + u_{i+1}^{j-1}\right) + f(u^{j-1}, x_i, t_{j-1})\right] & \end{cases}$$

where

$$u_0^j = g_0(t_j), \quad u_0^{j-1} = g_0(t_{j-1}), \quad u_M^j = g_1(t_j), \quad u_M^{j-1} = g_1(t_{j-1}).$$

Consistency – Cnt'd

Using classical Taylor analysis, we obtain the following consistency result.

Theorem

For $\vartheta \in [0,1]$, the ϑ-method for heat equation is consistent of order 1 in Δt and order 2 in Δx w.r.t. $\|\cdot\|_\infty$-norm, i.e.,

$$\|T^h(\bar{u}_h)\|_\infty = O(\Delta t + \Delta x^2)$$

for any classical solution $\bar{u}$ satisfying $\frac{\partial^\nu \bar{u}}{\partial t^\nu} \in C(\bar{G})$, $\nu = 1,2$, $\frac{\partial^\nu \bar{u}}{\partial x^\nu} \in C(\bar{G})$, $\nu = 0,1,2,3,4$.
For the Crank & Nicolson scheme, i.e., $\vartheta = \frac{1}{2}$, the error is even of order $O(\Delta t^2 + \Delta x^2)$ if one additionally has $\frac{\partial^3 \bar{u}}{\partial t^3} \in C(\bar{G})$.

Stability

Recall that the scheme is stable if

$$\|u - v\| \leq C\|T^h(u) - T^h(v)\|, \quad \forall u, v \in \mathbb{R}^{G_h}, \quad 0 < h \leq h_0, \quad h = (\Delta x, \Delta t).$$

We skip all stability proofs here. The idea is maily based on Hille & Yosida-type estimates for the associated discrete semigroups.

Theorem (L^∞-stability)

Under the condition

$$\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2(1-\vartheta)},$$

the ϑ-scheme for the heat equation is $\|\cdot\|_\infty$-stable on $\mathbb{R}^{G_h}$, i.e.,

$$\|u - v\|_\infty \leq (1 + T)\|T^h(u) - T^h(v)\|_\infty, \quad \forall u, v \in \mathbb{R}^{G_h}, \quad h = (\Delta x, \Delta t).$$

