MA4255 Numerical Methods in Differential Equations

Chapter 6: Introduction to the theory of finite difference (FD) schemes

6.1 Elliptic boundary-value problems

6.2 Methodology of FD schemes

6.3 FD approximation of a two-point boundary-value problem

6.4 Key steps of a general error analysis for FD approximations of elliptic PDEs

6.1 Elliptic boundary-value problems

Linear second-order elliptic PDEs

Elliptic PDEs are typified by the **Laplace equation**

$$\Delta u := \partial^2_{x_1x_1}u + \cdots + \partial^2_{x_nx_n}u = 0 \quad \text{in } \Omega,$$

and its nonhomogeneous counterpart, the **Poisson equation**

$$-\Delta u = f \quad \text{in } \Omega,$$

posed on a bounded open domain $\Omega \subset \mathbb{R}^n$. More generally, we consider the (linear) second-order PDE

$$-\sum_{i,j=1}^{n} \partial_{x_j}(a_{ij}\partial_{x_i}u) + \sum_{i=1}^{n} b_i\partial_{x_i}u + cu = f \quad \text{in } \Omega, \tag{1}$$

where $a_{ij} \in C^1(\overline{\Omega})$, $b_i, c, f \in C(\overline{\Omega})$ for $i, j \in \{1, \dots, n\}$, and additionally

$$\exists \tilde{c} > 0 : \quad \sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq \tilde{c}|\xi|^2 \quad \forall x \in \overline{\Omega},\ \xi \in \mathbb{R}^n. \tag{2}$$

Condition (2) is called **uniform ellipticity**. We call (1) an **elliptic PDE**.

Rk: Poisson's eqn is of the form (1) with $a_{ij} \equiv \delta_{ij}$, $b_i \equiv 0$, $c \equiv 0$, and the uniform ellipticity condition holds with $\tilde{c} = 1$.

An equivalent way of writing the PDE

Recall the general linear second-order elliptic PDE:

$$-\sum_{i,j=1}^{n} \partial_{x_j}(a_{ij}\partial_{x_i}u) + \sum_{i=1}^{n} b_i\partial_{x_i}u + cu = f \quad \text{in } \Omega$$

with uniform ellipticity condition

$$\exists \tilde{c} > 0 : \quad \sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq \tilde{c}|\xi|^2 \quad \forall x \in \overline{\Omega},\ \xi \in \mathbb{R}^n.$$

The PDE can equivalently be written as

$$-\mathrm{div}(A\nabla u) + \mathbf{b}\cdot\nabla u + cu = f \quad \text{in } \Omega,$$

where $A(x) = (a_{ij}(x))_{1\leq i,j\leq n}$ and $\mathbf{b}(x) = (b_1(x), \ldots, b_n(x))^{\mathrm{T}}$, and the uniform ellipticity condition can equivalently be written as

$$(A(x)\xi)\cdot\xi \geq \tilde{c}|\xi|^2 \quad \forall x \in \overline{\Omega},\ \xi \in \mathbb{R}^n.$$

Notation: $v \cdot w := v^{\mathrm{T}}w$ for $v, w \in \mathbb{R}^n$. Recall the divergence of a vector field $\mathbf{p}(x) = (p_1(x), \ldots, p_n(x))^{\mathrm{T}}$ is defined as $\mathrm{div}(\mathbf{p}) := \sum_{i=1}^{n} \partial_{x_i}p_i$.

Types of boundary conditions

The PDE

$$-\text{div}(A\nabla u) + \mathbf{b} \cdot \nabla u + cu = f \quad \text{in } \Omega,$$

is supplemented with one of the following **boundary conditions (b.c.)**:

- $u = g$ on $\partial\Omega$ (**Dirichlet b.c.**);
  (if $g \equiv 0$, this b.c. is called **homogeneous Dirichlet b.c.**)

- $\partial_\nu u = g$ on $\partial\Omega$, where ν denotes the unit outward normal vector to the boundary $\partial\Omega$ of Ω, and where the derivative in the direction of ν is defined by $\partial_\nu u := \nabla u \cdot \nu$ (**Neumann b.c.**);

- $\partial_\nu u + \sigma u = g$ on $\partial\Omega$, where $\sigma(x) \geq 0 \ \forall x \in \partial\Omega$ (**Robin b.c.**).

The PDE together with a b.c. is called **boundary-value problem (BVP)**.

What do we mean by a solution to a given BVP?

Let us consider the **homogeneous Dirichlet BVP**

$$\begin{aligned} -\mathrm{div}(A\nabla u) + \mathbf{b}\cdot\nabla u + cu &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega. \end{aligned}$$

Q: What do we mean by a solution u to this BVP?

The classical/seemingly obvious answer: A solution to this BVP is a fct $u \in C^2(\Omega) \cap C(\overline{\Omega})$ satisfying the eqn and the b.c. pointwise, i.e.,

$$\begin{aligned} -\mathrm{div}(A(x)\nabla u(x)) + \mathbf{b}(x)\cdot\nabla u(x) + c(x)u(x) &= f(x) \qquad \forall x \in \Omega, \\ u(x) &= 0 \qquad \forall x \in \partial\Omega. \end{aligned}$$

Such a function u is called a **classical solution** to the BVP.

The theory of PDEs tells us that the BVP has a unique classical solution, provided that a_{ij}, b_i, c, f and $\partial\Omega$ are sufficiently smooth.
However, in real-life applications one encounters BVPs where these smoothness requirements are violated, and a classical soln might not exist.

An elliptic BVP with no classical solution

Consider the Poisson equation on $\Omega := (-1,1)^n$, subject to a homogeneous Dirichlet b.c.:

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega, \end{aligned}$$

where

$$f : \overline{\Omega} \to \mathbb{R}, \quad f(x) := \mathrm{sgn}(\tfrac{1}{2} - |x|).$$

This problem does not have a classical solution $u \in C^2(\Omega) \cap C(\overline{\Omega})$. Indeed, if there were, then $\Delta u \in C(\Omega)$, which is impossible as $f \notin C(\Omega)$.

However, we will see later that this problem has a so-called **weak solution**.

