Class Notes for CAAM452

Beatrice Riviere *
INTRODUCTION

Scientists and engineers try to understand physical phenomena by developing models and by obtaining solutions to those models. Often, the mathematical models are characterized by a system of coupled nonlinear partial differential equations. As it is impossible in many cases to obtain analytical forms of the solutions to those PDEs, the main tool for obtaining those solutions is the use of numerical methods. This process is done in several steps, described below.

Step one: the starting point is a system of PDEs. We will consider the following:

\[-\Delta u = f\]  \hspace{1cm} (1)
\[\frac{\partial u}{\partial t} - \Delta u = f\]  \hspace{1cm} (2)

Equation (1) is an elliptic equation, also called diffusion equation. The exact solution \( u \) only depends on the spatial variable \( x \in \mathbb{R}^n \). Equation (2) is a parabolic equation, also referred to as the heat equation. The exact solution depends on both space \( x \) and time \( t \).

Step two: The computational domain (denoted by \( \Omega \)) is partitioned into a grid (or mesh) of cells (or elements) of a given size (usually denoted by \( h \)). In 1D, we will have a partition \( x_0 < x_1 < \cdots < x_N \) with \( x_i = ih \) for a uniform grid. In 2D, the cells can be triangles or quadrilaterals. In 3D, the elements can be prisms, tetrahedra, hexahedra... For time-dependent problem (such as the heat equation above), the time interval \([0,T]\) is also discretized, simply by \( t^i = i\Delta t \), where \( \Delta t \) measures the fineness of the time partition. As \( h \) goes to zero, the number of cells increases and similarly as \( \Delta t \) goes to zero, the number of time steps increases. Overall, the size of the numerical problem to solve increases as both \( h, \Delta t \) decrease.

Step three: Using a given numerical method, we obtain a linear system of the form \( AU_{h,\Delta t} = b \) where the components of the vector \( U_{h,\Delta t} \) approximate the exact solution \( u \) in some way. A good numerical method is such that

\[
\lim_{h \to 0, \Delta t \to 0} \|U_{h,\Delta t} - u\| = 0
\]

in some norm. We say that the numerical error decreases and the numerical method converges.

Step four: Using numerical linear algebra, we solve the linear system \( AU_{h,\Delta t} = b \) as efficiently as possible. For instance, one can use LU factorization, or conjugate gradient, or GMRES... In this class, we will simply use the Matlab command \( U_{h,\Delta t} = A \backslash b \). Another possibility is to use Lapack, or routines from the package Petsc.

Step five: Finally, we postprocess the vector solution \( U_{h,\Delta t} \) to obtain the quantity of interest. For instance, we want to plot the numerical solution, or we might want to compute its average over the whole domain.
During the semester we will see different numerical methods, focusing on finite difference and finite elements. We will compare those methods by pointing out their advantages/disadvantages. In particular, we will evaluate the cost versus accuracy property.
I. FINITE DIFFERENCE APPROXIMATIONS

An important tool is the Taylor series of a function $u$ about a point $x$. Assume that $u \in C^{k+1}(a,b)$ and let $h > 0$.

$$u(x + h) = u(x) + h u'(x) + \frac{h^2}{2} u''(x) + \cdots + \frac{h^k}{k!} u^{(k)}(x) + \frac{h^{k+1}}{(k+1)!} u^{(k+1)}(\xi)$$

for some $\xi \in (x, x + h)$.

Using the Taylor series, we can approximate derivatives by finite differences.

1. First-order derivatives

A one-sided approximation uses the function $u$ evaluated at $x$ and $x + h$:

$$D_+ u(x) = \frac{u(x + h) - u(x)}{h}$$

If we use the function evaluated at $x$ and $x - h$, we obtain another one-sided approximation:

$$D_- u(x) = \frac{u(x) - u(x - h)}{h}$$

We can evaluate the function at both sides of $x$, and we obtain the centered approximation:

$$D_0 u(x) = \frac{u(x + h) - u(x - h)}{2h}$$

The quantities $D_+ u(x), D_- u(x), D_0 u(x)$ all approximate the derivative $u'(x)$.

How accurate are those approximations? We need to compute the errors $D_+ u(x) - u'(x), D_- u(x) - u'(x)$ and $D_0 u(x) - u'(x)$. This is done by using Taylor series.

$$u(x + h) = u(x) + h u'(x) + \frac{h^2}{2} u''(\xi)$$

Equivalently, we have:

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

The error goes to zero with $h$. We say that $D_+$ is a first-order finite difference approximation of $u'(x)$. Similarly, we can show that $D_- u(x) - u'(x) = \mathcal{O}(h)$. For $D_0 u(x)$, we write two Taylor series:

$$u(x + h) = u(x) + h u'(x) + \frac{h^2}{2} u''(x) + \frac{h^3}{3!} u'''(\xi_1)$$

$$u(x - h) = u(x) - h u'(x) + \frac{h^2}{2} u''(x) - \frac{h^3}{3!} u'''(\xi_2)$$
So, we have

\[ D_0u(x) - u'(x) = \frac{h^2}{3!}(u'''(\xi_1) + u'''(\xi_2)) = \mathcal{O}(h^2) \]

The centered finite difference approximation is of second order, it is more accurate than the one-sided approximations.

2. Second-order and higher order derivatives

Consider the centered finite difference approximation:

\[ D^2u(x) = \frac{u(x-h) - 2u(x) + u(x+h)}{h^2} \]

One can show that

\[ D^2u(x) - u''(x) = \frac{h^2}{12}u^{(4)}(x) + \mathcal{O}(h^4) \]

We also remark that:

\[ D^2u = D_+D_-u \]

So to obtain a second order finite difference of \( u''(x) \) we have applied twice a first order finite difference of \( u'(x) \).

Can we always obtain a high order finite difference by applying several times a lower order finite difference? The answer is no, as shown by the following example.

**Ex 1:** let us compute \( D_+D^2u \), which is a combination of one first order and one second order approximations. Let \( f(x) = D^2u(x) \).

\[
D_+f = \frac{1}{h}(f(x+h) - f(x)) \\
= \frac{1}{h}\left(\frac{1}{h^2}(u(x) - 2u(x+h) + u(x+2h) - u(x-h)) - \frac{1}{h^2}(u(x+h) - u(x+h))\right) \\
= \frac{1}{h^3}(3u(x) - 3u(x+h) + u(x+2h) - u(x-h)) \\
= u'''(x) + \frac{h}{2}u^{(4)}(x) + \mathcal{O}(h^2)
\]

where the last equality is obtained by writing several Taylor series. This example yields a first order finite difference approximation of \( u'''(x) \).

**Ex 2:** Let us compute \( D_0D_+D_-u \).

\[
D_0D_+D_-u(x) = \frac{1}{2h^3}(u(x+2h) - 2u(x+h) + u(x) - u(x-2h)) = u'''(x) + \frac{h^2}{4}u^{(4)}(x) + \mathcal{O}(h^4)
\]

We have obtained a second order finite difference approximation of \( u'''(x) \).

3. A general method: undetermined coefficients
Assume we want to approximate \( u^{(k)}(x) \) by a finite difference involving function evaluations at \( n \) given points: \( x_1, \ldots, x_n \).

\[
u^{(k)}(x) \approx c_1 u(x_1) + c_2 u(x_2) + \cdots + c_n u(x_n)\]

We assume that \( n \geq k + 1 \) and that

\[
\max_{1 \leq i \leq n} |x - x_i| \leq Ch
\]

for some positive constant \( C \).

\textit{Ex 3:} Assume we want

\[
u'(x) \approx c_1 u(x) + c_2 u(x - h) + c_3 u(x - 2h)
\]

Here we have 3 given points: \( x_1 = x, x_2 = x - h, x_3 = x - 2h \). We would like to set up a linear system \( Ac = b \) where the vector \( c \) has the \( c_i \)'s for components. We write the Taylor series at each point.

\[
u(x - h) = u(x) - hu'(x) + \frac{h^2}{2} u''(x) - \frac{h^3}{6} u'''(x) + \ldots
\]

\[
u(x - 2h) = u(x) - 2hu'(x) + 2h^2 u''(x) - \frac{8h^3}{6} u'''(x) + \ldots
\]

Multiply by the coefficients \( c_i \)'s and add:

\[
c_1 u(x) + c_2 u(x - h) + c_3 u(x - 2h) = (c_1 + c_2 + c_3) u(x) - h(c_2 + 2c_3) u'(x) + h^2 \left( \frac{1}{2} c_2 + 2c_3 \right) u''(x) - \frac{1}{6} (c_2 + 8c_3) h^3 u'''(x) + \mathcal{O}(h^4)
\]

We want to obtain the highest possible order approximation of \( u'(x) \). Since we have 3 unknowns, we need 3 equations.

\[
c_1 + c_2 + c_3 = 0
\]

\[
-h(c_2 + 2c_3) = 1
\]

\[
h^2 \left( \frac{1}{2} c_2 + 2c_3 \right) = 0
\]

Solving for the coefficients yields

\[
c_1 = \frac{3}{2h}, \quad c_2 = -\frac{2}{h}, \quad c_3 = \frac{1}{2h}
\]

The remaining term is \( -\frac{1}{6} (c_2 + 8c_3) h^3 u'''(x) = \frac{1}{12} h^2 u'''(x) \). So we have obtained:

\[
u'(x) = \frac{3}{2h} u(x) - \frac{2}{h} u(x - h) + \frac{1}{2h} u(x - 2h) + \mathcal{O}(h^2).
\]
It is a second order finite difference approximation.

The method of undetermined coefficients is based on the Taylor series written at each point $x_i$.

$$u(x_i) = u(x) + (x_i - x)u'(x) + \cdots + \frac{1}{j!}(x_i - x)^j u^{(j)}(x) + \cdots$$

Multiply this equation by the coefficient $c_i$ and sum over all $i$.

$$\sum_{i=1}^{n} c_i u(x_i) = (\sum_{i=1}^{n} c_i) u(x) + u'(x) (\sum_{i=1}^{n} c_i (x_i - x)) + \cdots + u^{(j)}(x) \frac{1}{j!} (\sum_{i=1}^{n} c_i (x_i - x)^j) + \cdots$$

We want this expression to be equal to $u^{(k)}(x) + \mathcal{O}(h^p)$ for $p$ as high as possible. In fact, we expect the order of the method $p \geq n - k$. We set for $0 \leq j \leq n - 1$ (to get a square system)

$$\frac{1}{j!} (\sum_{i=1}^{n} c_i (x_i - x)^j) = \begin{cases} 0 & \text{if } j \neq k \\ 1 & \text{if } j = k \end{cases}$$

Let $\xi_i = x_i - x$. The system of equations is equivalent to $Ac = b$ with

$$A = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \xi_1 & \xi_2 & \cdots & \xi_n \\ \xi_1^2 & \xi_2^2 & \cdots & \xi_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ \xi_1^{n-1} & \xi_2^{n-1} & \cdots & \xi_n^{n-1} \end{pmatrix}$$

We obtain a VanderMonde matrix, which has a large condition number for $n$ large.
II. TWO-POINT BOUNDARY VALUE PROBLEM

1. Dirichlet problem

We want to solve the second order problem:

\[ u''(x) = f(x) \quad 0 < x < 1 \]
\[ u(0) = \alpha \]
\[ u(1) = \beta \]

The domain is the unit interval \((0, 1)\) that we first discretize by considering a partition:
\[ 0 = x_0 < x_1 < \cdots < x_N < x_{N+1} = 1 \].