Stability – Std'd

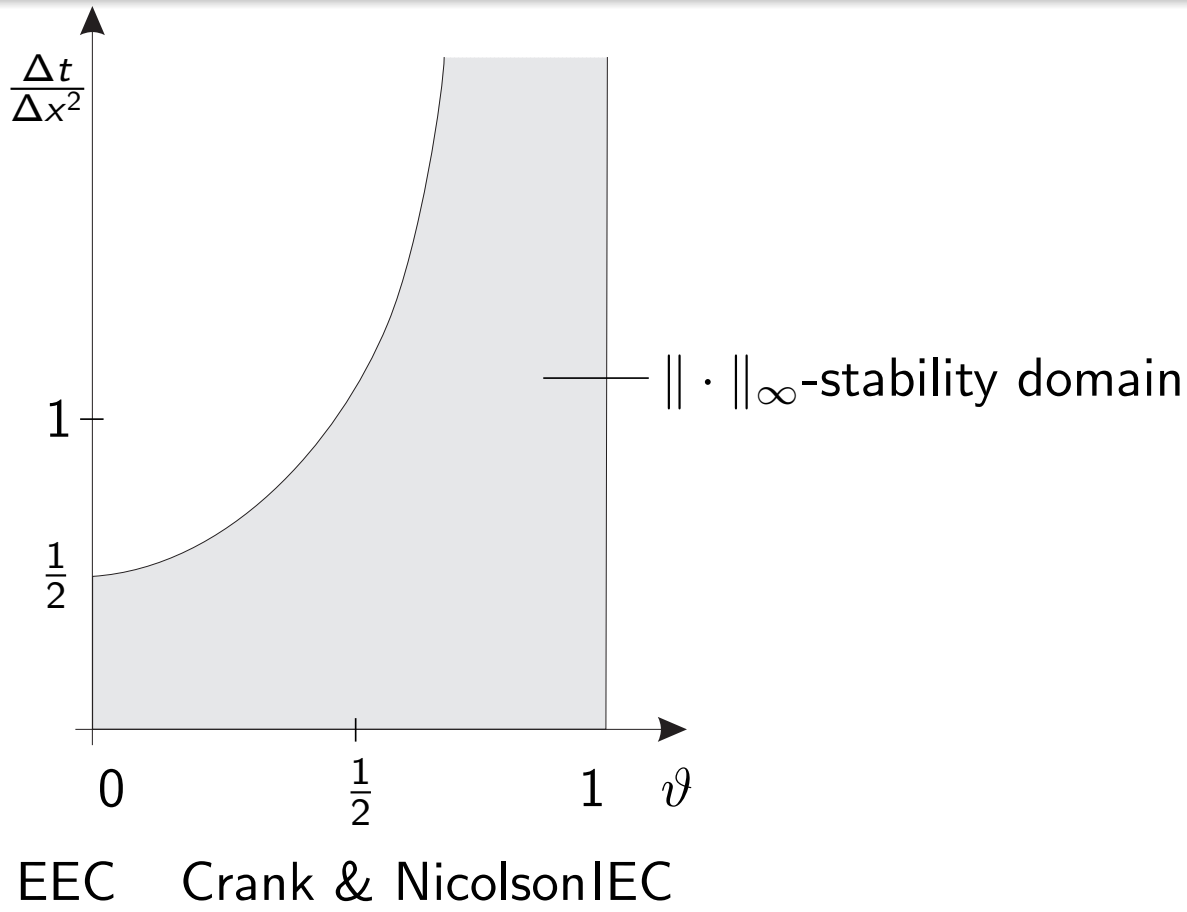


Figure: $\|\cdot\|_\infty$-stability

Convergence

Now, using Lax' principle, we get the L^∞-convergence.

Theorem

Under the conditions $\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2(1-\vartheta)}$ *and* $\vartheta\mu\Delta < 1$, *the* ϑ*-method with* $0 \leq \vartheta \leq 1$ *is convergent with respect to the* L^∞*-norm of order 1 in* Δt *and order 2 in* Δx *for any classical solution* $\bar{u}$ *to the heat equation satisfying* $\frac{\partial^\nu}{\partial t^\nu}\bar{u} \in C(\overline{G})$, $\nu = 1, 2$, $\frac{\partial^\nu}{\partial x^\nu}\bar{u} \in C(\overline{G})$, $\nu = 0, 1, 2, 3, 4$, $G = (0, 1) \times (0, T)$, *i.e.,*

$$\max\{|\bar{u}(x_i, t_j) - u^h(x_i, t_j)| \,|\, i = 1, \ldots, M-1, j = 0, \ldots, N\} \leq C \cdot (\Delta t + \Delta x^2),$$

where u^h *stands for the solution of* $T^h(u) = 0$. *For the Crank & Nicolson scheme,*

$$\max\{|\bar{u}(x_i, t_j) - u^h(x_i, t_j)| \,|\, i = 1, \ldots, M-1, j = 0, \ldots, N\} \leq C(\Delta t^2 + \Delta x^2)$$

if we additionally have $\frac{\partial^3}{\partial t^3}\bar{u} \in C(\bar{G})$.

Remark (L^2-stability & convergence)

If we consider the L^2*-norm instead of the* L^∞*-one, one can show that the Crank & Nicolson scheme is unconditionally stable and convergent for* $\vartheta \in [\frac{1}{2}, 1]$, *i.e., the time and space steps can be selected independently.*

Physical Model

Recall the Lamé system at the beginning of the talk. In the absense of external force terms, its dynamic version reads as

$$\rho U_{tt} - \operatorname{div} \sigma(\nabla U) = 0 \text{ in } (0,\infty) \times G.$$

Here, $U\colon (0,\infty) \times G \to \mathbb{R}^d$ is the unknown displacement field and $\rho > 0$ denotes the material density. Assuming the strain tensor to be linear, we obtain the classical equations of homogenenous isotropic elastodynamics

$$\rho U_{tt} - \mu \triangle U - (\lambda + \mu) \nabla \operatorname{div} U = 0.$$

Now, if we assume the vector field to be solenoidal (i.e., $\operatorname{div} U = 0$), we get

$$\rho U_{tt} - \mu \triangle U = 0.$$

Taking any of the d-components of this PDE system, we arrive at the classical scalar wave equation.