What is a weak solution? The idea:

Goal: generalize the notion of solution by weakening the differentiability requirements on u. Suppose that u is a classical solution of

$$\begin{aligned} -\mathrm{div}(A\nabla u) + \mathbf{b}\cdot\nabla u + cu = f \qquad & \text{in } \Omega, \\ u = 0 \qquad & \text{on } \partial\Omega. \end{aligned}$$

Then, for any $v \in C_c^1(\Omega)$ we have

$$-\int_\Omega \mathrm{div}(A\nabla u)\, v \,\mathrm{d}x + \int_\Omega \mathbf{b}\cdot\nabla u\, v\,\mathrm{d}x + \int_\Omega cu\, v\,\mathrm{d}x = \int_\Omega f\, v\,\mathrm{d}x.$$

Integration by parts (div. thm) and noting that $v = 0$ on $\partial\Omega$, we obtain

$$\int_\Omega (A\nabla u)\cdot\nabla v\,\mathrm{d}x + \int_\Omega \mathbf{b}\cdot\nabla u\, v\,\mathrm{d}x + \int_\Omega cu\, v\,\mathrm{d}x = \int_\Omega f\, v\,\mathrm{d}x \quad \forall\, v \in C_c^1(\Omega).$$

In order for this equality to make sense we no longer need to assume $u \in C^2(\Omega)$: it is sufficient that $u \in L^2(\Omega)$ and $\partial_{x_i} u \in L^2(\Omega)$ for $i \in \{1, \ldots, n\}$. Thus, it is natural to seek u in the space $H_0^1(\Omega)$ instead. We note that $C_c^1(\Omega) \subset H_0^1(\Omega)$, and observe that when $u \in H_0^1(\Omega)$ and $v \in H_0^1(\Omega)$, (instead of $v \in C_c^1(\Omega)$), the expressions on the left-hand side and right-hand side of this equality are both still meaningful.

Definition of a weak solution

Consider the homogeneous Dirichlet BVP

$$-\mathrm{div}(A\nabla u) + \mathbf{b}\cdot\nabla u + cu = f \qquad \text{in } \Omega, \tag{3}$$

$$u = 0 \qquad \text{on } \partial\Omega. \tag{4}$$

Definition (Weak solution)

Let $a_{ij} \in C(\overline{\Omega})$ for $i, j \in \{1, \ldots, n\}$, $b_i \in C(\overline{\Omega})$ for $i \in \{1, \ldots, n\}$, $c \in C(\overline{\Omega})$, $f \in L^2(\Omega)$. Let $A : \overline{\Omega} \to \mathbb{R}^{n\times n}$, $A(x) = (a_{ij}(x))_{1\le i,j\le n}$, and $\mathbf{b} : \overline{\Omega} \to \mathbb{R}^n$, $\mathbf{b}(x) = (b_1(x), \ldots, b_n(x))^{\mathrm{T}}$. A fct $u \in H_0^1(\Omega)$ satisfying

$$\int_\Omega (A\nabla u)\cdot\nabla v\,\mathrm{d}x + \int_\Omega \mathbf{b}\cdot\nabla u\,v\,\mathrm{d}x + \int_\Omega cu\,v\,\mathrm{d}x = \int_\Omega f\,v\,\mathrm{d}x \quad \forall\, v \in H_0^1(\Omega)$$

is called a **weak solution** of (3)–(4).
(All partial derivatives should be understood as weak derivatives.)

Existence and uniqueness of weak solutions: The key tool

The key tool in proving the existence and uniqueness of a weak solution is the **Lax–Milgram theorem**:

Theorem (Lax–Milgram)

Let V be a real Hilbert space with norm $\|\cdot\|_V$. Let $a : V \times V \to \mathbb{R}$ and $l : V \to \mathbb{R}$ be maps with the following properties:

- *l is linear and a is bilinear, i.e., $v \mapsto a(v, w)$ is linear for any fixed w, and $w \mapsto a(v, w)$ is linear for any fixed v,*
- *$\exists\, c_0 > 0$ s.t. $a(v, v) \geq c_0 \|v\|_V^2 \;\; \forall v \in V$ (coercivity of a),*
- *$\exists\, c_1 \geq 0$ s.t. $|a(v, w)| \leq c_1 \|v\|_V \|w\|_V \;\; \forall v, w \in V$ (boundedness of a),*
- *$\exists\, c_2 \geq 0$ s.t. $|l(v)| \leq c_2 \|v\|_V \;\; \forall v \in V$ (boundedness of l).*

Then, there exists a unique $u \in V$ such that $a(u, v) = l(v) \;\; \forall v \in V$.

Ex: Existence & uniqueness of weak soln via Lax–Milgram

Let $\Omega := (0,1)$, $f \in L^2(\Omega)$, $p : \overline{\Omega} \to \mathbb{R}$, $p(x) := 2e^x$. Consider the BVP

$$-(pu')' = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega.$$

Claim: This problem has a unique weak solution $u \in H_0^1(\Omega)$, i.e., there exists a unique $u \in H_0^1(\Omega)$ such that

$$\int_0^1 p\, u'\, v' \,\mathrm{d}x = \int_0^1 f\, v \,\mathrm{d}x \qquad \forall v \in H_0^1(\Omega).$$

Proof: Step 1: Define a Hilbert space V with norm $\|\cdot\|_V$, a bilinear map $a : V \times V \to \mathbb{R}$, and a linear map $l : V \to \mathbb{R}$ such that $u \in V$ is a weak solution iff $a(u,v) = l(v)\ \forall v \in V$.

We consider the Hilbert space $V := H_0^1(\Omega)$ with norm $\|\cdot\|_V := \|\cdot\|_{H^1(\Omega)}$. We define $a : V \times V \to \mathbb{R}$ and $l : V \to \mathbb{R}$ by

$$a(v,w) := \int_0^1 p\, v'\, w' \,\mathrm{d}x, \qquad l(v) := \int_0^1 f\, v \,\mathrm{d}x$$

for $v, w \in V$. Note that a is bilinear, l is linear, and we have that $u \in V$ is a weak solution iff $a(u,v) = l(v)\ \forall v \in V$.