For simplicity, we assume the partition is uniform. Let
\[ h = \frac{1}{N+1}, \quad x_j = jh \]

For each interior node, we have:
\[ u''(x_j) = f(x_j), \quad 1 \leq j \leq N \]

Then we replace the second derivative by a finite difference approximation:

\[ u''(x_j) \approx \frac{u(x_j + h) - 2u(x_j) + u(x_j - h)}{h^2} \]

We next introduce the approximations \(U_j^h\) defined by:

\[ \frac{U_{j+1}^h - 2U_j^h + U_{j-1}^h}{h^2} = f(x_j), \quad 1 \leq j \leq N \]
\[ U_0^h = \alpha \]
\[ U_{N+1}^h = \beta \]

The finite difference solution is the vector \(U^h\) with components \(U_j^h\). We may or may not include the components \(U_0^h\) and \(U_{N+1}^h\). We claim that
\[ U_j^h \approx u(x_j) \quad \forall j \]

Clearly for \(j = 0\) and \(j = N + 1\) we have exactly \(U_j^h = u(x_j)\). We say the Dirichlet boundary conditions are enforced strongly (or exactly) in the finite difference method.

Assume that the vector \(U^h\) has components \((U_1^h, U_2^h, \ldots, U_N^h)\). We obtain a linear system

\[ A^h U^h = b^h \] (3)
where
\[
A^h = \frac{1}{h^2} \begin{pmatrix}
-2 & 1 & 0 & 0 \\
1 & -2 & 1 & 0 \\
0 & 1 & -2 & 1 \\
& & \ddots & \ddots \\
0 & 1 & -2 & 1 \\
\end{pmatrix}
\]
\[
b^h = \begin{pmatrix}
f(x_1) - \frac{\alpha}{h^2} \\
f(x_2) \\
f(x_3) \\
\vdots \\
f(x_{N-1}) \\
f(x_N) - \frac{\beta}{h^2}
\end{pmatrix}
\]

The matrix $A^h$ is symmetric, tridiagonal. It is very easy to solve for the vector of unknowns $U^h$.

Question: How accurate is the finite difference solution? We define the error $e^h$, that is the vector with components:
\[
e^h_j = u(x_j) - U^h_j, \quad 1 \leq j \leq N
\]

We wish that
\[
\lim_{h \to 0} \|e^h\| = \lim_{N \to \infty} \|e^h\| = 0
\]
for a given norm of interest (for instance $\| \cdot \|_{\infty}$ or $\| \cdot \|_2$).

Def: **Local truncation error** at each interior node:
\[
\tau^h_j = \frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2} - f(x_j), \quad 1 \leq j \leq N
\]

The local truncation error measures the error made by replacing the numerical solution by the exact solution in the finite difference method. The local truncation error only depends on the exact solution. We already know by Taylor series:
\[
\tau^h_j = u''(x_j) + \frac{h^2}{12} u^{(4)}(\xi_j) - f(x_j) = \frac{h^2}{12} u^{(4)}(\xi_j)
\]
since $u$ is the exact solution. Therefore if $\tau^h$ is the vector with components $\tau^h_j$, we have:
\[
\|\tau^h\|_{\infty} \leq \frac{h^2}{12} \|u^{(4)}\|_{\infty}
\]
This implies that
\[
\lim_{h \to 0} \|\tau^h\|_{\infty} = 0
\]
We say that the finite difference method is **consistent**. To show consistency of any finite difference method, it suffices to show that $\|\tau^h\|_{\infty} = O(h^q)$ for some $q > 0$.

We now want to relate the local truncation error with the error $e^h$. For this we introduce a vector $W$ with components $u(x_j)$, where $u$ is the exact solution. From the definition of the local truncation error, one obtain:
\[
\tau^h = A^h W - b^h
\]
Subtracting (4) from (3) yields:

\[ A^h e^h = \tau^h \]

or equivalently if the matrix \( A^h \) is invertible:

\[ e^h = (A^h)^{-1} \tau^h \]

This implies:

\[ \|e^h\| \leq \|(A^h)^{-1}\|\|\tau^h\| \quad (5) \]

Def: Let \( A^h U^h = b^h \) be the linear system obtained by applying a finite difference method to a boundary value problem. The method is stable if \( A^h \) is invertible for all \( h \) sufficiently small and if there is a constant \( C \) independent of \( h \) such that

\[ \|(A^h)^{-1}\| \leq C \]

If the method is stable, (5) implies

\[ \|e^h\| \leq C\|\tau^h\| \]

If in addition, the method is consistent, we have \( \lim_{h \to 0} \|e^h\| = 0 \), which means that the method is convergent.

Therefore we have the fundamental result:

**CONSISTENCY AND STABILITY IMPLIES CONVERGENCE.**

Clearly if the method is stable and \( \|\tau^h\|_\infty = O(h^q) \) for some \( q > 0 \), then the method is convergent of order \( q \).

Remark: proving consistency is easy and requires the use of Taylor series. Proving stability of the method can be challenging. We will give an outline of the proof for stability in \( \|\cdot\|_2 \).

We recall that for any matrix \( A \) that is symmetric, \( \|A\|_2 = \rho(A) \), where \( \rho(A) \) is the spectral radius. Namely \( \rho(A) \) is the maximum \( |\lambda| \) over all eigenvalues \( \lambda \) of \( A \). Let \( \sigma(A) \) denote the set of all eigenvalues of \( A \). Since the matrix \( (A^h)^{-1} \) is symmetric, we have:

\[ \|(A^h)^{-1}\|_2 = \rho((A^h)^{-1}) = \frac{1}{\min_{\lambda \in \sigma(A^h)} |\lambda|} \]

Therefore we need to compute the eigenvalues of \( A^h \) and check that its minimum in absolute value is bounded away from zero by a constant independent of \( h \).

Claim: the eigenvalues of \( A^h \) are

\[ \lambda_k = \frac{2}{h^2}(\cos(k\pi h) - 1), \quad k = 1, \ldots, N \]
A simple figure shows that the minimum of $|\lambda_k|$ is obtained for $k = 1$. Using an expansion of $\cos$ around 0, we have

$$\lambda_1 \approx -\pi^2$$

This implies that

$$\|(A^h)^{-1}\| \leq \pi^2$$

We conclude that the finite difference (3) is stable, and convergent of second order.

Remark: the convergence rate (or order) of the method can be verified numerically by computing the ratios

$$\frac{\log(\|e^h\|)}{\log(\|e^{h/2}\|)}$$

on successively refined grids (i.e. grids obtained by dividing $h$ by 2 successively).

Remark: One can also show that the finite difference method is stable with respect to $\|\cdot\|_{\infty}$.

2. Neumann problem

We want to solve the second order problem:

$$u''(x) = f(x) \quad 0 < x < 1$$
$$u'(0) = \sigma$$
$$u(1) = \beta$$

The boundary condition $u'(0) = \sigma$ is called a Neumann boundary condition. As in Section 1, we start by defining a grid $x_j = jh$ for some $h > 0$ (say $h = 1/(N+1)$). There are several ways to modify the finite difference method defined above to handle the Neumann boundary condition.

2.1. Using a first order one-sided finite difference approximation

We replace $u'(0) = \sigma$ by:

$$\frac{U_h^1 - U_h^0}{h} = \sigma$$

We consider the augmented vector of unknowns $\tilde{U}^h = (U_0^h, U_1^h, \ldots, U_N^h, U_{N+1}^h)$. If we use the same second order centered finite difference approximation of $u''(x_j)$ as in Section 1, we obtain a new linear system:

$$\tilde{A}^h \tilde{U}^h = \tilde{b}^h$$
with

\[
\tilde{A}^h = \frac{1}{h^2} \begin{pmatrix}
-h & h & 0 & 0 \\
1 & -2 & 1 & 0 \\
0 & 1 & -2 & 1 \\
.. & .. & .. & .. \\
0 & 1 & -2 & 1 \\
0 & 0 & h^2
\end{pmatrix},
\tilde{b}^h = \begin{pmatrix}
\sigma \\
f(x_1) \\
f(x_2) \\
. \\
f(x_{N-1}) \\
f(x_N) \\
\beta
\end{pmatrix}
\]

The matrix \( \tilde{A}^h \) is tridiagonal of size \((N + 2) \times (N + 2)\). The resulting accuracy of this method is first order only. This is expected as we have used a first order approximation of the Neumann boundary condition.

2.2. Using a second order centered finite difference approximation
We replace \( u'(0) = \sigma \) by:

\[
\frac{U^h_1 - U^h_{-1}}{2h} = \sigma
\]

where \( U^h_{-1} \) is an intermediate value, that would approximate \( u(-h) \). We need another equation to eliminate this additional unknown. We write the finite difference approximation of \( u''(x_0) = f(x_0) \):

\[
\frac{U^h_1 - 2U^h_0 + U^h_{-1}}{h^2} = f(x_0)
\]

Combining the two equations above yields:

\[
-\frac{U_0 + U_1}{h} = \sigma + \frac{h}{2}f(x_0)
\]

The resulting matrix \( \tilde{A}^h \) is the same as in Section 2.1, but the right-hand side changes:

\[
\tilde{b}^h = \begin{pmatrix}
\sigma + \frac{h}{2}f(x_0) \\
f(x_1) \\
f(x_2) \\
. \\
f(x_{N-1}) \\
f(x_N) \\
\beta
\end{pmatrix}
\]

This simple change is enough to obtain a second order method.

2.3. Using a second order one-sided finite difference approximation
This technique is preferred over the previous one as it is the easiest to generalize for higher order finite difference methods. One replace 
\[
\frac{-\frac{3}{2} U_0^h + 2 U_1^h - \frac{1}{2} U_2^h}{h} = \sigma
\]

The linear system becomes:
\[
\tilde{A}^h = \frac{1}{h^2} \begin{pmatrix}
-\frac{3}{2} h & -2 h & -\frac{1}{2} h & 0 \\
1 & -2 & 1 & 0 & 0 \\
0 & 1 & -2 & 1 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 1 & -2 & 1 & 0 & 0 \\
\end{pmatrix}
\tilde{b}^h = \begin{pmatrix}
\sigma \\
f(x_1) \\
f(x_2) \\
f(x_3) \\
\vdots \\
f(x_{N-1}) \\
f(x_N) \\
\beta \\
\end{pmatrix}
\]

The accuracy of this method is \(O(h^2)\).

2.4. Pure Neumann problem

Assume we want to solve
\[
\begin{align*}
u''(x) &= f(x) \quad 0 < x < 1 \\
u'(0) &= \sigma_0 \\
u(1) &= \sigma_1
\end{align*}
\]

We first integrate the PDE:
\[
u'(1) - u'(0) = \int_0^1 u''(x) dx = \int_0^1 f(x) dx
\]
or equivalently
\[
\sigma_1 - \sigma_0 = \int_0^1 f(x) dx
\]

Thus the pure Neumann problem requires the data to satisfy a compatibility condition. Assuming this is satisfied, it is easy to see that if \(u\) is an exact solution, then \(u + c\) is another exact solution, for any constant \(c\). The pure Neumann problem does not have a unique solution, we say that it is not well-posed. This is also true for the numerical solution. The matrix obtained in Section 2.1 (or 2.2) is in this case:
\[
\tilde{A}^h = \frac{1}{h^2} \begin{pmatrix}
-h & h & 0 & 0 \\
1 & -2 & 1 & 0 & 0 \\
0 & 1 & -2 & 1 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 1 & -2 & 1 & 0 & 0 \\
\end{pmatrix}
\]
Clearly if $v$ is the vector with components all equal to 1, we see that $\tilde{A}^h v = 0$. Thus the matrix $\tilde{A}^h$ is not invertible.
III. ELLIPTIC PROBLEMS IN 2D

3.1 Definition

Consider the problem on the unit square $\Omega = (0, 1)^2$ with boundary denoted by $\partial \Omega$.

$$
\Delta u = f, \quad \text{in} \quad (0, 1)^2 = \Omega
$$

$$
u = g, \quad \text{on} \quad \partial \Omega
$$

We recall that for a function $u = u(x, y)$:

$$
\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}
$$

The finite difference grid is a tensor-product grid with size $\Delta x > 0$ in the $x-$direction and $\Delta y > 0$ in the $y-$direction.