Model Problem and Discretization

We consider the following model problem

$$\begin{aligned} u_{tt} &= c^2 u_{xx} + f(u_x, u, x, t), \quad 0 \le x \le 1, 0 \le t \le T, \\ u(x,0) &= u_0(x), \quad u_t(x,0) = u_1(x), \quad 0 \le x \le 1, \\ u(0,t) &= g_0(t), \quad u(1,t) = g_1(t), \quad 0 \le t \le T. \end{aligned}$$

Introducing an equidistant lattice on the space-time cyllinder $[0,1] \times [0,T]$

$$\begin{aligned} G_h &= \{(i\Delta x, j\Delta t) \,|\, i = 1, \ldots, M-1, j = 0, \ldots, N\}, \\ \Delta x &= \frac{1}{M}, \quad \Delta t = \frac{T}{N}, \quad h = (\Delta x, \Delta t) \end{aligned}$$

and applying the method of lines to the finite difference semidiscretization in space, we obtain the following numerical scheme similar to the ϑ-scheme for heat equation

$$\frac{1}{\Delta t^2}\left(u_i^{j+1} - 2u_i^j + u_i^{j-1}\right) = \vartheta\sigma_i^{j+1} + (1-2\vartheta)\sigma_i^j + \vartheta\sigma_i^{j-1}, \quad \vartheta \in \left[0, \tfrac{1}{2}\right]$$

with

$$\sigma_i^j = \frac{c^2}{\Delta x^2}\left(u_{i-1}^j - 2u_i^j + u_{i+1}^j\right) + f\left(\frac{1}{2\Delta x}\left(u_{i+1}^j - u_{i-1}^j\right), u_i^j, x_i, t_j\right),$$

where $x_i = i\Delta x$, $t_j = j\Delta t$.

Model Problem and Discretization – Cnt'd

For $j = 0$, we trivially set

$$u_i^0 = u_0(x_i), \quad i = 1, \ldots, M-1.$$

Further, we let

$$u_0^j = g_0(t_j), \quad u_M^j = g_1(t_j), \quad j = 0, \ldots, N.$$

To obtain an equation for u_i^1, we use the following second order approximation

$$\frac{u(x, \Delta t) - u(x, 0)}{\Delta t} = u_t(x, 0) + \frac{1}{2}\Delta t \cdot u_{tt}(x, 0) + O(\Delta t^2).$$

Taking into account

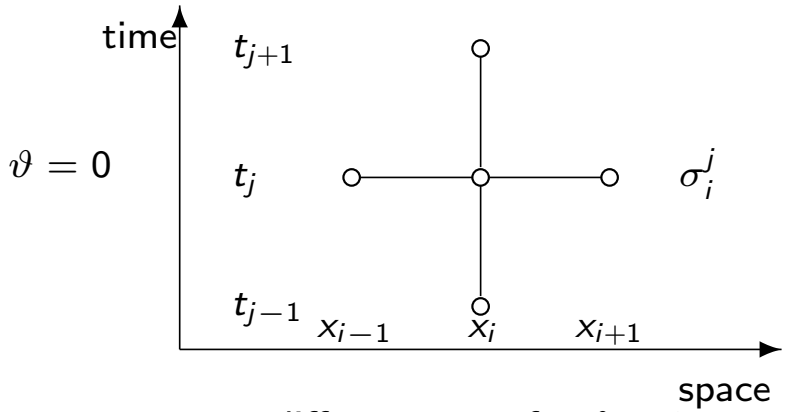
$$\bar{u}_{tt}(x, 0) = c^2 \bar{u}_{xx}(x, 0) + f(\bar{u}_x, \bar{u}, x, 0) = c^2 u_0''(x) + f(u_0', u_0, x, 0)$$

with the classical solution $\bar{u}$ to the wave equation, we can write

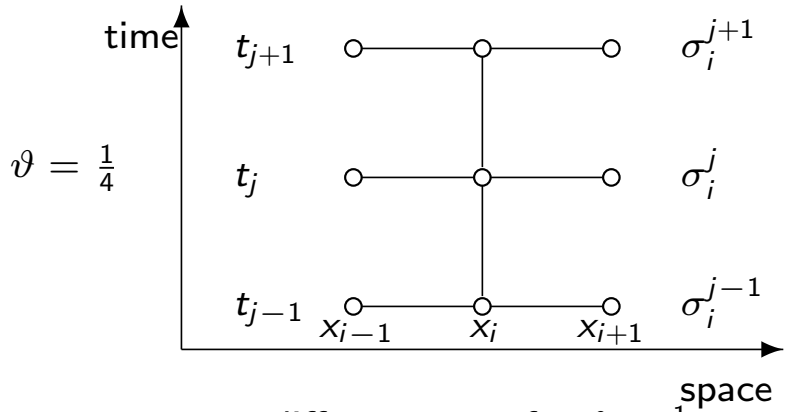
$$\frac{1}{\Delta t}\left(u_i^1 - u_i^0\right) = u_1(x_i) + \frac{\Delta t}{2}\left(c^2 u_0''(x_i) + f(u_0'(x_i), u_0(x_i), x_i, 0)\right)$$

for $i = 1, \ldots, M-1$.