Recall:

$$a : V \times V \to \mathbb{R}, \quad a(v,w) := \int_0^1 p\, v'\, w' \,\mathrm{d}x.$$

Step 2: Show coercivity of a, i.e., $\exists c_0 > 0$ s.t. $a(v,v) \geq c_0 \|v\|_V^2 \; \forall v \in V$.

Using that $p(x) = 2e^x \geq 2$ for all $x \in [0,1]$, we have for any $v \in V$ that

$$\begin{aligned} a(v,v) = \int_0^1 p\,|v'|^2 \,\mathrm{d}x &\geq 2 \int_0^1 |v'|^2 \,\mathrm{d}x \\ &\geq \frac{2}{1+c_\star} \left(\int_0^1 |v'|^2 \,\mathrm{d}x + \int_0^1 |v|^2 \,\mathrm{d}x \right) \\ &= \frac{2}{1+c_\star} \|v\|_{H^1(\Omega)}^2 = c_0 \|v\|_V^2, \end{aligned}$$

where $c_0 := \frac{2}{1+c_\star} > 0$ with $c_\star$ being the constant from the Poincaré–Friedrichs inequality

$$\int_0^1 |v|^2 \,\mathrm{d}x \leq c_\star \int_0^1 |v'|^2 \,\mathrm{d}x \quad \forall v \in V.$$

Recall:

$$a : V \times V \to \mathbb{R}, \quad a(v,w) := \int_0^1 p\, v'\, w' \, \mathrm{d}x.$$

Step 3: We show boundedness of a, i.e., that $\exists c_1 \geq 0$ such that $|a(v,w)| \leq c_1 \|v\|_V \|w\|_V \ \forall v, w \in V$.

Using that $|p(x)| = 2e^x \leq 2e \ \forall x \in [0,1]$, and using the Cauchy–Schwarz inequality, we have for any $v, w \in V$ that

$$\begin{aligned} |a(v,w)| \leq \int_0^1 |p|\,|v'|\,|w'| \, \mathrm{d}x \leq 2e \int_0^1 |v'|\,|w'| \, \mathrm{d}x &= 2e(|v'|, |w'|)_{L^2(\Omega)} \\ &\leq 2e \|v'\|_{L^2(\Omega)} \|w'\|_{L^2(\Omega)} \\ &\leq 2e \|v\|_{H^1(\Omega)} \|w\|_{H^1(\Omega)} = c_1 \|v\|_V \|w\|_V, \end{aligned}$$

where $c_1 := 2e \geq 0$ and we used that $\|v'\|_{L^2(\Omega)} \leq \|v\|_{H^1(\Omega)} \ \forall v \in H^1(\Omega)$ (and hence, in particular, $\|v'\|_{L^2(\Omega)} \leq \|v\|_{H^1(\Omega)} \ \forall v \in V$ as $V \subset H^1(\Omega)$).

Recall:

$$l : V \to \mathbb{R}, \quad l(v) := \int_0^1 f\, v \,\mathrm{d}x.$$

Step 4: Show boundedness of l, i.e., $\exists c_2 \geq 0$ s.t. $|l(v)| \leq c_2 \|v\|_V \ \forall v \in V$.

Using the Cauchy–Schwarz inequality, we have for any $v \in V$ that

$$\begin{aligned} |l(v)| = \left| \int_0^1 f\, v \,\mathrm{d}x \right| &= \left| (f, v)_{L^2(\Omega)} \right| \\ &\leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \\ &\leq \|f\|_{L^2(\Omega)} \|v\|_{H^1(\Omega)} = c_2 \|v\|_V, \end{aligned}$$

where $c_2 := \|f\|_{L^2(\Omega)} \geq 0$, and we used $\|v\|_{L^2(\Omega)} \leq \|v\|_{H^1(\Omega)} \ \forall v \in H^1(\Omega)$ (and hence, in particular, $\|v\|_{L^2(\Omega)} \leq \|v\|_{H^1(\Omega)} \ \forall v \in V$ as $V \subset H^1(\Omega)$).

Conclude: Altogether, by the Lax–Milgram theorem there exists a unique $u \in V$ such that $a(u, v) = l(v)$ for all $v \in V$, i.e., there exists a unique weak solution $u \in V$ to the given problem. □

In addition, we find that

$$c_0 \|u\|_{H^1(\Omega)}^2 \leq a(u, u) = l(u) \leq c_2 \|u\|_{H^1(\Omega)} = \|f\|_{L^2(\Omega)} \|u\|_{H^1(\Omega)},$$

i.e.,

$$\|u\|_{H^1(\Omega)} \leq \frac{1}{c_0} \|f\|_{L^2(\Omega)}.$$

The general existence and uniqueness result

Consider the homogeneous Dirichlet BVP

$$-\mathrm{div}(A\nabla u) + \mathbf{b}\cdot\nabla u + cu = f \qquad \text{in } \Omega, \tag{5}$$

$$u = 0 \qquad \text{on } \partial\Omega. \tag{6}$$

One can show the following existence and uniqueness result for weak solns:

Theorem (Existence and uniqueness of weak solutions)

Suppose that $a_{ij} \in C(\overline{\Omega})$ for $i, j \in \{1, \ldots, n\}$, $b_i \in C^1(\overline{\Omega})$ for $i \in \{1, \ldots, n\}$, $c \in C(\overline{\Omega})$, $f \in L^2(\Omega)$. Let $A : \overline{\Omega} \to \mathbb{R}^{n\times n}$, $A(x) = (a_{ij}(x))_{1\le i,j\le n}$, and $\mathbf{b} : \overline{\Omega} \to \mathbb{R}^n$, $\mathbf{b}(x) = (b_1(x), \ldots, b_n(x))^{\mathrm{T}}$. Assume that the uniform ellipticity condition is satisfied, and assume that $c - \frac{1}{2}\mathrm{div}(\mathbf{b}) \ge 0$ in $\overline{\Omega}$. Then, the BVP (5)–(6) possesses a unique weak solution $u \in H_0^1(\Omega)$. In addition, $\exists\, c_0 > 0$ s.t.