$$
x_i = i \Delta x, \quad y_j = j \Delta y
$$

We replace $\Delta u(x_i, y_j)$ by a five-point stencil that is second order accurate

$$
\frac{1}{(\Delta x)^2} (u(x_{i-1}, y_j) - 2u(x_i, y_j) + u(x_{i+1}, y_j)) + \frac{1}{(\Delta y)^2} (u(x_i, y_{j-1}) - 2u(x_i, y_j) + u(x_i, y_{j+1}))
$$

Denote by $U_{ij}$ the finite difference solution, that approximates $u(x_i, y_j)$. Then the FD scheme is:

$$
\frac{1}{(\Delta x)^2} (U_{i-1,j} - 2U_{ij} + U_{i+1,j}) + \frac{1}{(\Delta y)^2} (U_{i,j-1} - 2U_{ij} + U_{i,j+1}) = f(x_i, y_j), \quad \forall i, j
$$

For simplicity, assume that $\Delta x = \Delta y = h = \frac{1}{N+1}$ for some integer $N \geq 1$. The scheme becomes:

$$
\frac{1}{h^2} (U_{i-1,j} + U_{i+1,j} - 4U_{ij} + U_{i,j-1} + U_{i,j+1}) = f(x_i, y_j), \quad \forall 1 \leq i, j \leq N
$$

The boundary conditions become:

$$
U_{0,j} = g(0, y_j), \quad U_{N+1,j} = g(1, y_j), \quad U_{i,0} = g(x_i, 0), \quad U_{i,N+1} = g(x_i, 1)
$$

The vector of unknowns contains the FD solution at the interior nodes only, since the values on the boundary of the domain are known. We obtain a matrix of size $N^2 \times N^2$. The matrix is sparse: at most five entries per row are nonzero. The structure of the matrix strongly depends on the ordering of the unknowns.

We will consider two types of ordering: a natural rowwise ordering and a checkerboard ordering.

*Natural rowwise ordering*
We start with the bottom row and number the unknowns from left to right, then move to the second row, and repeat. This gives the vector

\[ U^h = (U_{11}, U_{21}, \ldots, U_{N1}, U_{12}, U_{22}, \ldots, U_{N2}, \ldots, U_{1N}, U_{2N}, \ldots, U_{NN}) \]

The resulting matrix is a block tridiagonal matrix:

\[
A^h = \frac{1}{h^2} \begin{pmatrix}
T & I & 0 & 0 \\
I & T & I & 0 \\
0 & I & T & I \\
& \ddots & \ddots & \ddots \\
0 & I & T & I \\
0 & 0 & I & T \\
\end{pmatrix}
\]

where each block \( T \) or \( I \) is a matrix of size \( N \times N \). The matrix \( I \) is simply the identity matrix, and the matrix \( T \) is a tridiagonal matrix:

\[
T = \begin{pmatrix}
-4 & 1 & 0 & 0 \\
1 & -4 & 1 & 0 \\
0 & 1 & -4 & 1 \\
& \ddots & \ddots & \ddots \\
0 & 1 & -4 & 1 \\
0 & 1 & -4 \\
\end{pmatrix}
\]

**Checkerboard ordering**

This type of ordering is also called black-red ordering. We assign each interior node a color (red or black) so that colors are alternated. For each red node, all neighbors (in the five-point stencil) are black nodes. Vice-versa for each black node, all neighbors are red ones. The unknowns are ordered so that all unknowns corresponding to nodes of the same color are grouped together. The resulting matrix takes the form

\[
A^h = \begin{pmatrix}
D & H \\
H^T & D \\
\end{pmatrix}
\]

The matrix \( D = (-4/h^2)I \) is of size \((N^2/2) \times (N^2/2)\). The matrix \( H \) has at most 4 nonzero entries per row.

### 3.2 Examples of stencils

The local truncation error for the five-point stencil given above is:

\[
\tau_{ij} = \frac{1}{h^2} \left( u(x_{i-1}, y_j) - 4u(x_i, y_j) + u(x_{i+1}, y_j) + u(x_i, y_{j-1}) + u(x_i, y_{j+1}) \right) - f(x_i, y_j)
\]

Using a Taylor expansion in both \( x \)- and \( y \)- directions, we obtain

\[
\tau_{ij} = \frac{h^2}{12} \left( \frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4} \right) + O(h^4)
\]
Another example of five-point stencil is:

\[
\frac{1}{h^2} \left( \frac{1}{2} U_{i-1,j-1} + \frac{1}{2} U_{i+1,j+1} - 2U_{ij} + \frac{1}{2} U_{i+1,j-1} + \frac{1}{2} U_{i-1,j+1} \right)
\]

which has a local truncation error \( O(h^2) \).

The following is a nine-point stencil

\[
\frac{1}{3h^2} \left( \frac{1}{2} U_{i-1,j+1} + 2U_{i,j+1} + \frac{1}{2} U_{i+1,j+1} + 2U_{i,j-1} - 10U_{ij} + 2U_{i+1,j} + \frac{1}{2} U_{i-1,j-1} + 2U_{i-1,j+1} + \frac{1}{2} U_{i+1,j-1} \right)
\]

which has a local truncation error of the form

\[
\tau = \frac{h^2}{12} \left( \frac{\partial^4 u}{\partial x^4} + 2 \frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4} \right) + O(h^4)
\]

### 3.3 Analysis of five-point stencil

A first approach is to follow the argument given in 1D. Let \( U^h \) be the vector of unknowns and let \( W \) be the vector with components \( u(x_i, y_j) \) in the same order than the \( U_{ij} \)'s. We have:

\[
A^h (W - U^h) = \tau
\]

where \( \tau \) is the vector with components \( \tau_{ij} \) in the same order than the \( U_{ij} \)'s. So the error vector \( e^h \) is obtained by solving:

\[
A^h e^h = \tau
\]

From this we obtain, if \( A^h \) is invertible:

\[
\|e^h\| \leq \|(A^h)^{-1}\| \|\tau\|
\]

If we have stability (i.e. \( \|(A^h)^{-1}\| \) bounded by a constant independent of \( h \)), then the FD method converges with the same order as the local truncation error. To prove stability, we need to look at the specific norm, and this can be tedious.

A more powerful proof of convergence of the FD method is now presented. The proof uses ideas of maximum principle for elliptic PDEs. For this, we slightly modify our original problem and consider the PDE:

\[
-\Delta u = f, \quad \text{in} \quad (0,1)^2 = \Omega
\]

\[
u = g, \quad \text{on} \quad \partial \Omega
\]

Let \( S_h \) denote the set of grid nodes: it contains both interior nodes and nodes on the boundary. We consider the five-point stencil and we denote by \( N, S, W, E \) the neighbors of any given point \( P \). Define the function \( B \) that acts on the nodes by:

\[
\forall P \in S_h, \quad B(P, Q) = \begin{cases} \frac{4}{h^2} & \text{if } Q = P \\ \frac{1}{h^2} & \text{if } Q \in \{ N, S, W, E \} \\ 0 & \text{otherwise} \end{cases}
\]
The finite difference method can be written as:

\[
\sum_{Q \in S_h} B(P,Q)\hat{U}(Q) = f(P), \quad \forall P \in S_h \setminus \partial \Omega
\]

where \( \hat{U} \) is a discrete function defined only at the nodes in \( S_h \). The boundary condition becomes:

\[
\hat{U}(P) = g(P), \quad \forall P \in S_h \cap \partial \Omega
\]

Let us define the operator \( L_h \) that acts on the set of such discrete functions:

\[
L_h \hat{V}(P) = \sum_{Q \in S_h} B(P,Q)\hat{V}(Q)
\]

If \( P \) has for coordinates \((x_i, y_j)\), we can define the local truncation error and the error as:

\[
\tau(P) = \tau_{ij}, \quad e(P) = u(P) - \hat{U}(P) = u(x_i, y_j) - U_{ij}
\]

**Theorem** Assume that \( \Omega \) is contained in the strip \( \{(x, y) : a < x < b\} \) for some \( a, b \in \mathbb{R} \). Then, for any point \( P \in S_h \) we have

\[
|e(P)| \leq \frac{1}{2} \max(a^2, b^2) \max_{Q \in S_h \setminus \partial \Omega} |\tau(Q)|
\]

**Proof** Define the function \( \Phi \) for all nodes

\[
\Phi(P) = \frac{1}{2} (\max(a^2, b^2) - x_P^2) \max_{Q \in S_h \setminus \partial \Omega} |\tau(Q)|
\]

where \( x_P \) is the \( x \)-coordinate of \( P \). Because of the assumption on \( \Omega \), we see that:

\[
\Phi(P) \geq 0, \quad \forall P \in S_h
\]

We now state two claims that the proof of the theorem uses.

**Lemma 1**

\[
\forall P \in S_h \setminus \partial \Omega, \quad L_h \Phi(P) \geq \max_{Q \in S_h \setminus \partial \Omega} |\tau(Q)|
\]

**Lemma 2** Let \( \hat{V} \) be a discrete function defined on the nodes such that

\[
\forall P \in S_h \setminus \partial \Omega, \quad L_h(\hat{V})(P) \leq 0
\]

\[
\forall P \in S_h \cap \partial \Omega, \quad \hat{V}(P) \leq 0
\]

Then we have

\[
\max_{P \in S_h} \hat{V}(P) \leq 0
\]
We will prove these claims at the end. The error function satisfies:

\[ L_h(e)(P) = \tau(P) \leq \max_{Q \in S_h \setminus \partial\Omega} |\tau(Q)| \leq L_h \Phi(P) \]

by the first lemma. Therefore, the function \( \hat{W} = e - \Phi \) is a discrete function that satisfies:

\[ \forall P \in S_h \setminus \partial\Omega, \quad L_h(\hat{W})(P) \leq 0 \]

Furthermore on the boundary we have

\[ e(P) = 0 \leq \Phi(P) \]

or equivalently

\[ \hat{W}(P) \leq 0 \]

So from Lemmma 2, we conclude that

\[ \max_{P \in S_h} \hat{W}(P) \leq 0 \]

which implies the result.

Now it remains to prove the two lemmas. For Lemma 1, we want to show that for all interior node \( P \)

\[ \sum_{Q \in S_h} B(P, Q) \hat{\Phi}(Q) \geq \max_{Q \in S_h \setminus \partial\Omega} |\tau(Q)| \]

which is equivalent to

\[ \sum_{Q \in S_h} B(P, Q)(\max(a^2, b^2) - x_Q^2) \geq 2 \]

which is again equivalent to

\[ B(P, P)(\max(a^2, b^2) - x_P^2) + \sum_{Q \in S_h, Q \neq P} B(P, Q)(\max(a^2, b^2) - x_Q^2) \geq 2 \]

Since the five-point stencil satisfies the property:

\[ B(P, P) = - \sum_{Q \in S_h, Q \neq P} B(P, Q) \]

we can rewrite the inequality as:

\[ \sum_{Q \in S_h, Q \neq P} B(P, Q)(x_P^2 - x_Q^2) \geq 2 \]

The left-hand side of the inequality reduces to:

\[ B(P, W)(x_P^2 - x_W^2) + B(P, E)(x_P^2 - x_E^2) = \frac{1}{h^2}(x_P^2 - x_W^2 + x_P^2 - x_E^2) = 2 \]
So in fact, we have equality. This proves the result.