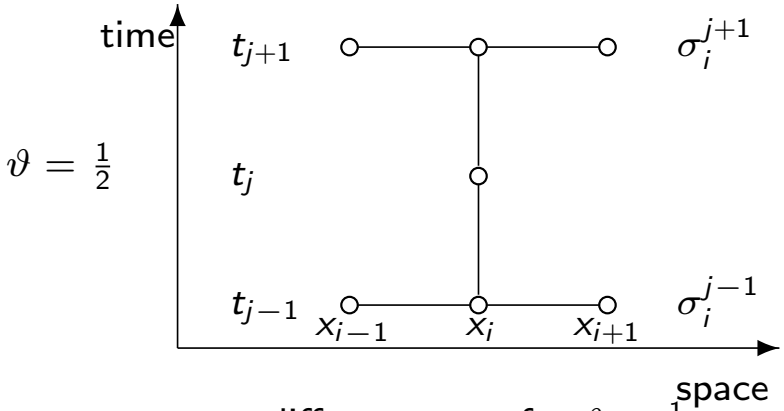
Graphical Illustration



difference star for $\vartheta = 0$



difference star for $\vartheta = \frac{1}{4}$



difference star for $\vartheta = \frac{1}{2}$

Compact Notation

In a compact form, our numerical scheme can be written as

$$T^h(u) = 0$$

with $T^h : \mathbb{R}^{G_h} \longrightarrow \mathbb{R}^{G_h}$,

$$(T^h(u))_i^j = \begin{pmatrix} u_i^0 - u_0(x_i),\ j = 0,\ i = 1, \ldots, M-1 \\ \frac{1}{\Delta t}\left(u_i^1 - u_i^0\right) - u_1(x_i) - \frac{\Delta t}{2}\left[c^2 u_0''(x_i) + f(u_0'(x_i), u_0(x_i), x_i, 0)\right], \\ j = 1,\ i = 1, \ldots, M-1 \\ \frac{1}{\Delta t^2}\left(u_i^j - 2u_i^{j-1} + u_i^{j-2}\right) - \left[\vartheta\sigma_i^j + (1-2\vartheta)\sigma_i^{j-1} + \vartheta\sigma_i^{j-2}\right], \\ j = 2, \ldots, M,\ i = 1, \ldots, M-1 \end{pmatrix}$$

where

$$\sigma_i^j = \frac{c^2}{\Delta x^2}\left(u_{i-1}^j - 2u_i^j + u_{i+1}^j\right) + f\left(\frac{1}{2\Delta x}\left(u_{i+1}^j - u_{i-1}^j\right), u_i^j, x_i, t_j\right).$$

Consistency

Standard Taylor analysis yields:

Theorem

For any solution $\bar{u} \in C^4([0,1] \times [0,T])$ to the wave equation, the parameter-dependent scheme is consistent of order $O(\Delta t^2 + \Delta x^2)$ if f ist smooth and Lipschitzian w.r.t. x.

Iteration Form

For the sake of simplicity, let $f \equiv 0$. Introducing the following notation

$$v(t) = (u(x_1, t), \ldots, u(x_{M-1}, t)), \quad v^j = v(t_j) = v(j\Delta t),$$

$$\Gamma = \frac{c^2}{\Delta x^2} \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 2 & -1 & 0 \\ 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{pmatrix}, \quad r^j = \frac{c^2}{\Delta x^2} \begin{pmatrix} g_0(t_j) \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ g_1(t_j) \end{pmatrix},$$

our iterative scheme can be written as

$$\begin{aligned} \frac{1}{\Delta t^2}\left(v^{j+1} - 2v^j + v^{j-1}\right) &= \vartheta\left(-\Gamma v^{j+1} + r^{j+1}\right) + (1 - 2\vartheta)(-\Gamma v^j + r^j) \\ &\quad + \vartheta(-\Gamma v^{j-1} + r^{j-1}), \quad j = 1, \ldots, N-1. \end{aligned}$$