$$\|u\|_{H^1(\Omega)} \le \frac{1}{c_0}\|f\|_{L^2(\Omega)}.$$

Stability of the solution w.r.t. perturbations in f

Consider the problem

$$\begin{aligned} -\mathrm{div}(A\nabla u) + \mathbf{b}\cdot\nabla u + cu &= f \qquad \text{in } \Omega,\\ u &= 0 \qquad \text{on } \partial\Omega, \end{aligned}$$

and suppose that the assumptions of the existence and uniqueness thm hold. Suppose $f_1, f_2 \in L^2(\Omega)$ are two different right-hand sides, with corresponding solutions $u_1, u_2 \in H_0^1(\Omega)$, and consider the problem

$$\begin{aligned} -\mathrm{div}(A\nabla \tilde{u}) + \mathbf{b}\cdot\nabla \tilde{u} + c\tilde{u} &= \tilde{f} \qquad \text{in } \Omega,\\ \tilde{u} &= 0 \qquad \text{on } \partial\Omega, \end{aligned}$$

where $\tilde{f} := f_1 - f_2 \in L^2(\Omega)$. By the existence and uniqueness thm, there exists a unique weak soln $\tilde{u} \in H_0^1(\Omega)$ to this problem, and there holds $\|\tilde{u}\|_{H^1(\Omega)} \leq \frac{1}{c_0}\|\tilde{f}\|_{L^2(\Omega)}$. Observing that $\tilde{u} = u_1 - u_2$, we find that

$$\|u_1 - u_2\|_{H^1(\Omega)} \leq \frac{1}{c_0}\|f_1 - f_2\|_{L^2(\Omega)}.$$

$\Longrightarrow$ "small" changes in f give rise to "small" changes in the corresponding solution u.

The maximum principle

We consider the BVP

$$
\begin{aligned}
-\Delta u &= f \qquad \text{in } \Omega, \\
u &= g \qquad \text{on } \partial\Omega,
\end{aligned}
$$

where $\Omega \subset \mathbb{R}^n$ is a bounded open set, $f \in C(\Omega)$ and $g \in C(\partial\Omega)$.

Theorem (Maximum principle)

Suppose $f(x) \leq 0 \ \forall x \in \Omega$, and $u \in C^2(\Omega) \cap C(\overline{\Omega})$ is a classical soln to the above BVP. Then,

$$
\max_{x \in \overline{\Omega}} u(x) = \max_{x \in \partial\Omega} u(x),
$$

i.e., the maximum value of u over $\overline{\Omega}$ is attained on $\partial\Omega$.

Proof of the maximum principle

Let $\Omega \subset \mathbb{R}^n$ bounded and open, $f \in C(\Omega)$ with $f(x) \leq 0$ for all $x \in \Omega$, $g \in C(\partial\Omega)$, and $u \in C^2(\Omega) \cap C(\overline{\Omega})$ a classical soln to the BVP

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = g \quad \text{on } \partial\Omega.$$

Claim: $\max_{x\in\overline{\Omega}} u(x) = \max_{x\in\partial\Omega} u(x)$.
Proof: *Step 1: First, suppose that $f < 0$ in Ω.* Suppose that u attains its maximum value at some interior point $x_0 \in \Omega$. Then,

$$\partial_{x_i} u(x_0) = 0, \quad \partial^2_{x_i x_i} u(x_0) \leq 0 \qquad \forall i \in \{1, \dots, n\}.$$

Hence, $-\Delta u(x_0) = -\sum_{i=1}^n \partial^2_{x_i x_i} u(x_0) \geq 0$; contradicting $f < 0$. Thus,

$$\max_{x\in\overline{\Omega}} u(x) = \max_{x\in\partial\Omega} u(x).$$

Proof of the maximum principle

Let $\Omega \subset \mathbb{R}^n$ bounded and open, $f \in C(\Omega)$ with $f(x) \leq 0$ for all $x \in \Omega$, $g \in C(\partial\Omega)$, and $u \in C^2(\Omega) \cap C(\overline{\Omega})$ a classical soln to the BVP

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = g \quad \text{on } \partial\Omega.$$

Claim: $\max_{x\in\overline{\Omega}} u(x) = \max_{x\in\partial\Omega} u(x)$.

Proof: *Step 2: Now, suppose only that $f \leq 0$ in Ω.* We define the fct $v \in C^2(\Omega) \cap C(\overline{\Omega})$ given by $v(x) := u(x) + \frac{\varepsilon}{2n}|x|^2$, where $\varepsilon > 0$. Then,

$$-\Delta v(x) = -\Delta u(x) - \varepsilon = f(x) - \varepsilon < 0 \qquad \forall x \in \Omega.$$

$\implies$ By Step 1, $\max_{x\in\overline{\Omega}} v(x) = \max_{x\in\partial\Omega} v(x)$. Consequently,

$$\begin{aligned}
\max_{x\in\partial\Omega} u(x) &= \max_{x\in\partial\Omega}\left[v(x) - \frac{\varepsilon}{2n}|x|^2\right] \\
&\geq \max_{x\in\partial\Omega} v(x) - \max_{x\in\partial\Omega}\left[\frac{\varepsilon}{2n}|x|^2\right] = \max_{x\in\overline{\Omega}} v(x) - \frac{\varepsilon}{2n}\max_{x\in\partial\Omega}|x|^2 \\
&\geq \max_{x\in\overline{\Omega}} u(x) - \frac{\varepsilon}{2n}\max_{x\in\partial\Omega}|x|^2.
\end{aligned}$$

$\varepsilon \searrow 0$: $\max_{x\in\partial\Omega} u(x) \geq \max_{x\in\overline{\Omega}} u(x)$. Converse inequality trivial. $\square$

The minimum principle

We consider the BVP

$$\begin{aligned} -\Delta u = f \qquad & \text{in } \Omega, \\ u = g \qquad & \text{on } \partial\Omega, \end{aligned}$$

where $\Omega \subset \mathbb{R}^n$ is a bounded open set, $f \in C(\Omega)$ and $g \in C(\partial\Omega)$.