Now for Lemma 2, we denote by $R$ the node for which
\[
\hat{V}(R) = \max_{P \in S_h} \hat{V}(P)
\]

If $R$ belongs to the boundary, we are done. Let us assume that $R$ is an interior node, and let $P_1, P_2, \ldots, P_n$ be a finite number of points that connect $R$ to a point $P_n$ on the boundary. We now show that $\hat{V}(R) = \hat{V}(P_1) = \hat{V}(P_2) = \cdots = \hat{V}(P_n)$ by looking at successive pairs. By contradiction, assume that $\hat{V}(P_1) < \hat{V}(R)$. Since $L_h \hat{V}(R) \geq 0$ we have:
\[
B(R, R) \hat{V}(R) + \sum_{Q \neq R} B(P, Q) \hat{V}(Q) \leq 0
\]

Using again the property
\[
0 \leq B(P, P) = -\sum_{Q \in S_h, Q \neq P} B(P, Q)
\]
we have
\[
\hat{V}(R) \leq -\sum_{Q \neq R} \frac{B(R, Q)}{B(R, R)} \hat{V}(Q) < -\sum_{Q \neq R} \frac{B(R, Q)}{B(R, R)} \hat{V}(R)
\]

since $\hat{V}(P_1) < \hat{V}(R)$ and $\hat{V}(R)$ is the maximum value. But the ratio $-\sum_{Q \neq R} \frac{B(R, Q)}{B(R, R)}$ is equal to 1, so we have
\[
\hat{V}(R) < \hat{V}(R)
\]
which is impossible. We consider next the following pair and show $\hat{V}(P_1) = \hat{V}(P_2)$. 

IV. FINITE ELEMENT METHOD FOR ONE-DIMENSIONAL BOUNDARY VALUE PROBLEMS

1. Definition

We consider the two-point boundary value problem:

\[-\frac{d}{dx} \left( k(x) \frac{du}{dx}(x) \right) + c(x)u(x) = f(x), \quad 0 < x < 1 \]

\[u(0) = u(1) = 0\]

where \(k\) and \(c\) are known functions satisfying

\[0 < k_0 \leq k(x) < k_1, \quad 0 \leq c(x)\]

Examples of such PDEs arise from many applications. For instance, in solid mechanics, the solution \(u\) is the displacement of an elastic bar fixed at both ends and subject to a tangential load \(f\). In fluid mechanics, \(u\) is the fluid pressure.

The first step (step 0) is to discretize the domain. We consider the partition

\[0 = x_0 < x_1 < \cdots < x_N < x_{N+1} = 1\]

and we define

\[h_i = x_{i+1} - x_i, \quad h = \max_{0 \leq i \leq N} h_i\]

Let \(V_h\) be the space of continuous piecewise polynomials of degree \(p\) defined on this partition. In other words, if \(v \in V_h\), then \(v \in \mathcal{C}(0,1)\) and \(v|_{(x_i,x_{i+1})}\) is a polynomial of degree \(p\). To obtain a finite element approximation of \(u\), we apply the following steps.

Step 1: multiply PDE by \(v \in V_h\) and integrate over domain

\[\int_0^1 \left( -\frac{d}{dx} \left( k(x) \frac{du}{dx}(x) \right) v(x) + c(x)u(x)v(x) \right) = \int_0^1 f(x)v(x)\]

Step 2: integrate by parts the first term

\[\int_0^1 k(x) \frac{du}{dx}(x) \frac{dv}{dx}(x) - k(1) \frac{du}{dx}(1)v(1) + k(0) \frac{du}{dx}(0)v(0) + \int_0^1 c(x)u(x)v(x) = \int_0^1 f(x)v(x)\]

Step 3: Assume that functions in \(V_h\) also satisfy the Dirichlet boundary conditions: \(v(0) = v(1) = 0\). The boundary terms drop.

\[\int_0^1 k(x) \frac{du}{dx}(x) \frac{dv}{dx}(x) + \int_0^1 c(x)u(x)v(x) = \int_0^1 f(x)v(x)\]
or equivalently
\[ \int_0^1 (ku'v' + cuv) = \int_0^1 f v \]
Let \( a(\cdot, \cdot) \) be a bilinear form and \( \ell(\cdot) \) a linear form defined by
\[ a(u, v) = \int_0^1 (ku'v' + cuv), \quad \ell(v) = \int_0^1 f v \]
The finite element space is
\[ V_h = \{ v \in \mathcal{C}(0,1) : v|_{(x_i, x_{i+1})} \in \mathbb{P}_p, \quad v(0) = v(1) = 0 \} \]
and the finite element method is defined as: find \( u_h \in V_h \) such that
\[ \forall v \in V_h, \quad a(u_h, v) = \ell(v) \]
This type of problem is called a variational formulation; the function \( v \) is called a test function and the function \( u_h \) a trial function.

By varying the polynomial degree, we obtain several finite element solutions of different orders.

2 Finite element solutions of first order

2.1 Basis functions
Let \( p = 1 \) (piecewise linears). \( V_h \) is a finite dimensional space of dimension equal to \( N \). We define standard basis functions for \( V_h \).

\[ \Phi_i(x) = \begin{cases} 0, & 0 \leq x \leq x_{i-1} \\ \frac{x-x_{i-1}}{h_i}, & x_{i-1} \leq x \leq x_i \\ \frac{x_{i+1}-x}{h_i}, & x_i \leq x \leq x_{i+1} \\ 0, & x_{i+1} \leq x \leq 1 \end{cases} \]

The function \( \Phi_i \) is called a hat function (or chapeau), One can show that \( \Phi_1, \Phi_2, \ldots, \Phi_N \) form a basis for \( V_h \). We also note that
\[ \Phi_i(x_j) = \delta_{ij} \]
Because of this property we say the \( \Phi_i \)'s form a nodal basis.

2.2 Linear system
We expand the solution \( u_h \) in the nodal basis:
\[ u_h = \sum_{j=1}^N \alpha_j \Phi_j \]
with $\alpha_j \in \mathbb{R}$. The coefficients $\alpha_j$’s are the unknowns. They satisfy:

$$\forall v \in V_h, \quad \sum_{j=1}^{N} \alpha_j a(\Phi_j, v) = \ell(v)$$

which is equivalent to

$$1 \leq i \leq N, \quad \sum_{j=1}^{N} \alpha_j a(\Phi_j, \Phi_i) = \ell(\Phi_i)$$

Define the matrix $A = (A_{ij})_{ij}$ with $A_{ij} = a(\Phi_j, \Phi_i)$ and the vectors $\mathbf{\alpha} = (\alpha_j)_j$, $\mathbf{b} = (b_j)_j$ with $b_j = \ell(\Phi_j)$. The matrix $A$ is sometimes called the 

\textit{stiffness} matrix and the right-hand side vector is called the 

\textit{load} vector. The linear system is:

$$A\mathbf{\alpha} = \mathbf{b}$$

Let us now write the entries of the matrix in the case where $k = 1, c = 0$. Because the nodal basis functions have compact support, each row of $A$ has at most three nonzero entries:

$$A_{ij} = 0, \quad \text{if } j \notin \{i - 1, i, i + 1\}$$

After some computations, we have

$$A_{ii} = \frac{1}{h_{i-1}} + \frac{1}{h_{i}}, \quad A_{i,i+1} = -\frac{1}{h_{i}}, \quad A_{i,i-1} = -\frac{1}{h_{i-1}}$$

If in addition, the mesh is uniform and $h = h_i$ for all $i$, we can write simply:

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

To compute the entries of $\mathbf{b}$, we need to use a quadrature rule since we may not be able to compute the integral for any $f$. We choose to use the trapezoidal rule on each interval $(x_i, x_{i+1})$.

$$b_i = \int_{x_{i-1}}^{x_i} f \Phi_i = \int_{x_{i-1}}^{x_i} f \Phi_i + \int_{x_i}^{x_{i+1}} f \Phi_i \\ \approx \frac{h_{i-1}}{2} (f(x_i)\Phi_i(x_i) + f(x_{i-1})\Phi_i(x_{i-1})) + \frac{h_{i}}{2} (f(x_i)\Phi_i(x_i) + f(x_{i+1})\Phi_i(x_{i+1})) \\
= \frac{1}{2} f(x_i)(h_{i-1} + h_{i})$$

23
3 Finite element solutions of second order

Let $p = 2$ (piecewise quadratics). $V_h$ is a finite dimensional space of dimension equal to $2N + 1$. The usual basis functions for $V_h$ are of two types. First we define a function $\Phi_i$ that satisfies $\Phi_i(x_j) = \delta_{ij}$. Second we define for each interval a function $\Psi_i$ that vanishes at all nodes and takes the value one at the midpoint of $(x_{i-1}, x_i)$.

4 Existence and uniqueness of finite element solution

The finite element solution $u_h \in V_h$ satisfies

$$\forall v \in V_h, \quad a(u_h, v) = \ell(v)$$

Since this is a linear problem in a finite-dimensional space, proving existence of the solution is equivalent to proving uniqueness of the solution. We will prove uniqueness. Assume that there are two solutions $u_h^1$ and $u_h^2$ and let $w_h = u_h^1 - u_h^2$. Using linearity of $a$ and $\ell$, we have

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

Pick $v = w_h$. This gives

$$\int_0^1 k(w_h')^2 + \int_0^1 c(w_h)^2 = 0$$

Since $k \geq k_0 > 0$, we conclude that $w_h$ is constant on each interval $(x_i, x_{i+1})$. Because $w_h$ is continuous and $w_h(0) = 0$, we have that $w_h = 0$ everywhere.

5 A priori error estimate

We want to bound the error between the numerical solution and the true solution with respect to the following norm:

$$\|v\|_E = (a(v, v))^{1/2}$$

One can check that indeed it is a norm. It is called the energy norm.

The first important result is the orthogonality equation:

$$\forall v \in V_h, \quad a(u - u_h, v) = 0$$
The following lemma reduces the problem of finding a bound of the error to an approximation problem.

**Lemma**

\[ \forall v \in V_h, \quad \| u - u_h \|_E \leq \| u - v \|_E \]

The proof uses Cauchy-Schwarz’s inequality, that we now state. For any functions \( f, g \) that are square-integrable, we have

\[
\int_0^1 fg \leq \left( \int_0^1 f^2 \right)^{1/2} \left( \int_0^1 g^2 \right)^{1/2}
\]

This is Cauchy-Schwarz’s inequality in the \( L^2 \) space (see definition in the next chapter). Cauchy-Schwarz’s inequality in \( \mathbb{R}^N \) is:

\[
\forall a_i, b_i, \quad \sum_{i=1}^N a_i b_i \leq \left( \sum_{i=1}^N a_i^2 \right)^{1/2} \left( \sum_{i=1}^N b_i^2 \right)^{1/2}
\]

We are now ready to prove the lemma. Using the orthogonality equation, we can write for all \( v \in V_h \)

\[
\| u - u_h \|_E^2 = a(u-u_h, u-u_h) = a(u-u_h, u-u_h) + a(u-u_h, u_h-v) = a(u-u_h, u-v) \tag{7}
\]

We expand

\[
a(u - u_h, u - v) = \int_0^1 k(u' - u'_h)(u' - v') + \int_0^1 c(u - u_h)(u - v)
\]

Using Cauchy-Schwarz’s inequality we have

\[
\int_0^1 k(u' - u'_h)(u' - v') \leq \left( \int_0^1 k(u' - u'_h)^2 \right)^{1/2} \left( \int_0^1 k(u' - v')^2 \right)^{1/2}
\]

\[
\int_0^1 c(u - u_h)(u - v) \leq \left( \int_0^1 c(u - u_h)^2 \right)^{1/2} \left( \int_0^1 c(u - v)^2 \right)^{1/2}
\]

Therefore we have

\[
a(u - u_h, u - v) \leq \| u - u_h \|_E \| u - v \|_E
\]

Combining with (7) gives the result. □

Taking infimum over all \( v \in V_h \) and noting that \( u_h \in V_h \), we have:

\[
\| u - u_h \|_E = \inf_{v \in V_h} \| u - v \|_E
\]
So the finite element solution yields the minimum energy norm. Next we need an approximation result (that is independent of the finite element solution). One can show that if $V_h$ contains polynomials of degree $p$ and if $u$ is "smooth" enough, there is a constant $C$ independent of $h$ such that

$$\inf_{v \in V_h} \| u - v \|_E \leq C h^k$$

This implies

$$\| u - u_h \|_E \leq C h^k$$

The piecewise linear finite element method is a first order method (in the energy norm).