Iteration Form – Cnt'd

Equivalently, the scheme reads as

$$\left(\frac{1}{\Delta t^2} I + \vartheta \Gamma\right) v^{j+1} = \frac{1}{\Delta t^2}\left(2v^j - v^{j-1}\right) + \vartheta r^{j+1} + (1-2\vartheta)(-\Gamma v^j + r^j) + \vartheta(-\Gamma v^{j-1} + r^{j-1}), \quad j = 1, \ldots, N-1$$

with starting values

$$\begin{aligned} v^0 &= (u_0(x_1), \ldots, u_0(x_{M-1})), \\ \frac{1}{\Delta t}(v^1 - v^0) &= (u_1(x_1), \ldots, u_1(x_{M-1})) + \frac{\Delta t}{2} c^2 (u_0''(x_1), \ldots, u_0''(x_{M-1})). \end{aligned}$$

The iteration structure suggests that the invertibility of matrix

$$A = \frac{1}{\Delta t^2} I + \vartheta \Gamma,$$

is crucial for the iteration to be resolvable at every time step. Note that A is obviously an M-matrix.

Stability

For stability analysis, we only consider the case $\vartheta = 0$ corresponding to the explicit Euler & Cauchy scheme.

The numerical solution u_i^j approximating $\bar{u}(x_i, t_j)$ depends on the initial data $u_{i-j}^0, u_{i-j+1}^0, \ldots, u_{i+j}^0$. We call the interval $[(i-j)\Delta x, (i+j)\Delta x] = [x_{i-j}, x_{i+j}]$ the numerical dependence interval at point $(x^*, t^*) = (x_i, t_j)$. (In multiple dimensions, one usually speaks of a numerical dependence ball.)

The figure below shows the so-called numerical hyperbolic cone containing all points on the space-time grid which affect the numerical approximation u_i^j.

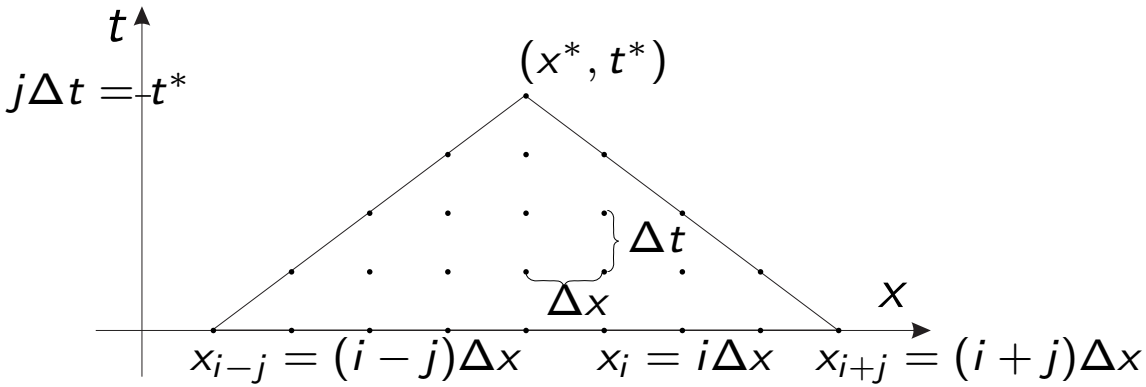


Figure: Numerical hyperbolic cone

Stability

Due to the final signal propagation speed correponding to solutions of the wave equation, analytic counterparts of the numerical dependence interval and the numerical hyperbolic cone can similarly be defined for the original hyperbolic PDE.
Obviously, if the numerical hyperbolic cone is smaller than the analytic one, the data outside of the numerical hyperbolic cone have no influence on the numerical solution whereas the latter is the case for the continuous PDE. Therefore, no consistency and/or stability can be expected.
This motivates the following:

Courant-Friedrichs-Levy or CFL condition

The numerical dependence interval must be contained in the analytic dependece interval, which is equivalent to the step condition

$$c\Delta t \leq \Delta x.$$

Stability

Under the CFL condition, one can use discrete Fourier analysis or energy techniques to prove:

Theorem (Stability)

The explicit Euler scheme for the wave equation is stable in the L^2-norm w.r.t x and L^∞-norm w.r.t. t.

Corollary

Assume the wave equation possesses a unique solution

$$\bar{u} \in C^4([0,1] \times [0,T]).$$

Then the explicit Euler scheme ($\vartheta = 0$) is convergent w.r.t. to the norm

$$\|u\|_{2,\infty} = \max\left\{ \left| \sum_{i=1}^{M-1} (u_i^j)^2 \right|^{1/2} \middle| j = 0, \ldots, N \right\}.$$

of order 2 both in x and t if the CFL condition is satisfied.