Theorem (Minimum principle)

Suppose $f(x) \geq 0\ \forall x \in \Omega$, and $u \in C^2(\Omega) \cap C(\overline{\Omega})$ is a classical soln to the above BVP. Then,

$$\min_{x \in \overline{\Omega}} u(x) = \min_{x \in \partial\Omega} u(x),$$

i.e., the minimum value of u over $\overline{\Omega}$ is attained on $\partial\Omega$.

Proof: The fct $\tilde{u} := -u \in C^2(\Omega) \cap C(\overline{\Omega})$ is a classical soln to

$$-\Delta \tilde{u} = -f \quad \text{in } \Omega, \qquad \tilde{u} = -g \quad \text{on } \partial\Omega.$$

$-f \leq 0 \implies_{\text{(Max.P.)}} \max_{\overline{\Omega}} \tilde{u} = \max_{\partial\Omega} \tilde{u}$, i.e., $-\min_{\overline{\Omega}} u = -\min_{\partial\Omega} u$. $\square$

6.2 Methodology of FD schemes

Motivation

Let $\Omega \subset \mathbb{R}^n$ bounded and open. Suppose we wish to solve the BVP

$$\begin{aligned} \mathscr{L}u &= f && \text{in } \Omega, \\ \mathscr{B}u &= g && \text{on } \Gamma := \partial\Omega, \end{aligned}$$

where $\mathscr{L} : u \mapsto \mathscr{L}u$ is a linear partial differential operator, and $\mathscr{B} : u \mapsto \mathscr{B}u$ is a linear operator which specifies the b.c.. For example,

$$\mathscr{L}u := -\mathrm{div}(A\nabla u) + \mathbf{b} \cdot \nabla u + cu,$$

and $\mathscr{B}u := u$ (Dirichlet b.c.), or $\mathscr{B}u := \partial_\nu u$ (Neumann b.c.).

In general, not possible to determine the true soln in closed form. Thus, the goal is to describe a simple and general numerical technique for the approximate soln of the BVP, called the **finite difference (FD) method**.

Methodology of FD schemes

The construction of a FD scheme consists of two basic steps:

1) the computational domain is approximated by a finite set of points, called the FD mesh,
2) the derivatives appearing in the PDE (and possibly also in the b.c.) are approximated by divided differences on the FD mesh.

1) Approximate $\overline{\Omega} = \Omega \cup \Gamma$ (where $\Gamma := \partial\Omega$) by a finite set of points

$$\overline{\Omega}_h = \Omega_h \cup \Gamma_h,$$

where $\Omega_h \subset \Omega$ and $\Gamma_h \subset \Gamma$. We call $\overline{\Omega}_h$ the **mesh**, Ω_h the **set of interior mesh-points** and Γ_h the **set of boundary mesh-points**. The parameter $h = (h_1, \dots, h_n)$ measures "fineness" of mesh (h_i denotes mesh-size in direction x_i): the smaller $\max_{1 \le i \le n} h_i$, the finer the mesh.

2) Replacing derivatives by divided differences yields FD scheme

$$\mathscr{L}_h U(x) = f_h(x) \quad , x \in \Omega_h, \qquad \mathscr{B}_h U(x) = g_h(x) \quad , x \in \Gamma_h,$$

where f_h and g_h are suitable approximations of f and g. This is a system of linear algebraic equations involving the values of U at the mesh-points. The values $\{U(x) : x \in \overline{\Omega}_h\}$ are approximations to $\{u(x) : x \in \overline{\Omega}_h\}$.

Two classes of problems associated with FD schemes

- The first, and more fundamental, is the problem of approximation. That is, whether the FD scheme approximates the BVP in some sense, and whether its solution $\{U(x) : x \in \overline{\Omega}_h\}$ approximates $\{u(x) : x \in \overline{\Omega}_h\}$, the values of the exact solution at the mesh-points.

- The second problem concerns the effective solution of the discrete problem (the resulting linear system) using techniques from numerical linear algebra.

In this course, our focus is on the first of these two problems.
(See MA4230 for the second problem.)

6.3 FD approximation of a two-point BVP

The BVP and the mesh

Let $\Omega := (0,1)$. We consider the BVP:

$$-u'' + cu = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega,$$

where $f, c \in C(\overline{\Omega})$ and $c(x) \geq 0 \ \forall x \in \overline{\Omega}$. This problem has a unique weak solution $u \in H_0^1(\Omega)$. We make the assumption that $u \in C^4(\overline{\Omega})$.

First, define the mesh: Let $N \in \mathbb{N}_{\geq 2}$ and set $h := \frac{1}{N}$ (mesh-size). The mesh-points are $x_i := ih$, $i \in \{0, \dots, N\}$.

Define the set of interior mesh-points

$$\Omega_h := \{x_1, \dots, x_{N-1}\},$$

the set of boundary mesh-points

$$\Gamma_h := \{x_0, x_N\},$$

and the mesh, i.e., the set of all mesh-points,

$$\overline{\Omega}_h := \Omega_h \cup \Gamma_h.$$

Divided difference operators

Using Taylor expansion, we see that

$$u(x_{i\pm 1}) = u(x_i \pm h) = u(x_i) \pm h u'(x_i) + \frac{h^2}{2} u''(x_i) \pm \frac{h^3}{6} u'''(x_i) + \mathcal{O}(h^4).$$

For the approximation of $u'(x_i)$ we introduce the **first divided difference operators** D_x^+, D_x^-, and $D_x^0 := \frac{1}{2} D_x^+ + \frac{1}{2} D_x^-$ given by

$$D_x^+ u(x_i) := \frac{u(x_{i+1}) - u(x_i)}{h} \quad = u'(x_i) + \mathcal{O}(h) \quad \textbf{(forward first d.d.o.)},$$

$$D_x^- u(x_i) := \frac{u(x_i) - u(x_{i-1})}{h} \quad = u'(x_i) + \mathcal{O}(h) \quad \textbf{(backward first d.d.o.)},$$

$$D_x^0 u(x_i) := \frac{u(x_{i+1}) - u(x_{i-1})}{2h} = u'(x_i) + \mathcal{O}(h^2) \quad \textbf{(central first d.d.o.)}.$$