Another convergence result in a weaker norm is:

$$\| u - u_h \|_{L^2(0, 1)} \leq C h^{k+1}$$

where $\| v \|_{L^2(0, 1)} = (\int_0^1 v^2)^{1/2}$. So the piecewise linear finite element method is a second order method in the $L^2$ norm.

6 Non-homogeneous Dirichlet boundary conditions

We consider the two-point boundary value problem with modified boundary conditions

$$-\frac{d}{dx} \left( k(x) \frac{du}{dx}(x) \right) + c(x) u(x) = f(x), \quad 0 < x < 1$$

$$u(0) = \alpha, \quad u(1) = \beta$$

We decompose the finite element solution $u_h = u_0 + w_h$, with $u_0$ a piecewise polynomial function satisfying the Dirichlet boundary conditions, and $w_h$ satisfying $w_h(0) = w_h(1) = 0$. We now construct $u_0$ for the case of piecewise linear finite element method. Let $\Phi_0$ and $\Phi_{N+1}$ be defined as:

$$\Phi_0(x) = \begin{cases} \frac{x_1-x}{h_0}, & x_0 \leq x \leq x_1 \\ 0, & x_1 \leq x \leq 1 \end{cases} \quad \Phi_{N+1}(x) = \begin{cases} 0, & 0 \leq x \leq x_N \\ \frac{x-x_N}{h_N}, & x_N \leq x \leq x_{N+1} \end{cases}$$

Then

$$u_0(x) = \alpha \Phi_0(x) + \beta \Phi_{N+1}(x)$$

Second, we write $w_h$ in the usual nodal basis function:

$$w_h(x) = \sum_{i=1}^N \alpha_i \Phi_i(x)$$

To solve for $w_h$, we solve the variational problem:

$$\forall v \in V_h, \quad a(w_h, v) = \ell(v) - a(u_0, v)$$
We see that the resulting matrix $A$ is the same as for zero Dirichlet boundary conditions, and only the right-hand side changes. If we compute $a(u_0, \Phi_i)$ for $1 \leq i \leq N$, we obtain a new right-hand side with $\alpha/h$ added to the first component, and $\beta/h$ added to the last component.
V. FUNCTIONAL SPACES

Convergence of finite element methods is obtained for problems with solutions not necessarily smooth. These solutions belong to particular Hilbert spaces, namely the Sobolev spaces.

1. Inner-product spaces

Definition An inner-product space is a real vector space $X$, in which an inner-product $(\cdot, \cdot)$ has been defined.

1. $\forall u, v \in X, \ (u, v)$ is a real number.
2. $\forall u, v \in X, \ (u, v) = (v, u)$
3. $\forall u, v \in X, \forall \alpha \in \mathbb{R}, \ (\alpha u, v) = \alpha (u, v)$
4. $\forall u \in X, \ (u, u) > 0$ if $u \neq 0$
5. $\forall u, v, w \in X, \ (u + v, w) = (u, w) + (v, w)$

An example of inner-product spaces is $X = \mathbb{R}^n$ with

$$\forall u = (u_i)_i, v = (v_i)_i, \ (u, v) = \sum_{i=1}^{n} u_i v_i$$

Another important example is the space $X = L^2(\Omega)$ for any domain $\Omega \subset \mathbb{R}^n$.

$$L^2(\Omega) = \{ u \text{ Lebesgue measurable} : \int_{\Omega} u^2 < \infty \}$$

Lemma: An inner-product space is a normed space with the norm:

$$\forall u \in X, \ ||u|| = (u, u)^{1/2}$$

This norm satisfies the following properties

1. $||u|| > 0$ if $u \neq 0$
2. $||\alpha u|| = |\alpha|||u||, \ \alpha \in \mathbb{R}$
3. $||u + v|| \leq ||u|| + ||v||$
4. Pythagorean law: If $(u, v) = 0$ then $||u + v||^2 = ||u||^2 + ||v||^2$
5. Cauchy-Schwarz's inequality:

$$| (u, v) | \leq ||u|| ||v||$$
Let us prove Cauchy-Schwarz’s inequality, which is a classic result. Pick \( \lambda \in \mathbb{R} \) and write
\[
0 \leq (u - \lambda v, u - \lambda v) = \|u\|^2 - 2\lambda(u, v) + \lambda^2\|v\|^2
\]
Now choose \( \lambda = (u, v) \):
\[
0 \leq \|u\|^2 - 2(u, v)^2 + (u, v)^2\|v\|^2
\]
First assume that \( \|v\| = 1 \). From the inequality above we have
\[
0 \leq \|u\|^2 - (u, v)^2
\]
Rearranging this inequality and taking square root gives:
\[
|(u, v)| \leq \|u\|
\]
which is Cauchy-Schwarz’s inequality in that particular case. Next, for a general \( v \), define \( w = v/\|v\| \) (possible if \( v \neq 0 \)), which has norm equal to one. Apply the result just obtained:
\[
|(u, v/\|v\|)| \leq \|u\|
\]
Multiply both sides by \( \|v\| \) to obtain the result.

The space \( L^2 \) is equipped with the inner-product and norm:
\[
(u, v) = \int_{\Omega} uv, \quad \|u\|_{L^2(\Omega)} = \left( \int_{\Omega} u^2 \right)^{1/2}
\]

**Definition** If \( (u, v) = 0 \), we say that \( u \) and \( v \) are orthogonal and we write \( u \perp v \). Let \( V \subset X \). If \( u \in X \) is orthogonal to all the elements in \( V \), we simply write \( u \perp V \).

It is easy to show the following:
\[
u = 0 \text{ if and only if } (u, v) = 0 \forall v \in X
\]

2. **Hilbert spaces**

**Definition** A Hilbert space is a complete inner-product space.

Let us recall the definition of Cauchy sequences, and complete space.

**Definition** A sequence \( (u_n)_{n \geq 0} \) in a normed space is said to be Cauchy if
\[
\forall \epsilon > 0, \quad \exists N_0 \quad \forall n \geq N_0, \forall m \geq N_0, \quad \|u_n - u_m\| \leq \epsilon
\]
In other words
\[
\lim_{n \to \infty} \sup_{i, j \geq n} \|u_i - u_j\| = 0
\]
**Definition** If every Cauchy sequence in the space $X$ is convergent to a point in $X$, then the space is said to be complete.

The space $L^2(\Omega)$ introduced in the previous section is an Hilbert space.

### 3. Dual spaces

**Definition** Let $X$ be an inner-product space and let $\Phi : X \to \mathbb{R}$ be a linear mapping. We say the $\Phi$ is a linear functional. In addition, $\Phi$ is called a bounded linear functional if

$$\sup_{u \in X, \|u\|=1} |\Phi(u)| < \infty$$

The norm of $\Phi$ is defined as

$$\|\Phi\| = \sup_{u \in X, \|u\|=1} |\Phi(u)|$$

The space of all bounded linear functionals is called the dual space of $X$ and it is denoted by $X'$.

**Lemma** A linear functional is continuous if and only if it is bounded.

Let $X$ be an inner-product space. Fix $u \neq 0$ in $X$. Define the mapping:

$$\forall v \in X, \; \Phi(v) = (u, v)$$

Then $\Phi \in X'$ and one can show that $\|\Phi\| = \|u\|$.

An important classic result for bounded linear functionals on Hilbert spaces is the Riesz representation theorem.

**Riesz representation theorem** Let $X$ be a Hilbert space and let $\Phi \in X'$. Then there is a unique $u \in X$ such that

$$\forall v \in X, \; \Phi(v) = (u, v)$$

So we can identify $X$ with its dual space $X'$.

As a consequence, if $X$ is a Hilbert space, its dual space $X'$ is also a Hilbert space.

### 4. Weak derivatives

**Definitions** A subset $K \subset \mathbb{R}^n$ is compact if and only if $K$ is closed and bounded. Let $\Omega \subset \mathbb{R}^n$ be an open domain. Let $u : \Omega \to \mathbb{R}$. The support of $u$ is

$$\text{supp}(u) = \{x \in \Omega : u(x) \neq 0\}$$

Let $C_0^\infty(\Omega)$ be the set of all functions that are infinitely differentiable in $\Omega$ and that have compact support. We note that all functions in $C_0^\infty(\Omega)$ and their partial derivatives of any order are zero on the boundary of $\Omega$. 
An example of such \( C_0^\infty(\Omega) \) function is given by \( \phi_{x_0} \), given a point \( x_0 \in \Omega \) and \( r > 0 \) such that the ball of radius \( r \) and origin \( x_0 \), denoted by \( B_r(x_0) \) belongs to \( \Omega \).

\[
\phi_{x_0}(x) = \begin{cases} 
    e^{-(r^2-|x-x_0|^2)^{-1}}, & \text{in } B_r(x_0) \\
    0 & \text{elsewhere}
\end{cases}
\]

**Definition** We say that \( u \) is locally integrable on \( \Omega \) if for any \( K \subset \Omega \), with \( K \) compact, \( \int_K u < \infty \). We write \( u \in L^1_{\text{loc}}(\Omega) \).

Next, we define the partial weak derivatives, that generalize the definition of classical (strong) partial derivatives. Assume \( u \in C^1(\Omega) \) and pick \( \phi \in C_0^\infty(\Omega) \). Using integration by parts and the fact that \( \phi = 0 \) on the boundary \( \partial \Omega \), we have:

\[
\int_{\Omega} \frac{\partial u}{\partial x} \phi = - \int_{\Omega} u \frac{\partial \phi}{\partial x}
\]

**Definition** Let \( u \) be locally integrable on \( \Omega \). The weak partial derivative of \( u \) with respect to \( x \) is the function \( g \in L^1_{\text{loc}}(\Omega) \) that satisfies:

\[
\forall \phi \in C_0^\infty(\Omega), \quad \int_{\Omega} g \phi = - \int_{\Omega} u \frac{\partial \phi}{\partial x}
\]

If, in addition \( u \in C^1(\Omega) \), then the weak derivative \( g \) is equal to the strong derivative \( \frac{\partial u}{\partial x} \). Therefore we abuse the notation and denote \( g \) by \( \frac{\partial u}{\partial x} \). Similarly if \( n, m \) are integers, the weak partial derivative of order \( n + m \) of \( u \) is the function \( g \in L^1_{\text{loc}}(\Omega) \) that satisfies:

\[
\forall \phi \in C_0^\infty(\Omega), \quad \int_{\Omega} g \phi = (-1)^{n+m} \int_{\Omega} u \frac{\partial^{n+m} \phi}{\partial x^n \partial y^m}
\]

We also abuse the notation and write:

\[
g = \frac{\partial^{n+m} u}{\partial x^n \partial y^m}
\]

These definitions exist for all partial derivatives of any order.

**Example** Let us compute the weak partial derivative of \( u \) with respect to \( x \) for \( u \) defined as:

\[
\forall (x, y) \in (0, 1)^2, \quad u(x, y) = \begin{cases} 
    x & 0 < x \leq 0.5 \\
    1 - x & 0.5 < x < 1
\end{cases}
\]

Clearly \( u \) is not differentiable w.r.t. \( x \) in the classical sense since the partial derivative for \( x = 0.5 \) and any \( y \) is not uniquely defined. Let \( \phi \in C_0^\infty(\Omega) \). We compute

\[
- \int_{\Omega} u \frac{\partial \phi}{\partial x} = - \int_0^1 \int_0^{1/2} x \frac{\partial \phi}{\partial x} - \int_0^1 \int_{1/2}^1 (1 - x) \frac{\partial \phi}{\partial x}
\]

\[
= \int_0^1 \int_0^{1/2} \phi - \int_0^1 \int_0^{1/2} 0.5 \phi(0.5, y) - \int_0^1 \int_{1/2}^1 \phi + \int_0^1 \int_0^{1/2} 0.5 \phi(0.5, y) = \int_{\Omega} g \phi
\]
with $g$ defined by:

$$g(x, y) = \begin{cases} 
1 & 0 < x \leq 0.5 \\
-1 & 0.5 < x < 1 
\end{cases}$$

Clearly $g \in L^1_{\text{loc}}(\Omega)$ and thus we conclude that $g$ is the weak partial derivative of $u$ with respect to $x$.