References

- Ciarlet, P. G.; Lions, J.L. (eds.), Handbook of Numerical Analysis. Vol. I, II, ..., Amsterdam – North-Holland, 1990+
- Kelly, L. G. Handbook of numerical methods and applications. Reading, Mass.: Addison-Wesley, 1967
- Schiesser, W. E. The numerical methods of lines, 1991
- Sparrow, E. M., and Murthy, J. (eds.) Handbook of numerical heat transfer. New York etc.: Wiley, 1988
- Tadmor, E. A review of numerical methods for nonlinear partial differential equations. Bulletin of the American Mathematical Society 49.4, 507-554, 2012

Merton's Portfolio Problem

Consider an investor who is planing his and her investment strategy over a finite time horizon $[0, T]$. His or her wealth S_t at time t is assumed to be random. At time t, the investor must decide the amount c_t of his or her wealth to consume and the fraction π_t to invest in a stock market, whereas the remaining fraction $1 - \pi_t$ is invested in a risk-free asset (e.g., a bond).
With r denoting the interest rate from the risk-free asset, μ and σ standing for the trend/exptected return and volatility of the financial market and $(W_t)_{t\geq 0}$ being a standard Wiener process, the equation for investor's wealth development reads as

$$\mathrm{d}S_t = \big((r + \pi_t(\mu - r))S_t - c_t\big)\,\mathrm{d}t + S_t\pi_t\sigma\,\mathrm{d}W_t, \qquad (*)$$

where ξ is investor's initial wealth (which can be random).

Merton's Portfolio Problem – Cnt'd

Equation $(*)$ is a stochastic differential equation (SDE). The differentials appearing in this equation cannot be interpreted classically. Indead, neither S_t nor W_t are classical differentiable (at most Hölder-continuous of degree $\alpha \in (0, \frac{1}{2})$). Thus, after applying Itô integration procedure, an ODE is usually interpreted as a stochastic integral equation. Unlike ODE, the solution one is looking for is a random process. Therefore, measurability and adaptedness issues are very important here.

Merton's Portfolio Problem – Cnt'd

Typical questions that arrise are:

- Find the random process solving the SDE or compute some function(al)s depending on the solution process, e.g., $\mathrm{E}\big[f(S_T)\big]$, etc.
- For a given utility function $u\colon [0,\infty) \to [0,\infty)$, under appropriate restrictions on consumption and investment, find an optimal consumption/investment strategy ("portfolio") to maximize some functional, e.g.,

$$\mathrm{E}\left[\int_0^T e^{-\rho s} u(S_s)\,\mathrm{d}s + e^{-\rho T} u(S_T)\right]$$

- etc.

Merton, R. C. Optimum consumption and portfolio rules in a continuous-time model. Journal of Economic Theory 3 (4), 1971

Itô Integral

We assume the audience to be familiar with basic stochastic process theory. We only remind of the two following seminal concepts. Let $(\Omega, \mathcal{F}, \mathrm{P})$ be a probability space.

Let $(W_t)_{t\geq 0}$ be a Wiener process adapted to its natural filtration $(\mathcal{F})_{t\geq 0}$ and let H be a random process adapted to the same filtration such that $H_t \in L^p$ w.r.t. t P-a.s. for some $p \in [1, \infty)$. The Itô integral is then given as

$$\int_0^T H_t \mathrm{d}W_t := \lim_{n\to\infty} \sum_{[t_{i-1},t_i]\in\pi_n} H_{t_{i-1}}(W_{t_i} - W_{t_{i-1}}),$$

where π_n denotes a partition of $[0, T]$ with its diameter going to zero and the approximation of H_t by a sequence of step processes converges in probability in L^p w.r.t. to t. Note that in contrast to classical Stiltjes integral, W_t has unbounded variation P-a.s.

One can show that the Itô integral is well-defined as an $\mathcal{F}_T$-measurable random variable. In L^2, one has the so-called Ito isometry

$$\mathrm{E}\left[\left(\int_0^t H_s \,\mathrm{d}W_s\right)^2\right] = \mathrm{E}\left[\int_0^t H_s^2 \,\mathrm{d}s\right].$$

Itô's Rule

For a Brownian motion $(W_t)_{t\geq 0}$, consider the random process given by

$$X_t = X_0 + \int_0^t \sigma_s \,\mathrm{d}W_s + \int_0^t \mu_s \,\mathrm{d}s,$$

where σ is predictable and $(W_t)_{t\geq 0}$-integrable and μ is predictable and Lebesgue-integrable.
Then we have the famous Itô's lemma.