For the approximation of $u''(x_i)$ we use the **(symmetric) second divided difference operator** $D_x^+ D_x^-$ $(= D_x^- D_x^+)$. Note that

$$\begin{aligned} D_x^+ D_x^- u(x_i) = \frac{D_x^- u(x_{i+1}) - D_x^- u(x_i)}{h} &= \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} \\ &= u''(x_i) + \mathcal{O}(h^2). \end{aligned}$$

The FD scheme

We replace the second derivative u'' in the DE at a mesh point x_i by the second divided difference $D_x^+ D_x^- u(x_i)$:

$$
\begin{aligned}
-D_x^+ D_x^- u(x_i) + c(x_i)u(x_i) \approx f(x_i), \quad i \in \{1, \ldots, N-1\}, \\
u(x_0) = u(x_N) = 0.
\end{aligned}
$$

$\Longrightarrow$ the approximate solution U should be sought as the solution of the following system of difference equations:

$$
\begin{aligned}
-D_x^+ D_x^- U_i + c(x_i)U_i = f(x_i), \quad i \in \{1, \ldots, N-1\}, \\
U_0 = U_N = 0,
\end{aligned}
$$

or equivalently,

$$
\begin{aligned}
-\frac{U_{i+1} - 2U_i + U_{i-1}}{h^2} + c(x_i)U_i = f(x_i), \quad i \in \{1, \ldots, N-1\}, \\
U_0 = U_N = 0.
\end{aligned}
$$

The values $U_0, U_1, \ldots, U_{N-1}, U_N$ obtained from this are our approximations to the values of the true soln at $x_0, x_1, \ldots, x_{N-1}, x_N$.

The FD scheme as linear system $AU = F$

The FD scheme

$$-\frac{U_{i+1} - 2U_i + U_{i-1}}{h^2} + c(x_i)U_i = f(x_i), \quad i \in \{1, \dots, N-1\},$$

$$U_0 = U_N = 0$$

can be written as

$$\underbrace{\begin{bmatrix} \frac{2}{h^2} + c(x_1) & -\frac{1}{h^2} & & & \mathbf{0} \\ -\frac{1}{h^2} & \frac{2}{h^2} + c(x_2) & -\frac{1}{h^2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{1}{h^2} & \frac{2}{h^2} + c(x_{N-2}) & -\frac{1}{h^2} \\ \mathbf{0} & & & -\frac{1}{h^2} & \frac{2}{h^2} + c(x_{N-1}) \end{bmatrix}}_{=:A} \underbrace{\begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_{N-2} \\ U_{N-1} \end{bmatrix}}_{=:U} = \underbrace{\begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{N-2}) \\ f(x_{N-1}) \end{bmatrix}}_{=:F}$$

$\Longrightarrow$ Solving the FD scheme is equivalent to solving the linear system

$$AU = F.$$

1. (ExUn) Existence & uniqueness of solns to FD scheme

Observation: The FD scheme has a unique solution U iff the matrix

$$A := \begin{bmatrix} \frac{2}{h^2} + c(x_1) & -\frac{1}{h^2} & & & \mathbf{0} \\ -\frac{1}{h^2} & \frac{2}{h^2} + c(x_2) & -\frac{1}{h^2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{1}{h^2} & \frac{2}{h^2} + c(x_{N-2}) & -\frac{1}{h^2} \\ \mathbf{0} & & & -\frac{1}{h^2} & \frac{2}{h^2} + c(x_{N-1}) \end{bmatrix} \in \mathbb{R}^{(N-1)\times(N-1)}$$

is invertible.

We are going to prove that A is indeed invertible. Recall $c(x) \geq 0\ \forall x \in \overline{\Omega}$.
Note: If we had that $c(x) > 0\ \forall x \in \overline{\Omega}$, this would be very simple:

Rk: If $c(x) > 0\ \forall x \in \overline{\Omega}$, then A is strictly diagonally dominant, i.e.,

$$|a_{ii}| > \sum_{j \in \{1,\dots,N-1\} \setminus \{i\}} |a_{ij}| \qquad \forall i \in \{1, \dots, N-1\},$$

and hence, A is invertible.

1. (ExUn) Proof of invertibility of A: the idea

Observe: A invertible iff the only soln to $AV = 0$ is $V = 0 \in \mathbb{R}^{N-1}$.

The argument which we develop is based on mimicking, at the discrete level, the following procedure based on integration-by-parts: (recall $u(0) = u(1) = 0$ and $c(x) \geq 0\ \forall x \in [0,1]$)

$$\int_0^1 (-u''(x) + c(x)u(x))\, u(x)\, \mathrm{d}x = \int_0^1 |u'(x)|^2\, \mathrm{d}x + \int_0^1 c(x)|u(x)|^2\, \mathrm{d}x$$
$$\geq \int_0^1 |u'(x)|^2\, \mathrm{d}x.$$

Thus, if $-u'' + cu = 0$, then $u' = 0$ which gives $u = 0$ (by b.c.).

For two functions V and W defined at the interior mesh-points $x_1, \ldots, x_{N-1}$, we define the inner product

$$(V, W)_h := \sum_{i=1}^{N-1} hV_iW_i,$$

resembling the L^2-inner product $(v, w)_{L^2((0,1))} := \int_0^1 v(x)w(x)\, \mathrm{d}x$.

1. (ExUn) Proof of invertibility of A: the key tool

Our key technical tool is the following summation-by-parts identity, which is the discrete counterpart of the integration-by-parts identity $(-u'', u)_{L^2((0,1))} = (u', u')_{L^2((0,1))} = \|u'\|^2_{L^2((0,1))}$ satisfied by the fct u, obeying the homogeneous b.c. $u(0) = u(1) = 0$.

Lemma (summation-by-parts)

Suppose that V is a function defined at the mesh-points x_i, $i \in \{0, \dots, N\}$, and $V_0 = V_N = 0$. Then, there holds

$$(-D_x^+ D_x^- V, V)_h = \sum_{i=1}^{N} h |D_x^- V_i|^2.$$

(Recall $(V, W)_h := \sum_{i=1}^{N-1} h V_i W_i$.)