5. Sobolev spaces

We recall the $L^2$ space:

$$L^2(\Omega) = \{v \text{ measurable} \mid \int_\Omega v^2 < \infty\}$$

We define the Sobolev space $H^1(\Omega)$. Let us denote the gradient of $u$ by $\nabla u = (\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y})$ for two-dimensional domains. The partial derivatives are here \textbf{weak}.

$$H^1(\Omega) = \{v \in L^2(\Omega), \nabla u \in L^2(\Omega)\}$$

In other words, $\frac{\partial u}{\partial x} \in L^2(\Omega)$ and $\frac{\partial u}{\partial y} \in L^2(\Omega)$.

In one-dimensional domains, this simply gives:

$$H^1(a, b) = \{v \in L^2(a, b), u' \in L^2(a, b)\}$$

If we increase the order of weak partial derivatives, we obtain other Sobolev spaces.

$$H^2(\Omega) = \{v \in H^1(\Omega), \frac{\partial^2 u}{\partial x^n \partial y^m} \in L^2(\Omega) : 0 \leq n + m \leq 2\}$$

Similarly we define by $H^k(\Omega)$ the space of functions $v \in L^2(\Omega)$ such that all partial weak derivatives of order up to $k$ belong to $L^2(\Omega)$. Clearly we have:

$$k \geq 3, \quad H^k(\Omega) \subset H^{k-1}(\Omega) \subset H^2(\Omega) \subset H^1(\Omega) \subset L^2(\Omega)$$

Sobolev spaces are Hilbert spaces. For instance $H^1(\Omega)$ is equipped with the inner-product and norm:

$$(u, v) = \int_\Omega (\nabla u \cdot \nabla v + uv), \quad \|u\|_{H^1(\Omega)} = \left(\int_\Omega ((\nabla u)^2 + u^2)\right)^{1/2}$$

For PDEs with Dirichlet boundary conditions, we also need the Sobolev space:

$$H^1_0(\Omega) = \{v \in H^1(\Omega), v = 0 \text{ on } \partial \Omega\}$$
The equation \( v = 0 \) on \( \partial \Omega \) has to be understood in the sense of traces. This is a technical and delicate topic, that we do not cover in class. For the space \( H^1_0(\Omega) \), one can show that this space is Hilbert equipped with the inner-product and norm:

\[
(u, v) = \int_{\Omega} \nabla u \cdot \nabla v, \quad \|u\|_{H^1_0(\Omega)} = \left( \int_{\Omega} (\nabla u)^2 \right)^{1/2}
\]

6. Green’s theorem, Poincaré’s inequality

Let \( u \in H^2(\Omega) \) and \( v \in H^1(\Omega) \). We have the following Green’s theorem:

\[
-\int_{\Omega} (\Delta u)v = \int_{\Omega} \nabla u \cdot \nabla v - \int_{\partial \Omega} (\nabla u \cdot n)v
\]

where \( n \) is the unit outward normal vector to \( \Omega \). If \( K \) is a symmetric positive definite matrix, we also have:

\[
-\int_{\Omega} (\nabla \cdot (K \nabla u))v = \int_{\Omega} K \nabla u \cdot \nabla v - \int_{\partial \Omega} (K \nabla u \cdot n)v
\]

Poincaré’s inequality states that there is a constant \( C_P > 0 \) such that

\[
\forall v \in H^1_0(\Omega), \quad \|v\|_{L^2(\Omega)} \leq C_P \|\nabla v\|_{L^2(\Omega)} = \|v\|_{H^1_0(\Omega)}
\]
VI. VARIATIONAL PROBLEMS

Finite element methods belong to the class of discrete variational problems. In this chapter, we study the theory for general variational problems. Variational problems (or formulations) are also called weak problems.

1. Definition

Let $X, V$ be two Hilbert spaces. Let $a : X \times V \to \mathbb{R}$ be a bilinear form and let $\ell : V \to \mathbb{R}$ be a linear form. The variational problem is to find $u \in X$ such that

$$\forall v \in V, \quad a(u, v) = \ell(v)$$

Note that in class, most of the cases have $X = V$.

2. Examples

2.1. Elliptic problems with Dirichlet boundary

Let $\Omega \subset \mathbb{R}^d$ and assume its boundary is partitioned into a Neumann part and a Dirichlet part: $\partial \Omega = \Gamma_N \cup \Gamma_D$, with $\Gamma_N \cap \Gamma_D = \emptyset$. Assume that the Dirichlet part has positive measure: $|\Gamma_D| > 0$. The general elliptic problem is

$$-\nabla \cdot K \nabla u + b \cdot \nabla u + cu = f, \quad \text{in } \Omega$$

$$u = 0, \quad \text{on } \Gamma_D$$

$$K \nabla u \cdot n = g, \quad \text{on } \Gamma_N$$

Formally, we multiply the PDE by a function $v$, integrate over $\Omega$, and use Green’s theorem for the first term:

$$\int_{\Omega} K \nabla u \cdot \nabla v + \int_{\Omega} (b \cdot \nabla u)v + \int_{\Omega} cuv - \int_{\Gamma_N \cup \Gamma_D} K \nabla u \cdot n v = \int_{\Omega} fv$$

Next, we use the boundary conditions and assume that $v = 0$ on $\Gamma_D$. The equation reduces to:

$$\int_{\Omega} K \nabla u \cdot \nabla v + \int_{\Omega} (b \cdot \nabla u)v + \int_{\Omega} cuv = \int_{\Omega} fv + \int_{\Gamma_N} gv$$

For this equation to be well defined, each term should be finite. This implies that we need the following assumptions:

- The matrix $K = (K_{ij})_{i,j}$ is such that $K_{i,j} \in L^\infty(\Omega)$. In addition, we will assume that $K$ is symmetric positive definite.

- $b \in (L^\infty(\Omega))^d$
• $c \in L^\infty(\Omega)$
• $f \in L^2(\Omega)$, $g \in L^2(\Gamma_N)$
• $u, v \in H^1(\Omega)$

Therefore we consider the space

$$X = V = \{ v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D \}$$

and the forms

$$a(u, v) = \int_\Omega K \nabla u \cdot \nabla v + \int_\Omega (b \cdot \nabla u)v + \int_\Omega cuv$$

$$\ell(v) = \int_\Omega fv + \int_{\Gamma_N} gv$$

The weak solution $u \in V$ of the elliptic problem satisfies:

$$\forall v \in V, \quad a(u, v) = \ell(v)$$

Recall that the derivatives in the definition of $a$ are all weak derivatives. A strong solution of the elliptic problem is a function $u \in C^2(\Omega) \cup C^1(\bar{\Omega})$ satisfying the PDE and boundary conditions pointwisely for every point in $\Omega$. The input data $K, b, c, f, g$ are assumed to be smooth in the classical sense.

**Remark:** What is the relationship between the weak and the strong solutions? Clearly a strong solution is also a weak solution. The converse is not true in general. If $u$ is a weak solution and if we assume that $u \in H^2(\Omega)$, then $u$ satisfies the PDE and boundary conditions in the $L^2$ sense, namely almost everywhere in $\Omega$.

**Remark:** The boundary condition $u = 0$ is imposed in the space $V$; it is called an *essential* boundary condition. The Neumann boundary condition is then called *natural*.

### 2.2. Elliptic problems with pure Neumann boundary

Let us consider the simplified elliptic problem

$$-\Delta u = f, \quad \text{in } \Omega$$
$$\nabla u \cdot n = g, \quad \text{on } \partial \Omega$$

Using a similar approach as in the previous section, we can formally obtain a variational formulation defined by

$$V = X = H^1(\Omega)$$

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v$$

$$\ell(v) = \int_\Omega fv + \int_{\partial \Omega} gv$$
This problem does not have a solution for any \( f, g \). Indeed, if we choose \( v = 1 \), the weak problem reduces to

\[
\int_{\Omega} f + \int_{\partial\Omega} g = 0
\]

This condition is called the *compatibility condition* and it should be satisfied by the data in order for the problem to make sense. From now on, we assume this condition is satisfied.

Next, we observe that if \( u \) is a weak solution, then \( u + C \) where \( C \) is any constant, is also a solution. Thus, the weak solution is not unique and the weak formulation is not well-posed. To overcome this difficulty, we replace the space \( H^1(\Omega) \) by the quotient space

\[
V = H^1(\Omega)/\mathbb{R} = \{ [v], v \in H^1(\Omega) \}
\]

where \([v]\) is an equivalence class:

\[
[v] = \{ w \in H^1(\Omega) w - v = C \in \mathbb{R} \}
\]

In other words, two functions in \( V \) are equal if they differ by a constant. We then define new forms:

\[
a([u], [v]) = \int_{\Omega} \nabla u \cdot \nabla v
\]
\[
\ell([v]) = \int_{\Omega} f v + \int_{\partial\Omega} g v
\]

In this case, this new weak problem is well posed.

### 2.3. Stokes problem

Stokes equations characterize fluid flows that are not turbulent. Let \( u(x) \in \mathbb{R}^d \) be the fluid velocity and \( p(x) \in \mathbb{R} \) the fluid pressure at a point \( x \in \Omega \). Let \( \mu > 0 \) be the fluid viscosity. The Stokes equations are:

\[
-\mu \Delta u + \nabla p = f, \quad \text{in} \quad \Omega
\]
\[
\nabla \cdot u = 0, \quad \text{in} \quad \Omega
\]
\[
u = 0, \quad \text{on} \quad \partial\Omega
\]

Recall that \( \Delta u \) is a vector with components \( \Delta u_j \). The weak problem is defined by the following spaces

\[
X = V = (H^1(\Omega))^d
\]
\[
Q = L^2_0(\Omega) = \{ q \in L^2(\Omega) : \int_{\Omega} q = 0 \}
\]
and the forms

\[ \forall u, v \in V, \quad a(u, v) = \int_\Omega \mu \sum_{i,j} \frac{\partial u_i}{\partial x_j} \frac{\partial v_i}{\partial x_j} \]

\[ \forall v \in V, \forall q \in Q, \quad b(v, q) = \int_\Omega q \nabla \cdot v \]

\[ \forall v \in V, \quad \ell(v) = \int_\Omega f \cdot v \]

The weak solution \((u, p) \in V \times Q\) satisfies:

\[ \forall v \in V, \quad a(u, v) + b(v, p) = \ell(v) \]

\[ \forall q \in Q, \quad b(u, q) = 0 \]

Note that this can be rewritten as a standard variational problem given in Section 1.

3. Existence and uniqueness of weak solutions

Well-posedness of the weak problem can be obtained by applying the Lax-Milgram theorem, stated below.