Lemma (Itô's Rule or Stochastic Chain Rule)

There holds

$$\begin{aligned}\mathrm{d}f(t,X_t) &= \frac{\partial f}{\partial t}\,\mathrm{d}t + (\nabla f(X_t)^T\,\mathrm{d}X_t + \frac{1}{2}(\mathrm{d}X_t)^T(\nabla^2 f)\,\mathrm{d}X_t,\\ &= \left(\frac{\partial f}{\partial t} + (\nabla f)^T\mu_t + \frac{1}{2}\mathrm{tr}[\sigma_t^T(\nabla^2 f)\sigma_t]\right)\mathrm{d}t + (\nabla f)^T\sigma_t\,\mathrm{d}W_t\end{aligned}$$

for any $f \in C^2$.

Note that this equation differs from the standard chain rule due to the additional term involving the second derivative of f, which comes from the property that Brownian motion has non-zero quadratic variation.

Existence & Uniqueness for SDE

Let $A, B \in C([0,T] \times \mathbb{R}^d)$, $(\mathcal{F}_t)_{t\geq 0}$ be a Wiener process adapted to a σ-algebra filtration $(\mathcal{F}_t)_{t\geq 0}$ and ξ be an $\mathcal{F}_0$-measurable r.v. Consider the SDE $(\mathcal{F}_t)_{t\geq 0}$-adapted process $(X_t)_{t\geq 0}$

$$\mathrm{d}X_t = A(t, X_t)\mathrm{d}t + B(t, X_t)\mathrm{d}W_t, \quad X_0 = \xi$$

or, equivalently,

$$X_t = \xi + \int_0^t A(s, X_s)\mathrm{d}s + \int_0^t B(s, X_s)\mathrm{d}W_s \quad \mathrm{P} - \text{a.s. for any } t \in [0,T].$$

Note that the solution is P-a.s. path-continuous.

Existence & Uniqueness for SDE – Cnt'd

Theorem

Let

- A, B *are continuous on* $[0, T] \times \mathbb{R}^d$,
- $|A(t,x)| \leq K(1+|x|)$, $|B(t,x)| \leq K(1+|x|)$ *for* $t \in [0,T]$, $x \in \mathbb{R}^d$,
- A, B *are Lipschitzian w.r.t.* x *uniformly in* $t \in [0,T]$,
- ξ *has a finite variance.*

Then the SDE possesses a unique solution on $[0,T]$. *Moreover, the variance of* X_t *is finite for every* $t \in [0,T]$.

Euler-Maruyama Scheme

One of the most popular solution approaches for SDE are finite difference schemes. A typical example is given by Euler-Maruyama scheme being a probabilistic counterpart of Euler scheme. In contrast to Runge-Kutta schemes, etc. used for deterministic ODE, finite difference schemes must be compatible with Itô's rule.
Consider an SDE

$$\mathrm{d}X_t = A(t, X_t)\mathrm{d}t + B(t, X_t)\mathrm{d}W_t \text{ for } t \in [0, T], \quad X_0 = \xi.$$

For a step value $h = \frac{T}{N}$ with $N \in \mathbb{N}$, we define an equidistant lattice $\{t_0, \ldots, t_N\}$ with $t_n = nh$. Denoting with $\hat{X}_{t_n}$ the approximation for X_{t_n}, the Euler-Maruyama scheme reads as

$$\hat{X}_{t_{n+1}} = \hat{X}_{t_n} + A(t_n, \hat{X}_{t_n})h + B(t_n, \hat{X}_{t_n})\Delta W_{t_n}, \quad \hat{X}_0 = \xi$$

with the Brownian motion increments $\triangle W_{t_n} = W_{t_{n+1}} - W_{t_n} \sim \mathcal{N}(0, 1)$.

Strong Convergence

Applying Itô's rule along with a discrete Gronwall inequality, we get:

Theorem

There exists a constant $C > 0$ such that for any $h > 0$

$$\max_{0 \le t_n \le T} \mathrm{E}\left[|X_{t_n} - \hat{X}_{t_n}|\right] \le Ch^{1/2}$$

if A, B are C^2 functions which are bounded together with their derivatives up to order 2.