1. (ExUn) Proof of invertibility of A: the key tool

Lemma (summation-by-parts)

Suppose that V is a function defined at the mesh-points x_i, $i \in \{0, \dots, N\}$, and $V_0 = V_N = 0$. Then, there holds

$$(-D_x^+ D_x^- V, V)_h = \sum_{i=1}^{N} h |D_x^- V_i|^2.$$

Proof: We have that

$$\begin{aligned} -\sum_{i=1}^{N-1} h\,(D_x^+ D_x^- V_i) V_i &= -\sum_{i=1}^{N-1} \frac{V_{i+1} - V_i}{h} V_i + \sum_{i=1}^{N-1} \frac{V_i - V_{i-1}}{h} V_i \\ &= -\sum_{i=1}^{N} \frac{V_i - V_{i-1}}{h} V_{i-1} + \sum_{i=1}^{N} \frac{V_i - V_{i-1}}{h} V_i \\ &= \sum_{i=1}^{N} \frac{V_i - V_{i-1}}{h} (V_i - V_{i-1}) = \sum_{i=1}^{N} h |D_x^- V_i|^2. \end{aligned}$$

1. (ExUn) Proof of invertibility of A

Let $V = (V_1, \ldots, V_{N-1})^{\mathrm{T}} \in \mathbb{R}^{N-1}$ be s.t. $AV = 0$. We prove that $V = 0$.

We set $V_0 := V_N := 0$. Then, by summation-by-parts, and using that $c(x) \geq 0 \ \forall x \in \overline{\Omega}$, we have

$$
\begin{aligned}
0 = \underbrace{(AV, V)_h}_{:=\sum_{i=1}^{N-1} h(AV)_i V_i} &= \underbrace{(-D_x^+ D_x^- V + cV, V)_h}_{:=\sum_{i=1}^{N-1} h(-D_x^+ D_x^- V_i + c(x_i) V_i) V_i} \\
&= (-D_x^+ D_x^- V, V)_h + (cV, V)_h \\
&= \sum_{i=1}^{N} h|D_x^- V_i|^2 + \sum_{i=1}^{N-1} h\, c(x_i)|V_i|^2 \geq \sum_{i=1}^{N} h|D_x^- V_i|^2.
\end{aligned}
$$

$\Longrightarrow D_x^- V_i = \frac{V_i - V_{i-1}}{h} = 0 \ \forall i \in \{1, \ldots, N\}$ and thus, $V = 0$ (as $V_0 = V_N = 0$).
It follows that A is invertible. $\square$

Thus, the FD scheme has a unique solution:

$$U = A^{-1} F.$$

2. (Stab) Stability of the FD scheme

Goal: Prove a discrete version of the stability bound

$$\|u\|_{H^1(\Omega)} \leq \frac{1}{c_0}\|f\|_{L^2(\Omega)}.$$

Recall pf:

$c_0\|u\|^2_{H^1(\Omega)} \leq a(u,u) = (f,u)_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)}\|u\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)}\|u\|_{H^1(\Omega)}.$

Define the **discrete L^2-norm** $\|\cdot\|_h$ and the **discrete H^1-norm** $\|\cdot\|_{1,h}$ by

$$\|V\|_h := \sqrt{(V,V)_h} = \sqrt{\sum_{i=1}^{N-1} h|V_i|^2}, \qquad \|V\|_{1,h} := \sqrt{\|V\|_h^2 + \|D_x^- V]|_h^2},$$

where $\|V]|_h := \sqrt{(V,V]_h} = \sqrt{\sum_{i=1}^{N} h|V_i|^2}$ with $(V,W]_h := \sum_{i=1}^{N} hV_iW_i$.
Using this notation, we have shown on the previous slide that

$$(AU,U)_h \geq \|D_x^- U]|_h^2.$$

2. (Stab) Proof of stability of the FD scheme

Lemma (Discrete Poincaré–Friedrichs inequality)

Let V be a fct defined on the FD mesh $\{x_i := ih \,:\, i \in \{0, \dots, N\}\}$, where $h := \frac{1}{N}$ and $N \in \mathbb{N}_{\geq 2}$, and such that $V_0 = V_N = 0$. Then, $\exists$ a constant $c_\star > 0$, independent of V and h, s.t., for all such V,

$$\|V\|_h^2 \leq c_\star \|D_x^- V]|_h^2.$$

Rk: The constant $c_\star$ can be taken to be $c_\star = \frac{1}{2}$.

$\Longrightarrow$ $\|U\|_h^2 \leq \frac{1}{2}\|D_x^- U]|_h^2$. Using $(AU, U)_h \geq \|D_x^- U]|_h^2$, we find

$$\frac{2}{3}\|U\|_{1,h}^2 = \frac{2}{3}\|U\|_h^2 + \frac{2}{3}\|D_x^- U]|_h^2 \leq \|D_x^- U]|_h^2 \leq (AU, U)_h = (f, U)_h.$$

Noting that $(f, U)_h \leq \|f\|_h \|U\|_h \leq \|f\|_h \|U\|_{1,h}$, we proved that the FD scheme is stable with stability bound

$$\|U\|_{1,h} \leq \frac{3}{2}\|f\|_h.$$

3. (Conv) Convergence of the FD scheme

We define the **global error** e by

$$e_i := u(x_i) - U_i, \quad i \in \{0, \ldots, N\}.$$

Note that $e_0 = e_N = 0$. For $i \in \{1, \ldots, N-1\}$, we have

$$\begin{aligned} -D_x^+ D_x^- e_i + c(x_i) e_i &= -D_x^+ D_x^- u(x_i) + c(x_i) u(x_i) - f(x_i) \\ &= -D_x^+ D_x^- u(x_i) + c(x_i) u(x_i) - \big(-u''(x_i) + c(x_i) u(x_i)\big) \\ &= u''(x_i) - D_x^+ D_x^- u(x_i). \end{aligned}$$

Thus,

$$-D_x^+ D_x^- e_i + c(x_i) e_i = \varphi_i, \quad i \in \{1, \ldots, N-1\}, \qquad e_0 = e_N = 0,$$

where

$$\varphi_i := -D_x^+ D_x^- u(x_i) + c(x_i) u(x_i) - f(x_i) = u''(x_i) - D_x^+ D_x^- u(x_i)$$

is the **consistency error** (or **truncation error**). By the stability bound,

$$\|u - U\|_{1,h} = \|e\|_{1,h} \leq \frac{3}{2} \|\varphi\|_h.$$

$\Longrightarrow$ It remains to estimate the term $\|\varphi\|_h$.