**Lax-Milgram Theorem**

Let \(X\) be a Hilbert space with norm \(\| \cdot \|_X\). Let \(a : X \times X \to \mathbb{R}\) be a bilinear form and let \(\ell : X \to \mathbb{R}\) be a linear form. Assume that

1. The form \(a\) is bounded: there is a constant \(M > 0\) such that

\[ \forall u, v \in X, \quad a(u, v) \leq M \|u\|_X \|v\|_X \]

2. The form \(a\) is coercive: there is a constant \(\alpha > 0\) such that

\[ \forall v \in X, \quad a(v, v) \geq \alpha \|v\|_X^2 \]

3. The form \(\ell\) is continuous, i.e. \(\ell \in X'\)

Then there exists a unique \(u \in X\) satisfying

\[ \forall v \in X, \quad a(u, v) = \ell(v) \]

In addition, the bound holds:

\[ \|u\|_X \leq \frac{1}{\alpha} \|\ell\|_{X'} \]

In that case we say the weak problem is well-posed.
Coercivity of \( a \) is sometimes hard to obtain, or even impossible. A more general Lax-Milgram theorem replaces the coercivity condition by the following inf-sup condition (also called the LBB condition):

\[
\inf_{u \in X} \sup_{v \in X} \frac{|a(u, v)|}{\|u\|_X \|v\|_X} \geq \alpha > 0
\]

One can check that Lax-Milgram’s theorem can be applied to the weak problems defined in Section 2. We now give an example by considering the weak problem: find \( u \in H^1_0(\Omega) \) such that

\[
\forall v \in H^1_0(\Omega), \quad \int_{\Omega} \nabla u \cdot \nabla v = \int_{\Omega} fv
\]

First we check continuity of \( a \), by applying Cauchy-Schwarz’s inequality:

\[
a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v \leq \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)}
\]

We know that \( \|v\|_{H^1_0(\Omega)} = \|\nabla v\|_{L^2(\Omega)} \), so we have continuity of \( a \) with the constant \( M = 1 \).

Second, we check coercivity of \( a \). This is trivial:

\[
a(v, v) = \int_{\Omega} \nabla v \cdot \nabla v = \|v\|_{H^1_0(\Omega)}^2 \geq \|v\|_{H^1_0(\Omega)}^2
\]

and the coercivity constant is \( \alpha = 1 \).

Third, we check continuity of \( \ell \).

\[
\ell(v) = \int_{\Omega} fv \leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)}
\]

Using Poincaré’s inequality, we have:

\[
\|v\|_{L^2(\Omega)} \leq C_P \|v\|_{H^1_0(\Omega)}
\]

for some constant \( C_P > 0 \). Therefore, we have

\[
\ell(v) \leq C_P \|f\|_{L^2(\Omega)} \|v\|_{H^1_0(\Omega)}
\]

which proves boundedness of \( \ell \) and also proves that

\[
\|\ell\|_{X'} \leq C_P \|f\|_{L^2(\Omega)}
\]

We then conclude there is a unique weak solution \( u \). In addition, the bound holds:

\[
\|u\|_{H^1_0(\Omega)} \leq C_P \|f\|_{L^2(\Omega)}
\]
4. Discrete weak problems

Let $X_h \subset X$ be a finite dimensional subspace of the Hilbert space $X$. The discrete variational problem is: find $u_h \in X_h$ such that

$$\forall v_h \in X_h, \quad a(u_h, v_h) = \ell(v_h)$$

This is a finite-dimensional problem, and we can apply Lax-Milgram by replacing $X$ by $X_h$. In other words, if we have continuity and coercivity of $a$ in $X_h$ and continuity of $\ell$ in $X_h$, then the discrete weak problem is well-posed.

Of course, if the assumptions of Lax-Milgram’s theorem hold for $X$, then they also hold for $X_h$. In that case, we have an error bound:

$$\|u - u_h\|_X \leq \left(1 + \frac{M}{\alpha}\right) \inf_{w_h \in X_h} \|u - w_h\|_X$$
VII. FINITE ELEMENT METHODS FOR ELLIPTIC PROBLEMS

In this chapter, we study the finite element method for solving the problem:

$$- \nabla \cdot K \nabla u + cu = f, \quad \text{in } \Omega$$
$$u = 0, \quad \text{on } \partial \Omega$$

We assume \( \Omega \) is a 2D or 3D bounded polygonal domain. The discrete variational problem is: find \( u_h \in X_h \) such that:

$$\forall v_h \in X_h, \quad \int_{\Omega} K \nabla u_h \cdot \nabla v_h + \int_{\Omega} cu_h v_h = \int_{\Omega} f v_h$$

We will define the finite element spaces \( X_h \subset H^1_0(\Omega) \), their basis functions and their approximation properties. We will also describe the resulting linear system.

1. Finite element meshes

The starting point in the numerical solution of PDEs is the discretization of the domain \( \Omega \). Let \( \mathcal{T}_h \) be a partition (or mesh) of \( \Omega \). In other words, \( \mathcal{T}_h \) is made of disjoint triangles or quadrilaterals in 2D, and of hexaedra or tetrahedra or prisms in 3D.

The subscript \( h \) refers to the maximum diameter of the mesh elements.

$$\forall K \in \mathcal{T}_h, \quad h_K = \text{diam}(K) = \max_{x,y \in K} \|x - y\|, \quad h = \max_{K \in \mathcal{T}_h} h_K$$

As \( h \) tends to zero, we say the mesh is finer and finer. If \( N_h \) denotes the number of mesh elements, then we clearly have:

$$\lim_{h \to 0} N_h = \infty$$

**Definition** For any element \( K \), we denote by \( \rho_K \) the diameter of the largest ball inscribed in \( K \). A family of meshes \((\mathcal{T}_h)_h\) is said to be shape-regular if there is a constant \( \theta > 0 \) independent of \( h \) such that

$$\forall h > 0, \quad \forall K \in \mathcal{T}_h, \quad \frac{h_K}{\rho_K} \leq \theta$$

This means that as \( h \to 0 \), we do not allow for the mesh elements to become flat.

**Definition** The family of meshes \((\mathcal{T}_h)_h\) is said to be quasi-uniform if it is shape regular and if there exists a constant \( m \) independent of \( h \) such that

$$\forall h > 0, \quad \forall K \in \mathcal{T}_h, \quad mh \leq h_K$$
2. Finite element spaces

Let $p > 0$ denote the polynomial degree. For each element $K \in \mathcal{T}_h$, let $\mathbb{P}_p(K)$ denote the set of polynomials of degree less than or equal to $p$ defined on $K$. The finite element space is:

$$X_h = \{ v \in C(\Omega) : \forall K \in \mathcal{T}_h \ v|_K \in \mathbb{P}_p(K), \ v = 0 \text{ on } \partial \Omega \}$$

There exists a set of points in $\Omega$ such that any $v \in X_h$ is uniquely determined by its values at these points; these special points are called finite element nodes. Functions in $X_h$ which are zero at all nodes except one, are called nodal functions.

The standard finite element method is the continuous piecewise linear finite element method, namely $p = 1$.

2.1. Piecewise linear elements

Assume for simplicity that $\Omega \subset \mathbb{R}^2$ and that the mesh consists of either triangular elements or quadrilateral elements. Then, the finite element nodes are the interior vertices. Let $N^0_h$ denote the set of interior vertices and let $N^\partial_h$ denote the set of vertices that belong to the boundary. We can characterize the space $X_h$ by the following lemma.

**Lemma**

$$X_h = \{ v : \forall K \in \mathcal{T}_h \ v|_K \in \mathbb{P}_1(K), \ v \text{ continuous at } x_i \in N^0_h, \text{ and } v(x_i) = 0 \forall x_i \in N^\partial_h \}$$

This lemma says that in order to obtain a function that is continuous everywhere, it suffices to have continuity at the interior nodes. This is easy to check. Since $v \in X_h$ is continuous in the interior of each element, it suffices to check continuity on each edge. Fix an edge that is shared by two elements $K_1$ and $K_2$ and consider the difference $w = v|_{K_1} - v|_{K_2}$. The function $w$ is a linear polynomial defined on the edge, that takes the value zero at the endpoints of the edge. Thus, the function $w$ is zero along the edge.
From this, we deduce that the dimension of $X_h$ is equal to the number of interior vertices. For a given node $x_i \in \mathcal{N}_h^0$, let $\Phi_i$ be the piecewise linear function that takes the value $1$ at $x_i$ and the value $0$ at all the other nodes and boundary vertices. Clearly, the function $\Phi_i$ has local compact support that consists of all elements $K$ such that the point $x_i$ is a vertex of $K$. The $\Phi_i$ is called a hat function (as in 1D).

The $\Phi_i$'s form a nodal basis for $X_h$:

$$X_h = \text{span}\{\Phi_i : x_i \in \mathcal{N}_h^0\}$$

### 2.1. Piecewise quadratic elements

In this case, the functions $v \in X_h$ are continuous piecewise quadratic polynomials that vanish on the boundary. The set of finite element nodes varies differs is the elements are triangles or quadrilaterals. First, for triangular elements, $\mathcal{N}_h^0$ consists of the interior vertices plus the set of midpoints of interior edges. For quadrilateral elements, $\mathcal{N}_h^0$ consists of the interior vertices plus the set of midpoints of interior edges, plus the set of barycenters.
For triangular elements, we have two types of basis functions: the hat functions similar to those defined in Section 2.1 and the edge functions.

3. Linear system

The nodal basis functions introduced above are also called global basis functions. Using the nodal basis functions \( \Phi_i \)'s, we can write:

\[
    u_h = \sum_{j=1}^{N} \alpha_j \Phi_j
\]

where \( N \) is the number of nodes in \( \mathcal{N}_h^0 \), and \( \alpha_j \in \mathbb{R} \). Using the fact that \( \Phi_j(x_i) = \delta_{ij} \), we see that

\[
    \alpha_j = u_h(x_j), \quad \forall x_j \in \mathcal{N}_h^0
\]

As in 1D, we obtain the linear system by writing:

\[
    \forall 1 \leq i \leq N, \quad \int_{\Omega} K \nabla u_h \cdot \nabla \Phi_i + \int_{\Omega} c u_h \Phi_i = \int_{\Omega} f \Phi_i
\]

or equivalently

\[
    \forall 1 \leq i \leq N, \quad \sum_{j=1}^{N} \left( \int_{\Omega} K \nabla \Phi_j \cdot \nabla \Phi_i + \int_{\Omega} c \Phi_j \Phi_i \right) = \int_{\Omega} f \Phi_i
\]

Let \( A = (A_{ij})_{i,j}, \ b = (b_i)_i \) and \( \alpha = (\alpha_i)_i \) satisfying:

\[
    A \alpha = b
\]
with
\[
A_{ij} = \int_{\Omega} K \nabla \Phi_j \cdot \nabla \Phi_i + \int_{\Omega} c \Phi_j \Phi_i, \quad b_i = \int_{\Omega} f \Phi_i
\]
In order to compute the entries of the matrix and right-hand side vector, we write:
\[
A_{ij} = \sum_{K \in T_h} \int_K (K \nabla \Phi_j \cdot \nabla \Phi_i + c \Phi_j \Phi_i) = \sum_{K \in T_h} A^K_{ij}
\]
and similarly
\[
b_i = \sum_{K \in T_h} \int_K f \Phi_i = \sum_{K \in T_h} b^K_i
\]
Therefore, if we fix a triangle $K$, we need to compute the contribution $A^K_{ij}$ that corresponds to the interaction between nodes $x_i$ and $x_j$. The matrix $A$ is called the global matrix, and the matrices $A^K$ are called the local matrices.

Throughout this section, we assume that the polynomial degree is one.

3.1 Assembling the global matrix and right-hand side
Here we assume we know how to compute the local matrices $A^K$ and right-hand side vectors $b^K$. This will be explained later in the chapter. We start with the finite element mesh, which contains the following information:

- List of vertices with their coordinates. The interior vertices are numbered first, from 1 to $N$, then the boundary vertices are numbered from $N + 1$ to some number. For each vertex, we associate an integer that takes the value 0 if the vertex is interior, and the value $-1$ if the vertex is on the boundary. If a vertex is on the boundary, we say that it is constrained, otherwise it is called free.

- List of elements with their connectivities. If we have triangular elements, the list of vertices for each triangle is given, in the counterclockwise fashion. From the mesh, it is easy to create a two-dimensional array $\text{LGnum}$, such that $\text{LGnum}(i, k)$ is equal to the number of the $k^{th}$ vertex of element number $i$, for $1 \leq k \leq 3$, if the $k^{th}$ vertex is in the interior. Otherwise, if the $k^{th}$ vertex is on the boundary, $\text{LGnum}(i, k)$ is equal to $-1$. 