In contrast to the deterministic case, the convergence order is $\frac{1}{2}$ and not 1 as it is the case for explicit Euler scheme.
Since we do not know the actual expectation in practice, $\mathrm{E}[\cdot]$ is estimated by the sample mean. By the virtue of Central Limit Theorem, reliable confidence intervals can be obtained.

Weak Convergence

For the weak convergence, not the discrepancy between X_t and $\hat{X}_t$ but the one between some "test functionals" depending on these random variables is estimated.

Theorem

For any test function

$$g \in C^4(\mathbb{R}^d, \mathbb{R}^d) \text{ with polynomial growth,}$$

there exists a constant $C > 0$ *such that for any* $h > 0$

$$\max_{0 \le t_n \le T} \mathrm{E}\big[|g(X_{t_n}) - g(\hat{X}_{t_n})|\big] \le Ch$$

if A, B *are bounded* C^2*-functions will bounded derivatives up to order 2.*

Though the weak convergence order is 1 and thus is higher than the strong one, weak approximation does not imply any proximity between the numerical paths and the actual ones. It only guarantees that the distributions (on the space of continuous functions) are similar.

Weaker Convergence – Cnt'd

At the same time, if not the whole solution process $(X_t)_t$ but only, say, $g(X_T)$ or $\mathrm{E}[X_T]$ need to approximated, the weak convergence is appropriate.

Recall Merton's portfolio problem considered before. If we only want to compute the value of utility functional depending our investment strategy (and the functional is regular), we get a higher convergence rate than the strong one.

A typical example in financial mathematics is options princing. Given a pay-off function, we can use Euler-Maruyama scheme to compute the strike/excercise price for

- American put or call options
- European put or call options
- Asian put or call options
- etc.

or evaluate various portfolio values, etc.

Option Pricing

Given a simple market model

$$\mathrm{d}S_t = rS_t + \sigma S_t \mathrm{d}S_t, \quad S_t = s \text{ nonrandom}$$

for the stock price S_t as well as a strike function, we want to compute the discounted option price at time T

$$\mathrm{E}\big[\exp(-r(T-t)S_t\big].$$

As we already pointed out, one can use the Euler-Maruyama scheme to perform a Monte-Carlo simulation to estimate the option price. An alternative idea is based on Fokker-Planck equation establishing a connection between SDE and parabolic PDE.
Note that solutions to SDE are Markovian processes. Therefore, under appropriate regularity assumptions, they possess infitesimal generators and can be interpreted in terms of semigroup or Kato theory.

Fokker-Planck Equation

Assume we are given (for simplicity) a 1D SDE

$$\mathrm{d}X_t = a(t, X_t)\mathrm{d}t + b(t, X_t)\mathrm{d}W_t \text{ for } t \in [0, T], \quad X_0 = \xi.$$

The solution process is assumed to possess a time-dependent density function $\psi = \psi(t, x)$. Then we can write

$$\mathrm{E}[g(X_t)] = \int_{\mathbb{R}} g(x)\psi(t, x)\mathrm{d}x.$$

Applying Itô's rule, one can show under appropriate conditions that

$$\psi_t + (a\psi)_x - \tfrac{1}{2}(b^2\psi)_{xx} \text{ in } (0, T) \times \mathbb{R}, \quad \psi(0, \cdot) = f_\xi,$$

where f_ξ is the pdf of ξ (f_ξ is a delta-function if ξ is fixed.)
Thus, instead of applying Euler-Maruyama scheme, we can use a numerical procedure to solve the parabolic Fokker-Planck PDE and then estimate $\mathrm{E}[g(X_t)]$ based on the numerical approximation of density function. Severe problems ("curse of dimensionality") arise though when the process $(X_t)_{t\geq 0}$ is multidimensional.

References

- Akinbo, B. J., Faniran, T., and Ayoola, E. O. Numerical Solution of Stochastic Differential Equations.
- Chang, C.-C. Numerical solution of stochastic differential equations. Lawrence Berkeley Lab., CA (USA), 1985
- Evans, L. C. An introduction to stochastic differential equations. Vol. 82. American Mathematical Soc., 2012
- Gard, T. C. Introduction to stochastic differential equations. M. Dekker, 1988
- Kloeden, P. E., and Platen, E. Numerical Solution of Stochastic Differential Equations, Stochastic Modelling and Applied Probability, Vol. 23, Springer, 1992
- Mackevičius, V. Numerical Solution of Stochastic Differential Equations. Introduction to Stochastic Analysis, Wiley, 2011

Thank you for your attention!