3. (Conv) Convergence of FD scheme: Pf of error bound

Taylor expansion yields

$$\varphi_i = u''(x_i) - D_x^+ D_x^- u(x_i) = u''(x_i) - \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} = -\frac{h^2}{12} u^{(4)}(\xi_i)$$

for some $\xi_i \in (x_{i-1}, x_{i+1})$. Thus, $|\varphi_i| \leq \frac{h^2}{12} \|u^{(4)}\|_{C([0,1])}$. Thus,

$$\|\varphi\|_h = \sqrt{\sum_{i=1}^{N-1} h|\varphi_i|^2} \leq \frac{h^2}{12} \|u^{(4)}\|_{C([0,1])} \sqrt{\sum_{i=1}^{N-1} h} \leq \frac{h^2}{12} \|u^{(4)}\|_{C([0,1])}.$$

(Note $(N-1)h \leq Nh = 1$.) Combining with $\|u - U\|_{1,h} \leq \frac{3}{2} \|\varphi\|_h$, we have proved the following convergence theorem/error bound:

Theorem (Convergence of the soln U of the FD scheme to the true soln u)

Let $f, c \in C([0,1])$ with $c(x) \geq 0$ $\forall x \in [0,1]$, and suppose that the unique weak soln $u \in H_0^1((0,1))$ to the BVP satisfies $u \in C^4([0,1])$. Then,

$$\|u - U\|_{1,h} \leq \frac{h^2}{8} \|u^{(4)}\|_{C([0,1])} = \mathcal{O}(h^2).$$

6.4 Key steps of a general error analysis for FD approximations of elliptic PDEs

General error analysis of FD schemes for elliptic PDEs

Consider the general FD scheme

$$\mathscr{L}_h U(x) = f_h(x) \quad , x \in \Omega_h, \qquad \mathscr{B}_h U(x) = g_h(x) \quad , x \in \Gamma_h$$

for the numerical soln of the BVP $\mathscr{L}u = f$ in Ω, $\mathscr{B}u = g$ on $\partial\Omega$.

Step 1: Prove stability of scheme in appropriate mesh-dependent norm:

$$|||U|||_{\Omega_h} \le C_1(\|f_h\|_{\Omega_h} + \|g_h\|_{\Gamma_h}),$$

where $|||\cdot|||_{\Omega_h}$, $\|\cdot\|_{\Omega_h}$, $\|\cdot\|_{\Gamma_h}$ are mesh-dependent norms involving mesh-points of Ω_h (or $\overline{\Omega}_h$) and Γ_h, and $C_1 > 0$ is a constant indep. of h.

Step 2: Estimate the size of the **consistency error**,

$$\varphi_{\Omega_h} := \mathscr{L}_h u - f_h \quad \text{in } \Omega_h, \qquad \varphi_{\Gamma_h} := \mathscr{B}_h u - g_h \quad \text{on } \Gamma_h.$$

If $\|\varphi_{\Omega_h}\|_{\Omega_h} + \|\varphi_{\Gamma_h}\|_{\Gamma_h} \to 0$ as $h \to 0$ for a sufficiently smooth soln u of the BVP, we say that the FD scheme is **consistent**. If $p \in \mathbb{N}$ is the largest natural number such that, for all sufficiently smooth u,

$$\|\varphi_{\Omega_h}\|_{\Omega_h} + \|\varphi_{\Gamma_h}\|_{\Gamma_h} = \mathcal{O}(h^p),$$

the scheme has **order of accuracy** (or **order of consistency**) p.

The FD scheme is said to be **convergent** in the norm $|||\cdot|||_{\Omega_h}$, if

$$|||u-U|||_{\Omega_h} \longrightarrow 0 \qquad \text{as } h \to 0.$$

If $q \in \mathbb{N}$ is the largest natural number such that $|||u-U|||_{\Omega_h} = \mathcal{O}(h^q)$ as $h \to 0$, then the scheme is said to have **order of convergence** q.

Theorem (Stability + Consistency $\Longrightarrow$ Convergence)

If the FD scheme is stable, i.e., $|||U_{f_h,g_h}|||_{\Omega_h} \leq C_1(\|f_h\|_{\Omega_h} + \|g_h\|_{\Gamma_h})$ for all f_h, g_h, where U_{f_h,g_h} denotes the corresponding soln of the FD scheme, and consistent, i.e., $\|\varphi_{\Omega_h}\|_{\Omega_h} + \|\varphi_{\Gamma_h}\|_{\Gamma_h} \to 0$ as $h \to 0$, then it is convergent, and its order of convergence $q \geq$ its order of accuracy p.

Proof: Define the **global error** $e := u - U$. Then, as $\mathscr{L}_h$ is linear, we have

$$\mathscr{L}_h e = \mathscr{L}_h(u-U) = \mathscr{L}_h u - \mathscr{L}_h U = \mathscr{L}_h u - f_h = \varphi_{\Omega_h}$$

in Ω_h. Similarly, as $\mathscr{B}_h$ is linear, we have $\mathscr{B}_h e = \varphi_{\Gamma_h}$ on Γ_h. Since the FD scheme is assumed to be stable, it then follows that

$$|||e|||_{\Omega_h} \leq C_1(\|\varphi_{\Omega_h}\|_{\Omega_h} + \|\varphi_{\Gamma_h}\|_{\Gamma_h}) \longrightarrow 0$$

as $h \to 0$. Further, if $\|\varphi_{\Omega_h}\|_{\Omega_h} + \|\varphi_{\Gamma_h}\|_{\Gamma_h} = \mathcal{O}(h^p)$, then $|||e|||_{\Omega_h} = \mathcal{O}(h^p)$, which shows that the order of convergence q is at least p.

End of “Chapter 6: Introduction to the theory of FD schemes”.