The assembling of the global matrix and right-hand side vector is given below.

- Initialize $A$ and $b$ to zero
- Loop over elements $K \in \mathcal{T}_h$
  
  Compute local matrix $A^K$ and right-hand side $b^K$
  
  Loop over vertices of $K$: for $li = 1$ to 3
    
    If $\text{LGnum}(K, li)$ is positive:
      
      Set $GI = \text{LGnum}(K, li)$
      
      $b(GI) = b(GI) + b^K(li)$
      
      Loop over vertices of $K$: for $lj = 1$ to 3
        
        If $\text{LGnum}(K, lj)$ is positive:
          
          Set $GJ = \text{LGnum}(K, lj)$
          
          $A(GI, GJ) = A(GI, GJ) + A^K(li, lj)$
        
      End of if statement
    
  End of loop $lj$
  
  End of if statement
  
  End of loop $li$

- End of loop over $K$

### 3.2 Computing the local matrix and right-hand side

Fix an element $K$. The entry $(i, j)$ of the local matrix is

$$A^K_{i,j} = \int_K (K \nabla \Phi_j \cdot \nabla \Phi_i + c \Phi_j \Phi_i)$$

The global basis function $\Phi_i$ has local compact support, so for most of the elements, the restriction of $\Phi_i$ on these elements is zero.

If we fix now the nodes $x_i$ and $x_j$, the quantity $A^K_{i,j}$ is zero if either the supports of $\Phi_i$ and $\Phi_j$ do not intersect (or intersect on a set of measure zero), or if the element $K$ is not included in the intersection of the supports of $\Phi_i$ and $\Phi_j$.

For a given node $x_i$ and element $K$, if $K$ is included in the support of $\Phi_i$, we say that the restriction of $\Phi_i$ on $K$ is a local basis function.

$$\Phi_i|_K = \Psi_\ell, \quad \text{for some } \ell \in \{1, 2, 3\}$$
where $\Psi_1, \Psi_2, \Psi_3$ are linear polynomials defined on $K$. If the vertices of $K$ are denoted by $V_1, V_2, V_3$, then we have the nodal property:

$$\Psi_\ell(V_k) = \delta_{\ell k}$$

The $\Psi_i$’s form a local basis on $K$.

From now on, we consider only the case where $K$ is contained in the intersection of the supports of $\Phi_i$ and $\Phi_j$. Assume that $\Phi_i|_K = \Psi_\ell$ and $\Phi_j|_K = \Psi_k$ for some $\ell, k$ in the set $\{1, 2, 3\}$. Then we can write

$$A^K_{i,j} = \int_K (K \nabla \Psi_\ell \cdot \nabla \Psi_k + c \Psi_\ell \Psi_k)$$

We need to compute an integral on the mesh element $K$. This is done very efficiently by mapping the element $K$ onto a reference element, denoted by $\hat{K}$.

### 3.2.1 Reference elements versus physical elements

A mesh element $K \in T_h$ is also referred to as a **physical element**. We now define the **reference element** $\hat{K}$.

If $K$ is a triangle, the reference element $\hat{K}$ is the triangle with vertices $\hat{V}_1 = (0, 0)$, $\hat{V}_2 = (1, 0)$, $\hat{V}_3 = (0, 1)$.

There exists an affine mapping $F_K$ that maps $\hat{K}$ onto $K$, such that the vertex $\hat{V}_i$ is mapped onto the vertex $V_i$. If $(x_i, y_i)$ are the coordinates of the vertex $V_i$, then the mapping $F_K$ is defined by

$$\forall (\hat{x}, \hat{y}) \in \hat{K}, \quad F_K(\hat{x}, \hat{y}) = \left( \sum_{i=1}^3 x_i \hat{\Psi}_i(\hat{x}, \hat{y}), \sum_{i=1}^3 y_i \hat{\Psi}_i(\hat{x}, \hat{y}) \right)$$
where the linear functions $\hat{\Psi}_i$ are given by:

\[
\begin{align*}
\hat{\Psi}_1(\hat{x}, \hat{y}) &= 1 - \hat{x} - \hat{y}, \\
\hat{\Psi}_2(\hat{x}, \hat{y}) &= \hat{x}, \\
\hat{\Psi}_3(\hat{x}, \hat{y}) &= \hat{y}
\end{align*}
\]

It is easy to check that:

\[
\forall (\hat{x}, \hat{y}) \in \hat{K}, \quad F_K(\hat{x}, \hat{y})^T = B_K(\hat{x}, \hat{y})^T + b_K
\]

where $B_K$ is a $2 \times 2$ matrix with constant coefficients, and $b_K$ is a $2 \times 1$ constant vector. One can also compute the determinant of $B_K$ and check that

\[
\det(B_K) = 2 \operatorname{area}(K)
\]

Via the mapping $F_K$, we can rewrite the integrals on $K$ as:

\[
\int_K \nabla \Psi_\ell \cdot \nabla \Psi_k = \det(B_K) \int_{\hat{K}} (B_K^T)^{-1} \hat{\nabla} \hat{\Psi}_\ell \cdot (B_K^T)^{-1} \hat{\nabla} \hat{\Psi}_k
\]

and

\[
\int_K \Psi_\ell \Psi_k = \det(B_K) \int_{\hat{K}} \hat{\Psi}_\ell \hat{\Psi}_k
\]

The notation $\hat{\nabla} \hat{\Psi}_\ell$ simply means the gradient of $\hat{\Psi}_\ell$ with respect to $(\hat{x}, \hat{y})$. We also remark that it is very easy to compute the transpose and inverse of $B_K$, which is a $2 \times 2$ matrix. Finally, as was done in class, one can show that there are constants $\beta_i$ that depend only on the coordinates of the vertices of $K$, and $3 \times 3$ constant matrix $M_i$ such that

\[
A^K = \sum_{j=1}^{3} \beta_j M_j
\]

The computation of the matrices $M_j$ can be done exactly by hand (see homework) and these matrices should be defined in the code, at the initialization step. Then, it suffices to compute the numbers $\beta_j$ inside the loop over the elements $K$. The computation of $A^K$ is then very fast.

If $K$ is a quadrilateral, the reference element $\hat{K}$ is the rectangle with vertices $\hat{V}_1 = (-1, -1)$, $\hat{V}_2 = (1, -1)$, $\hat{V}_3 = (1, 1)$, $\hat{V}_3 = (-1, 1)$.
There exists a mapping $F_K$ that maps $\hat{K}$ onto $K$, such that the vertex $\hat{V}_i$ is mapped onto the vertex $V_i$. But the mapping is not necessarily affine. If $(x_i, y_i)$ are the coordinates of the vertex $V_i$, then the mapping $F_K$ is defined by

$$\forall (\hat{x}, \hat{y}) \in \hat{K}, \quad F_K(\hat{x}, \hat{y}) = \left( \sum_{i=1}^{4} x_i \hat{\Psi}_i(\hat{x}, \hat{y}), \sum_{i=1}^{4} y_i \hat{\Psi}_i(\hat{x}, \hat{y}) \right)$$

where the linear functions $\hat{\Psi}_i$ are given by:

$$\hat{\Psi}_1(\hat{x}, \hat{y}) = \frac{1}{4}(1 - \hat{x})(1 - \hat{y}),$$
$$\hat{\Psi}_2(\hat{x}, \hat{y}) = \frac{1}{4}(1 + \hat{x})(1 - \hat{y}),$$
$$\hat{\Psi}_3(\hat{x}, \hat{y}) = \frac{1}{4}(1 + \hat{x})(1 + \hat{y}),$$
$$\hat{\Psi}_4(\hat{x}, \hat{y}) = \frac{1}{4}(1 - \hat{x})(1 + \hat{y})$$

Let $J_K$ be the Jacobian of $F_K$. The determinant of $J_K$ is not necessarily a constant, and we have:

$$\int_K \nabla \Psi_\ell \cdot \nabla \Psi_k = \int_{\hat{K}} \det(J_K)(J_K^T)^{-1} \nabla \hat{\Psi}_\ell \cdot (J_K^T)^{-1} \nabla \hat{\Psi}_k$$

and

$$\int_K \Psi_\ell \Psi_k = \int_{\hat{K}} \det(J_K)\hat{\Psi}_\ell \hat{\Psi}_k$$

We note that the determinant of $J_K$ is not zero if all corner angles in the quadrilateral are strictly between 0 and $\pi$.

### 3.2.2 Quadrature rules on reference element

It remains to compute the local right-hand side $b^K$. Assume that $K$ is a triangle, by the mapping $F_K$, we have

$$\int_K f \Psi_\ell = \det(B_K) \int_{\hat{K}} (f \circ F_K)\hat{\Psi}_\ell$$

Since $f$ is a general function, we cannot compute this integral analytically in most cases. It is sufficient to use quadrature rules. We replace the integral by a finite sum, that involves quadrature weights $w_j$ and quadrature points $(\xi_j, \eta_j) \in \hat{K}$:

$$\int_K f \Psi_\ell \approx \det(B_K) \sum_{j=1}^{m} (f \circ F_K)(\xi_j, \eta_j)\hat{\Psi}_\ell(\xi_j, \eta_j)$$
There are various quadrature rules of various degrees. An example is given by
\[(\xi_1, \eta_1) = (\frac{1}{2}, 0), (\xi_2, \eta_2) = (\frac{1}{2}, \frac{1}{2}), (\xi_3, \eta_3) = (0, \frac{1}{2}), w_1 = w_2 = w_3 = \frac{1}{6}\]
This rule is exact for quadratic polynomials.

4. Convergence rates
From the general theory of variational problems, we already know that there is a constant \(C > 0\) independent of \(h\) such that
\[
\|u - u_h\|_\mathcal{E} \leq C \inf_{w_h \in X_h} \|u - w_h\|_\mathcal{E}
\]
where the energy norm is given by:
\[
\|u - u_h\|_\mathcal{E} = \left(\|K^{1/2}\nabla(u - u_h)\|^2_{L^2(\Omega)} + \|c^{1/2}(u - u_h)\|^2_{L^2(\Omega)}\right)^{1/2}
\]
So the problem of finite element error reduces to an approximation problem where we want the element \(w_h\) in \(X_h\) that minimizes the energy norm.

4.1 h-version of FEM
In the classical finite element method, the polynomial degree is fixed and equal to \(p\). In most cases, in fact, \(p = 1\). The mesh is successively refined, so that \(h\) decreases to zero. In this case, one can show there is a constant independent of \(h\) such that
\[
\|u - u_h\|_\mathcal{E} \leq Ch^p\|u\|_{H^{p+1}(\Omega)}
\]
By using the Aubin-Nitsche lift, one can show an error estimate in the \(L^2\) norm:
\[
\|u - u_h\|_{L^2(\Omega)} \leq Ch^{p+1}\|u\|_{H^{p+1}(\Omega)}
\]

4.2 p-version of FEM
In the p-version of finite element method, the mesh is fixed \((h = h_0)\) and the polynomial degree gradually increases. Let us denote by \(u_{h_0}^p\) the finite element solution. In this case, one can show there is a constant independent of \(p\) such that
\[
\|u - u_{h_0}^p\|_\mathcal{E} \leq C_p\|u\|_{H^{p+1}(\Omega)}
\]

4.3 hp-version of FEM
A more popular technique than the p-version, is the hp-version of finite element method. Both the mesh size and the polynomial degree vary. Let us denote by \(u_{h}^p\) the finite element solution. There is a constant independent of \(h\) and \(p\) such that
\[
\|u - u_{h}^p\|_\mathcal{E} \leq C_{hp}\frac{h^p}{p^p}\|u\|_{H^{p+1}(\Omega)}
\]
By carefully choosing \(h\) and \(p\), one may even obtain exponential rate of convergence in this case.